SBCAS25

XXV SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO APLICADA À SAÚDE

9 A 13 DE JUNHO DE 2025 PORTO ALEGRE , RS

# LIVRO DE MINICURSOS


**ORGANIZAÇÃO**

Mariana Recamonde-Mendoza

Lina Garcés

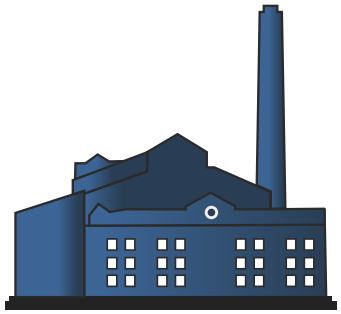

**REALIZAÇÃO**

INF INSTITUTO DE INFORMÁTICA UFRGS

ICMC USP SÃO CARLOS

UNISINOS

SBC Sociedade Brasileira de Computação




25º Simpósio Brasileiro de Computação Aplicada à Saúde

Porto Alegre, RS, 09 a 13 de junho de 2025

# Livro de Minicursos do XXV Simpósio Brasileiro de Computação Aplicada à Saúde

**Organização do Livro**
Mariana Recamonde Mendoza (UFRGS)
Lina Garcés (USP)

**Coordenação Geral do Simpósio**
Rodrigo da Rosa Righi (UNISINOS)

**Realização**
Universidade do Vale do Rio dos Sinos – UNISINOS
Sociedade Brasileira de Computação – SBC

Porto Alegre
Sociedade Brasileira de Computação – SBC
2025

Dados Internacionais de Catalogação na Publicação (CIP)

S612 Simpósio Brasileiro de Computação Aplicada à Saúde (25. : 06 – 13 junho 2025 : Porto Alegre)
Minicursos do SBCAS 2025 [recurso eletrônico] / organização: Mariana Recamonde Mendoza; Lina Garcés. – Dados eletrônicos. – Porto Alegre: Sociedade Brasileira de Computação, 2025.
354 p. : il. : PDF ; 23 MB

Modo de acesso: World Wide Web.
ISBN 978-85-7669-631-5 (e-book)

1. Computação – Brasil – Evento. 2. Saúde aplicada. 3. Tecnologia em saúde. I. Mendoza, Mariana Recamonde. II. Garcés, Lina. III. Sociedade Brasileira de Computação. IV. Título.

CDU 004(063)

Ficha catalográfica elaborada por Annie Casali – CRB-10/2339
Biblioteca Digital da SBC – SBC OpenLib

# Sociedade Brasileira de Computação – SBC



**Contato**
Av. Bento Gonçalves, 9500
Setor 4 - Prédio 43.412 - Sala 219
Bairro Agronomia
91.509-900 – Porto Alegre RS

CNPJ: 29.532.264/0001-78
http://www.sbc.org.br

## Comitê de Programa de Minicursos

Alexandre Sztajnberg (UERJ)
Ana Carolina Inocêncio (UFG)
Anderson Rocha Tavares (UFRGS)
Antonio Tadeu Azevedo Gomes (LNCC)
Danielo G. Gomes (UFC)
Débora Christina Muchaluat Saade (UFF)
Dianne Scherly Varela de Medeiros (UFF)
Diogo Menezes Ferrazani Mattos (UFF)
Eduardo Simões Albuquerque (UFG)
Gabriel de Oliveira Ramos (UNISINOS)
Jorge Barbosa (UNISINOS)
Karina S. Machado (FURG)
Lucas Ferrari de Oliveira (UFPR)
Márcia Ito (FATEC-SP)
Nicollas Rodrigues de Oliveira (UFF)
Renato de Freitas Bulcão Neto (UFG)
Rossana Maria de Castro Andrades (UFC)
Sérgio Carvalho (UFG)

## Comissão Especial de Computação Aplicada à Saúde (CE-CAS) - SBC

Cristiano André da Costa (UNISINOS) - Coordenador
Rodrigo de Melo Souza Veras (UFPI) - Vice-Coordenador
Débora Christina Muchaluat Saade (UFF)
Lina Garcés (USP)
Lucas Ferrari de Oliveira (UFPR)
Márcia Ito (FATEC-SP)
Natalia Castro Fernandes (UFF)
Paulo Eduardo Ambrósio (UESC)
Sergio Teixeira de Carvalho (UFG)

# Mensagem da Coordenação Geral

É com grande satisfação que apresentamos este livro de minicursos, um material que reflete a essência do nosso evento: a intersecção vibrante entre tecnologia e saúde na era da Saúde 4.0. Entendemos que o futuro da medicina e do cuidado ao paciente é indissociável da inovação tecnológica, e é por isso que reunimos especialistas para compartilhar conhecimentos de ponta que moldarão as próximas décadas. Os temas abordados aqui variam desde a exploração aprofundada e recuperação de grandes volumes de informações biomédicas, cruciais para pesquisa e aplicação clínica avançada, passando pela avaliação da experiência do usuário em ambientes tecnológicos imersivos, onde a biofeedback revela-se uma ferramenta poderosa para aprimorar a interação humana. Mergulhamos também no universo da inteligência artificial (IA), explorando como a Geração Aumentada por Recuperação (RAG) está revolucionando a forma de responder a perguntas de clínica médica, fornecendo suporte vital à decisão. Abordamos as complexidades da integração de dados em larga escala no contexto brasileiro, com foco em aspectos metodológicos e práticos do Big Data Linkage, essencial para gerar insights valiosos em saúde pública. Não menos importante, este volume detalha a imperativa necessidade de proteger as infraestruturas de saúde contra ameaças cibernéticas, apresentando estratégias robustas de avaliação de riscos e investimentos em cibersegurança. Além disso, oferece os primeiros passos na preparação e análise de dados para quem deseja iniciar sua jornada na ciência de dados aplicada à saúde, discute a construção de sistemas inteligentes éticos e equitativos, fundamentais para evitar vieses e promover a inclusão na IA em saúde, e explora técnicas inovadoras para monitoramento não invasivo de sinais vitais. Essa é uma edição especial, pois estamos comemorando 25 anos do nosso SBCAS. Os oito minicursos apresentados nesse livro foram frutos de uma seleção muito rigorosa; portanto, agradecemos não somente aos autores aqui, mas também àqueles que submeteram a sua proposta. Por fim, esperamos que este material sirva como um guia valioso para aprimorar seus conhecimentos e habilidades, impulsionando a inovação e o desenvolvimento de soluções e projetos que transformarão a saúde no Brasil e no mundo. Tenham um excelente evento e aproveitem cada aprendizado!

Rodrigo da Rosa Righi (UNISINOS)

Coordenador Geral do SBCAS 2025

# Mensagem da Coordenação de Minicursos

O Livro de Minicursos do XXV Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2025) apresenta uma coletânea de oito capítulos correspondentes aos minicursos selecionados para esta edição do evento. Recebemos um total de 16 propostas, todas de excelente qualidade e inquestionável relevância, avaliadas por meio de um criterioso processo de revisão por pares do tipo blind. Com o apoio do coordenador geral do evento, Prof. Rodrigo Righi, conseguimos ampliar a programação desta trilha e incorporar um número expressivo de minicursos, alcançando uma taxa de aceitação de 50%. O objetivo foi contemplar o máximo possível das contribuições recebidas, com atenção à diversidade temática e ao potencial de impacto na área da computação em saúde. Agradecemos imensamente aos revisores pelo comprometimento e rigor, que possibilitaram a seleção de oito propostas representativas de áreas emergentes, com grande potencial para transformar e melhorar o cuidado em saúde. Estendemos também nosso agradecimento ao Prof. Rodrigo Righi pelo apoio essencial durante a organização do evento e da programação dos minicursos, e pelo convite para atuarmos como coordenadoras desta trilha.

Os capítulos deste livro cobrem estratégias de cibersegurança para o setor da saúde; o acesso e análise de dados clínicos no banco MIMIC-IV; técnicas fundamentais para preparação e análise exploratória de dados em saúde; fundamentos e estratégias para o desenvolvimento de modelos de inteligência artificial justos e equitativos; o uso de biofeedback em ambientes imersivos; a Geração Aumentada por Recuperação (RAG) para responder perguntas clínicas com maior precisão; aplicações da fotopletismografia por imagem para extração de sinais vitais; e os aspectos metodológicos e práticos do Big Data Linkage no Brasil. Sem dúvidas, este material cobre diferentes aspectos da computação aplicada à saúde, oferecendo uma visão abrangente e prática das tendências e desafios da área.

Foi uma honra coordenar a trilha de minicursos em um ano tão simbólico, que marca os 25 anos do SBCAS. Ao longo dessas décadas, o simpósio consolidou-se como um espaço fundamental para o intercâmbio de ideias, a disseminação de conhecimento, a inovação e a formação de recursos humanos qualificados em uma área de caráter profundamente interdisciplinar. O elevado número de submissões de minicursos reforça a importância do evento para a comunidade acadêmica e profissional. Esses resultados só são possíveis devido à dedicação e ao trabalho de qualidade dos colegas que nos precederam nas edições anteriores do evento.

Esperamos que todos os estudantes, pesquisadores e profissionais das áreas da computação e da saúde aproveitem esta oportunidade de atualização de conhecimento, troca de experiências e interação com a comunidade. Este livro pretende ser não apenas um registro do conhecimento compartilhado durante os minicursos, mas também uma contribuição concreta à formação de profissionais conscientes dos desafios técnicos, éticos e sociais da computação aplicada à saúde. Que a leitura inspire novas pesquisas, colaborações e soluções que ampliem o impacto positivo da tecnologia no cuidado com a vida.

Mariana Recamonde-Mendoza (UFRGS)

Lina Garcés (USP)

Coordenadoras de Minicursos do SBCAS 2025

# Sumário

# Capítulo 1

# Saúde Sob Ataque: Da Avaliação de Riscos ao Desenvolvimento de Estratégias de Investimentos em Cibersegurança na Área da Saúde

Muriel Figueredo Franco[1], Laura Rodrigues Soares[2], Jéferson Campos Nobre[2]

[1]Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA)
Departamento de Ciências Exatas e Sociais Aplicadas, Porto Alegre, Brasil
[2]Universidade Federal do Rio Grande do Sul (UFRGS)
Instituto de Informática (INF), Porto Alegre, Brasil

***Abstract.*** *Cybersecurity is one of the essential pillars of the digital revolution. Cybercriminals are increasingly targeting governments and companies from different sectors. These attacks have different technical impacts on their infrastructures and services, which result in direct and indirect economic impacts on their business models. In addition, the social and reputational impacts have been increasingly frequent as society increasingly depends on digital services. The health sector has been most affected by cybercriminals' financial incentives to obtain sensitive data and make critical services unavailable. The World Health Organization (WHO) has emphasized the profound impact of cyberattacks on hospitals and health services, calling for urgent and collective global action to tackle this growing crisis. Therefore, in this work, we will analyze and understand the risks and particularities of the health sector, as well as define the main steps for efficient planning of cybersecurity strategies for the health sector. In addition, we will map the primary cybersecurity efforts of academia and industry aimed at the healthcare sector.*

***Resumo.*** *A cibersegurança é um dos pilares essenciais durante a revolução digital. Governos e empresas de diferentes setores têm sido, cada vez mais, alvos de ataques de cibercriminosos. Tais ataques têm diferentes impactos técnicos em suas infraestruturas e serviços, que resultam em impactos econômicos diretos e indiretos em seus modelos de negócios. Além disso, os impactos sociais e de reputação têm sido cada vez mais frequentes, já que a sociedade depende cada vez mais de serviços digitais. O setor da saúde tem sido um dos mais afe-*

*tados, principalmente devido aos incentivos financeiros que os cibercriminosos possuem para obter dados sensíveis e tornar serviços críticos indisponíveis. A Organização Mundial da Saúde (OMS) vem enfatizando o grave impacto dos ataques cibernéticos em hospitais e serviços de saúde, exigindo uma ação global urgente e coletiva para enfrentar essa crise crescente. Portanto, neste trabalho, iremos analisar e compreender os riscos e as particularidades do setor da saúde, bem como definir os principais passos para um planejamento eficiente de estratégias de cibersegurança para o setor da saúde. Além disso, serão mapeados os principais esforços em cibersegurança da academia e da indústria direcionados ao setor da saúde.*

## 1.1. Introdução

A atenção e preocupação com a cibersegurança têm aumentado na última década devido à sua importância para manter sistemas digitais e serviços interdependentes disponíveis. Incidentes de cibersegurança têm sido noticiados pela mídia de forma cada vez mais constante, já que governos, empresas e a sociedade se tornaram dependentes de sistemas computacionais [Singer and Friedman 2013]. Com isso, a cibersegurança emerge como um pilar essencial para a sociedade. Investimentos em cibersegurança têm-se tornado cada vez mais comuns; porém, a cibersegurança ainda é vista como um custo [Gordon et al. 2018] e não como a prioridade necessária para manter a disponibilidade de serviços, a operação de negócios e garantir a proteção de dados de usuários.

A rápida evolução tecnológica tem criado um cenário fértil para inovações e permitido acesso a serviços cada vez mais complexos e automatizados. Tal evolução permite benefícios diretos para a sociedade, como serviços para comunicação, gestão financeira e monitoramento de saúde, e também oportunidades para empresas. Porém, com a dependência tecnológica, também existe um aumento crescente de ciberataques (por exemplo, ransomware, phishing e negação de serviço) a sistemas e usuários, com diversos impactos técnicos, econômicos, legais e sociais [Franco et al. 2023a]. Tais impactos reforçam a ideia de que a cibersegurança não deve ser pensada apenas sob uma perspectiva técnica.

Os ciberataques podem gerar prejuízos significativos, independentemente do setor de atuação ou do tamanho da organização. Em situações como a interrupção de serviços ou o vazamento de dados, os impactos econômicos são imediatos, envolvendo perda de clientes, danos à reputação e eventuais ações judiciais decorrentes da exposição de informações sensíveis. Além das consequências financeiras, esses ataques também podem provocar efeitos sociais diretos, especialmente quando atingem infraestruturas críticas, como os sistemas de transporte [Islam et al. 2023], energia [Beerman et al. 2023] e saúde [Javaid et al. 2023], afetando diretamente a vida e o bem-estar da população.

De acordo com relatórios de vazamento de dados, dados médicos ainda são os mais vazados e com maiores custos [IBM Security 2024], seguidos de dados bancários e dados pessoais. Em um estudo conduzido com 22.052 incidentes de vazamentos de dados [Verizon Business 2025], foi observado que o setor da saúde permanece como um dos principais alvos de ciberataques. Tal motivação se deve ao alto valor econômico e social dos dados e sistemas. O setor é um dos mais visados, por exemplo, por ataques de ransomware devido à urgência de acesso a dados e sistemas, onde uma disrupção nos serviços pode ocasionar impactos críticos em um curto período de tempo. Ciberataques

como negação de serviço e phishing permanecem, junto com ransomware, como as maiores ameaças em todos os setores.

O setor da saúde se caracteriza pelo uso da tecnologia em diferentes frentes para possibilitar serviços otimizados para gestores, profissionais da saúde e cuidados ao paciente. O uso de diferentes tecnologias pode ser observado, como, por exemplo, equipamentos de diagnóstico e monitoramento, aplicações para gerenciamento de consultas e exames e sensores. Tais sistemas e equipamentos possuem integrações e geram dados sensíveis que podem ser acessados por diferentes partes interessadas. Por exemplo, médicos precisam ter acesso às informações mais recentes sobre a saúde de um paciente durante um procedimento. Portanto, ciberataques que visam o setor podem explorar desde a necessidade de disponibilidade de serviços e equipamentos até mesmo a confidencialidade das informações médicas e sensíveis geradas e armazenadas sobre procedimentos e pacientes.

Devido à complexidade e importância do ecossistema do setor da saúde, existem diversas oportunidades também para atacantes. Por exemplo, ataques de ransomware podem tornar indisponível o acesso a sistemas de internação, bloquear o acesso a equipamentos de exames e afetar até mesmo cirurgias. Ataques dessa magnitude podem causar impactos em atendimentos em todo um país, como o caso reportado no Reino Unido, em 2017 (*cf.* Seção 1.4.2). Além disso, ciberataques de phishing com foco em pacientes ou profissionais da saúde podem possibilitar a obtenção de acesso a sistemas críticos. Além do componente humano, diversas aplicações existentes para o setor (por exemplo, mHealth, prontuários eletrônicos, telessaúde e sistemas de apoio à decisão clínica) são, muitas vezes, desenvolvidas ou utilizadas sem os cuidados necessários com a cibersegurança [Aljedaani and Babar 2021], permitindo assim que ciberataques possam comprometer sistemas críticos e informações sensíveis.

A cibersegurança consolidou-se como um dos principais desafios tecnológicos da atualidade para diversos setores. Entre os obstáculos mais relevantes destacam-se: *(i)* a carência de educação e treinamento adequados, o que torna o fator humano um dos principais vetores de ataque; *(ii)* a escassez de investimentos e a ausência de estratégias e planejamentos eficazes; *(iii)* a dificuldade de quantificar os riscos e impactos decorrentes de ciberataques; e *(iv)* a falta de conscientização sobre a importância da cibersegurança por parte de organizações, governos e sociedade. Diante desse cenário, este capítulo abordará os principais riscos e impactos de ciberataques, apresentando também práticas para o planejamento de estratégias de cibersegurança. O capítulo terá como foco o setor da saúde, que é historicamente vulnerável a incidentes de segurança cibernética, e cujas consequências vão além da perspectiva técnica, gerando impactos econômicos, jurídicos e sociais significativos.

A organização deste capítulo está apresentada na Figura 1.1, auxiliando na compreensão dos principais tópicos que serão abordados nas diferentes seções deste capítulo. Na Seção 1.2 são apresentados os conceitos básicos sobre cibersegurança relacionados à confidencialidade, integridade e disponibilidade. Também, é apresentada nesta seção uma breve descrição e modus operandi dos principais ciberataques (por exemplo, ransomware, negação de serviço e phishing). Na Seção 1.3 é conduzido um estudo do estado da arte, incluindo soluções disponíveis na indústria e academia, regulamentações e documentos

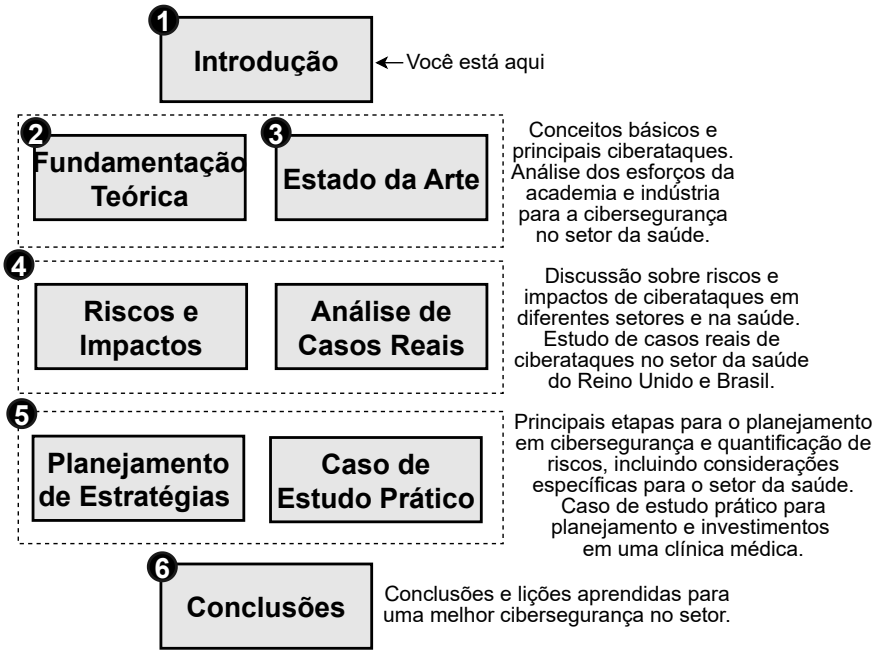



**Figura 1.1. Organização do Capítulo**

de boas práticas. A seção é concluída com uma discussão sobre tendências, desafios e oportunidades na área de cibersegurança na saúde.

Já na Seção 1.4 são apresentados e discutidos os riscos e impactos de ciberataques em diferentes setores, tendo como foco apresentar as nuances do setor da saúde. Também, nesta seção, são apresentados dois estudos de caso de ciberataques passados no mundo real: ataque de ransomware no Reino Unido e vazamento de dados no sistema de saúde do Brasil. A Seção 1.5 introduz as principais etapas para o planejamento em cibersegurança e também para a quantificação de riscos, discutindo as tarefas críticas para o setor da saúde. Nesta seção também é conduzido um caso de estudo utilizando uma plataforma educacional para planejamento e simulação de riscos em cibersegurança. Por fim, na Seção 1.6 são apresentadas as conclusões e lições aprendidas para uma cibersegurança mais robusta no setor da saúde em curto, médio e longo prazo.

## 1.2. Fundamentação Teórica

Para melhor compreender o impacto que ataques cibernéticos têm em organizações de saúde, é necessário primeiro abordar alguns conceitos básicos de cibersegurança. Essa seção explora as propriedades de segurança desejáveis em sistemas de informação, bem como a maneira com que atacantes (ou seja, adversários) podem tentar comprometer essas propriedades para obter acesso indevido a dados. Por fim, são apresentados exemplos de cada tipo de ataque dentro do setor da saúde e suas consequências técnicas e financeiras para indivíduos e organizações.

Um dos principais objetivos de ferramentas de segurança é fornecer as proprie-

dades de Confidencialidade, Integridade e Disponibilidade (Confidentiality, Integrity, and Availability, CIA), que são pilares da cibersegurança. Garantir a confidencialidade da informação é garantir que ela não está disponível e nem pode ser descoberta por indivíduos, entidades ou processos que não têm autorização para acessá-la [Beckers et al. 2015]. Manter a integridade dos dados significa ter a garantia de que os dados estão corretos e completos durante todo o seu ciclo de vida, assim assegurando que os dados não foram modificados de forma não autorizada [Boritz 2005]. Por fim, a disponibilidade é a propriedade que garante que a informação pode ser acessada no momento em que ela é necessária, ou seja, de que os sistemas usados para seu armazenamento e processamento estão funcionando corretamente, assim como as medidas de segurança usadas para protegê-la e os canais de comunicação necessários para acessá-la.

Em sua maioria, os mecanismos de segurança usados em sistemas computacionais buscam garantir as propriedades da tríade CIA em alguma medida. Para garantir a confidencialidade de uma informação, por exemplo, é necessário garantir que um usuário possua a autenticação necessária e esteja autorizado a acessá-la. Mecanismos de autenticação servem também para determinar se uma tentativa de acesso é legítima, ou seja, se o usuário ou dispositivo é quem diz ser. Já um mecanismo de autorização serve para garantir que o usuário tem as permissões necessárias para realizar o acesso em questão. Um exemplo de mecanismo de segurança que garante a integridade dos dados é o cálculo de somas de verificação (*checksum*) de arquivos. A disponibilidade, por sua vez, é o principal objetivo de diversos mecanismos de segurança que buscam impedir a interrupção de serviços, como mecanismos de redundância, *backups* e planos de recuperação.

Durante ataques cibernéticos, adversários tentam contornar a estrutura de cibersegurança que garante a confidencialidade, integridade e a disponibilidade de informações e serviços. Para isso, atacantes contam com oportunidades como, por exemplo, vulnerabilidades em sistemas de informação e erro humano, ou usam força bruta e grande quantidade de recursos (por exemplo, processamento e rede) para sobrecarregar os sistemas. No restante dessa seção, são introduzidos os principais ataques que afetam o setor da saúde e também discutido como tais ataques conseguem contornar estruturas de segurança de uma organização para obter acesso indevido ou impedir o funcionamento de sistemas.

### 1.2.1. Ransomware

Ransomware, palavra em inglês derivada de sequestro (*ransom*) e *software*, é um tipo específico de malware que, ao ter acesso a um sistema, criptografa os dados existentes (por exemplo, dados pessoais ou sistemas da vítima) e impede seu acesso até que um valor de resgate seja pago [Young and Yung 1996]. Se tratando de organizações, bancos de dados inteiros podem ficar inacessíveis. Na maioria das ocasiões, o atacante também ameaça divulgar publicamente os dados. Caso a organização não tenha as medidas de cibersegurança necessárias para se proteger desse tipo de ataque, as consequências são a disrupção de serviços, perda de reputação e perda financeira. Tais impactos ocorrem devido ao tempo de interrupção do serviço, à perda permanente dos dados caso o resgate não seja pago, e a potenciais multas relacionadas com infrações de *compliance* caso esses dados sejam vazados pelos criminosos. Os responsáveis pelo ataque geralmente pedem o pagamento do resgate através de moedas digitais, como Bitcoin e Monero, de modo que o valor pago se torna indetectável e, na maioria das vezes, não possa ser rastreado. Um

ataque de ransomware geralmente tem origem quando um usuário do sistema recebe e executa um arquivo malicioso que aparenta ser legítimo, por exemplo, no anexo de um e-mail de phishing. Depois de executado, esse tipo de malware se espalha rapidamente dentro do sistema, infectando outras máquinas conectadas à mesma rede e criptografando todos os dados os quais consegue acesso.

Não é recomendado que organizações façam o pagamento dos resgates na tentativa de retomar o acesso aos seus dados. Como em qualquer tipo de sequestro, o pagamento assume a boa fé do atacante que pode optar por não entregar a chave criptográfica que precisa ser usada pra decodificar os arquivos. Com o pagamento, a empresa também corre o risco de que o ataque se repita, e de que os atacantes tentem replicar o ataque bem-sucedido em outras empresas [Healthcare Information and Management Systems Society 2024]. Porém, não é raro que os responsáveis optem pelo pagamento na tentativa de diminuir o impacto do ataque. De acordo com relatório da Claroty, em 2024, 78% das empresas do setor de saúde entrevistadas relataram pagamentos. Em 39% delas, o valor dos resgates alcançou entre US$ 1 milhão e US$ 5 milhões. Cerca de um quarto dos participantes reportou perdas que chegaram a US$ 1 milhão ou mais, entre perda de receita, gastos com recuperação de sistemas e gastos legais [Claroty 2025].

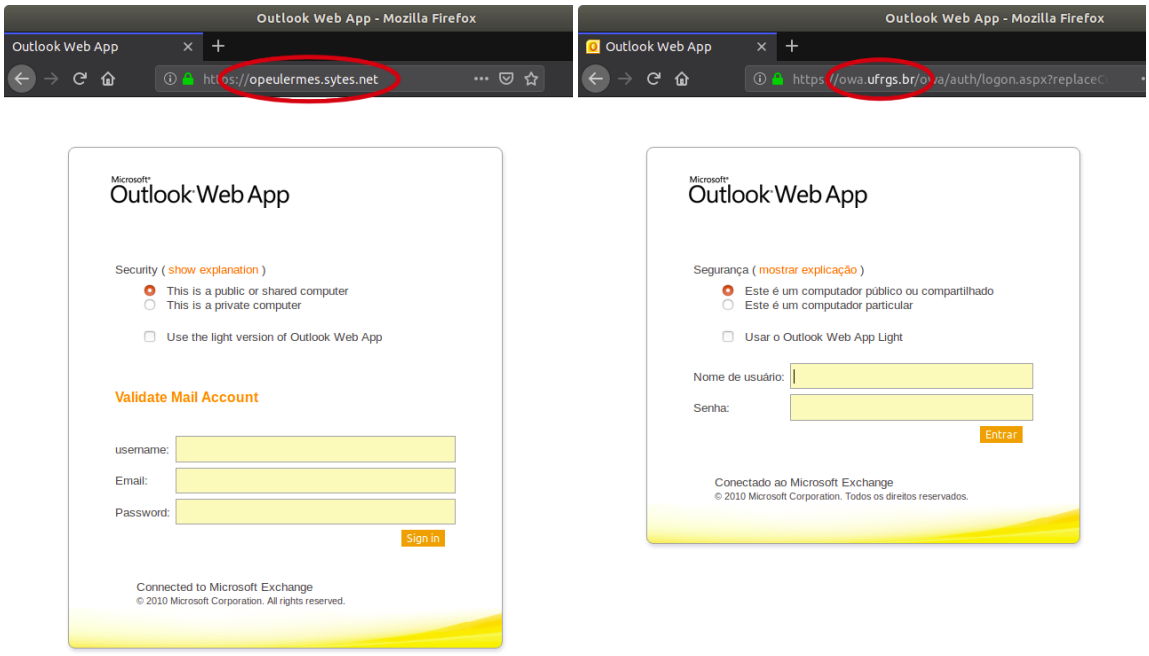
Um dos ataques de ransomware mais proeminentes no setor da saúde ocorreu em 2024 e afetou a empresa estadunidense Change Healthcare. A Change Healthcare é uma subsidiária do grupo UnitedHealth, uma das maiores empresas de gerenciamento de receita, processamento de pagamentos e compartilhamento de informações de saúde do país. A empresa admitiu ter pago US$ 22 milhões em Bitcoins na tentativa de fazer o resgate dos dados [Claroty 2025], que não foram devolvidos pelos atacantes. Os dados expostos incluem informações de contato, detalhes de apólices de seguros de saúde, informações médicas e dados financeiros de pacientes. A empresa ficou fora de operação pelo período do ataque, o que ocasionou atrasos no pagamento de serviços prestados por profissionais da saúde. O gasto total da empresa com o ataque, somando implicações legais, técnicas e o pagamento do resgate, chegou a exorbitantes US$ 3,1 bilhões [Olsen 2025].

### 1.2.2. Phishing

Ataques de phishing são uma das maiores causas de vazamentos de dados no setor da saúde [Alder 2024], e um dos ataques de engenharia social mais efetivos contra organizações no setor [United States Department of Health and Human Services 2024a]. Engenharia social é o nome dado à técnica de manipular indivíduos a divulgar informações de acesso restritas ou a qualquer outra ação prejudicial a dispositivos e sistemas aos quais se tem acesso. Existem inúmeras estratégias de engenharia social usadas em golpes de phishing, e seus alvos são amplos e variados. De funcionários a clientes e fornecedores, qualquer informação divulgada por uma vítima de phishing tem potencial de causar danos a sistemas e expor dados sigilosos. Devido a isso, o phishing muitas vezes é o vetor de outros ataques [Adebukola et al. 2022]. Por exemplo, um usuário pode ser enganado a clicar em um link que realiza o download de um malware ou credenciais obtidas através de phishing podem ser usadas para obter acesso não autorizado a dados e sistemas. Também por esse motivo, é difícil precisar a real escala do impacto de ataques de phishing no setor da saúde, já que eventos de cibersegurança de impacto considerável podem ter tido origem em uma interação que começou com um simples e-mail de phishing.

As técnicas de engenharia social usadas em ataques de phishing são variadas e estão em constante evolução, e seus alvos em potencial são amplos e diversificados. Um e-mail de phishing pode ser encaminhado, por exemplo, para todos os funcionários de uma organização. Na Figura 1.2, temos o exemplo de uma campanha de phishing contra funcionários e alunos de uma universidade (ou seja, UFRGS). O ataque começa com um e-mail aparentemente legítimo demandando urgência, por exemplo, alguma situação precisa ser regularizada ou o usuário corre o risco de perder e-mails importantes ou ter sua conta suspensa. O corpo do e-mail de phishing geralmente apresenta um link malicioso que o alvo deve clicar para ser redirecionado e regularizar a situação. No caso da Figura 1.2, o cliente de e-mail Outlook é um dos recomendados para o acesso ao e-mail institucional. Ao clicar no link, o usuário é redirecionado a uma página falsa (ver Figura 1.2a). Caso o usuário não perceba o erro e informe suas credenciais de acesso, o atacante vai obter um *login* e senha que poderão ser usados tanto para prejudicar o indivíduo quanto para acessar áreas restritas do sistema da universidade. Dependendo do nível de permissão do usuário, inúmeros outros ciberataques podem ser executados uma vez obtido acesso.




**(a) Ataque de phishing imitando o cliente de e-mails Outlook Web, hospedado em um domínio malicioso.**

**(b) A pagina verdadeira do cliente de e-mails Outlook Web, com o domínio legítimo da UFRGS.**

**Figura 1.2. Exemplo de um ataque de phishing direcionado ao e-mail institucional da UFRGS, para os usuários do cliente Outlook Web. Acervo pessoal.**

Campanhas de phishing genéricas, como o exemplo da Figura 1.2, muitas vezes contêm erros de escrita e não são muito convincentes. Um tipo diferente de ataque de phishing é o phishing direcionado (ou *spear* phishing, em inglês). Esses ataques têm como alvo indivíduos específicos sobre os quais os atacantes têm alguma informação relevante, o que adiciona veracidade às informações falsas apresentadas no ataque e aumenta as chances de atrair as vítimas. Por ser menos genérico e elaborado de forma direcionada,

esse tipo de ataque tem uma eficácia muito mais alta e pode ser a primeira parte de um ataque mais complexo. Segundo um relatório sobre cibersegurança no setor da saúde [Healthcare Information and Management Systems Society 2024], o phishing permanece sendo o maior vetor de ataque e o principal meio pelo qual sistemas são comprometidos. Uma tendência preocupante que vem sendo observada por empresas no setor é o uso de *deepfakes*, uma técnica de Inteligência Artificial usada para gerar imagens falsas realistas, para aumentar a eficácia de golpes de phishing.

A principal forma de combater o phishing é através da conscientização e treinamento. Para isso, é necessária uma estratégia de treinamento implementada de forma contínua aplicada a funcionários, clientes, e quem mais tiver acesso a sistemas restritos de uma organização. Essa estratégia pode incluir cursos tanto online quanto palestras sobre tópicos de cibersegurança, boas práticas de cibersegurança e, principalmente, apresentar métodos de engenharia social usados em golpes de phishing e suas tendências. Devem ser abordados tópicos como, por exemplo, a falta de segurança em redes públicas, o uso de gerenciadores de senha, a necessidade de trocas de senha regulares, entre outros. Outra estratégia interessante é a simulação de cenários de phishing pelo próprio setor de Tecnologia da Informação (TI) da organização [Cartwright 2023]. Dessa forma, é possível manter o estado de atenção entre os usuários e também identificar os indivíduos que precisam de reforço em seus conhecimentos de cibersegurança. Esses métodos de combate ao phishing devem estar inseridos na estratégia de cibersegurança de empresas, o que demanda investimentos e a atenção de especialistas.

### 1.2.3. Negação de Serviço

Os ataques de negação de serviço (Denial of Service, DoS) são ataques onde um serviço é sobrecarregado com uma quantidade de tráfego ou requisições tão alta que resulta em um estado no qual os usuários legítimos não conseguem acessar o serviço [De Neira et al. 2023]. Assim, os ataques DoS não necessariamente exploram vulnerabilidades específicas na infraestrutura ou sistema. Em sua versão mais poderosa e distribuída (Distributed Denial of Service, DDoS), um grande número de hosts é utilizado para gerar o volume de tráfego ou requisições necessário ao ataque. Imagine uma estação de trem onde milhares de pessoas estão paradas em frente à porta do trem, fingindo estarem aguardando para entrar. Porém, na verdade, tais pessoas estão bloqueando que usuários legítimos (ou seja, quem realmente quer entrar no trem) acessem o serviço. Essa analogia nos permite compreender como funciona um ataque de tal magnitude.

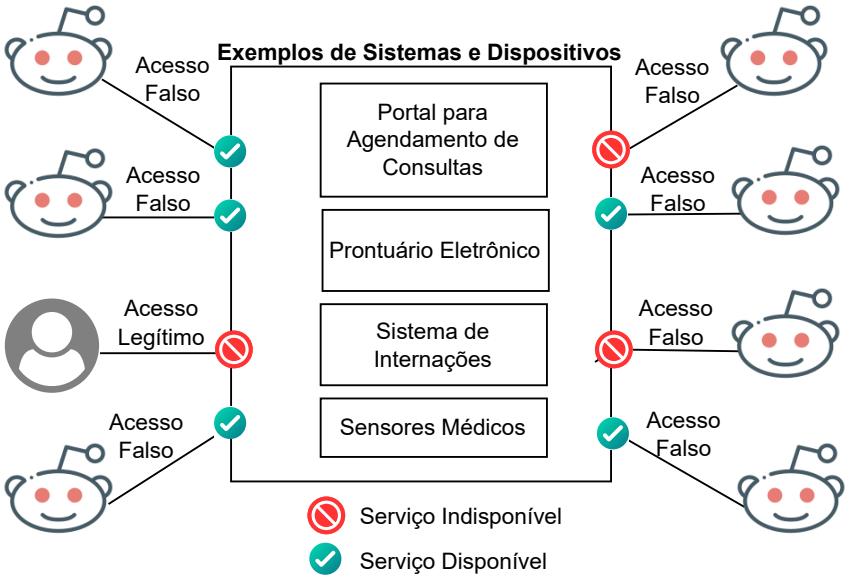



**Figura 1.3. Exemplo de um DDoS, com um acesso legítimo a serviços sendo negado devido a uma grande quantidade de acessos falsos.**

Para obter acesso a um grande número de dispositivos, outros tipos de ataque são empregados, como a distribuição de malware por e-mail (ver Seção 1.2.2). Decisões de projeto e processos de administração de sistemas que fornecem segurança fraca tornam ainda mais fácil comprometer uma grande quantidade de dispositivos [Navruzov and Kabulov 2022]. Tendências como a Internet das Coisas (*Internet of Things*, IoT) levaram a um aumento no número de dispositivos conectados à Internet. Muitas vezes, esses dispositivos não possuem mecanismos de segurança implementados, de modo que a simples verificação manual de serviços expostos publicamente é suficiente para obter acesso a eles. A existência de bilhões desses dispositivos com segurança fraca tem permitido que invasores criem ataques cada vez mais poderosos a cada ano. Por exemplo, o malware Mirai permitiu controlar cerca de 600 mil equipamentos infectados, criando assim uma rede de computadores zumbis conhecida como botnet [Gallopeni et al. 2020]. Dentre os dispositivos infectados estão câmeras de segurança, roteadores, geladeiras e até mesmo dispositivos médicos.

No setor da saúde, os ataques de DDoS podem impactar serviços com conectividade na Internet, causando a indisponibilidade para usuários legítimos. A Figura 1.3 apresenta uma representação de um ataque coordenado com zumbis enviando requisições falsas e impedindo que um usuário legítimo possa acessar os recursos (ou seja, os sistemas e dispositivos).

Por exemplo, sistemas de agendamento de consultas podem receber diversas requisições até que sejam sobrecarregados, de modo que fiquem muito lentos e os pacientes desistam de agendar uma consulta. Em outro exemplo, ataques a sistemas de prontuários eletrônicos podem sobrecarregar os servidores responsáveis pelo armazenamento e acesso às informações clínicas, tornando-os inacessíveis. Isso impede médicos e profissionais de

saúde de acessarem dados essenciais, como históricos de pacientes e resultados de exames, comprometendo a qualidade do atendimento e colocando vidas em risco. Portanto, tais ataques não só interrompem a operação dos serviços médicos, como também geram confusão e aumentam a insatisfação dos pacientes, com impacto direto na eficiência e na reputação das instituições de saúde.

## 1.3. Estado da Arte da Cibersegurança no Setor da Saúde

É inegável que a conectividade fornecida pela Internet traz inúmeros benefícios para o setor da saúde, como a interação em tempo real entre sistemas e o aumento na disponibilidade de serviços. A adoção de dispositivos de IoT para tarefas de diagnóstico e monitoramento de pacientes traz maior comodidade para equipes médicas e administrativas em unidades de saúde [He and Zeadally 2014]. Tecnologias como Inteligência Artificial (IA) têm grande potencial em diversas áreas relacionadas à saúde, podendo atuar, por exemplo, como ferramentas de diagnóstico [Chamberlaim et al. 2023], auxiliar no processo de tomada de decisões estratégicas hospitalares, e na cibersegurança [Nankya et al. 2024]. As mudanças nos padrões de trabalho possibilitadas pela Internet permitem que setores das organizações trabalhem, inclusive, de forma remota. Contudo, expor sistemas e serviços à Internet também pode levar a ataques cibernéticos, o que, por sua vez, pode causar disrupção de serviços, vazamento de Informações de Saúde Protegidas (*Protected Health Information*, PHI), perdas financeiras e ameaça até mesmo à saúde de pacientes.

Inúmeras estratégias para proteção cibernética de organizações de saúde são propostas todos os anos na indústria e na academia. Sistemas usando dispositivos IoT em contextos médicos (*Internet of Medical Things*, IoMT, ou Internet das Coisas Médicas) têm se mostrado particularmente vulneráveis, por se tratar de uma quantidade substancial de equipamentos continuamente gerando grandes volumes de dados, vindos de fabricantes diferentes e com pouca padronização em seus protocolos e atualizações de segurança [Soares et al. 2023]. Essas características levam ao aumento da superfície de ataque de organizações médicas e representam um desafio extra para as soluções de segurança no setor. Essas soluções precisam ser capazes de lidar com acessos não autorizados, perda, extravio e/ou descarte inapropriado de dispositivos IoMT [Cartwright 2023], além dos principais ciberataques discutidos na Seção 1.2.

Uma linha de pesquisa em alta na área de cibersegurança para o setor da saúde é o uso de soluções baseadas em blockchain. Existe um interesse crescente na literatura em usar tecnologias distribuídas de registro para resolver problemas tradicionais do setor, como falta de interoperabilidade, dificuldades de auditoria e vazamentos de dados [Santos et al. 2021]. Em particular, sistemas de saúde podem se beneficiar da descentralização, transparência e imutabilidade inerentes a sistemas de blockchain [Scheid et al. 2021b]. Esses sistemas podem facilitar o acesso e compartilhamento de registros médicos, além de contribuir para os esforços de padronização entre diferentes instituições [Abou Jaoude and Saade 2019]. Contudo, seu emprego também pode trazer desvantagens e novas complicações para sistemas de saúde. Além das preocupações com custo computacional e armazenamento, existe a preocupação com *compliance* e regulamentação, já que todos os dados de uma transação escritos em blockchain são permanentes e não podem ser apagados. Portanto, soluções empregando sistemas de blockchain para a área da saúde precisam empregar outras técnicas em conjunto para o armazenamento de dados pessoais e PHI.

Também em alta no mercado e na literatura estão soluções de segurança usando técnicas de Aprendizado de Máquina (*Machine Learning*, ML) e IA. Dentro do escopo de segurança cibernética, essas soluções contribuem principalmente para a detecção de ataques através do monitoramento de comportamento anômalo, vulnerabilidades de dia-zero e respostas automatizadas [Nankya et al. 2024]. Desafios para essa linha de pesquisa incluem principalmente o tratamento dos dados usados para treinamento, que podem incluir informações pessoais de pacientes. Técnicas de Aprendizado Federado (*Federated Learning*), em particular, permitem que vários clientes façam o treinamento de um modelo enquanto mantêm os dados usados de forma descentralizada, ou seja, sem compartilhar os dados entre si [Wen et al. 2023]. Também vale destacar o aumento do uso de ferramentas de IA para realizar ataques ou mesmo adicionar novas vulnerabilidades no setor. O uso de *deepfakes* (vídeos ou imagens falsas geradas artificialmente usando *deep learning*) tem o potencial de aumentar a eficácia de golpes de *phishing*, uma das principais ameaças do setor [Healthcare Information and Management Systems Society 2024]. A falta de regulamentação do uso de ferramentas de IA generativa por funcionários também põe organizações de saúde em risco, já que as principais ferramentas para esse fim (por exemplo, ChatGPT, Gemini e DeepSeek) são proprietárias e os dados são compartilhados com organizações externas.

O vazamento e/ou compartilhamento não autorizado de PHIs e dados pessoais de pacientes está entre os principais riscos do setor da saúde. De acordo com o relatório publicado pela IBM em 2024, o setor tem o maior custo associado a vazamento de dados: em média US$ 9,77 milhões por ataque [IBM Security 2024]. A maioria dos países possui ferramentas de regulação e *compliance* que lidam com empresas responsáveis por vazar dados pessoais de seus usuários. Na Europa, a *General Data Protection Regulation* (GDPR) se aplica a todos os indivíduos e organizações que lidam com dados pessoais de cidadãos europeus [GDPR.EU Horizon 2020 2021]. Em 2024, a GDPR aplicou 202 multas a hospitais, farmacêuticas, profissionais da saúde e fornecedores de equipamento médico em 26 países, totalizando 16,5 milhões de euros [Runte 2024]. Nos Estados Unidos, a *Health Insurance Portability and Accountability Act* (HIPAA) traz as diretrizes que organizações e provedores de seguros de saúde devem seguir para garantir que PHIs e informações pessoais identificáveis sejam protegidas de fraude ou roubo [U.S. Government 1996]. Em janeiro de 2025, um fornecedor de monitores de glicose, bombas de insulina e outros equipamentos para pessoas com diabetes foi multado em US$ 3 milhões [United States Department of Health and Human Services 2024b] depois de um vazamento que teve origem em um ataque de *phishing* direcionado. No Brasil, a Lei Geral de Proteção de Dados (LGPD) e a Autoridade Nacional de Proteção de Dados (ANPD) controlam a privacidade, uso e tratamento de dados pessoais. Em contramão ao restante das entidades regulatórias no mundo, a ANPD não aplicou multas em função de vazamento de dados em 2024 [Calegari 2025].

Diante de um cenário tão complexo quanto a computação no setor da saúde, as ferramentas de segurança precisam evoluir rapidamente para acompanhar inovações aceleradas, regulamentações complexas e, principalmente, demandas de sistemas críticos. Para isso, é necessário forte investimento em cibersegurança, já que a falta de orçamento é citada como um dos maiores desafios para pequenas empresas no setor da saúde nos Estados Unidos [HIMSS, FinThrive 2025]. Uma vez assegurado o investimento, é ne-

cessário que executivos, gestores e técnicos priorizem quais ameaças são mais urgentes em sua organização ou setor. Dispositivos IoMT têm grande potencial de serem explorados por adversários. Segundo relatório da Claroty, IoMT com vulnerabilidades críticas de segurança estão presentes em 89% das organizações de saúde pesquisadas [Claroty 2025]. Outra necessidade é a conscientização constante de funcionários sobre cibersegurança, através da aplicação de cursos, treinamento e até mesmo incidentes simulados. Por exemplo, em 2020, um funcionário do Hospital Albert Einstein publicou no GitHub de forma acidental uma planilha com senhas de funcionários do Ministério da Saúde (ver Seção 1.4.3), levando à exposição de dados de pelo menos 16 milhões de pacientes de Covid-19 [Cambricoli 2020]. Levando em consideração esses e outros fatores, essa seção apresenta um panorama da cibersegurança no setor da saúde, considerando a extensão dos desafios do setor.

### 1.3.1. Soluções

Alguns tópicos se destacam no estado da arte com aplicações de cibersegurança para a área da saúde. Dentre eles, soluções de controle de acesso e autenticação merecem atenção por garantir que dados sigilosos de saúde serão acessados apenas por pessoas autorizadas, sejam elas pacientes, profissionais de saúde ou de gestão. Em particular, sistemas utilizando dispositivos IoT em grande escala precisam de cuidados redobrados com seus mecanismos de segurança e privacidade. Também em alta estão soluções de segurança baseadas em blockchain, por fornecerem propriedades de segurança como autenticação e integridade de dados. Outra tecnologia emergente no estado da arte são algoritmos de ML atuando na detecção de ataques e respostas automatizadas a incidentes.

#### 1.3.1.1. Segurança de Dispositivos de IoT em Saúde

IoT é uma área ampla que compreende diversas indústrias, como cidades inteligentes, monitoramento industrial, agricultura, entre várias outras. Dentro do escopo de saúde, dispositivos de IoMT têm potencial para trazer melhor qualidade de vida para pacientes com necessidade de monitoramento contínuo e maior comodidade para equipes médicas e de gerenciamento [He and Zeadally 2014]. Contudo, o emprego desses dispositivos aumenta consideravelmente os desafios de segurança enfrentados por organizações médicas e unidades de saúde. Algumas consequências do uso inadequado de dispositivos IoMT são, por exemplo, o vazamento de informações pessoais de pacientes e o atraso na detecção de eventos de saúde importantes devido a interrupções de serviço, entre outros [Sun et al. 2019]. Técnicas de segurança padrão em outros sistemas computacionais nem sempre podem ser aplicadas em sistemas IoMT, devido a restrições na capacidade computacional desses dispositivos, sendo que dispositivos sensores e vestíveis costumam ter ainda menos recursos.

Soluções de segurança visando sistemas de IoMT necessitam de técnicas de autenticação confiáveis, assim como esquemas de verificação e validação para manter a confiabilidade dos participantes, sejam eles pacientes, profissionais de saúde ou organizações médicas [Adil et al. 2024]. A autenticação é o processo que visa confirmar a identidade de um usuário e garantir que ele tem as permissões necessárias para acessar aquelas informações ou realizar determinada tarefa. No caso de dispositivos IoMT, por exemplo,

é necessário garantir que os dados de medição de sensores sejam acessados apenas pela equipe médica responsável e não por pessoas em outros cargos dentro da rede hospitalar. Um esquema de autenticação robusto garante vários requisitos de segurança desejáveis, como, por exemplo, controle de acesso, disponibilidade das informações e integridade de dados. Também existe a preocupação com a transmissão de dados médicos entre os dispositivos IoMT e seus respectivos *gateways* de acesso, já que esses dispositivos muitas vezes não têm poder computacional o suficiente para empregar técnicas de criptografia robustas que garantam que um adversário observando o tráfego de rede não será capaz de espionar os dados em trânsito [Sun et al. 2019].

[Gupta et al. 2019] propõe um mecanismo de autenticação para dispositivos IoT utilizando Disjunção Exclusiva (XOR) e funções *hash* criptográficas de mão única, com o objetivo de proteger a comunicação entre dispositivos de forma eficiente e com baixo custo computacional. Também lidando com o processo de autenticação, [Ostad-Sharif et al. 2019] propõe um sistema híbrido onde um algoritmo de criptografia mais leve é utilizado na primeira parte do processo, quando os participantes estão trocando as chaves criptográficas que serão usadas na comunicação, e, posteriormente, um algoritmo de Criptografia de Curva Elíptica (ECC) é usado para criptografar as mensagens contendo os dados. Já [Ding et al. 2019], por sua vez, propõe distribuir o custo computacional de algoritmos mais custosos usando dispositivos de borda (conhecidos em inglês como *Edge*), que têm capacidade computacional intermediária entre sensores IoMT e computadores tradicionais. Com isso, eles desenvolvem algoritmos de verificação de integridade para os dados armazenados, além dos mecanismos de controle de acesso, autenticação e privacidade.

#### 1.3.1.2. Soluções Baseadas em Blockchain

Blockchain é uma tecnologia de registro distribuído que surgiu inicialmente para uso em sistemas de criptoativos [Scheid et al. 2021b]. Suas características incluem fornecer um ambiente descentralizado onde transações podem ser feitas sem a necessidade de supervisão por um terceiro. Em sistemas financeiros tradicionais, a supervisão é necessária para garantir não apenas a identidade dos participantes, mas também a ordem das transações realizadas, assim assegurando que, caso uma operação seja repetida, o mesmo valor não será debitado duas vezes de um dos participantes. Na blockchain, essa funcionalidade é replicada por meio de registros que são encadeados e mantidos por uma rede de nós que compartilham tarefas e arquivos entre si. Após sua proposta inicial em criptoativos, sistemas de blockchain sofreram um processo de generalização e passaram a ser aplicados nas mais diversas áreas [Abou Jaoude and Saade 2019]. Alguns dos benefícios em potencial do uso de blockchain são particularmente relevantes em sistemas de saúde [Scheid et al. 2021a]. A estrutura de blocos encadeados garante intrinsecamente a imutabilidade de registros, o que contribui para a integridade, confiabilidade e garantia de acesso aos dados. Além disso, a necessidade de autenticação dos participantes e a transparência das operações realizadas dentro da rede também são características de interesse dentro da área da saúde [Santos et al. 2021].

Com base em suas propriedades de cibersegurança em potencial, sistemas para gerenciamento de dados de saúde baseados em blockchain são amplamente estudados na

literatura acadêmica. A maioria das soluções foca nos desafios relacionados ao armazenamento e compartilhamento seguro de dados médicos [Arbabi et al. 2023]. Isso acontece pois não é possível remover ou apagar quaisquer informações uma vez armazenadas na blockchain. Ao mesmo tempo que essa característica tem potencial para contribuir com a integridade de registros, ela também representa um desafio para a implementação de sistemas que manipulam PHI. Essas informações estão sujeitas a leis de regulamentação que exigem, por exemplo, que seja possível remover informações pessoais e de saúde a pedido do usuário. Portanto, soluções usando blockchain precisam garantir que ou as informações armazenadas estão criptografadas de tal forma que não são recuperáveis sem a chave secreta, ou usam uma solução híbrida que armazena na blockchain apenas metadados e similares, enquanto os dados de saúde são armazenados em outro lugar.

MedShare [Wang et al. 2021], por exemplo, propõe um esquema usando criptografia baseada em atributos (Attribute-Based Encryption, ABE) para assegurar que os dados de saúde armazenados na blockchain são acessíveis apenas por partes autorizadas. De forma similar, [Zhang et al. 2022] também utiliza uma variante de ABE para permitir que seja feita a busca por palavras-chave nos dados criptografados, enquanto a blockchain é usada para garantir a imutabilidade das chaves e dos registros criptografados. Utilizando uma técnica de armazenamento híbrido, o EdgeMediChain [Akkaoui et al. 2020] usa uma rede de blockchain em dois níveis que salva os registros médicos criptografados separadamente, fora da blockchain. Uma blockchain privada intermediária gerencia a autenticação e os dados gerados por dispositivos geograficamente próximos. Depois dessa análise inicial, os dados relevantes são armazenados separadamente e seu endereço é colocado em uma blockchain pública global, garantindo o acesso a outros participantes do sistema. Isso permite que o sistema de blockchain em camadas funcione como um índice, oferecendo integridade de registros e gerenciabilidade de direitos de acesso.

#### 1.3.1.3. Usos de ML e IA

Talvez um dos tópicos mais discutidos na atualidade, soluções de segurança usando IA e ML precisam de cuidado especial ao serem empregadas em sistemas de saúde. Além de atuar em funções diagnósticas e de monitoramento de saúde de pacientes, essas soluções também são usadas para detecção de ameaças e anomalias no tráfego de rede, o que pode indicar um incidente de segurança em andamento. As técnicas usadas nessa área podem ser divididas em modelos de aprendizado supervisionado e não-supervisionado. Modelos supervisionados necessitam de uma base de dados para treinamento, onde as vulnerabilidades e ameaças já conhecidas historicamente estejam devidamente identificadas e sinalizadas. Depois do treinamento, os modelos são capazes de reconhecer esses padrões com precisão em dados inéditos, como, por exemplo, durante o monitoramento da rede em tempo real. Já nos modelos de aprendizado não-supervisionado, não é feita essa sinalização prévia das vulnerabilidades conhecidas, e fica a cargo do modelo identificar sozinho as anomalias que se desviam do padrão de tráfego normal. Dessa forma, os modelos não-supervisionados são usados para identificar ameaças novas e desconhecidas, conhecidas como ameaças de dia-zero [Nankya et al. 2024].

Além da contribuição para a segurança de sistemas e redes de comunicação em organizações de saúde, técnicas de IA também têm aplicação na hora de proteger a pri-

vacidade de pacientes. No caso de ferramentas de IA que são usadas para diagnóstico, o treinamento dos modelos demanda grandes quantidades de informações médicas e de saúde de pacientes para que possam ser reconhecidos os padrões de doenças e anomalias. Nesses casos, é vital que os dados médicos usados no treinamento sejam anonimizados de alguma forma antes de passados para o modelo, visando manter a conformidade com as leis de proteção de dados vigentes. Para isso, técnicas como Encriptação Homomórfica e Privacidade Diferencial auxiliam na hora de garantir que os dados individuais de um paciente não possam ser identificados dentro de um conjunto maior de dados semelhantes. Dentro desse tópico, estão em destaque técnicas de IA como Aprendizado Federado. Ele permite que várias entidades façam o treinamento de modelos de IA localmente usando apenas os dados aos quais têm acesso, e depois compartilhem entre si apenas as atualizações do modelo [Aouedi et al. 2023]. Em outras palavras, elas compartilham "apenas o que o modelo aprendeu", e não os dados brutos de saúde de pacientes.

É importante ressaltar que ferramentas de IA também têm surgido como vetores e amplificadores de ameaças cibernéticas, chamando atenção negativamente na literatura e na mídia. Ataques de *phishing*, conhecidamente uma das maiores ameaças ao setor da saúde, dependem de técnicas de engenharia social para enganar funcionários com acesso a sistemas restritos com o objetivo de roubar credenciais válidas. *Deepfakes* e ferramentas de IA generativa têm contribuído para aumentar a eficácia desses ataques [Healthcare Information and Management Systems Society 2024], o que exige esforço e investimento cada vez maior na conscientização e educação cibernética de funcionários e colaboradores. Alguns exemplos das consequências desse tipo de ataque são o acesso não autorizado de adversários aos sistemas de organizações de saúde, podendo levar a vazamentos de dados pessoais e danos à estrutura de segurança do sistema. Ferramentas de *deep learning* também podem ser usadas para adulterar o resultado de diagnósticos por imagem, com o objetivos variando entre cometer sabotagem direcionada a indivíduos, fraude de seguros de saúde, ou até mesmo atentados [Mirsky et al. 2019].

### 1.3.2. Regulamentação, Compliance e Boas Práticas

Organizações em todo o mundo são obrigadas a obedecer à legislação vigente em suas áreas de atuação no que diz respeito ao tratamento de dados pessoais e identificáveis de seus usuários. No setor da saúde, a regulamentação e *compliance* são especialmente importantes por se tratar de uma área que manipula quase que integralmente dados sensíveis e de saúde de pacientes e usuários de serviços médicos. Diferentemente de outras categorias de dados pessoais, uma exposição de dados médicos pode revelar condições de saúde de um indivíduo [Sun et al. 2019]. Alguns dos principais regulamentos no tópico são o GDPR, da União Europeia, e o HIPAA atuante nos Estados Unidos. No Brasil, embora a LGPD tenha passado a valer em agosto de 2020, a autoridade responsável por vistoriar e aplicar sanções a organizações em descumprimento da lei ainda encontra dificuldades em sua atuação [O Globo 2024].

Se tratando da confidencialidade de informações médicas, a GDPR estipula que as informações médicas armazenadas devem ser apagadas depois de processadas e não mais necessárias, e que organizações devem obter o consentimento explícito dos pacientes para compartilhar seus dados com terceiros. A LGPD, fortemente baseada na GDPR, também estipula que os dados pessoais armazenados devem ser eliminados após o término

de seu uso, salvo circunstâncias específicas. Ela também exige que o usuário autorize o compartilhamento de dados com terceiros. Já a HIPAA não tem nenhuma restrição quanto a um período máximo de armazenamento dos dados, nem quanto à possibilidade de compartilhamento de dados entre diferentes provedores de saúde [Sun et al. 2019].

Outra característica relevante no âmbito do armazenamento de dados pessoais e de saúde é o direito do paciente de requisitar que seus dados armazenados junto a organizações de saúde sejam apagados, a qualquer momento. O Art. 17 do GDPR garante esse direito, conhecido como "direito ao esquecimento". Na LGPD, o mesmo direito é conhecido como "direito à eliminação de dados"e é previsto no inciso XIV do Art. 5º. Ele prevê que o titular dos dados pode solicitar a eliminação de suas informações pessoais armazenadas em banco de dados, independentemente do procedimento empregado para obtê-las. Essa regulamentação é particularmente relevante se tratando de sistemas de armazenamento de dados em blockchain, devido à imutabilidade dos registros. Portanto, qualquer implementação deve garantir que os dados armazenados (ou sua possibilidade de acesso) possam ser removidos da blockchain a pedido do paciente, garantindo assim a conformidade com a legislação. A HIPAA, por sua vez, não prevê nenhum equivalente do direito ao esquecimento.

Os regulamentos de proteção de dados pessoais também têm normativas tratando do vazamento de informações e suas consequências. A GDPR estipula que qualquer incidente que ocasione no vazamento de dados pessoais de saúde deve ser reportado dentro de no máximo 72 horas. As multas estipuladas para as infrações à GDPR são baseadas em sua gravidade. Para as menos severas, o valor da multa será o maior valor entre 10 milhões de euros ou 2% do faturamento anual da companhia. Para as infrações consideradas severas, o valor pode chegar a 20 milhões de euros ou 4% do faturamento [Wolford 2020]. Valores semelhantes de multas são estipulados pela LGPD. O Artigo 52 prevê uma multa de 2% do faturamento da pessoa jurídica, grupo ou conglomerado, limitada no total a 50 milhões de reais. Em julho de 2023, a Telekall Infoservice foi multada pela ANPD em 2% de seu faturamento anual, R$14.400,00 no total, por vender uma listagem de contatos de WhatsApp de eleitores para disseminação de material de campanha eleitoral [Autoridade Nacional de Proteção de Dados 2023]. Em contrapartida à GDPR, LGPD não estipula um prazo máximo para reportar um vazamento de dados, deixando essa definição a cargo da ANPD [Koch 2020]. Se tratando da HIPAA, organizações são obrigadas a reportar um vazamento de dados em no máximo 60 dias, caso ele afete mais de 500 pessoas [United States Department of Health and Human Services 2013]. As multas previstas na HIPAA são categorizadas por gravidade, variando de infrações civis a criminais. O valor máximo de cada categoria varia de US$ 25 mil a US$ 1,5 milhão, sendo aplicado a cada violação individual cometida [Edemekong et al. 2024].

### 1.3.3. Tendências, Desafios e Oportunidades

Segundo relatório [Claroty 2025], dispositivos IoMT são o ponto onde hospitais e organizações na área da saúde estão mais expostos a ciberataques, especialmente considerando dispositivos operando em sistemas operacionais legados que não recebem mais atualizações de segurança. 96% das organizações pesquisadas apresentavam nesses dispositivos vulnerabilidades relacionadas com campanhas de *ransomware* que podem comprometer a disponibilidade de serviços e, consequentemente, a saúde de pacientes. É importante

que dispositivos IoMT sejam atualizados com frequência para garantir que eles sejam capazes de lidar com as ameaças mais recentes. Em adição a isso, é urgente a necessidade de colaboração entre a indústria, a academia e agências de padronização para garantir a intercomunicabilidade, segurança e regulamentação de tecnologias emergentes no escopo de dispositivos IoMT.

Outra tendência importante diz respeito ao uso de IA de modo geral. Em todo o mundo, países se movimentam para desenvolver arcabouços regulatórios visando especificamente a IA, estimulados pela sua ampla adoção entre indivíduos e corporações. Essa movimentação é necessária tanto pelos benefícios da tecnologia, quanto pelos potenciais riscos de seu uso. Na União Europeia, o *EU AI Act* é um dos primeiros atos regulatórios visando especificamente o uso de IA [Lewis et al. 2025], adotado em junho de 2024 e efetivo a partir de fevereiro de 2025 . Ele usa uma estratégia baseada no risco da tecnologia aos usuários para estipular os requisitos necessários para que o uso de IA seja permitido. Por exemplo, tecnologias de identificação e categorização de características biométricas em larga escala são proibidas, enquanto o uso em serviços públicos essenciais é considerado de alto risco. Caso a adoção de uma estrutura regulatória pela União Europeia fortaleça a tendência de outras regiões a também adotarem regulações mais estritas no âmbito da IA, sistemas usando essas tecnologias precisarão se adaptar para garantir sua conformidade.

Outro desafio para o setor é a implementação de estratégias de treinamento, conscientização e preparação dos funcionários. Um dos maiores impactos da pandemia de COVID-19 foi a mudança nos padrões de trabalho nas mais diversas áreas. No setor da saúde em particular, essa mudança fez com que estratégias de segurança desenvolvidas ao longo dos anos deixassem de ser aplicadas no momento em que funcionários passaram a trabalhar de casa com nenhum ou pouco conhecimento em cibersegurança [Cartwright 2023]. Essa falta de treinamento torna uma organização especialmente vulnerável a ataques de phishing. Ataques de phishing direcionados a funcionários podem expor credenciais e levar a vazamentos de dados substanciais, como foi o caso da empresa Solara [United States Department of Health and Human Services 2024b], multada em US$ 3 milhões em 2025 nos Estados Unidos. Para prevenir esse tipo de ataque, as organizações precisam garantir o treinamento de cibersegurança de funcionários independentemente de papéis, o que requer investimentos consideráveis na área.

É vital que as ferramentas de cibersegurança sejam capazes de acompanhar o ritmo das inovações aceleradas que acontecem no setor da saúde. Porém, esse nem sempre é o caso. Os investimentos em cibersegurança costumam ser negligenciados no setor de forma global, o que resulta, por exemplo, na ausência de estratégias eficientes de proteção e no uso continuado de equipamentos obsoletos que nem sempre recebem *patches* de segurança e suporte [Cartwright 2023]. Devido às características próprias do setor, como o fato de os dados manipulados serem pessoais e sensíveis em sua maioria, tem-se que o custo médio de um vazamento de dados no setor da saúde ultrapassou o valor de US$ 10 milhões em 2024 [IBM Security 2024]. Historicamente, o setor é o que tem o custo mais elevado associado a esse tipo de ataque.

De acordo com o relatório produzido pela IBM, o setor de finanças está em segundo lugar com um custo médio por vazamento chegando a aproximadamente 53% do

custo do setor de saúde. Esse cenário demanda um investimento considerável do setor da saúde em ferramentas de cibersegurança para proteger equipamentos, sistemas e principalmente, dados. O valor de investimento necessário em cibersegurança é citado como um dos maiores desafios para pequenas empresas no setor da saúde nos Estados Unidos [HIMSS, FinThrive 2025]. Ainda nos Estados Unidos, o número de ataques cibernéticos no setor afetando 500 ou mais indivíduos está projetado para alcançar quase 700 incidentes em 2025. Esse número ultrapassa em mais de 10 vezes a média de ataques entre 2016 e 2022 [United States Department of Health and Human Services ]. Nas próximas seções, serão abordados os impactos de ciberataques no setor da saúde e será apresentada uma metodologia para um planejamento eficiente de investimentos em cibersegurança, com casos de uso voltados para o setor da saúde.

## 1.4. Riscos e Impactos de Ciberataques

Organizações possuem diferentes riscos associados, como, por exemplo, os riscos de falha nos sistemas, ciberataques, incêndios e vazamentos de dados. Tais riscos podem ocasionar impactos com diferentes dimensões e magnitudes, que podem afetar diretamente as organizações, seus funcionários e usuários. A Figura 1.4 apresenta uma visão geral dos diferentes domínios de impacto de um ataque cibernético nas empresas. Primeiro, o domínio *Econômico* envolve todos os custos diretos e indiretos relacionados a um ataque cibernético. Como a perda financeira é uma das principais preocupações das empresas [Franco et al. 2023b], o foco das campanhas de segurança cibernética pode usar isso como um argumento poderoso para justificar a preocupação com a segurança cibernética. Em seguida, há o impacto *Legal* dos ataques cibernéticos, que transferem os casos de segurança cibernética para a esfera jurídica, as regulamentações e os aspectos de governança. Além disso, a esfera jurídica pode afetar diretamente os fatores econômicos, pois os efeitos colaterais envolvem os custos com advogados, compliance e multas aplicadas pelos órgãos reguladores.

Além disso, há diferentes impactos *Social*, pois os ataques cibernéticos podem interferir diretamente na vida das pessoas e nas estruturas sociais. Por exemplo, os ataques cibernéticos podem ser responsáveis por um colapso no sistema de saúde de um país, como no caso do Sistema Nacional de Saúde do Reino Unido [National Audit Office 2018], ou afetar a vida das pessoas, interrompendo serviços essenciais, como o fornecimento de alimentos [R. Mccrimmon and M. Matishak 2021] e a infraestrutura essencial dos países [J. R. Reeder, P. F. McQuade, S. A. Schipma 2021]. Além disso, o grande número de ataques cibernéticos que exploram a boa-fé dos seres humanos (por exemplo, técnicas de engenharia social e diferentes tipos de phishing) afeta a mudança de comportamentos sociais, o que faz com que as pessoas tenham muito mais medo, mesmo quando estão realizando interações legítimas [Parsons et al. 2013]. Por fim, o domínio *Técnico* de ciberataques descreve as principais disrupções e falhas de infraestrutura que podem também ocasionar um ou mais dos demais impactos descritos.

Em 2024, o custo médio global de uma violação de dados atingiu US$ 4,88 milhões, o maior valor já registrado, representando um aumento de 10% em relação a 2023 [IBM Security 2024]. Para pequenas e médias empresas (PMEs), os custos variaram entre US$ 120.000 e US$ 1,24 milhão, dependendo da gravidade do incidente [BigID 2024]. Empresas podem reduzir significativamente esses custos por meio de práticas efi-

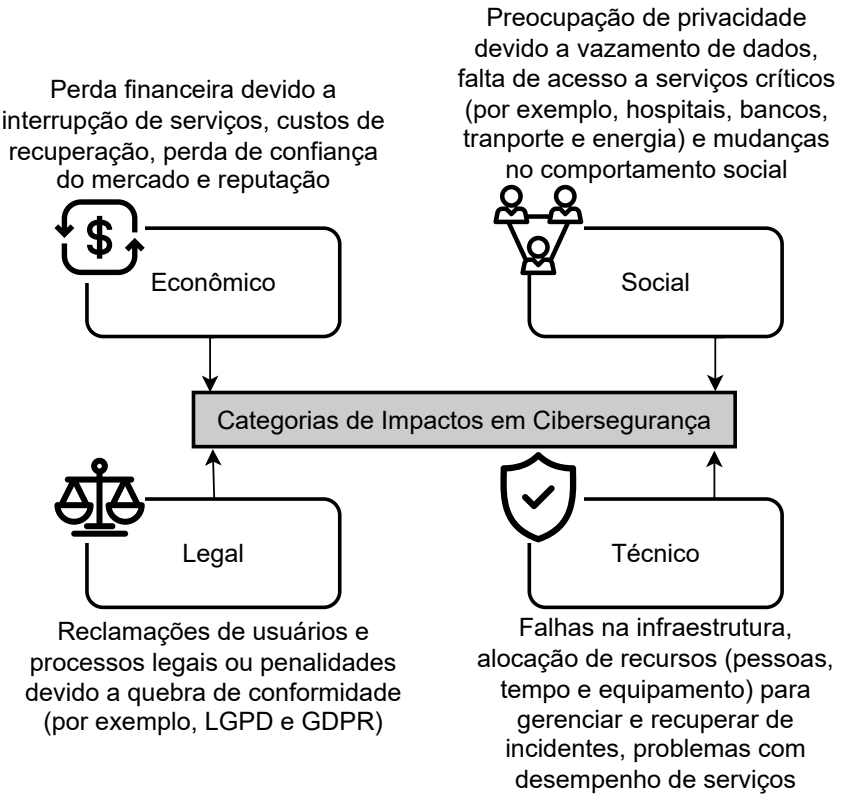



**Figura 1.4. Impactos de Ciberataques**

cazes de resposta a incidentes. Por exemplo, uma detecção rápida de violações pode reduzir substancialmente as perdas financeiras, e uma divulgação proativa aos clientes e partes interessadas pode atenuar os danos financeiros [IBM Security 2024]. As previsões mais recentes indicam que os custos globais do crime cibernético devem atingir US$ 9,5 trilhões por ano em 2024 [Secureworks 2024], enquanto os danos causados por ataques de ransomware podem ultrapassar US$ 275 bilhões até 2031 [Cybersecurity Ventures 2024].

Em relação aos custos, é importante mencionar que eles podem variar com base em diferentes características das empresas, como o país, setor, tamanho da organização e também configurações técnicas e proteções implementadas. Um estudo recente analisou diversos relatórios publicamente disponíveis de empresas de consultoria em cibersegurança para identificar os fatores que estão relacionados aos custos de um ciberataque e sua magnitude [Franco et al. 2024a]. Foi observado que o país, setor e o tamanho da organização estão diretamente relacionados aos custos de um ciberincidente. Fatores técnicos relevantes nos custos de um ciberataque também incluem o acesso remoto de funcionários, a utilização de computação em nuvem e a ausência de medidas de proteção básicas (por exemplo, antivírus, firewalls e autenticação multifator). A Figura 1.5 apresenta uma análise de cada fator e o seu impacto na variação dos custos de ciberataques. Por exemplo, organizações de um *País* específico podem reportar impactos financeiros até 100% maiores do que a média, enquanto organizações de outros países podem reportar impactos

75% menores que a média.

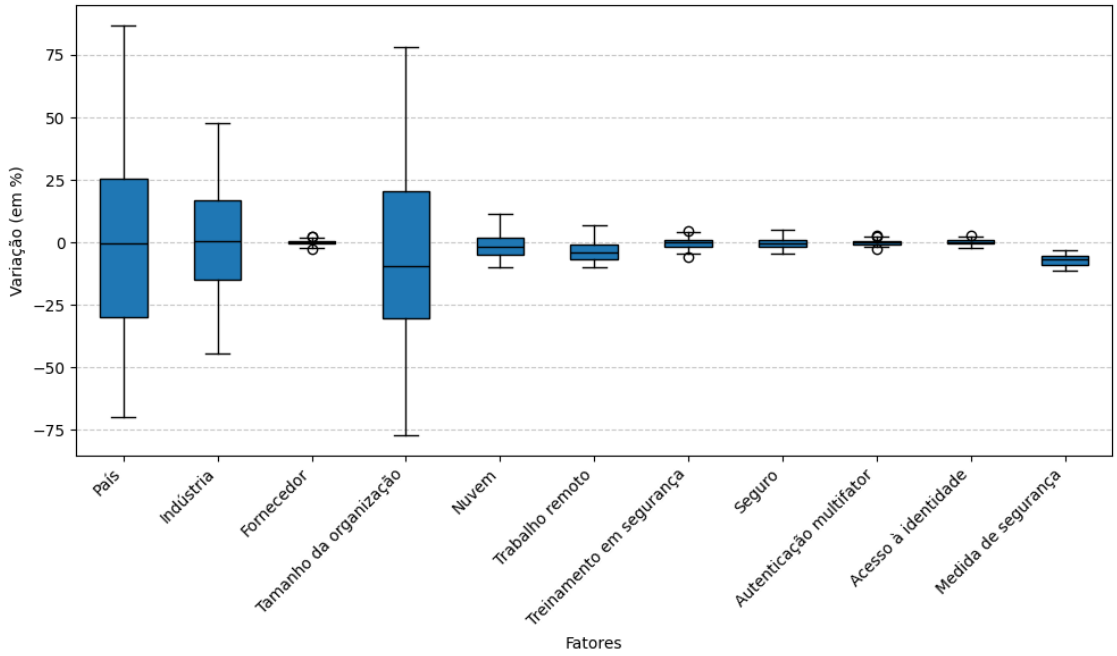



**Figura 1.5. Distribuição de Variação de Impactos por Fatores**

### 1.4.1. Ameaças e Impactos de Ciberataques no Setor da Saúde

De acordo com a Organização das Nações Unidas (ONU) e a Organização Mundial da Saúde (OMS), a quantidade de ciberataques no setor da saúde é uma ameaça global que não pode ser ignorada [Mishra 2024]. Relatórios de 2025 mostram que mais de dois terços das instituições de saúde entrevistadas sofreram ao menos um ataque de ransomware nos últimos anos [Arctic Wolf Labs 2025]. É possível observar o crescimento do interesse de cibercriminosos pelo setor da saúde, resultando em um aumento de ciberataques e colocando o setor como o segundo em número de ciberataques reportados [CheckPoint 2025]. Tal movimento vai também de encontro à falsa ideia de que cibercriminosos evitam o setor por questões éticas. Na verdade, a maior motivação de ciberataques ainda é econômica, tendo como alvo setores que possuem a maior quantidade de sistemas e informações críticas, o que significa uma maior propensão a obter lucros com o ciberataque.

O mercado de cibersegurança na área da saúde poderá atingir US$ 125 bilhões entre 2020 e 2025 [Cybersecurity Ventures, Herjavec Group 2021]. Esse crescimento é impulsionado pelo aumento acelerado dos ataques cibernéticos no setor, intensificados pela pandemia de COVID-19, iniciada no começo de 2020. A crise sanitária global desencadeou uma corrida não apenas pelo desenvolvimento de tratamentos, mas também por tecnologias de monitoramento de contato com infectados [Franco et al. 2021]. Como consequência, o setor de saúde tornou-se um dos principais alvos de criminosos cibernéticos como nunca antes na história. Por exemplo, de acordo com o relatório da Cybersecurity Ventures, 62% dos administradores hospitalares entrevistados se sentem inadequadamente preparados para planejar ou reagir a incidentes de cibersegurança que possam afetar suas instituições.

Embora o setor de saúde possua agilidade e interesse na adoção de novas tecnologias, o mesmo não se pode dizer em relação à ações para protegê-las contra ameaças de segurança cibernética. Apesar da reconhecida importância da cibersegurança nessa área, os dados sobre a situação atual são alarmantes. A escassez de profissionais de cibersegurança não é um problema exclusivo da área da saúde, mas, nesse setor, a dimensão do problema é particularmente preocupante diante dos riscos envolvidos. Segundo dados da pesquisa conduzida em [Thyagarajan et al. 2020], três em cada quatro hospitais não contam com um profissional designado especificamente para tratar de questões de cibersegurança.

Em um estudo de 2018 [Fuentes and Huq 2018], é possível observar diversos vetores de ataque em dispositivos médicos, inclusive nos protocolos para Comunicação de Imagens Digitais na Medicina (Digital Imaging and Communications in Medicine, DICOM) e Sistema de Arquivamento e Comunicação de Imagens (Picture Archiving and Communication System, PACS). Ataques simples, ainda hoje, são possíveis de serem executados, ocasionando vazamento de informações relevantes, como informações de pacientes e imagens de exames. Por exemplo, a Figura 1.6 nos mostra o total de 2.204 dispositivos (por exemplo, servidores PACS e equipamentos de imagens), encontrados através de um sistema de busca especializado que mapeia e rastreia dispositivos e sistemas conectados à internet, que estão expostos publicamente na Internet e respondendo a requisições do protocolo DICOM. Ao realizar requisições legítimas para tais dispositivos, podemos, por exemplo, ter acesso a dados pessoais e médicos de pacientes (por exemplo, quais exames foram realizados, data dos exames e até mesmo imagens dos exames). Portanto, ainda que as organizações de saúde invistam recursos significativos na integração de sistemas, os investimentos para manter os softwares atualizados e os sistemas protegidos ainda são insuficientes. Esse problema é agravado pela escassez generalizada de especialistas em cibersegurança, além das dificuldades enfrentadas para manter os poucos profissionais qualificados que existem, cujo custo é elevado e cuja demanda no mercado é intensa [Coventry and Branley 2018, US Health Care Industry Cybersecurity Task Force 2017].

As principais causas das violações de segurança no setor são malwares e as ameaças internas (por exemplo, auxílio de funcionários para campanhas de phishing e pacientes). Dado que malwares (por exemplo, como o ransomware [Neprash et al. 2022] e ataques direcionados a equipamentos médicos [Mirsky et al. 2019]) são recorrentes nesse contexto, o setor tem direcionado investimentos específicos para proteger-se dessas ameaças, especialmente aquelas que afetam dispositivos de IoT, cuja relevância tende a crescer significativamente nos próximos anos, como a utilização de sensores em monitores de sinais vitais, localização e monitoramento de equipamentos, controle de salas de emergência e cuidados ao paciente. Outro ponto importante é o processo de desenvolvimento e de inovação dentro do setor. Por exemplo, em uma análise de vulnerabilidades [Knight 2021] em 30 aplicativos de saúde, foi identificado que 77% possuíam exposição de dados sensíveis e potenciais ataques em suas APIs de comunicação. Esses números mostram que a velocidade de inovação e a necessidade de desenvolvimento de soluções têm sido realizadas sem o cuidado necessário com a cibersegurança desde a sua concepção.

Recentemente, em janeiro de 2025, a União Europeia definiu um plano para reforçar a cibersegurança no setor da saúde, incluindo iniciativas para prevenção, detecção

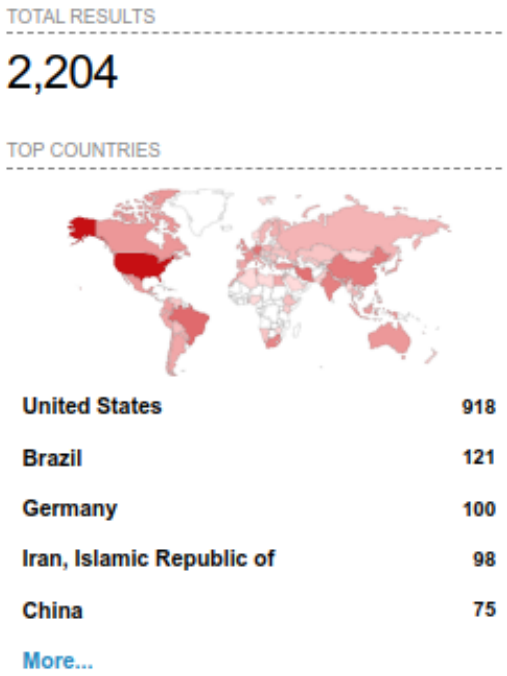



**Figura 1.6. Resultado de Busca no Shodan por Dispositivos Expostos na Internet e Respondendo a Requisições DICOM**

e resposta às ameaças [European Commission 2024]. Essa é uma ação necessária devido ao crescente aumento de ciberataques no setor. Somente os 27 países membros da União Europeia reportaram, juntos, 309 incidentes de grande magnitude no setor da saúde, resultando em atraso de procedimentos médicos, bloqueios de salas de emergência e disrupção de serviços essenciais para a gestão de serviços hospitalares. No Brasil, a situação ainda é alarmante, com uma grande quantidade de empresas do setor sendo alvos de ransomware e phishing, tornando o setor com a maior quantidade de ataques e com dados valendo até 50 vezes mais que os demais setores no mercado ilegal [Fonseca 2025, IBM Security 2024]. Além disso, legislações como a Lei Geral de Proteção de Dados (LGPD) podem afetar todos os setores com multas de até 2% do faturamento.

Os impactos de ciberataques no setor são extremamente preocupantes, já que envolvem dimensões que afetam diretamente a vida de pessoas, como é o exemplo da primeira morte resultante de um ciberataque, conforme reportado em [Associated Press 2020]. Além disso, existem também os impactos financeiros que causam para as organizações e profissionais que prestam serviços de saúde, o que pode levar até à falência de empresas que prestam serviços essenciais para a sociedade. No restante dessa seção, serão apresentados exemplos de dois casos de ciberataques que ocasionaram impactos sociais e econômicos relevantes no setor da saúde. Tais ciberataques tiveram como foco os dados e também sistemas críticos para a operação das organizações, resultando em interrupção de serviços e vazamentos de dados sensíveis. Embora os cenários apresentados sejam de grande magnitude, é importante que o setor compreenda que diversos ciberataques acontecem no setor, porém, em instituições menores, mas com um alto impacto para suas operações, funcionários e pacientes.

### 1.4.2. Estudo de Caso #1: Ataque de Ransomware no NHS-UK

Neste cenário, serão analisadas as causas e os impactos de um dos casos mais conhecidos de ciberataques no setor da saúde. O ataque do ransomware WannaCry, ocorrido em maio de 2017 no National Health System (NHS) do Reino Unido (United Kingdom - UK) [National Audit Office 2018] foi um marco em cibersegurança para o setor da saúde, resultando em impactos diretos na saúde de milhares de pacientes no Reino Unido. Discutiremos também boas práticas e o que poderia ter sido feito para evitar o ataque, bem como analisaremos as lições aprendidas, de nível técnico e administrativo, pós-ciberataque [NHS Foundation Trust 2023].

O ransomware conhecido como WannaCry infectou mais de 200 mil dispositivos em cerca de 156 países, tendo como foco a encriptação de arquivos de sistemas, planilhas eletrônicas e documentos importantes para a operação de negócios no mundo todo. Após a encriptação e indisponibilidade de sistemas, um valor em Bitcoin era solicitado como taxa de resgate dos dados. Tal ataque tinha como alvo dispositivos rodando o sistema operacional Microsoft Windows e explorava uma falha de segurança em sistemas de compartilhamento de arquivos para contaminar a maior quantidade de dispositivos possíveis. O NHS-UK, que possuía milhares de sistemas vulneráveis (por exemplo, com Windows XP e 7 sem atualizações recomendadas de segurança) foi alvo do ataque, tendo como resultado a disrupção de serviços de 80 das 236 unidades organizacionais que oferecem serviços de saúde à população (por exemplo, hospitais gerais, especializados ou serviços de ambulância).

Baseado em auditoria realizada [National Audit Office 2018], o ataque infectou e deixou totalmente fora de operação, ao menos, 34 unidades de saúde, enquanto outras 46 unidades reportaram disrupção de serviços devido à falta de recursos oferecidos pelas demais unidades infectadas. Por exemplo, funcionários dos serviços de saúde não puderam acessar dispositivos computacionais, gerando atrasos no processamento de informações de pacientes e nos resultados de exames médicos. Além disso, equipamentos médicos foram bloqueados ou isolados como forma de prevenção ao ciberataque, resultando em disrupção nos serviços de radiologia e análises clínicas que dependem de equipamentos digitais (por exemplo, diagnóstico por imagem e testes de sangue). O cronograma do ataque, baseado no relatório realizado pelo National Audit Office (2018), é apresentado abaixo.

- **12 de maio de 2017 (Início do Incidente):**
  - $\approx$11:00 horas: Primeiras unidades de saúde relatam problemas em sua operação.
  - 13:06 horas: Primeira notificação para as equipes de resposta a incidentes.
  - 16:00 horas: NHS-UK declara o ciberataque um incidente nacional.
  - 18:45 horas: Decisões estratégicas e coordenação para resposta ao incidente.
  - $\approx$22:00 horas: Especialista descobre como interromper (kill switch) o ransomware e consegue parar a propagação do ciberataque.
- **De 13 a 15 de maio de 2017:** Equipes do NHS-UK implementam soluções manuais para manter serviços essenciais ativos. A Microsoft divulga atualizações de segurança imediatas.

- **De 15 a 18 de maio de 2017:** O NHS-UK declara recuperação parcial dos sistemas, mas alerta sobre possíveis novos ataques. Atualizações de segurança são realizadas e proteções adicionais são implementadas (por exemplo, antivírus e atualização de sistemas).
- **19 de maio de 2017 (Fim do Incidente):** O incidente é contido pelo NHS-UK e os sistemas são restaurados.

Ao fim do ciberataque, foi identificado que, ao menos, 1220 equipamentos de diagnóstico foram infectados (≈1% de todos os equipamentos do NHS-UK), além dos computadores das unidades. Milhares de sistemas computacionais não foram infectados devido ao isolamento, o que evitou a propagação do ransomware, mas também ocasionou a interrupção dos serviços. Ao menos cinco hospitais necessitaram redirecionar todos os serviços de emergência e ambulância para outros hospitais, incluindo hospitais de referência como o Royal London Hospital e o Lister Hospital.

Os impactos do ciberataque podem ser divididos entre impactos aos pacientes e custos financeiros. A Tabela 1.1 apresenta um resumo dos principais impactos identificados e mensurados. Em relação aos pacientes, o ciberataque resultou em 6.912 consultas canceladas diretamente no período em que o NHS-UK estava enfrentando o ciberataque (ou seja, de 12 a 19 de maio de 2017). Tal número não inclui os impactos em consultas agendadas para o pós-incidente, o que pode chegar a cerca de 19.494 consultas canceladas. O NHS-UK reportou que, ao menos, 139 pacientes necessitando de diagnósticos urgentes em relação a câncer tiveram exames cancelados. Cirurgias e demais operações após o incidente foram afetadas, porém os números exatos não foram mensurados e reportados pelo NHS-UK. Em relação ao impacto financeiro, o NHS-UK não realizou estudos mais profundos em relação à redução de atendimentos no período do ataque. Porém, pesquisadores estimaram em torno de £ 5.9 milhões o impacto financeiro ao analisarem a redução das atividades durante o período [Ghafur et al. 2019]. Análises anteriores estimaram que o ataque WannaCry causou um prejuízo de £ 92 milhões ao NHS-UK, com base na suposição de que o ataque afetou 1% de todos os serviços do NHS, incluindo os cuidados primários (como os atendimentos em consultórios de médicos de família) [The Telegraph 2018]. No entanto, dados sobre cuidados primários não foram coletados na época. A pesquisa realizada, portanto, focou apenas nos cuidados secundários (hospitalares), utilizando mudanças reais observadas nas atividades.

As lições aprendidas com o ciberataque geraram diferentes reflexões e ações para embasar novas estratégias de cibersegurança e também ações diretas para evitar ciberataques futuros [Smart 2018, National Audit Office 2018]. Ficou claro para o NHS-UK e para o setor da saúde que a questão não é `se` mas sim `quando` o próximo ciberataque irá acontecer. Portanto, os principais desafios incluem estar preparado e capaz de responder rapidamente em caso de incidente. Como primeira lição, ficou evidente a necessidade de desenvolver um plano de resposta em caso de ciberataques, bem como definir papéis e responsabilidades em nível local e nacional dentro do NHS-UK. Além disso, é necessário garantir que todas as organizações processem e implementem alertas de cibersegurança, incluindo atualização de software para correção de vulnerabilidades e antivírus atualizados.

Porém, sem que organizações, líderes e equipes tratem ameaças digitais como um problema real, não será possível mitigar riscos de forma eficaz. Assim, é importante que

**Tabela 1.1. Resumo e Principais Impactos em Pacientes e Econômicos do Ciberataque de Ransomware no NHS-UK**

| Tipo | Quantidade | Descrição |
|---|---|---|
| Equipamentos de diagnósticos infectados | 1229 equipamentos | Atraso em exames e atendimentos |
| Unidades organizacionais de saúde afetadas | 80 unidades (24.000 funcionários) | Impacto direto nos serviços de rotina e de urgência em diversas regiões do país |
| Consultas canceladas | 6.912 consultas (19.494 pós-ciberataque) | Impossibilidade de realizar consultas por falta de acesso aos sistemas de agenda, exames e informações de pacientes |
| Diagnósticos urgentes afetados | 139 pessoas | Pacientes em investigação de neoplasia foram diretamente afetados pela interrupção de serviços de patologia clínica e exames de imagem |
| Perda financeira devido a redução das atividades | £ 5.9 milhões | Redução na entrada de novos pacientes e nos atendimentos, bem como cancelamento de pacientes agendados |
| Impacto financeiro total estimado | £ 92 milhões | Custos relacionados ao impacto direto devido a redução das atividades, custos de TI e investimentos em infraestrutura adicional |

os diferentes atores envolvidos no setor da saúde (por exemplo, políticos, gestores e os profissionais que lidam diretamente com pacientes) estejam cientes de riscos diretos aos serviços críticos e trabalhem proativamente para maximizar a resiliência da infraestrutura e minimizar o impacto ao cuidado aos pacientes. Por fim, todas as unidades organizacionais foram comunicadas para resolver e implementar todos os alertas de cibersegurança emitidos pelo NHS Digital entre março e maio de 2017. Também foram tomadas ações para garantir a proteção local através de firewalls.

Como consequência do ciberataque, houve uma priorização do orçamento de TI para aprimorar a cibersegurança nos principais centros traumatológicos e melhorias no sistema de alertas de segurança[1], além de uma lista com 21 recomendações de autoria da chefia do Departamento de Saúde e Assistência Social do UK [Smart 2018]. Desde então, diversos novos ciberataques aconteceram ao NHS-UK, e também ao redor do mundo, incluindo um recente vazamento de dados de pacientes e exames realizados em um laboratório de patologia que processa exames de sangue em nome de várias organizações do NHS[2].

### 1.4.3. Estudo de Caso #2: Vazamento de Dados no SUS

Em 2019, um atacante afirmou possuir dados de identificação de 205 milhões de usuários do Cartão Nacional de Saúde (CADSUS), incluindo nome, nome da mãe, endereço, CPF

[1]https://digital.nhs.uk/cyber-alerts

[2]https://www.england.nhs.uk/synnovis-cyber-incident/

e data de nascimento. Como prova, foram vazados 2 milhões dos dados em um website chamado *www.leaksus.com.br*. A Figura 1.7 apresenta um screenshot do website, que foi retirado do ar após alguns dias. O vazamento de dados ocorreu através de uma falha em uma API disponibilizada para consulta de dados de um usuário através do seu número do cartão SUS e senha. Porém, após realizar uma requisição legítima (ou seja, com um número de cartão e senha de um usuário real), foi possível realizar milhões de solicitações apenas alterando o número do CPF na chamada para a API. Por exemplo, a chamada *"consulta.php?cpf=xxx.xxx.xxx.xx"* retornaria todos os dados do CPF *xxx.xxx.xxx.xx* e poderia ser feita para qualquer CPF após realizar uma primeira consulta legítima. Tal vazamento ocorreu, portanto, não devido a um ciberataque sofisticado, mas sim como resultado de uma falha de implementação do sistema, conforme discutido na Seção 1.4.1 e reforçado por estudos sobre a segurança no desenvolvimento de aplicações para o setor da saúde [Knight 2021].


**Figura 1.7. Website Criado em 2019 para Compartilhar Informações Vazadas do SUS devido a Exposição de API**

Cerca de um ano após o vazamento devido à exposição de API, ocorreu um fato ainda mais curioso de exposição de dados. Informações sensíveis (por exemplo, CPF, endereço, telefone e doenças pré-existentes) de cerca de 16 milhões de pacientes ficaram expostas devido à exposição de senhas de usuários do Ministério da Saúde que possuíam acesso a tais informações. O vazamento das senhas aconteceu por erro humano e demorou cerca de 1 mês para a identificação da falha. De acordo com o levantamento do Jornal O Estadão de São Paulo [Cambricoli 2020], uma planilha com as senhas foi compartilhada no Github, uma plataforma para compartilhamento de códigos e trabalho colaborativo [3], juntamente com o código de um modelo estatístico sendo desenvolvido em uma parceria

[3]https://www.github.com

do Hospital Albert Einstein e o Ministério da Saúde; porém, o responsável pelo desenvolvimento não removeu o arquivo com as senhas do repositório público. Com as senhas publicadas, era possível acessar registros relacionados à COVID-19, incluindo casos suspeitos e internações por síndrome respiratória aguda grave. Tal falha mostra a importância da proteção de dados e também de políticas bem definidas para o gerenciamento de informações sensíveis em projetos na área da saúde [Todde et al. 2020].

Segundo o painel de Registro de Incidentes com Dados Pessoais [4], mantido como forma de conformidade com a LGPD, o Ministério da Saúde registrou três incidentes relatados desde a vigência da lei. O primeiro incidente relata o vazamento de credenciais do sistema CADSUS, que expôs dados demográficos e sensíveis de usuários, durante o período de abril de 2019 até junho de 2022. Também existe um incidente, em 2022, de venda ilegal de bases de dados administrativas vindas dos sistemas de saúde. Por fim, foi reportado o incidente referente a falha de API discutida anteriormente nesta seção, sendo comunicado o incidente aos titulares dos dados. Como lições aprendidas, foram adotadas estratégias de segurança adicionais, como autenticação multifator e a troca de senha a cada três meses como política obrigatória. Além disso, foi realizada verificação de vulnerabilidades utilizando ferramentas comerciais e teste de penetração. Os incidentes também foram comunicados às autoridades, como a Polícia Federal e a Secretaria de Governo Digital do Ministério da Economia.

Embora ações tenham sido tomadas em relação aos vazamentos, não é possível remover os dados já expostos de bases ilegais. Portanto, mesmo que informações como senhas e número de usuários possam ser alterados, os dados pessoais e sensíveis vazados continuam a ser válidos, afetando a vida de milhões de brasileiros e podendo resultar em discriminações, crimes financeiros ou mesmo exposição e chantagem. É importante adotar medidas de notificação aos titulares dos dados, como recomendado pela Autoridade Nacional de Proteção de Dados, mas também é fundamental que as organizações adotem medidas que evitem os vazamentos e não apenas medidas para remediar um incidente.

Diferentemente de impactos apenas técnicos ou econômicos, os vazamentos de dados podem ter um impacto contínuo e impossível de mensurar nas vidas das pessoas durante anos. Além disso, tais dados podem (e são) utilizados para fomentar o cibercrime, com campanhas de phishing cada vez mais eficientes, já que possuem dados que validam diversos cenários para induzirem usuários ao erro. Por exemplo, imagine um cenário onde um médico de um hospital próximo à sua residência entre em contato e solicite que você acesse um link para verificar possíveis tratamentos de uma doença crônica que você possui. Se os dados estiverem corretos, a chance de você clicar será aumentada. Essa ação pode resultar em potenciais riscos para a sua segurança cibernética e da empresa onde você trabalha. Tal cenário também pode acontecer de forma contrária: alguém entrando em contato com o hospital ou profissionais da saúde. Portanto, esse cenário analisado mostra a importância de estratégias de cibersegurança que auxiliem na proteção de dados sensíveis e, principalmente, na prevenção de vazamentos, incluindo possíveis falhas em aplicações, dispositivos móveis, sensores e sistemas de comunicação que são amplamente utilizados no setor da saúde para trazer inovação tecnológica e melhor acesso ao tratamento de pacientes.

[4]https://www.gov.br/saude/pt-br/acesso-a-informacao/lgpd/registro-de-incidentes-com-dados-pessoais

## 1.5. Planejamento em Cibersegurança: Análise, Priorização de Riscos e Investimentos

O planejamento e o investimento em cibersegurança devem ser encarados como componentes estratégicos essenciais à sustentabilidade operacional e econômica de organizações, independentemente do setor de atuação [Franco et al. 2023b]. À medida que as ameaças digitais se tornam mais sofisticadas e frequentes, torna-se indispensável adotar abordagens proativas de análise e gestão de riscos (sejam eles técnicos, econômicos, legais ou sociais), alinhadas às necessidades de proteção de ativos críticos, conformidade regulatória e resiliência cibernética. Essa necessidade é ainda mais pronunciada em setores que lidam com dados sensíveis ou operam serviços críticos à sociedade. A saúde, nesse contexto, destaca-se como uma área particularmente desafiadora [Thyagarajan et al. 2020], tanto pela criticidade das informações e sistemas envolvidos quanto pela complexidade de seus ambientes tecnológicos em constante evolução [Levina et al. 2022]. Assim, embora os fundamentos do planejamento em cibersegurança sejam aplicáveis de forma transversal, sua aplicação na área da saúde exige atenção adicional devido à sua importância para a sociedade, complexidade operacional e valor para cibercriminosos.

O planejamento de estratégias de cibersegurança no setor da saúde demanda uma abordagem com atenção em certos pontos, em virtude da natureza sensível dos dados tratados, da alta dependência tecnológica das atividades clínicas e das rígidas exigências impostas por marcos regulatórios. Dados clínicos, como prontuários eletrônicos, laudos diagnósticos e registros históricos de pacientes, são considerados informações pessoais sensíveis e, por isso, estão sujeitos a normativas como a LGPD, HIPAA e GDPR. Tais normativas impõem a implementação de mecanismos robustos de controle de acesso, rastreabilidade, governança e gestão do consentimento [Aragão and Schiocchet 2020]. Além disso, devido à sua importância e valor no mercado ilegal, tais dados têm sido alvo de ciberataques nos últimos anos [IBM Security 2024]. Portanto, devido à natureza crítica dos serviços oferecidos, é indispensável a elaboração de estratégias que garantam a resiliência de serviços críticos e sejam capazes de garantir a disponibilidade ininterrupta dos sistemas e resposta eficaz frente a incidentes, como ataques de ransomware e negação de serviço. O ambiente hospitalar, por exemplo, apresenta uma heterogeneidade tecnológica marcada pela convivência de sistemas legados, dispositivos médicos conectados e infraestruturas críticas de TI, muitas vezes sem atualizações regulares ou com vulnerabilidades conhecidas.

A avaliação de riscos e impactos nesse contexto deve considerar a singularidade de cada ativo digital e sua interdependência com processos clínicos, a fim de mitigar possíveis vetores de ataque. Além disso, o engajamento de profissionais da saúde no uso seguro das tecnologias exige a formulação de políticas e uma cultura organizacional alinhadas aos serviços de saúde que são a atividade principal, mas também que reduzam os riscos e potenciais vetores de ataques. Assim, a evolução constante das ameaças digitais, aliada às demandas por conformidade regulatória, auditorias e certificações institucionais, reforça a necessidade de um planejamento estratégico que integre dimensões técnicas, regulatórias, humanas e econômicas de forma coordenada. Tais esforços incluem, por exemplo, desde o treinamento em cibersegurança para evitar ciberataques com foco nos profissionais e usuários até a implementação de proteções robustas para mitigar os riscos em sistemas e dados críticos. Discutiremos como construir estratégias eficientes ao longo

desta seção, reforçando as nuances de setores como o da saúde frente a setores menos críticos.

### 1.5.1. Metodologia para Planejamento

A metodologia proposta em [Franco et al. 2023b] compreende cinco etapas que representam as tarefas sequenciais que os tomadores de decisão devem considerar ao planejar uma nova estratégia de segurança cibernética (ou atualizar uma estratégia já existente) [Franco et al. 2022]. A Figura 1.8 mostra a metodologia, incluindo todas as fases (de A a E) e exemplos de etapas críticas que devem ser executadas em cada uma dessas fases. Essa metodologia foi definida com base em uma análise aprofundada da literatura, em entrevistas com especialistas em segurança cibernética e tomadores de decisão do setor, das pequenas e médias empresas e do meio acadêmico, e com base em todo o conhecimento obtido e nas discussões realizadas pelo autor. É importante mencionar que as etapas destacadas para cada etapa são exemplos de etapas gerais comuns à maioria das empresas, mas não são exaustivas. A metodologia pode ser ampliada e adaptada para atender às demandas específicas de uma determinada empresa ou setor. Também, é fundamental observar que existem etapas que devem ser consideradas com maior cautela quando se considera a cibersegurança em setores críticos, como é o caso do setor da saúde.

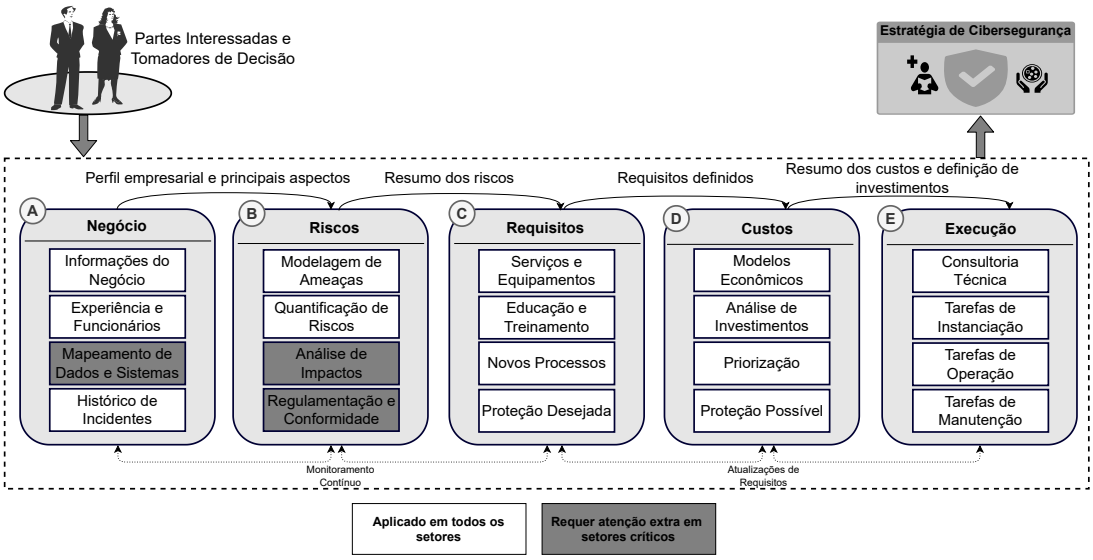



**Figura 1.8. Etapas para Planejamento de Estratégias de Cibersegurança**

O planejamento começa na **Etapa A** (ou seja, Negócio), na qual todas as informações relacionadas ao negócio devem ser coletadas e um briefing deve ser conduzido, considerando todas as partes interessadas envolvidas. Exemplos de partes interessadas no setor da saúde incluem a direção executiva e o conselho de hospitais e clínicas, profissionais de saúde, órgãos reguladores e governamentais, equipes de segurança da informação e os próprios pacientes. Para essa fase, as informações sobre o negócio são fundamentais, como a atuação da empresa, as tecnologias utilizadas, o número de funcionários, a receita e o portfólio. No caso da saúde, é essencial compreender os serviços fornecidos e sua criticidade.

Em seguida, a experiência do pessoal é um indicador importante para entender os possíveis desafios ou pontos fracos técnicos a serem considerados durante o planejamento de uma estratégia de segurança cibernética. Se a organização não possuir um elevado nível de conscientização sobre segurança cibernética, poderá tornar-se um vetor para diversos tipos de ataque (por exemplo, phishing e ransomware). Profissionais da saúde sem treinamento básico em cibersegurança, por exemplo, que tenham acesso a sistemas ou informações sensíveis, podem representar um elo fraco durante um ciberataque.

Além disso, o mapeamento de dados e sistemas é um aspecto que requer atenção especial no setor da saúde. Devido à necessidade de gerenciar dados sensíveis de pacientes, bem como ao uso da tecnologia como meio para cuidar da saúde, é fundamental mapear e compreender como cada dado e sistema está interconectado e configurado. Essa atividade, aliada à análise do histórico de incidentes anteriores, contribuirá para a avaliação de riscos e dos possíveis impactos decorrentes de ciberataques.

Na **Etapa B** (ou seja, Riscos), o foco é a análise de segurança e a modelagem de ameaças da organização. Para isso, podem ser consideradas ferramentas e soluções voltadas à avaliação de riscos, incluindo plataformas consolidadas no mercado E utilizadas para analisar dados relevantes à segurança e realizar testes de penetração (por exemplo, Nmap, Metasploit e Shodan). Além disso, durante essa fase, a modelagem de ameaças pode ser conduzida utilizando abordagens específicas, como a modelagem de ameaças utilizando STRIDE [Microsoft 2022] e a estrutura MITRE ATT&CK[5] para mapeamento de técnicas utilizadas por ciberataques. Um exemplo de mapeamento utilizando as metodologias STRIDE e DREAD [EC-Councl 2022] no setor da saúde está disponível em [Fuentes and Huq 2018], juntamente com uma análise dos principais sistemas médicos expostos. A análise de impactos também é uma etapa altamente relevante, especialmente em setores críticos. No caso do setor da saúde, além dos impactos técnicos e econômicos, há uma relação direta com a vida dos pacientes e com o bem-estar social. Além disso, devem ser consideradas as regulamentações e boas práticas específicas. Portanto, nessa etapa, é necessário realizar uma análise detalhada para compreender os riscos e os impactos de um ataque de forma individualizada, incluindo análises específicas para sistemas e dados sensíveis. Por exemplo, um DDoS em uma infraestrutura de cirurgia remota tem um impacto crítico, enquanto um ataque de ransomware em um sistema de pagamentos de uma clínica representa um impacto econômico relevante. Compreender tais riscos e impactos é essencial para, então, planejar estratégias eficazes de mitigação.

Já na **Etapa C** deve ser utilizado todo o conhecimento adquirido nas etapas anteriores para decidir quais requisitos de proteções são necessários. Por exemplo, caso, baseado na análise do negócio, riscos e impactos associados, tenha sido definido que ataques de phishing possuem alto risco de acontecer no corpo clínico de um hospital com o objetivo de propagar malwares ou acessar sistemas com informações sensíveis, é importante investir em treinamento para que os profissionais não sejam afetados. Além disso, proteções de endpoint (por exemplo, clientes de e-mails e estações de trabalho) podem ser contratadas para mitigar os riscos. Portanto, é nessa etapa que serão definidas quais proteções, treinamentos e novos processos podem ser implementados para mitigar os riscos e os impactos no contexto do negócio.

[5]https://attack.mitre.org/

Após a definição dos requisitos, na **Etapa D**, serão considerados os custos reais para implantação e operação das proteções e a definição ou ajuste do orçamento existente. Nessa etapa temos um dos principais desafios da cibersegurança: o baixo orçamento e os custos elevados das soluções. Considerando tal desafio, podemos utilizar modelos econômicos para a cibersegurança [L. A. Gordon, M. P. Loeb, L. Zhou 2021] que auxiliem na definição do orçamento e também na análise dos investimentos, seja de modo a otimizar o investimento ou fornecer fatos mensuráveis para angariar fundos adicionais para a cibersegurança junto à gestão (por exemplo, caso não seja investido X, podemos perder até 10 vezes o valor de X em um ano). Após a definição do orçamento, é necessário definir as prioridades (ou seja, reduzir os riscos de situações que possuam maior impacto) e, então, selecionar as proteções possíveis de serem implementadas dentro do orçamento e que sejam condizentes com o nível de cibersegurança almejado. Por fim, na **Etapa E**, a estratégia de cibersegurança definida será implementada e operada conforme especificada. Lembrando que toda estratégia deve ser testada, monitorada e atualizada com uma frequência apropriada (também definida baseado na análise do setor e ativos), já que os riscos e possíveis impactos são dinâmicos.

Para priorização, podemos utilizar desde métricas técnicas como o Exploit Prediction Scoring System (EPSS) [Jacobs et al. 2021], que auxilia na compreensão de quais vulnerabilidades possuem maior probabilidade de serem exploradas, ou abordagens que permitam a análise dos possíveis impactos financeiros, sociais e legais em caso de um ciberataque. Tais abordagens podem ser específicas para o setor da saúde ou adaptações de abordagens generalistas, mas levando em consideração a realidade de cada setor.

### 1.5.2. Quantificação e Priorização de Riscos

Ao realizar a análise de riscos, é importante quantificar os riscos de forma mensurável (por exemplo, compreender os reais riscos e seus impactos) de forma a possuir as informações necessárias para uma priorização baseada nos riscos, impactos e orçamento disponível. No entanto, essa quantificação é desafiadora, pois exige conhecimento profundo sob perspectivas técnica, econômica e jurídica em relação a uma empresa e ao cenário de ameaças existente [Franco et al. 2024b]. Além disso, em setores críticos como o da saúde, tal quantificação envolve também os impactos na sociedade e na vida humana. Portanto, as abordagens para quantificação de riscos e impactos devem lidar com esses desafios e encontrar maneiras eficazes de contornar as limitações existentes, incluindo a capacidade de lidar com *(i)* assimetria de informações entre as empresas, *(ii)* falta de comunicação entre os níveis de diretoria e *(iii)* falta de mapeamento quantitativo entre as ameaças e seus impactos reais.

Para a quantificação de riscos podemos utilizar modelos e simulações, como por exemplo o proposto em [Franco et al. 2024a] e [Nunes et al. 2024]. Ambos modelos utilizam dados estatísticos disponíveis em relatórios públicos de empresas de consultoria em cibersegurança para simular e prever possíveis riscos de ataques acontecerem e também seus impactos econômicos. Para isso, são utilizadas informações como o setor, tipo de ataque, localização geográfica e informações específicas dos serviços oferecidos. Esse tipo de abordagem permite reduzir a assimetria de informações e compreender quais riscos devem ser observados com maior atenção.

Por exemplo, uma clínica de exames médicos com sede no Brasil e uma filial na Alemanha deve ter em mente os principais ciberataques e riscos que têm como foco o setor da saúde (por exemplo, phishing, ransomware e vazamentos de dados). Além dos riscos dos sistemas específicos, precisamos estar atentos a informações estatísticas de forma global para encontrarmos um planejamento local eficiente. Por exemplo, o Brasil possui uma das maiores quantidades de ataques de phishing e a Alemanha possui impactos econômicos de ciberataques 4% acima do que a média global. Além disso, o impacto de vazamento de dados no Brasil é de ≈R$ 8 e na Alemanha é de ≈R$ 30 por cada registro vazado. Esse valor é uma média obtida por estudos realizados em 2024 [IBM Security 2024]. Porém, no setor da saúde, esses valores podem ser muito maiores, chegando a uma média global de ≈R$ 50. Apenas com essas informações, já seria possível compreender e quantificar alguns riscos e impactos que podem auxiliar na priorização.

Além disso, métricas técnicas podem ser utilizadas para compreender quais ciberataques e vulnerabilidades possuem a maior probabilidade de acontecer no mundo real. O EPSS, por exemplo, estima a probabilidade de que uma vulnerabilidade seja explorada na prática nos próximos 30 dias. Ele combina dados públicos, como Common Vulnerabilities and Exposures (CVE)[6] e histórico de exploração, para ajudar organizações a priorizarem a correção de falhas com maior risco real. Imagina que a clínica de exames mencionada acima possui diferentes sistemas com possíveis impactos técnicos, econômicos, legais e sociais em caso de um ciberincidente. É importante mapear os riscos de ciberataques específicos em cada sistema, além dos impactos para cada ciberataque em cada sistema. Na Tabela 1.2 é apresentada uma análise inicial dos principais sistemas da clínica e as vulnerabilidades encontradas. Para isso, foi utilizada a métrica EPSS para definir a probabilidade de uma vulnerabilidade (ou seja, CVE) ser explorada. Com essa informação, podemos definir o risco de cada vulnerabilidade para o sistema. Porém, além do risco de uma vulnerabilidade acontecer, precisamos correlacionar também com os impactos. Por exemplo, uma vulnerabilidade com alto risco de ser explorada, mas com um impacto baixo não deveria possuir uma prioridade alta.

| **Vulnerabilidade** | **Sistema Impactado** | **EPSS** | **Probabilidade** |
|---|---|---|---|
| CVE-2025-Exemplo | Portal de Agendamento de Consultas | 85% | Alto |
| CVE-2022-Exemplo | Servidor com Dados de Pacientes e Exames (PACS/RIS) | 15% | Baixo |
| CVE-2023-Exemplo | Servidor de Laudos e Impressão de Exames | 65% | Alto |
| CVE-2024-Exemplo | Portal Web da Clínica | 5% | Baixo |
| CVE-2021-Exemplo | Página de Agendamento e Tabela de Exames | 30% | Médio |

**Tabela 1.2. Exemplos de Vulnerabilidades e Riscos Mapeados para os Sistemas da Clínica de Exames**

Ao analisar a Tabela 1.2, observamos que o Servidor com Dados de Pacientes e Exames (PACS/RIS) foi definido como risco Baixo, pois possui um EPSS de 15% (ou seja, possui 15% de chance de a vulnerabilidade ser explorada nos próximos 30 dias).

[6]https://www.cve.org/

Porém, como o servidor é um sistema crítico para a clínica, deveríamos alterar a prioridade para Alto, baseada nos possíveis impactos em caso de um incidente, enquanto a Página de Agendamento e Tabela de Exames poderiam ser alteradas para uma prioridade Baixa, por exemplo, já que o risco é Moderado, mas o impacto será Baixo nos sistemas e informações críticas da clínica. É possível usar outras métricas técnicas para compreender a gravidade dos riscos, porém, é fundamental ter em mente o contexto do setor e do negócio. No caso da saúde, por exemplo, precisamos priorizar a vida e também reduzir os possíveis impactos econômicos no negócio que podem surgir com falhas técnicas e questões de conformidade regulatória ou jurídicas.

Ao utilizar modelos econômicos, simulações e métricas técnicas disponíveis na indústria e academia, podemos priorizar a correção de problemas e a proteção de sistemas e informações críticas de forma a otimizar o planejamento e investimento em cibersegurança. Porém, tal tarefa não é simples, já que cada abordagem exige conhecimento técnico e possui diferentes curvas de aprendizagem. Tal fato é ainda mais crítico para setores que possuem usuários e profissionais com pouca experiência em TI, além de poucos profissionais dedicados à cibersegurança, como o caso do setor da saúde. Para isso, existem esforços para propor ferramentas que auxiliem no processo de compreensão dos riscos e aplicação automatizada de modelos, simulações e técnicas para quantificação de impactos e otimização de investimentos em cibersegurança. No resto desta seção, será conduzido um caso de estudo para o setor da saúde, permitindo assim o planejamento de uma estratégia de cibersegurança seguindo as etapas definidas ao longo da seção.

### 1.5.3. Caso de Estudo

Inicialmente, seguindo a abordagem definida na Figura 1.8, devemos compreender o negócio. Portanto, suponha um Laboratório de Análises Clínicas (LAC) situado no Brasil, com 50 funcionários e que atue diretamente como prestador de serviços para clientes privados e hospitais. O LAC possui cerca de 10% de seus funcionários atuando remotamente e com experiência básica na operação de TI. Além disso, embora o LAC não tenha sofrido nenhum ciberataque no último ano, alguns de seus funcionários já foram vítimas de ataques de phishing executados com sucesso.

Os sistemas disponíveis no LAC incluem um servidor de banco de dados para armazenamento de informações, exames e procedimentos de rotina, além de uma página web para agendamento de exames. Como o LAC já segmenta suas atividades, iremos focar apenas nesses dois ativos apenas, sem considerar os equipamentos médicos e de exames que estão sob cuidados de outro setor da empresa. A Figura 1.9 apresenta uma visão geral das proteções já implementadas na empresa.

Na segunda etapa, é necessária a modelagem das ameaças e a quantificação de riscos. Primeiramente, precisamos compreender os riscos inerentes ao cenário onde a empresa está situada e, então, conduzir a análise de riscos. Para isso, podemos utilizar histórico de incidentes no setor e em parceiros. Ao verificar os relatórios oferecidos pela plataforma *IMPACTO* (ver Figura 1.10), que foi desenvolvida no contexto do programa Hackers do Bem[7] para capacitação e planejamento em cibersegurança [8], verificamos que o setor da saúde possui uma alta taxa de phishing e também possui uma tendência de ataques

[7]https://hackersdobem.org.br

[8]https://www.inf.ufrgs.br/gt-impacto/


**Figura 1.9. Proteções Implementadas de Forma Geral pelo LAC**

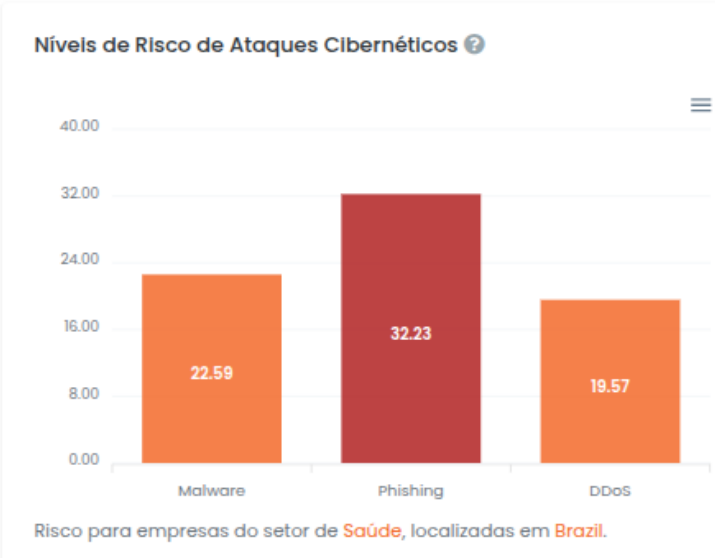



**Figura 1.10. Exemplo de Média de Ataques no Setor da Saúde e no Brasil Conforme Relatórios Publicamente Disponíveis**

de malware. Tais dados confirmam que precisamos ter uma estratégia clara para evitar problemas de malwares específicos (por exemplo, ransomware) e que possuem ataques de phishing como parte sua estratégia de propagação. Como diversos funcionários da empresa já foram vítimas de ataques de phishing com sucesso no passado, isso também

reforça a necessidade de proteção contra phishing. Por fim, observamos também uma incidência razoável de DDoS no setor, principalmente em empresas de saúde localizadas no Brasil.

Tais ameaças possuem impactos diretos nos ativos. Por exemplo, ataques de phishing e ransomware podem ser utilizados para tornar indisponíveis os sistemas de banco de dados e também utilizados como forma de vazamento de dados e extorsão. Já ataques de DDoS podem tornar inacessível o portal de agendamento de exames e consulta de resultados. Os possíveis impactos de cada ataque nos ativos são apresentados na Tabela 1.3.

**Tabela 1.3. Exemplo de Mapeamento de Impactos e Ataques por Ativo**

| Impacto | Ativo | Ataque(s) | Descrição |
|---|---|---|---|
| Médio | Portal de Agendamentos | DDoS, Ransomware | Faturamento de R$ 200 mil por mês, disrupção do negócio afeta diretamente os agendamentos, perda de reputação |
| Alto | Banco de Dados | Ransomware | Disrupção do negócio, Multas por vazamento de dados, perda de reputação e extorsão |
| Alto | Funcionários e Clientes | Phishing | Roubo de credenciais e Acesso a sistemas e informações sensíveis, perda de reputação, inicio de outros ataques, como, por exemplo, o ransomware |

Podemos quantificar os riscos econômicos utilizando estratégias de, por exemplo, verificar o quanto um ativo pode afetar os ganhos de uma empresa caso fique inacessível ou mesmo consultar relatórios que apresentam a média de perda financeira por cada dado vazado. Já os riscos aos pacientes e procedimentos precisam ser investigados com cautela, já que envolvem vidas e o apetite para risco não deve existir. Devemos também considerar nessa etapa demandas específicas da LGPD e do HIPAA, para garantir que não existem requisitos específicos não cumpridos. Nesse caso, identificamos que nem todo dado em trânsito está sendo criptografado, permitindo que ataques acessem dados de exames em alguns cenários, o que viola boas práticas.

Para identificar a real exposição aos riscos, é possível utilizar ferramentas para escanear a rede e sistemas, como o Nmap ou mesmo ferramentas pagas como o Nessus Tenable. Tais ferramentas auxiliam a mapear os ativos expostos e também possíveis vulnerabilidades que podem ser exploradas por atacantes. Além disso, métricas como o EPSS (ver Seção 1.5.2) auxiliam na identificação de quais vulnerabilidades são mais prováveis de serem exploradas e quais devem ser priorizadas.

Ao identificar os impactos e riscos, podemos definir quais proteções deverão ser priorizadas, levando em consideração aspectos técnicos (efetividade das proteções) e também econômicos (custos e orçamento disponível). Para definir o orçamento ideal, utilizaremos o modelo de Gordon-Loeb [Gordon et al. 2016], que pode ser utilizado como um benchmark para definir o investimento ótimo em cibersegurança. Ao adicionar os ativos, seu valor para o LAC e os riscos, o modelo de Gordon-Loeb define que o inves-

timento ótimo em cibersegurança deverá ser de R$ 87 mil para proteção contra malware, R$ 75 mil contra phishing e R$ 57 mil contra DDoS. Tais valores foram definidos através de simulações executadas utilizando a plataforma GT-IMPACTO e consideram como entrada valores hipotéticos como o lucro estimado da empresa e a importância de cada ativo para o faturamento e seus potenciais impactos financeiros. Além disso, são considerados os riscos de ataques com sucesso e potenciais impactos financeiros para definição do investimento ótimo. É importante lembrar que tal cálculo considera apenas os fatores econômicos e técnicos, desconsiderando especificamente os impactos sociais que são extremamente importantes no setor da saúde. Para isso, é importante levar em consideração os riscos para pacientes e vidas humanas durante a priorização dos investimentos. Encontrar um balanço entre o investimento ótimo e a redução de riscos críticos para a vida humana é fundamental para uma estratégia eficiente do ponto de vista técnico e econômico.

**Tabela 1.4. Exemplos de Proteções baseado nos Requisitos de Cibersegurança Mapeados, incluindo Custos e Justificativas para os Investimentos**

| Tipo de Proteção | Custo Anual (R$) | Justificativa do Investimento |
|---|---|---|
| Proteção contra DDoS | R$ 15.000 | Previne interrupções de serviço causadas por DDoS, garantindo disponibilidade dos sistemas críticos |
| Segurança de E-mail contra Phishing | R$ 6.000 | Reduz o risco de comprometimento de credenciais e infecção por malware via e-mails maliciosos, protegendo dados sensíveis |
| Anti-Virus e Anti-Malware (Endpoints) | R$ 4.500 | Garante a proteção dos dispositivos da empresa contra ameaças conhecidas e zero-day, reduzindo riscos de vazamento e interrupção |
| Gestão de Patches e Atualizações | R$ 10.000 | Automatiza a aplicação de atualizações de segurança, corrigindo vulnerabilidades conhecidas e melhorando a postura de segurança |
| Treinamento e Conscientização em Segurança | R$ 30.000 | Capacita os colaboradores para reconhecerem ameaças digitais, reduzindo riscos humanos e fortalecendo a cultura de segurança |
| Criptografia de Dados em Trânsito | R$ 20.000 | Protege dados sensíveis durante a comunicação entre sistemas e banco de dados, garantindo confidencialidade e integridade |
| Conformidade com LGPD e HIPAA | R$ 30.000 | Garante que os processos da empresa estejam alinhados com legislações de privacidade, evitando multas e prejuízos de reputação |
| **Total Estimado** | **R$ 115.500** | |

Com o orçamento definido, podemos verificar quais proteções podem ser aplicadas no LAC para *(i)* reduzir os riscos de ransomware que têm como alvo o banco de dados e serviços críticos, *(ii)* diminuir a chance de funcionários e pacientes serem vítimas de phishing, *(iii)* mitigar os riscos de DDoS em serviços essenciais para o negócio e *(iv)* evitar problemas de conformidade e processos jurídicos devido a incidentes e vazamentos de dados. A Tabela 1.4 apresenta um planejamento inicial de investimento, totalizando R$ 115,5 mil de investimentos para proteção, sendo R$ 15 mil contra DDoS, R$ 50,5 mil contra phishing e malware e R$ 50 mil para adequação e verificação de conformidade. Tais valores são exemplos e ainda existe a possibilidade de aumentar proteções já que,

por exemplo, o valor para proteção contra DDoS está usando apenas 25% do valor ótimo sugerido por modelos econômicos.

Por fim, a última fase envolve a execução e a implantação da estratégia de cibersegurança. Se ainda não existir a experiência necessária na empresa, o suporte técnico pode ser obtido por meio da contratação de consultores. Além disso, é preciso definir um cronograma claro de implementação, pois alguns setores da empresa podem precisar interromper suas operações por algumas horas para implementar totalmente as soluções e os novos processos. Com isso, é possível compreender os diferentes fatores que devem ser considerados durante o planejamento e investimento em cibersegurança. É importante ressaltar que diversas ferramentas da indústria e da academia podem ser utilizadas para apoiar o processo de decisão, sendo fundamental priorizar também os elementos críticos para o setor e considerar os equipamentos e protocolos legados, como acontece no setor da saúde.

## 1.6. Conclusões e Lições Aprendidas

Neste capítulo, analisamos o cenário de cibersegurança no setor da saúde, um dos mais vulneráveis e visados para ação de cibercriminosos. A convergência entre a alta criticidade dos serviços prestados, o elevado valor dos dados sensíveis e a adoção acelerada de tecnologias digitais torna esse setor um alvo recorrente de ataques cibernéticos. A análise das ameaças, vulnerabilidades e estudos de caso evidencia um ponto crítico: a cibersegurança na saúde ainda não acompanha a velocidade da inovação tecnológica, e isso gera um risco sistêmico de grande impacto técnico, econômico, social e humano.

Portanto, assim como em outros setores críticos, os desafios enfrentados por hospitais, clínicas e demais instituições de saúde vão além da dimensão técnica. Existem gargalos estruturais relacionados à governança, escassez de profissionais especializados, cultura organizacional despreparada e falta de investimentos em cibersegurança. Além disso, embora as regulamentações, normativas e fiscalizações tenham evoluído, as boas práticas de segurança da informação ainda estão longe de ser uma realidade consolidada no setor da saúde.

O aumento de dispositivos conectados no setor, como por exemplo sensores, *wearables* e equipamentos médicos inteligentes, tem ampliado significativamente a superfície de ataque. Por exemplo, podemos observar diversos dispositivos médicos, servidores PACS e equipamentos de imagens publicamente vulneráveis. Os vazamentos de dados, sejam de órgãos públicos ou privados, e os ataques de phishing diretamente ao paciente estão cada vez mais frequentes no setor.

É importante que a tecnologia da informação e todos os envolvidos no setor da saúde compreendam que, em um setor onde vidas humanas estão diretamente em jogo, não há espaço para colocarmos os sistemas e dados em risco. A cibersegurança na saúde é, antes de tudo, uma questão de responsabilidade ética, social e profissional. Assim, esse capítulo tem como objetivo servir de ponto de partida para gestores, pesquisadores e profissionais que buscam transformar a cibersegurança em um aliado estratégico para a proteção e a sustentabilidade dos serviços de saúde.

Como caminhos futuros, entende-se que, em curto prazo, a conscientização e treinamento no setor serão cruciais e o principal componente para melhorarmos a cibersegu-

rança. Em médio prazo, será necessário um plano estratégico do setor para otimizar os processos já existentes, como a configuração básica de serviços visando à cibersegurança e também uma maior atenção para a proteção de dados. Também, será necessário adicionar camadas de proteção adicional em serviços legados, já que existem equipamentos e serviços antigos operando que não foram projetados para as ameaças do mundo atual, mas que cumprem muito bem suas aplicações na área da saúde. Por fim, em longo prazo, precisamos de políticas rígidas para incentivar o setor a investir em cibersegurança, bem como propor mecanismos que auxiliem a tornar os sistemas legados menos ossificados do ponto de vista de cibersegurança.

## Referências

Abou Jaoude, J. and Saade, R. G. (2019). Blockchain applications–usage in different domains. *Ieee Access*, 7:45360–45381.

Adebukola, A., Navya, A., Jordan, F., Jenifer, N., and Begley, R. D. (2022). Cyber security as a threat to health care. *Journal of Technology and Systems*, 4(1):32–64.

Adil, M., Khan, M. K., Kumar, N., Attique, M., Farouk, A., Guizani, M., and Jin, Z. (2024). Healthcare internet of things: Security threats, challenges, and future research directions. *IEEE Internet of Things Journal*, 11(11):19046–19069.

Akkaoui, R., Hei, X., and Cheng, W. (2020). Edgemedichain: A hybrid edge blockchain-based framework for health data exchange. *IEEE access*, 8:113467–113486.

Alder, S. (2024). Healthcare Data Breaches Due to Phishing. https://www.hipaajournal.com/healthcare-data-breaches-due-to-phishing/.

Aljedaani, B. and Babar, M. A. (2021). Challenges With Developing Secure Mobile Health Applications: Systematic Review. *JMIR mHealth and uHealth*, 9(6):e15654.

Aouedi, O., Sacco, A., Piamrat, K., and Marchetto, G. (2023). Handling privacy-sensitive medical data with federated learning: Challenges and future directions. *IEEE Journal of Biomedical and Health Informatics*, 27(2):790–803.

Aragão, S. M. d. and Schiocchet, T. (2020). Lei Geral de Proteção de Dados: Desafio do Sistema Único de Saúde. *Revista Eletrônica de Comunicação, Informação e Inovação em Saúde*, 14(3).

Arbabi, M. S., Lal, C., Veeraragavan, N. R., Marijan, D., Nygård, J. F., and Vitenberg, R. (2023). A survey on blockchain for healthcare: Challenges, benefits, and future directions. *IEEE Communications Surveys Tutorials*, 25(1):386–424.

Arctic Wolf Labs (2025). 2025 Cybersecurity Predictions. https://arcticwolf.com/arctic-wolf-labs-2025-cybersecurity-predictions/.

Associated Press (2020). German hospital hacked, patient taken to another city dies. https://apnews.com/article/technology-hacking-europe-cf8f8eee1adcec69bcc864f2c4308c94.

Autoridade Nacional de Proteção de Dados (2023). ANPD aplica a primeira multa por descumprimento à LGPD. https:

//www.gov.br/anpd/pt-br/assuntos/noticias/anpd-aplica-a-primeira-multa-por-descumprimento-a-lgpd.

Beckers, K., Heisel, M., and Hatebur, D. (2015). Pattern and security requirements. *Pattern Secur. Requir. Eng. Establ. Secur. Stand*, pages 1–474.

Beerman, J., Berent, D., Falter, Z., and Bhunia, S. (2023). A review of colonial pipeline ransomware attack. In *IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW 2023)*, pages 8–15.

BigID (2024). A cost comparison of data breaches. https://bigid.com/blog/a-cost-comparison-of-data-breaches/.

Boritz, J. E. (2005). Is practitioners' views on core concepts of information integrity. *International Journal of Accounting Information Systems*, 6(4):260–279.

Calegari, L. (2025). ANPD não multou empresas por violação da LGPD em 2024. https://valor.globo.com/legislacao/noticia/2025/01/29/anpd-nao-multou-empresas-por-violacao-da-lgpd-em-2024.ghtml.

Cambricoli, F. (2020). Nova falha do Ministério da Saúde Expõe Dados Pessoais de mais de 200 Milhões de Brasileiros. https://tinyurl.com/falha-sus-200milhoes.

Cartwright, A. J. (2023). The Elephant in the Room: Cybersecurity in Healthcare. *Journal of Clinical Monitoring and Computing*, 37(5):1123–1132.

Chamberlaim, A., de Azevedo Flor, B., da Silva Pereira, E., Almeida, L. S., Martins, L. D., Silva, Y. S., Siqueira, G. G., Maiczak, T., and Bovo, F. (2023). Inteligência artificial (ia) e suas aplicações em exames de imagem: uma nova era para diagnósticos na área da saúde. *Cuadernos de Educación y Desarrollo*, 15(12):17605–17624.

CheckPoint (2025). The State of Cyber Security 2025. https://engage.checkpoint.com/security-report-2025.

Coventry, L. and Branley, D. (2018). Cybersecurity in Healthcare: a Narrative Review of Trends, Threats and Ways Forward . *International Journal of Midlife Health and Beyond (MATURITAS)*, (113):48–52.

Cybersecurity Ventures (2024). Global ransomware damage costs predicted to reach $275 billion by 2031. https://cybersecurityventures.com/global-ransomware-damage-costs-predicted-to-reach-250-billion-usd-by-

Cybersecurity Ventures, Herjavec Group (2021). The 2020-2021 Healthcare Cybersecurity Report. https://www.herjavecgroup.com/2021-healthcare-cybersecurity-report-cybersecurity-ventures/.

De Neira, A. B., Kantarci, B., and Nogueira, M. (2023). Distributed denial of service attack prediction: Challenges, open issues and opportunities. *Computer Networks*, 222:109553.

Ding, R., Zhong, H., Ma, J., Liu, X., and Ning, J. (2019). Lightweight privacy-preserving identity-based verifiable iot-based health storage system. *IEEE Internet of Things Journal*, 6(5):8393–8405.

EC-Councl (2022). DREAD Threat Modeling: An Introduction to Qualitative Risk Analysis. https://www.eccouncil.org/cybersecurity-exchange/threat-intelligence/dread-threat-modeling-intro/.

Edemekong, P., Annamaraju, P., Afzal, M., and Haydel, M. (2024). Health insurance portability and accountability act (hipaa) compliance. *StatPearls*.

European Commission (2024). EU Action Plan to Increase Healthcare Cybersecurity. https://healthcare-in-europe.com/en/news/eu-action-plan-increase-healthcare-cybersecurity.html.

Fonseca, F. (2025). Saúde é Setor que mais Sofre Ataque Cibernético. https://valor.globo.com/publicacoes/especiais/inovacao-na-medicina/noticia/2025/02/27/saude-e-setor-que-mais-sofre-ataque-cibernetico.ghtml.

Franco, M., Rodrigues, B., Killer, C., Scheid, E. J., De Carli, A., Gassmann, A., Schoenbaechler, D., and Stiller, B. (2021). WeTrace: a Privacy-preserving Tracing Approach. *Journal of Communications and Networks*, 1(1):1–16.

Franco, M. F., Granville, L. Z., and Stiller, B. (2023a). CyberTEA: a Technical and Economic Approach for Cybersecurity Planning and Investment. In *36th IEEE/IFIP Network Operations and Management Symposium (NOMS 2023)*, pages 1–6, Miami, USA.

Franco, M. F., Granville, L. Z., and Stiller, B. (2023b). CyberTEA: a Technical and Economic Approach for Cybersecurity Planning and Investment. In *36th IEEE/IFIP Network Operations and Management Symposium (NOMS 2023)*, pages 1–6, Miami, USA.

Franco, M. F., Künzler, F., von der Assen, J., Feng, C., and Stiller, B. (2024a). RCVaR: an Economic Approach to Estimate Cyberattacks Costs using Data from Industry Reports. *Computers & Security*, page 103737.

Franco, M. F., Lacerda, F. M., and Stiller, B. (2022). A framework for the planning and management of cybersecurity projects in small and medium-sized enterprises. *Journal of Business and Projects (Revista de Gestão e Projetos)*, 13(3):1–25.

Franco, M. F., Mullick, A. R., and Jha, S. (2024b). QBER: Quantifying Cyber Risks for Strategic Decisions. https://arxiv.org/abs/2405.03513.

Fuentes, M. R. and Huq, N. (2018). Securing Connected Hospitals: A Research on Exposed Medical Systems and Supply Chain Risks. https://documents.trendmicro.com/assets/rpt/rpt-securing-connected-hospitals.pdf.

Gallopeni, G., Rodrigues, B., Franco, M., and Stiller, B. (2020). A Practical Analysis on Mirai Botnet Traffic. In *2020 IFIP Networking Conference (Networking)*, pages 667–668. IEEE.

GDPR.EU Horizon 2020 (2021). Complete guide to GDPR compliance. https://gdpr.eu/.

Ghafur, S., Kristensen, S., Honeyford, K., Martin, G., Darzi, A., and Aylin, P. (2019). A retrospective impact analysis of the WannaCry cyberattack on the NHS. *npj Digital Medicine*, 2:98.

Gordon, L. A., Loeb, M. P., Lucyshyn, W., and Zhou, L. (2018). Empirical evidence on the determinants of cybersecurity investments in private sector firms. *Journal of Information Security*, 9(2):49–61.

Gordon, L. A., Loeb, M. P., and Zhou, L. (2016). Investing in Cybersecurity: Insights from the Gordon-Loeb Model. *Journal of Information Security*, 7:49–59.

Gupta, A., Tripathi, M., Shaikh, T. J., and Sharma, A. (2019). A lightweight anonymous user authentication and key establishment scheme for wearable devices. *Computer Networks*, 149:29–42.

He, D. and Zeadally, S. (2014). An analysis of rfid authentication schemes for internet of things in healthcare environment using elliptic curve cryptography. *IEEE internet of things journal*, 2(1):72–83.

IBM Security (2024). Cost of a data breach report 2024. `https://www.ibm.com/reports/data-breach`.

Islam, T., Sheakh, M. A., Jui, A. N., Sharif, O., and Hasan, M. Z. (2023). A review of cyber attacks on sensors and perception systems in autonomous vehicle. *Journal of Economy and Technology*, 1:242–258.

J. R. Reeder, P. F. McQuade, S. A. Schipma (2021). Cybersecurity's Pearl Harbor Moment: Lessons Learned from the Colonial Pipeline Ransomware Attack. *Greenberg-Traurig Data, Privacy  Cybersecurity*, 1:1–25.

Jacobs, J., Romanosky, S., Edwards, B., Adjerid, I., and Roytman, M. (2021). Exploit Prediction Scoring System (EPSS). *Digital Threats: Research and Practice*, 2(3):1–17.

Javaid, M., Haleem, A., Singh, R. P., and Suman, R. (2023). Towards insighting cybersecurity for healthcare domains: A comprehensive review of recent practices and trends. *Cyber Security and Applications*, 1:100016.

Knight, A. V. (2021). All That We Let In: Hacking 30 Mobile Health Apps and APIs . `https://approov.io/info/all-that-we-let-in-hacking-30-mobile-health-apps-and-apis`.

Koch, R. (2020). What is the LGPD? Brazil's version of the GDPR. `https://gdpr.eu/gdpr-vs-lgpd/`.

L. A. Gordon, M. P. Loeb, L. Zhou (2021). Information Segmentation and Investing in Cybersecurity. *Journal of Information Security*, 12:115–136.

Levina, A., Iliashenko, V. M., Kalyazina, S., and Overes, E. (2022). Smart hospital architecture: It and digital aspects. In Jahn, C., Ungvári, L., and Ilin, I., editors, *Algorithms and Solutions Based on Computer Technology*, pages 235–247, Cham. Springer International Publishing.

Lewis, D., Lasek-Markey, M., Golpayegani, D., and Pandit, H. J. (2025). Mapping the regulatory learning space for the eu ai act. *arXiv preprint arXiv:2503.05787*.

Microsoft (2022). Microsoft Threat Modeling Tool Threats. https://learn.microsoft.com/en-us/azure/security/develop/threat-modeling-tool-threats/.

Mirsky, Y., Mahler, T., Shelef, I., and Elovici, Y. (2019). CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning. In *28th USENIX Conference on Security Symposium*, SEC'19, page 461–478, USA. USENIX Association.

Mishra, V. (2024). Cyberattacks on healthcare: A global threat that can't be ignored. https://news.un.org/en/story/2024/11/1156751.

Nankya, M., Mugisa, A., Usman, Y., Upadhyay, A., and Chataut, R. (2024). Security and privacy in e-health systems: A review of ai and machine learning techniques. *IEEE Access*, 12:148796–148816.

National Audit Office (2018). Investigation: WannaCry Cyber Attack and the NHS. https://www.nao.org.uk/wp-content/uploads/2017/10/Investigation-WannaCry-cyber-attack-and-the-NHS.pdf.

Navruzov, E. and Kabulov, A. (2022). Detection and analysis types of ddos attack. In *2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pages 1–7. IEEE.

Neprash, H. T., McGlave, C. C., Cross, D. A., Virnig, B. A., Puskarich, M. A., Huling, J. D., Rozenshtein, A. Z., and Nikpay, S. S. (2022). Trends in Ransomware Attacks on US Hospitals, Clinics, and Other Health Care Delivery Organizations, 2016-2021. *JAMA Health Forum*, 3(12):e224873.

NHS Foundation Trust (2023). NHS England business continuity management toolkit case study: WannaCry attack. https://www.england.nhs.uk/long-read/case-study-wannacry-attack/.

Nunes, J., Franco, M., Scheid, E., Kozenieski, G., Lindemann, H., Soares, L., Nobre, J., and Granville, L. (2024). SIM-Ciber: Uma Soluçao Baseada em Simulaçoes Probabilısticas para Quantificaçao de Riscos e Impactos de Ciberataques Utilizando Relatórios Estatısticos. *Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg)*, pages 570–585.

O Globo (2024). Em um mês, autoridade brasileira de dados abre mais investigações contra empresas do que em quatro anos. Acessado em 8 de maio de 2025.

Olsen, E. (2025). UnitedHealth hikes number of Change cyberattack breach victims to 190 million. https://www.healthcaredive.com/news/change-healthcare-cyberattack-affects-190-million-unitedhealth/738351/.

Ostad-Sharif, A., Abbasinezhad-Mood, D., and Nikooghadam, M. (2019). A robust and efficient ecc-based mutual authentication and session key generation scheme for healthcare applications. *Journal of medical systems*, 43(1):10.

Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., and Jerram, C. (2013). Phishing for the Truth: A Scenario-Based Experiment of Users' Behavioural Response to Emails. In *Security and Privacy Protection in Information Processing Systems*, pages 366–378, Berlin, Heidelberg. Springer.

R. Mccrimmon and M. Matishak (2021). Cyberattack on Food Supply Followed Years of Warnings. https://www.politico.com/news/2021/06/05/how-ransomware-hackers-came-for-americans-beef-491936.

Claroty (2025). State of cps security: Healthcare exposures 2025. https://claroty.com/resources/reports/state-of-cps-security-healthcare-exposures-2025.

Healthcare Information and Management Systems Society (2024). 2024 himss healthcare cybersecurity survey. https://cdn.sanity.io/files/sqo8bpt9/production/4f1c1968050411b8bf9335a187301881f9153b9f.pdf.

HIMSS, FinThrive (2025). Survey Reveals Cybersecurity Funding is a Top Challenge for Smaller Hospitals. https://tinyurl.com/finthrive.

Runte, C. (2024). GDPR Enforcement Tracker Report. https://cms.law/en/gbr/publication/gdpr-enforcement-tracker-report.

Santos, J. A., Inacio, P. R., and Silva, B. M. (2021). Towards the use of blockchain in mobile health services and applications. *Journal of Medical Systems*, 45(2):17.

Scheid, E. J., Knecht, A., Strasser, T., Killer, C., Franco, M., Rodrigues, B., and Stiller, B. (2021a). Edge2BC: a Practical Approach for Edge-to-Blockchain IoT Transactions. In *IEEE International Conference on Blockchain and Cryptocurrency (ICBC 2021)*, pages 1–9.

Scheid, E. J., Rodrigues, B., Killer, C., Franco, M., Niya, S. R., and Stiller, B. (2021b). *Blockchains and Distributed Ledgers Uncovered: Clarifications, Achievements, and Open Issues*, pages 1–29. Number 1 in IFIP AICT Festschrifts. Springer, Cham, Switzerland.

Secureworks (2024). Boardroom cybersecurity report 2024. https://www.secureworks.com/centers/boardroom-cybersecurity-report-2024.

Singer, P. W. and Friedman, A. (2013). *Cybersecurity and Cyberwar: What Everyone Needs to Know®*. Oxford University Press.

Smart, W. (2018). Lessons learned review of the WannaCry Ransomware Cyber Attack.

Soares, L. R., Nobre, J. C., and Kerschner, G. (2023). Design of a blockchain-based secure storage architecture for resource-constrained healthcare. In *2023 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6.

Sun, Y., Lo, F. P.-W., and Lo, B. (2019). Security and privacy for the internet of medical things enabled healthcare systems: A survey. *IEEE Access*, 7:183339–183355.

The Telegraph (2018). WannaCry cyber attack cost NHS £92m as 19,000 appointments cancelled.

Thyagarajan, C., S.Suresh, Sathish, N., and Suthir, S. (2020). A Typical Analysis And Survey On Healthcare Cyber Security. *International Journal of Scientific Technology Research*, 9(3):1–5.

Todde, M., Beltrame, M., Marceglia, S., and Spagno, C. (2020). Methodology and Workflow to Perform the Data Protection Impact Assessment in Healthcare Information Systems. *Informatics in Medicine Unlocked*, 19:100361.

United States Department of Health and Human Services. Breach Portal: Notice to the Secretary of HHS Breach of Unsecured Protected Health Information. https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf.

United States Department of Health and Human Services (2013). Breach Notification Rule. https://www.hhs.gov/hipaa/for-professionals/breach-notification/index.html.

United States Department of Health and Human Services (2024a). Social Engineering Attacks Targeting the HPH Sector. https://www.hhs.gov/sites/default/files/social-engineering-targeting-the-hph-sector-tlpclear.pdf.

United States Department of Health and Human Services (2024b). Solara Medical Supplies, LLC Resolution Agreement and Corrective Action Plan. https://www.hhs.gov/hipaa/for-professionals/compliance-enforcement/agreements/solara-ra-cap/index.html.

US Health Care Industry Cybersecurity Task Force (2017). Report On Improving Cybersecurity in the Health Care Industry. https://www.phe.gov/Preparedness/planning/CyberTF/Documents/report2017.pdf.

U.S. Government (1996). Health insurance portability and accountability act of 1996. https://www.govinfo.gov/content/pkg/PLAW-104publ191/html/PLAW-104publ191.htm.

Verizon Business (2025). 2025 data breach investigations report. https://www.verizon.com/business/resources/reports/dbir/.

Wang, M., Guo, Y., Zhang, C., Wang, C., Huang, H., and Jia, X. (2021). Medshare: A privacy-preserving medical data sharing system by using blockchain. *IEEE Transactions on Services Computing*, 16(1):438–451.

Wen, J., Zhang, Z., Lan, Y., Cui, Z., Cai, J., and Zhang, W. (2023). A survey on federated learning: challenges and applications. *International Journal of Machine Learning and Cybernetics*, 14(2):513–535.

Wolford, B. (2020). What are the GDPR Fines? https://gdpr.eu/fines/.

Young, A. and Yung, M. (1996). Cryptovirology: Extortion-based security threats and countermeasures. In *Proceedings 1996 IEEE Symposium on Security and Privacy*, pages 129–140. IEEE.

Zhang, J., Yang, Y., Liu, X., and Ma, J. (2022). An efficient blockchain-based hierarchical data sharing for healthcare internet of things. *IEEE Transactions on Industrial Informatics*, 18(10):7139–7150.

Todos os links foram visitados em maio de 2025.

# Capítulo 2

# Acesso e Recuperação de Dados Biomédicos no MIMIC-IV

Willian de Vargas, André Gonçalves Jardim, Viviane Rodrigues Botelho, Thatiane Alves Pianoschi, Ana Trindade Winck

***Abstract***

*This chapter provides a practical and accessible guide for accessing and retrieving clinical data from the MIMIC-IV database, one of the leading open-source biomedical information sources. It covers the database's modular structure, local access strategies via PostgreSQL and SQLite, and the use of SQL and Python for exploratory analyses. The text includes comprehensive tutorials for data import, building derived views, and extracting relevant clinical subsets. The goal is to empower healthcare professionals and researchers to use real clinical data in reproducible studies, promoting technical autonomy and interdisciplinarity in data science applied to healthcare.*

***Resumo***

*Este capítulo apresenta um guia prático e acessível para acesso e recuperação de dados clínicos a partir do banco de dados MIMIC-IV, uma das principais fontes abertas de informações biomédicas. São abordadas a estrutura modular do banco, estratégias de acesso local via PostgreSQL e SQLite, e o uso de SQL e Python para análises exploratórias. O texto inclui tutoriais completos para importação dos dados, construção de visões derivadas e extração de subconjuntos clínicos relevantes. O objetivo é capacitar profissionais da saúde e pesquisadores a utilizarem dados clínicos reais em estudos reprodutíveis, promovendo a autonomia técnica e a interdisciplinaridade em ciência de dados aplicada à saúde.*

## 2.1. Introdução

Bancos de dados clínicos abertos têm desempenhado um papel fundamental no avanço da pesquisa em saúde, oferecendo acesso a informações para investigação científica e desenvolvimento de tecnologias para a área da saúde. Dentre os repositórios mais conhecidos, destaca-se o *PhysioNet*, mantido pelo Laboratório de Fisiologia Computacional do *Massachusetts Institute of Technology* (MIT). Essa iniciativa disponibiliza dados biomédicos anonimizados de atendimentos hospitalares reais, com o objetivo de fomentar o desenvolvimento de pesquisas científicas e soluções tecnológicas aplicadas à medicina.

Um dos conjuntos de dados mais completos e utilizados da plataforma *PhysioNet* é o MIMIC (*Medical Information Mart for Intensive Care*), atualmente em sua quarta versão, o MIMIC-IV. Publicada em 2024, essa versão reúne dados clínicos anonimizados de cerca de 315 mil pacientes atendidos no hospital *Beth Israel Deaconess Medical Center*, incluindo internações hospitalares e atendimentos em unidades de terapia intensiva entre 2008 e 2019. O banco está dividido em módulos, oferecendo acesso a informações detalhadas como sinais vitais, exames laboratoriais, prescrições, procedimentos e notas clínicas, além de permitir análises temporais e multivariadas.

Apesar de seu grande potencial, a complexidade estrutural do MIMIC-IV pode representar uma barreira significativa para profissionais da saúde e pesquisadores com pouca familiaridade em ciência de dados e bancos de dados relacionais. Este capítulo propõe um olhar prático e acessível sobre o processo de acesso, configuração e exploração dos dados do MIMIC-IV, utilizando ferramentas como PostgreSQL, SQLite, SQL e Python. São apresentados tutoriais para importar os dados, realizar análises exploratórias e extrair subconjuntos relevantes para estudos clínicos, com o objetivo de contribuir para a democratização do uso do MIMIC-IV na comunidade científica brasileira, promovendo a interdisciplinaridade e a autonomia técnica dos profissionais interessados em aplicar ciência de dados à saúde.

## 2.2. O Banco de Dados MIMIC

O *Medical Information Mart for Intensive Care* (MIMIC) é um banco de dados relacional disponibilizado pela plataforma *PhysioNet*. Desenvolvido a partir de uma colaboração entre o hospital *Beth Israel Deaconess Medical Center* e o laboratório de Fisiologia Computacional do MIT (MIT-LCP), o MIMIC tem como objetivo fornecer acesso gratuito e anonimizado a dados clínicos reais de pacientes que passaram por atendimento em unidades de terapia intensiva (UTIs) deste hospital [Johnson et al. 2024].

A iniciativa do MIMIC surgiu da necessidade de disponibilizar conjuntos de dados clínicos detalhados para pesquisadores, promovendo o avanço da medicina baseada em evidências e o desenvolvimento de ferramentas computacionais para suporte à decisão médica. Desde sua primeira versão, lançada em 2003, o projeto evoluiu até chegar à versão mais recente, o MIMIC-IV, que oferece uma estrutura de dados mais moderna e dividida em módulos [Johnson et al. 2024].

Entre os módulos disponibilizados pelo MIMIC-IV, destacam-se dois conjuntos principais de tabelas: o módulo HOSP e o módulo ICU. Esses módulos estruturam as informações de maneira complementar, permitindo analises tanto em nível geral de inter-

nação quanto em situações de cuidados intensivos.

O módulo HOSP abrange os dados hospitalares gerais, incluindo informações administrativas e clínicas ao longo da internação do paciente. Esse módulo contém tabelas com resultados de exames laboratoriais, prescrições médicas, procedimentos realizados, diagnósticos, notas clínicas, registros demográficos e administrativos, entre outros.

O módulo ICU é focado nos dados coletados especificamente durante a permanência dos pacientes em unidades de terapia intensiva (UTI). Ele inclui medições frequentes de sinais vitais, fluidos, intervenções realizadas, uso de dispositivos médicos, entre outros dados críticos que refletem o estado clínico dos pacientes em tempo quase real.

Essa divisão entre os módulos permite análises direcionadas e mais eficientes. Pesquisadores podem optar por explorar aspectos mais amplos da internação hospitalar utilizando os dados do módulo HOSP ou concentrar-se em episódios de maior gravidade, com base nos registros detalhados das UTIs presentes no módulo ICU. Essa flexibilidade é fundamental para o desenvolvimento de estudos clínicos com diferentes escopos.

Para facilitar o uso e promover a padronização das análises realizadas com o MIMIC-IV, foi desenvolvido o repositório MIMIC-IV *Concepts*, que disponibiliza uma coleção de consultas SQL pré-definidas e reutilizáveis, denominadas *concepts*. Essas consultas implementam definições clínicas comumente utilizadas em pesquisas, como ventilação mecânica, uso de vasopressores, comorbidades e critérios diagnósticos, permitindo que diferentes estudos adotem critérios uniformes. Um exemplo prático é a *view vitalSign*, que consolida medições frequentes de sinais vitais dos pacientes, como frequência cardíaca, pressão arterial, temperatura e saturação de oxigênio. Essa *view* organiza os dados de forma estruturada a partir de múltiplas tabelas do módulo ICU, simplificando o acesso a informações essenciais para a avaliação do estado clínico dos pacientes ao longo do tempo.

### 2.2.1. Histórico e Evolução do MIMIC

O banco de dados MIMIC teve sua primeira versão disponibilizada em 2003, inicialmente como uma iniciativa modesta, composta por dados clínicos de pacientes internados em UTIs do *Beth Israel Deaconess Medical Center* (BIDMC). Desde então, o projeto passou por diversas fases de expansão e refinamento, resultando nas versões MIMIC-II, MIMIC-III e, mais recentemente, no MIMIC-IV, lançado oficialmente em 2024 [Johnson et al. 2024].

O MIMIC-II trouxe melhorias significativas em relação à padronização dos dados e ao suporte para estudos retrospectivos com foco em desfechos clínicos. Já o MIMIC-III, amplamente adotado pela comunidade científica, consolidou o banco como uma das principais fontes públicas de dados clínicos [Johnson et al. 2016], abrangendo mais de 60.000 admissões em UTI entre os anos de 2001 e 2012. Ele introduziu melhorias na documentação, maior detalhamento nas tabelas e a inclusão de notas clínicas desidentificadas, ampliando o potencial de uso em pesquisas com Processamento de Linguagem Natural (PLN).

A versão mais atual do *MIMIC-IV* abrange hospitalizações e atendimentos no departamento de emergência do *Beth Israel Deaconess Medical Center* entre 2008 e 2019,

reunindo dados de aproximadamente 315 mil pacientes, totalizando mais de 524 mil admissões hospitalares. O *MIMIC-IV* é composto por registros eletrônicos de saúde (*Electronic Health Record* - EHRs) e sistemas de gestão hospitalar, como o *MetaVision*, e inclui informações demográficas, sinais vitais, resultados laboratoriais, administração de medicamentos, diagnósticos, procedimentos, observações clínicas e notas desidentificadas. Sua organização modular e a incorporação de dados do departamento de emergência ampliam o escopo da base para além das UTIs, permitindo estudos mais robustos e com maior representatividade, inclusive em análises de trajetórias clínicas desde a entrada no hospital até o desfecho final [Johnson et al. 2024].

Todos os registros são anonimizados, e os identificadores dos pacientes são substituídos por códigos, de modo a proteger a privacidade e atender às exigências éticas e legais para o uso dos dados em pesquisas. A utilização do MIMIC exige que os pesquisadores concluam um curso de proteção de dados e apresentem uma proposta de pesquisa aprovada [Johnson et al. 2024].

Uma das grandes vantagens do MIMIC é a riqueza e diversidade dos dados. Além dos dados estruturados, como exames laboratoriais e medicações de sinais vitais, ele também inclui notas clínicas em linguagem natural, o que permite pesquisas avançadas utilizando técnicas de Processamento de Linguagem Natural. Essa variedade viabiliza uma ampla gama de estudos, desde análises epidemiológicas e avaliações de desempenho hospitalar até aplicações de aprendizado de máquina para predição de desfechos clínicos.

A relevância do MIMIC para a pesquisa biomédica pode ser observada na quantidade crescente de estudos que o utilizam como base. Uma busca nas tendências de publicação do PubMed [National Library of Medicine (US) 2025] pelos termos "MIMIC-I", “MIMIC-II”, “MIMIC-III” ou “MIMIC-IV” revela um aumento consistente nas publicações ao longo da última década, refletindo o crescente interesse da comunidade científica nessa base de dados, conforme apresentado na Figura 2.1.

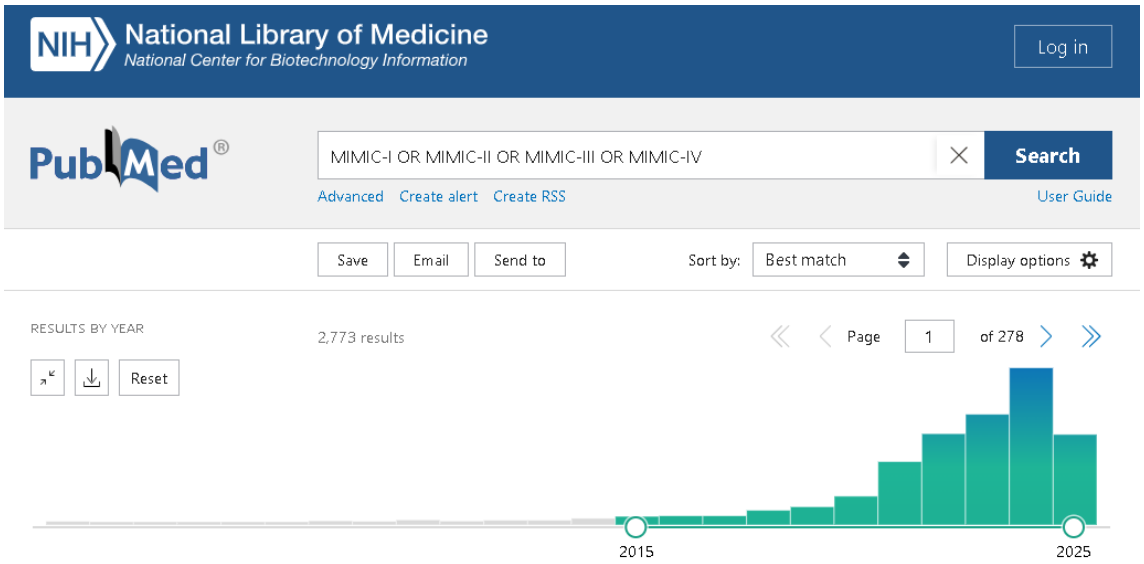



**Figura 2.1. Tendência de publicações no PubMed contendo os termos *MIMIC-I*, *MIMIC-II*, *MIMIC-III* ou *MIMIC-IV*, no período de 2015 a 2025. [National Library of Medicine (US) 2025]**

### 2.2.2. Exemplos da utilização do MIMIC em estudos científicos

A versatilidade do MIMIC é evidenciada pelo crescente número de estudos científicos que utilizam seus dados para desenvolver modelos preditivos, avaliar riscos clínicos e propor intervenções baseadas em evidências. A seguir, são apresentados exemplos de aplicações práticas recentes com base na literatura.

Em [Jung et al. 2024] foram explorados fatores preditivos para a progressão da insuficiência cardíaca em pacientes hipertensos, utilizando exclusivamente dados de diagnósticos prévios à hipertensão presentes no banco MIMIC-IV. O objetivo foi possibilitar a antecipação do risco de insuficiência cardíaca no momento do diagnóstico de hipertensão. Para isso, os autores aplicaram testes qui-quadrado e modelos baseados em *XGBoost* para analisar fatores preditivos específicos por faixa etária. Os resultados revelaram um conjunto de condições associadas ao agravamento da insuficiência cardíaca, incluindo fibrilação atrial, insuficiência renal, doença pulmonar obstrutiva crônica, anemia e uso de anticoagulantes. A abordagem adotada contribui para o monitoramento personalizado e a estratificação de risco desses pacientes.

No campo das doenças infecciosas, em [Pérez-Tome et al. 2024] foi desenvolvido um modelo preditivo de mortalidade por sepse utilizando técnicas de aprendizado de máquina, com foco específico no algoritmo *Random Forest*. O estudo foi conduzido com dados de pacientes internados em unidades de terapia intensiva (UTIs) de três hospitais espanhóis, além de pacientes provenientes da base de dados MIMIC-III, totalizando 4.739 pacientes (180 locais e 4.559 do MIMIC-III). O objetivo foi construir um classificador capaz de prever o risco de óbito em pacientes com sepse, a partir de variáveis fisiológicas e laboratoriais coletadas durante a internação. Os resultados demonstraram elevado desempenho do modelo, com acurácia de 96,77% e área sob a curva ROC (AUC) de 95% no conjunto de dados local, e ainda melhor desempenho com os dados do MIMIC-III: acurácia de 98,28% e AUC de 97,3%. Entre as variáveis mais relevantes para a predição da mortalidade por sepse, destacaram-se os níveis de lactato, o débito urinário e os parâmetros relacionados ao equilíbrio ácido-base. Curiosamente, os níveis de potássio foram mais determinantes na base MIMIC-III do que nos dados dos hospitais espanhóis, o que pode refletir particularidades clínicas ou laboratoriais específicas do contexto norte-americano. Esses achados demonstram não apenas a viabilidade de modelos de aprendizado de máquina na estratificação de risco em pacientes sépticos, como também reforçam a utilidade do MIMIC como um repositório confiável e robusto para o desenvolvimento e validação de ferramentas de apoio à decisão clínica em ambientes de terapia intensiva.

Em um estudo voltado para complicações renais em pacientes críticos, em [Lin et al. 2024] foram desenvolvidos e validados modelos preditivos de lesão renal aguda (LRA) em pacientes com pancreatite aguda (PA) utilizando dados do banco MIMIC-IV. Foram analisados 1.235 pacientes internados com PA, dos quais 54% desenvolveram LRA durante a hospitalização. Sete algoritmos de aprendizado de máquina foram aplicados para a construção dos modelos: *Random Forest*, *Support Vector Machine* (SVM), *K-Nearest Neighbors*(KNN), *Naive Bayes* (NB), *Neural Network* (NNET), *Generalized Linear Model* (GLM) e *Gradient Boosting Machine* (GBM). O modelo baseado em GBM apresentou o melhor desempenho, com AUC de 0.867 no conjunto de teste, indicando alta capacidade discriminativa para prever a ocorrência de LRA. Os autores destacam o potencial

desses modelos para apoiar decisões clínicas ao identificar precocemente pacientes em risco, promovendo intervenções oportunas e potencialmente reduzindo a mortalidade em unidades de terapia intensiva.

Ainda no contexto da sepse, em [Sun et al. 2024] foi desenvolvido e validado um nomograma preditivo (ferramenta gráfica que estima a probabilidade de um desfecho clínico com base em múltiplas variáveis) para estimar a mortalidade em 30 dias de pacientes com sepse associada a sangramento gastrointestinal (GIB), utilizando dados do banco MIMIC-IV. O estudo retrospectivo incluiu 1.435 pacientes, divididos aleatoriamente em coortes de treinamento e validação. Para a construção do modelo, os autores aplicaram regressão LASSO para seleção de variáveis e regressão logística multivariada para estimativa dos riscos. O desempenho do modelo foi avaliado por meio do índice de concordância (C-index), curvas ROC e análise de decisão (DCA), apresentando boa capacidade discriminativa tanto no conjunto de treinamento (C-index: 0.746) quanto de validação (C-index: 0.716). O nomograma incorporou variáveis como idade, tabagismo, glicose, ureia (BUN), lactato, escore SOFA, ventilação mecânica $\geq 48h$, nutrição parenteral e DPOC como fatores preditivos independentes. Os autores destacam o potencial clínico do modelo como uma ferramenta de apoio à tomada de decisões individualizadas no tratamento de pacientes críticos com sepse e GIB.

Em [Fan et al. 2025] foi realizado um estudo de coorte retrospectivo utilizando dados de 5.110 pacientes com diferentes tipos de acidente vascular cerebral (AVC), extraídos do banco de dados MIMIC-IV, abrangendo o período de 2010 a 2020. O objetivo do estudo foi identificar os principais fatores associados à mortalidade em curto, médio e longo prazos — especificamente em 30 dias, 90 dias, 1 ano e 3 anos após o evento. As análises estatísticas incluíram modelos de regressão de Cox, *Random Forest* e *Gradient Boosting*, que revelaram a importância de variáveis como tipo de AVC, índice de comorbidades de Charlson, escore SOFA, níveis de hemoglobina, idade avançada, tempo de internação e disfunções orgânicas. O estudo mostrou, por exemplo, que um aumento no escore SOFA está relacionado a maior risco de mortalidade em todos os períodos analisados. Esses achados oferecem informações valiosas para o aprimoramento da estratificação de risco, planejamento terapêutico e definição de estratégias de acompanhamento individualizado em pacientes com AVC.

Esses estudos ilustram não apenas a variedade de aplicações do MIMIC em diferentes domínios clínicos, mas também sua contribuição efetiva para o desenvolvimento de modelos de apoio à decisão médica baseados em dados reais. A capacidade de acessar dados clínicos estruturados e não estruturados, como exames laboratoriais, sinais vitais e notas clínicas, permite a construção de soluções que atendem a necessidades específicas em contextos de cuidados intensivos. Além disso, a ampla janela temporal e a riqueza demográfica da base possibilitam a análise de tendências, estratificação de risco e identificação precoce de eventos adversos, tornando o MIMIC uma ferramenta estratégica para a medicina de precisão.

Graças à sua qualidade, abrangência e acessibilidade, o MIMIC tornou-se um recurso bastante relevante para pesquisadores de diversas áreas, incluindo medicina, enfermagem, engenharia biomédica, ciência da computação e estatística. O acesso aberto e gratuito à base, aliado ao seu contínuo aprimoramento técnico e documental, facilita a re-

alização de estudos reprodutíveis, estimula colaborações interdisciplinares e democratiza o uso de dados clínicos em projetos acadêmicos e de inovação tecnológica. Essa abertura tem incentivado a formação de comunidades científicas em torno do uso do MIMIC, promovendo a troca de conhecimento e boas práticas no desenvolvimento de soluções orientadas por dados.

Portanto, a compreensão do funcionamento do MIMIC, suas origens, estrutura e potenciais de aplicação pode ser bastante útil para profissionais e pesquisadores interessados em explorar a interface entre dados clínicos e de saúde. O domínio dessa base de dados permite não apenas a realização de estudos retrospectivos com alto valor científico, mas também a criação de ferramentas com impacto direto na prática clínica. Com isso, o MIMIC se consolida como um elemento central na formação de uma nova geração de soluções em saúde, baseadas em ciência de dados e evidências clínicas reais.

## 2.3. Acesso ao MIMIC-IV

O acesso ao banco de dados MIMIC-IV pode ser realizado por diferentes meios, a depender das necessidades do usuário e dos recursos disponíveis. Esta seção apresenta as opções de acesso, os pré-requisitos para obtenção dos dados e um tutorial para configurar um ambiente local e importar os dados utilizando o PostgreSQL.

### 2.3.1. Formas de Acesso

Atualmente, existem duas formas principais de acessar os dados do MIMIC-IV:

- **Acesso Local**: envolve o download dos arquivos no formato CSV e posterior importação para um sistema gerenciador de banco de dados (SGBD), como PostgreSQL ou SQLite.
- **Acesso em Nuvem**: permite o uso de instâncias públicas hospedadas na Amazon Web Services (AWS) ou Google Cloud Platform (GCP), geralmente através de ferramentas como BigQuery.

Embora o acesso em nuvem seja conveniente para aplicações de produção ou análise de grandes volumes de dados, este tutorial terá como foco o acesso local, pois ele proporciona maior controle de custos, portabilidade e independência de infraestrutura externa.

### 2.3.2. Pré-requisitos para Acesso aos Dados

Para acessar a versão completa do MIMIC-IV, é necessário:

- Criar uma conta no portal PhysioNet (`https://physionet.org/`).
- Realizar um curso sobre ética em pesquisa com seres humanos, como o CITI Program (`https://physionet.org/about/citi-course/`).
- Submeter o certificado de conclusão e aceitar formalmente o termo de uso de dados (Data Use Agreement – DUA), disponível no próprio portal.

Se você está apenas explorando a estrutura do banco ou deseja seguir os exemplos deste tutorial, também é possível utilizar a versão Demo do MIMIC-IV, que é totalmente aberta e não exige cadastro ou certificação.

#### 2.3.2.1. Download dos Dados

Na página de download os dados são distribuídos no formato `.csv.gz`, organizados por domínios como `hosp` e `icu`. Existem duas opções de download, uma que é a versão completa dos dados (Figura 2.2) e outra que é a versão de demonstração dos dados (Figura 2.3):

**Versão Completa do MIMIC-IV**

1. Acesse `https://physionet.org/content/mimiciv/`
2. Clique em "Files"
3. Baixe o arquivo ZIP correspondente

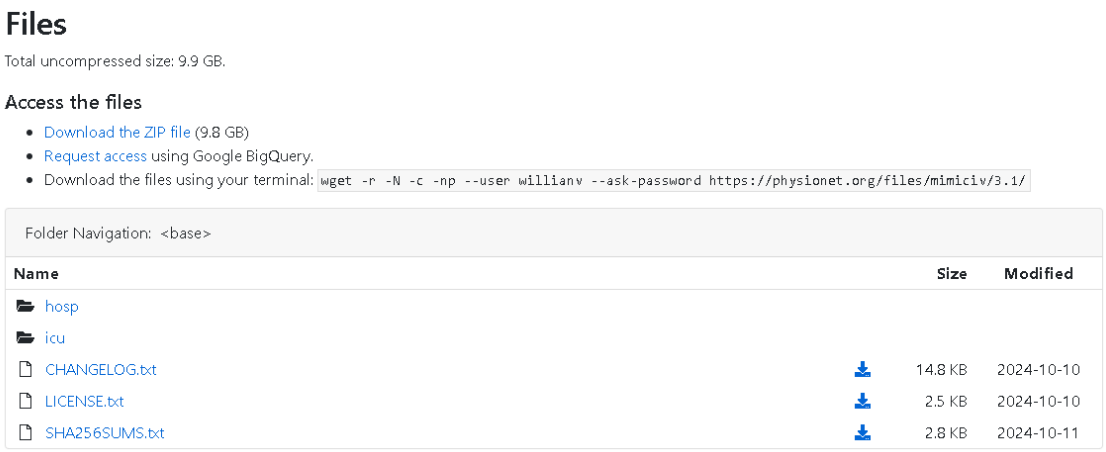



**Figura 2.2. Página de download da versão completa do MIMIC-IV.**

**Versão Demo do MIMIC-IV**

1. Acesse `https://physionet.org/content/mimic-iv-demo/`
2. Clique em "Files"
3. Baixe o arquivo ZIP correspondente

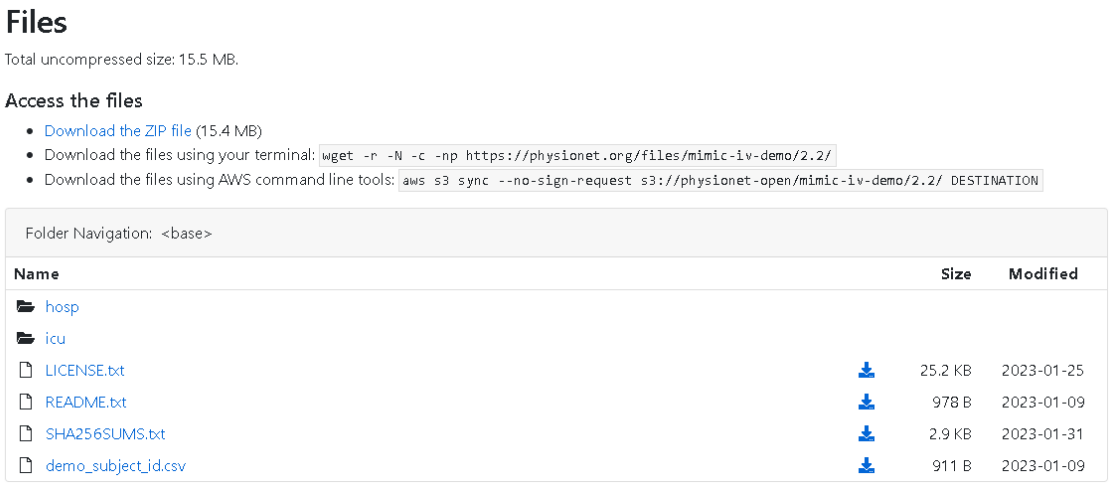



**Figura 2.3. Página de download da versão demo do MIMIC-IV.**

### 2.3.3. Configuração do Ambiente Local usando PostgreSQL *database*

A seguir, apresentamos o processo completo para configurar um ambiente local e importar os dados do MIMIC-IV utilizando o PostgreSQL. Os passos estão organizados para usuários de diferentes sistemas operacionais.

#### Passo 1. Instalar o PostgreSQL

#### Linux (Ubuntu):

```
sudo apt update
sudo apt install postgresql postgresql-contrib
```

Após a instalação, o serviço do PostgreSQL será iniciado automaticamente. Você pode verificar o status com:

```
sudo systemctl status postgresql
```

#### Windows/macOS:

Baixe o instalador em: `https://www.postgresql.org/download/`

Após instalar, verifique se tudo está instalado corretamente executando o seguinte comando no seu terminal:

```
psql --version
```

#### Passo 2. Criar um Banco de Dados

Antes de importar os dados do MIMIC-IV, precisamos criar um banco de dados vazio no PostgreSQL. Isso pode ser feito de duas formas:

- Via terminal (linha de comando)
- Via interface gráfica, como o pgAdmin (opcional, para quem preferir evitar o terminal)

Vamos primeiro explicar o método via terminal, que é o mais direto e compatível com os scripts de importação.

**A. Via Terminal:**

Primeiramente, abra seu terminal:

- Linux (Ubuntu): Pressione `Ctrl + Alt + T` ou procure por "Terminal" no menu de aplicativos
- Windows: Pressione `Win + R`, digite cmd ou powershell e pressione Enter
- macOS: Abra o Terminal pela busca (`Command + Espaço` → digite "Terminal")

Digite o seguinte comando para criar o banco de dados

```
createdb mimiciv
```

Esse comando cria um banco de dados chamado mimiciv, onde os dados do MIMIC-IV serão importados.

Se você receber um erro como:

```
createdb: command not found
```

ou

```
could not connect to database postgres: FATAL: role "seu_usuario
    " does not exist
```

isso pode indicar que o PostgreSQL não está configurado para o seu usuário do sistema.

**Alternativa: usando o usuário postgres**

O PostgreSQL, por padrão, cria um usuário chamado postgres. Você pode usar esse usuário para criar o banco com o comando abaixo (Linux/Mac):

```
sudo -u postgres createdb mimiciv
```

Esse comando pede a senha do seu sistema (não a do banco de dados).

Se estiver no Windows e quiser usar o usuário postgres, você pode: Abrir o terminal do psql (o cliente interativo do PostgreSQL):

```
psql -U postgres
```

Criar o banco dentro do terminal do PostgreSQL:

```
CREATE DATABASE mimiciv;
\q  -- para sair
```

**B. Via pgAdmin:**

Se você prefere evitar o terminal, também pode criar o banco pelo pgAdmin, a interface web oficial do PostgreSQL.

Quando você instala o PostgreSQL no seu computador, o pgAdmin geralmente é instalado automaticamente. Para abri-lo:

- Windows: Procure por pgAdmin no menu iniciar e clique para abrir.
- macOS/Linux: Você pode buscar pelo pgAdmin 4 na lista de aplicativos ou executá-lo a partir do terminal com o comando pgadmin4 (caso esteja instalado via terminal).

**Conectar ao Servidor PostgreSQL:**

Ao abrir o pgAdmin, a tela inicial será exibida. Caso você não tenha conectado a nenhum servidor, será solicitado para adicionar uma nova conexão.

Clique em **"Add New Server"** (Adicionar Novo Servidor).

Na janela que abrir, insira os seguintes dados:

- **General (Figura 2.4):**

  Name: Escolha um nome para a conexão, como PostgreSQL Local ou algo de sua escolha.

- **Connection (Figura 2.5):**

  Host: localhost (se estiver utilizando a instalação local).

  Port: 5432 (porta padrão do PostgreSQL).

  Username: postgres (ou o nome de usuário configurado no PostgreSQL).

  Password: A senha do seu usuário PostgreSQL.

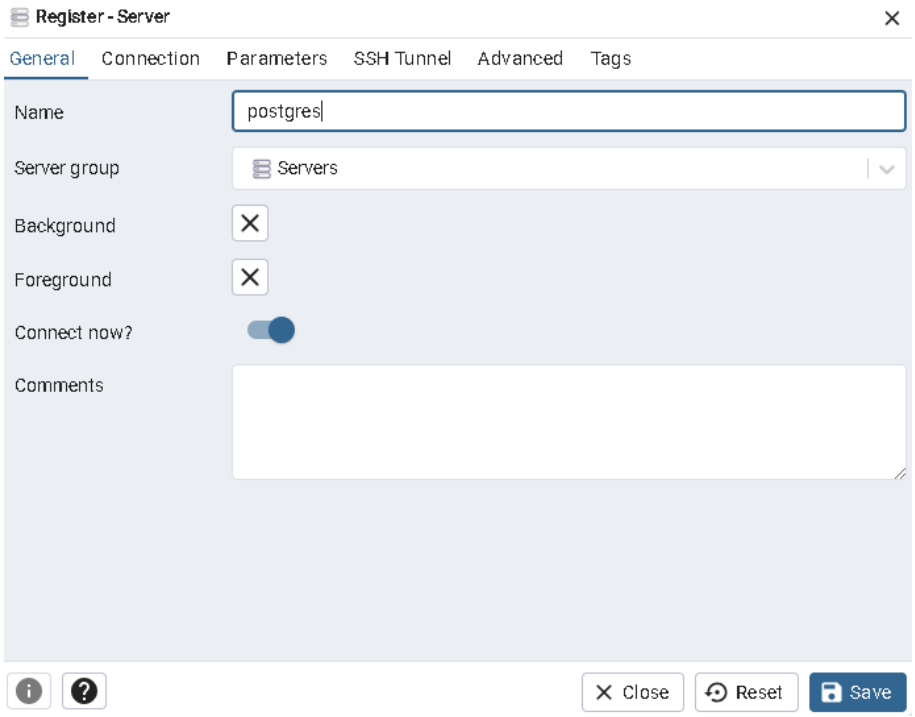



**Figura 2.4. Captura de tela da aba "General"da tela de criação de um novo Server.**

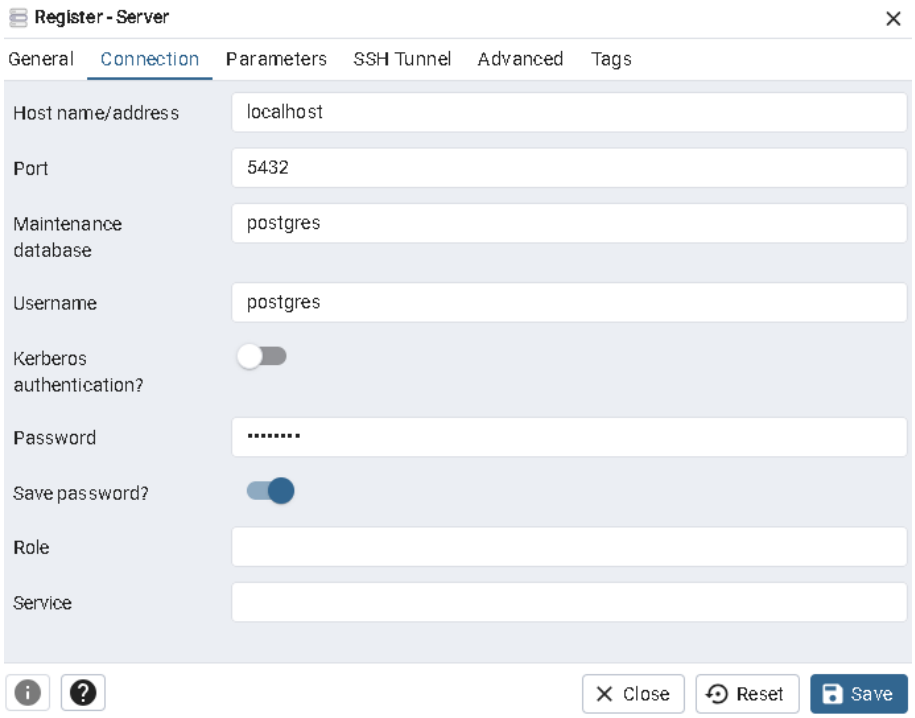



**Figura 2.5. Captura de tela da aba "Connection"da tela de criação de um novo Server.**

Após preencher os campos, clique em Save para estabelecer a conexão com o servidor.

**Criar um Banco de Dados:**

Após conectar-se ao servidor, o pgAdmin mostrará o painel de navegação à esquerda, com o nome do seu servidor na árvore.

Expanda o servidor recém-adicionado clicando sobre ele e, em seguida, sobre a pasta Databases.

Clique com o botão direito em Databases e selecione Create > Database..., como demonstrado na Figura 2.6

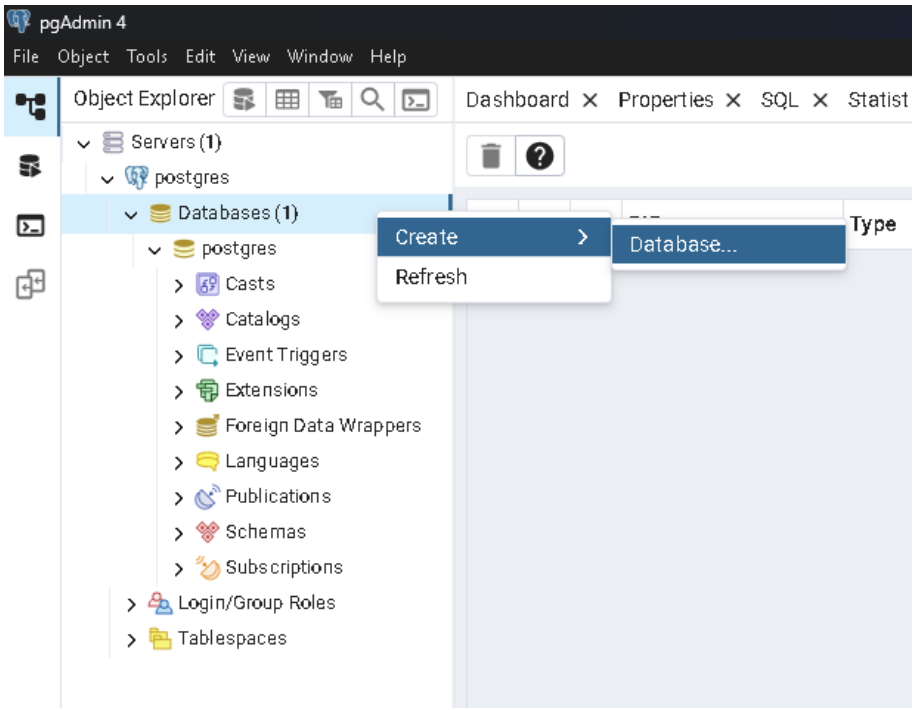



**Figura 2.6. Captura de tela mostrando a opção de criação de uma nova base de dados no pgAdmin.**

Na janela Create - Database, insira as seguintes informações:

- Database: Nomeie o banco de dados como mimiciv.
- Deixe as outras configurações no padrão.

Após preencher o nome do banco, clique em Save para criar o banco de dados.

**Verificar o Banco Criado:**

Após criar o banco, ele aparecerá na lista de bancos de dados dentro do servidor (Figura 2.7).

Clique sobre o banco de dados mimiciv para expandir sua estrutura e começar a trabalhar com ele.

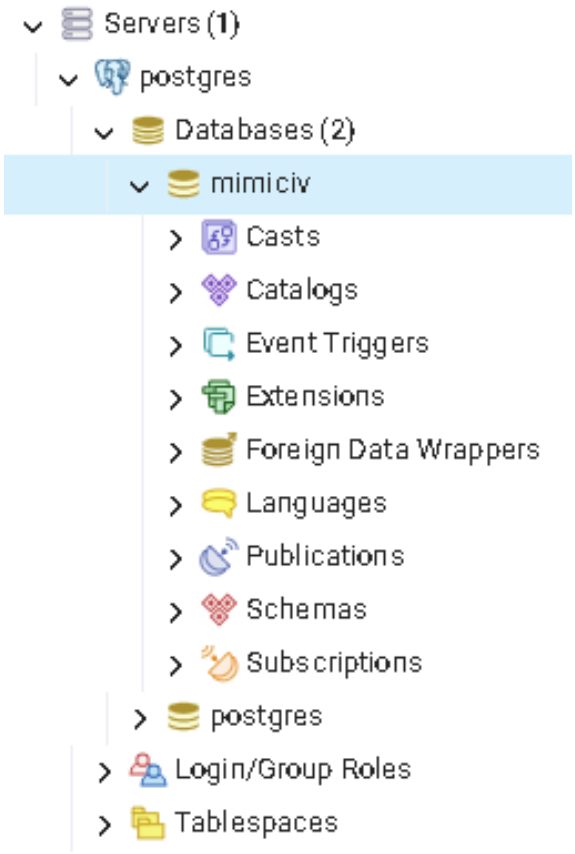



**Figura 2.7. Captura de tela mostrando o novo banco de dados criado.**

### Passo 3. Baixar os Scripts Oficiais

Para importar os dados corretamente no PostgreSQL, você precisa dos scripts SQL que criam as tabelas e realizam a carga dos arquivos CSV. Esses scripts são fornecidos oficialmente pelo grupo MIT-LCP no repositório GitHub:

Repositório: `https://github.com/MIT-LCP/mimic-code`

Você tem duas opções para obter os arquivos:

### Opção 1: Baixar ZIP

- Acesse o link acima.
- Clique no botão verde "Code".
- Escolha a opção "Download ZIP".
- Após o download, extraia o conteúdo do arquivo ZIP para uma pasta no seu computador. Por exemplo:

  `C:/mimic-code` no Windows

  `~/Documentos/mimic-code` no Linux/macOS

### Opção 2: Clonar com Git

Se você tem o Git instalado em seu computador, pode clonar o repositório diretamente com o comando:

```
git clone https://github.com/MIT-LCP/mimic-code.git
```

Esse comando criará uma pasta chamada mimic-code com todos os arquivos dentro. A vantagem dessa abordagem é que, caso os scripts sejam atualizados no futuro, você poderá baixar as atualizações simplesmente rodando:

```
git pull
```

#### Organização dos Arquivos

Depois de baixar ou clonar, a estrutura da pasta será como a seguir:

```
mimic-iv-3.1/
|-- mimic-iii/
|   \-- ...
|-- mimic-iv/
|   |-- buildmimic/
|      |-- postgres/
|         |-- constraint.sql
|         |-- create.sql
|         |-- index.sql
|         \-- load.sql
|      |-- sqlite/
|      \-- ...
|   |-- concepts/
|   \-- ...
\-- ...
```

Esses são os principais arquivos que usaremos para criar as tabelas e importar os dados para o banco de dados mimiciv.

### Passo 4. Descompactar os Arquivos CSV

Após o download do arquivo ZIP (realizado previamente), seja na versão completa ou na Demo, extraia o arquivo com um descompactador de arquivos de sua preferência.

**Estrutura dos Dados Locais**

Após o desempacotamento do arquivo ZIP, o diretório da base terá a seguinte estrutura:

```
mimic-iv-3.1/
|-- hosp/
|   |-- admissions.csv.gz
|   |-- labevents.csv.gz
|   |-- patients.csv.gz
|   |-- transfers.csv.gz
|   \-- ...
|-- icu/
|   |-- chartevents.csv.gz
|   |-- icustays.csv.gz
|   |-- procedureevents.csv.gz
|   \-- ...
```

Cada subpasta representa um "módulo"do banco de dados (core, hosp, icu), com arquivos CSV comprimidos. A importação converte esses arquivos em tabelas relacionais dentro do SGBD de sua escolha.

### Passo 5. Criar as Tabelas no PostgreSQL

O script create.sql do repositório oficial cria todos os esquemas e tabelas necessários no banco mimiciv. Ele está localizado no seguinte caminho dentro do repositório: mimic-code/mimic-iv/buildmimic/postgres/create.sql

Para executar o script, use o seguinte comando no seu terminal:

```
psql -U [seu_usuario] -d mimiciv -f [caminho_local]/mimic-code/
   mimic-iv/buildmimic/postgres/create.sql
```

Substitua:

- seu_usuario: seu nome de usuário do PostgreSQL (por padrão, pode ser postgres).
- caminho_local: caminho absoluto para o arquivo no seu sistema.

### Passo 6. Importar os Arquivos CSV

Os scripts load_gz.sql (caso os arquivos a serem importados sejam .gz) ou load.sql (caso os arquivos a serem importados sejam .csv) importam os arquivos diretamente no banco, por meio de comandos COPY. Estes scripts estão localizados no seguinte caminho dentro do repositório: mimic-code/mimic-iv/buildmimic/postgres/load.sql e mimic-code/mimic-iv/buildmimic/postgres/load_gz.sql

Antes de executar, edite esses scripts em um editor de texto e ajuste todos os caminhos dos arquivos (.csv.gz ou .csv) para apontar corretamente para o local onde você extraiu os arquivos em seu computador.

Exemplo de linha antes de ajustar:

```
\COPY mimiciv_hosp.admissions FROM admissions.csv DELIMITER ','
   CSV HEADER NULL '';
```

Exemplo após ajuste:

```
\COPY mimiciv_hosp.admissions FROM 'C:/Users/usuario/Downloads/
   mimic-iv-3.1/hosp/admissions.csv.gz' DELIMITER ',' CSV HEADER
    NULL '';
```

Para executar o script, use um dos seguintes comandos no seu terminal:

```
psql -U [seu_usuario] -d mimiciv -f [caminho_local]/mimic-code/
   mimic-iv/buildmimic/postgres/load.sql
```

```
psql -U [seu_usuario] -d mimiciv -f [caminho_local]/mimic-code/
   mimic-iv/buildmimic/postgres/load_gz.sql
```

No caso da importação da versão completa do MIMIC-IV, este processo pode levar muitas horas. Isso é completamente normal e esperado, considerando o alto volume de dados a serem importados. É importante manter com computador ligado durante todo o processo.

### Passo 7. Verificar a Importação

Após importar os dados com sucesso, é importante validar que:

- Os dados estão presentes em cada tabela.
- É possível fazer consultas simples com o SQL.

Para isso, utilizaremos o cliente de linha de comando do PostgreSQL, chamado psql.

No terminal, execute:

```
psql -U seu_usuario -d mimiciv
```

Substitua seu_usuario pelo seu usuário PostgreSQL, geralmente postgres.

Se tudo estiver funcionando corretamente, você verá algo assim:

```
mimiciv=#
```

Dentro do terminal interativo do psql, execute o seguinte comando para verificar se as tabelas realmente existem:

```
\dt mimiciv.*
```

Esse comando lista todas as tabelas criadas nos esquemas do MIMIC-IV (como hosp e icu).

Você deverá ver uma saída como esta (resumo):

```
 Schema |       Name        | Type  |  Owner
--------+-------------------+-------+----------
 hosp   | admissions        | table | postgres
 hosp   | diagnoses_icd     | table | postgres
 ...
 icu    | caregiver       | table | postgres
 icu    | chartevents       | table | postgres
 ...
```

Agora vamos executar consultas simples para contar os registros e verificar se as tabelas têm dados:

```
SELECT COUNT(*) FROM mimiciv.hosp.patients;
SELECT COUNT(*) FROM mimiciv.hosp.admissions;
SELECT COUNT(*) FROM mimiciv.hosp.diagnoses_icd;
SELECT COUNT(*) FROM mimiciv.icu.chartevents;
```

Os resultados variam entre a versão completa e a Demo, mas devem sempre ser maiores que zero.

### 2.3.4. Configuração do Ambiente Local usando SQLite

Além do uso de bancos relacionais completos como o PostgreSQL, o MIMIC-IV também pode ser importado para um banco de dados SQLite, o que pode ser útil para fins de exploração leve dos dados, testes rápidos ou situações em que uma instalação de servidor de banco de dados não é viável.

A equipe do PhysioNet fornece dois scripts para facilitar essa importação: import.sh, escrito em shell script POSIX, e import.py, escrito em Python. Ambos permitem gerar um arquivo SQLite contendo todas as tabelas do MIMIC-IV a partir dos arquivos CSV ou CSV.GZ disponibilizados.

#### Requisitos

Para utilizar o script import.sh, são necessários:

- Shell POSIX compatível (ex: `bash`, `zsh`, `dash`)
- SQLite[1]

[1] `https://sqlite.org/index.html`

- gzip (geralmente já instalado por padrão em sistemas Linux, BSD e macOS)

Para utilizar o script `import.py`, é necessário:

- Python 3 instalado
- Biblioteca pandas[2]

### Estrutura de diretórios

Os scripts devem estar localizados na mesma pasta onde estão os arquivos CSV do MIMIC-IV. A estrutura recomendada é:

```
path/to/mimic-iv/
|-- import.sh
|-- import.py
|-- hosp/
|   |- admissions.csv.gz
|   |- ...
|   \- transfers.csv.gz
\-- icu/
    |- chartevents.csv.gz
    |- ...
    \_ procedureevents.csv.gz
```

### Execução dos scripts

Para gerar o banco de dados SQLite, basta executar um dos scripts desejados diretamente no terminal:

```
$ ./import.sh
```

ou

```
$ python import.py
```

A execução completa pode demorar, especialmente durante o carregamento da tabela `chartevents`, que é volumosa.

Ao final do processo, será gerado um arquivo `mimic4.db`, que pode ser aberto com qualquer ferramenta compatível com SQLite, como `DB Browser for SQLite`[3] ou diretamente via linha de comando usando o utilitário `sqlite3`:

```
$ sqlite3 mimic4.db
```

[2] https://pandas.pydata.org/
[3] https://sqlitebrowser.org/

### Ajuste necessário no script `import.py`

Durante os testes com o script `import.py`, foi identificado um erro relacionado à conversão de colunas de data/hora para o tipo `datetime`. O trecho original do código utilizava o parâmetro `format='ISO8601'` na função `pandas.to_datetime`, conforme segue:

```
df[c] = pd.to_datetime(df[c], format='ISO8601')
```

Contudo, o argumento `format='ISO8601'` não é reconhecido diretamente pelo `pandas`, o que resultava no seguinte erro de execução:

```
ValueError: time data does not match format ISO8601
```

Para contornar esse problema e tornar a conversão mais robusta, a linha foi substituída por:

```
df[c] = pd.to_datetime(df[c], errors='coerce')
```

Essa alteração permite que o `pandas` identifique automaticamente os formatos de data/hora e, quando não for possível realizar a conversão, o valor inválido é substituído por `NaT` (*Not a Time*). Essa abordagem garante maior estabilidade durante o processamento dos arquivos CSV, especialmente em colunas como `charttime`, `admittime` e `dob`, que podem conter registros incompletos ou fora de padrão.

### Considerações

O script `import.sh` define todos os campos como texto, o que pode limitar certos tipos de consultas mais complexas. Já o script em Python pode oferecer maior controle sobre os tipos de dados, mas consome mais memória e tempo durante a importação.

## 2.4. Estrutura do MIMIC-IV

A estrutura do MIMIC-IV foi projetada para refletir a complexidade do ambiente hospitalar, mantendo ao mesmo tempo a integridade relacional e a escalabilidade do banco de dados. A organização dos dados tem como objetivo facilitar a análise clínica e a pesquisa biomédica por meio da separação lógica das fontes de dados e pela utilização de identificadores anonimizados consistentes ao longo das tabelas.

### 2.4.1. Visão Geral da Arquitetura de Dados

O MIMIC-IV está organizado em dois grandes módulos principais: *hosp* e *icu*, representando respectivamente os dados do prontuário eletrônico hospitalar (EHR – *Electronic Health Record*) e do sistema de monitoramento clínico da UTI (*MetaVision*) [Johnson et al. 2023]. Essa separação modular reflete diretamente as origens dos dados e permite ao pesquisador focar em contextos específicos (internações gerais versus cuidados críticos), com maior controle e clareza.

O banco é composto por 31 tabelas no total, distribuídas entre os dois módulos, além de tabelas auxiliares de dicionário e mapeamento de conceitos clínicos [Johnson et al. 2023].

### 2.4.2. Módulo hosp

O módulo hosp agrega as informações mais amplas sobre os pacientes, coletadas ao longo de toda a jornada hospitalar. Esses dados provêm diretamente do EHR, e incluem desde informações administrativas até dados clínicos mais detalhados, como demonstrado nas Tabelas 2.1, 2.2 e Figura 2.8 [Johnson et al. 2023].

**Tabela 2.1. Tabelas do módulo Hosp do MIMIC-IV e suas descrições (Colunas "MIMIC-IV v3.1 e MIMIC-IV Demo v2.2"representam o número de linhas preenchidas na tabela em cada base de dados) (Parte 1) [MIT Laboratory for Computational Physiology 2023]**

| Nome da Tabela | Descrição da Tabela | MIMIC-IV v3.1 | MIMIC-IV Demo v2.2 |
|---|---|---|---|
| admissions | Informações detalhadas sobre internações hospitalares. | 546.028 | 275 |
| diagnoses_icd | Diagnósticos codificados (ICD-9/ICD-10) faturados durante as hospitalizações. | 6.364.488 | 4.506 |
| drgcodes | Códigos de diagnóstico relacionados a grupos (DRG) faturados durante a internação. | 761.856 | 454 |
| d_hcpcs | Tabela de dimensão para *hcpcsevents*; descreve os códigos CPT (*Current Procedural Terminology*) de procedimentos médicos. | 89.208 | 89.200 |
| d_icd_diagnoses | Tabela de dimensão para *diagnoses_icd*; fornece a descrição de diagnósticos ICD-9/ICD-10. | 112.107 | 109.775 |
| d_icd_procedures | Tabela de dimensão para *procedures_icd*; fornece a descrição de procedimentos ICD-9/ICD-10. | 86.423 | 85.257 |
| d_labitems | Tabela de dimensão para *labevents*; contém a descrição dos exames laboratoriais. | 1.650 | 1.622 |
| emar | Registro eletrônico de administração de medicamentos (eMAR). | 42.808.593 | 35.835 |
| emar_detail | Informações suplementares das administrações registradas em *emar*. | 87.371.064 | 72.018 |
| hcpcsevents | Eventos faturados durante a hospitalização; inclui códigos CPT. | 186.074 | 61 |
| labevents | Resultados de exames laboratoriais a partir de amostras dos pacientes. | 158.374.764 | 107.727 |

**Tabela 2.2. Tabelas do módulo Hosp do MIMIC-IV e suas descrições (Colunas "MIMIC-IV v3.1 e MIMIC-IV Demo v2.2"representam o número de linhas preenchidas na tabela em cada base de dados) (Parte 2) [MIT Laboratory for Computational Physiology 2023]**

| Nome da Tabela | Descrição da Tabela | MIMIC-IV v3.1 | MIMIC-IV Demo v2.2 |
|---|---|---|---|
| microbiologyevents | Culturas microbiológicas. | 3.988.224 | 2.899 |
| omr | Contém informações diversas do prontuário eletrônico. | 7.753.027 | 2.964 |
| patients | Contém sexo, idade e data de óbito (se disponível) dos pacientes. | 364.627 | 100 |
| pharmacy | Informações sobre medicamentos prescritos: dosagem, fórmula, entre outros. | 17.847.567 | 15.306 |
| poe | Ordens médicas realizadas pelos profissionais de saúde. | 52.212.109 | 45.154 |
| poe_detail | Detalhes suplementares das ordens médicas registradas em *poe*. | 8.504.982 | 3.795 |
| prescriptions | Medicamentos prescritos. | 20.292.611 | 18.087 |
| procedures_icd | Procedimentos codificados realizados durante a hospitalização. | 859.655 | 722 |
| provider | Lista os identificadores de profissionais de saúde (deidentificados). | 42.244 | 40.508 |
| services | Serviço(s) hospitalar(es) responsáveis pelo cuidado do paciente. | 593.071 | 319 |
| transfers | Informações detalhadas sobre as transferências de unidade dos pacientes. | 2.413.581 | 1.190 |

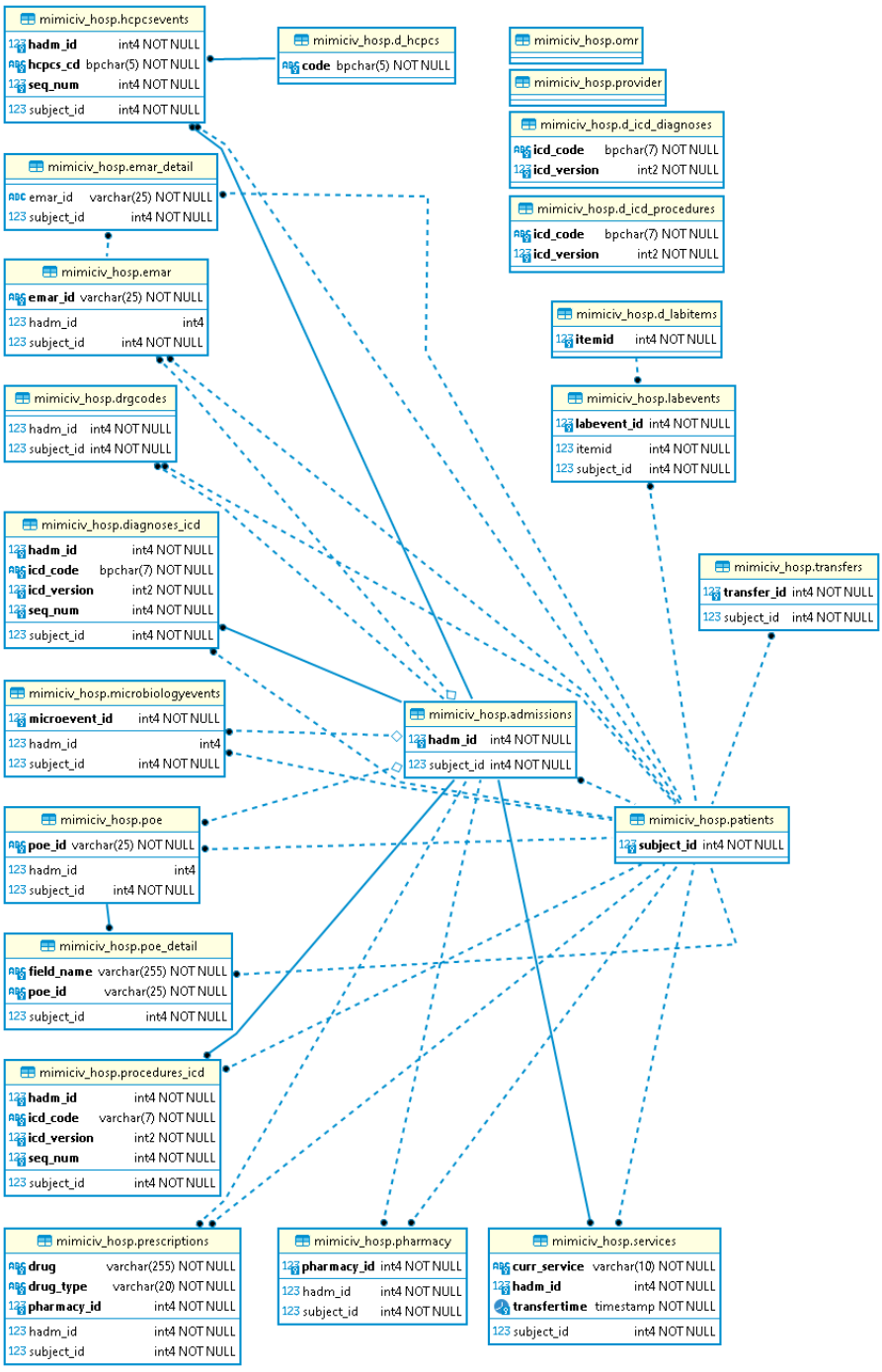



**Figura 2.8. Diagrama ER do módulo Hosp do MIMIC-IV.**

A granularidade dos dados varia: enquanto tabelas como *admissions* têm um registro por internação, outras como *labevents* podem conter centenas de entradas para um único paciente durante um único episódio hospitalar, refletindo medições realizadas várias vezes ao dia.

### 2.4.3. Módulo icu

Os dados contidos no módulo icu provêm do sistema de monitoramento contínuo da UTI (MetaVision), responsável por coletar medições detalhadas do estado clínico dos pacientes em cuidados intensivos, como demonstrado na Tabela 2.3 e Figura 2.9 [Johnson et al. 2023].

**Tabela 2.3. Tabelas do módulo ICU do MIMIC-IV e suas descrições (Colunas "MIMIC-IV v3.1 e MIMIC-IV Demo v2.2"representam o número de linhas preenchidas na tabela em cada base de dados) [MIT Laboratory for Computational Physiology 2023]**

| Nome da Tabela | Descrição da Tabela | MIMIC-IV v3.1 | MIMIC-IV Demo v2.2 |
|---|---|---|---|
| caregiver | Lista os identificadores de profissionais de saúde (deidentificados) utilizados no módulo de UTI. | 17.984 | 15.468 |
| chartevents | Itens registrados durante a permanência na UTI; contém a maior parte das informações clínicas da UTI. | 432.997.491 | 668.862 |
| datetimeevents | Informações registradas com formato de data (ex: data da última diálise). | 9.979.761 | 15.280 |
| d_items | Tabela de dimensão que descreve os *itemid*; define os conceitos registrados nas tabelas de eventos da UTI. | 4.095 | 4.014 |
| icu_stays | Informações sobre a permanência na UTI, incluindo horários de admissão e alta. | 94.458 | 140 |
| ingredientevents | Ingredientes de administrações contínuas ou intermitentes, incluindo conteúdo nutricional e hídrico. | 14.253.480 | 25.728 |
| inputevents | Informações sobre infusões contínuas ou administrações intermitentes documentadas. | 10.953.713 | 20.404 |
| outputevents | Informações sobre saídas do paciente, como urina, drenagem, entre outras. | 5.359.395 | 9.362 |
| procedureevents | Procedimentos documentados durante a internação na UTI (ex: ventilação), mesmo que realizados fora da UTI (ex: exames de imagem). | 808.706 | 1.468 |

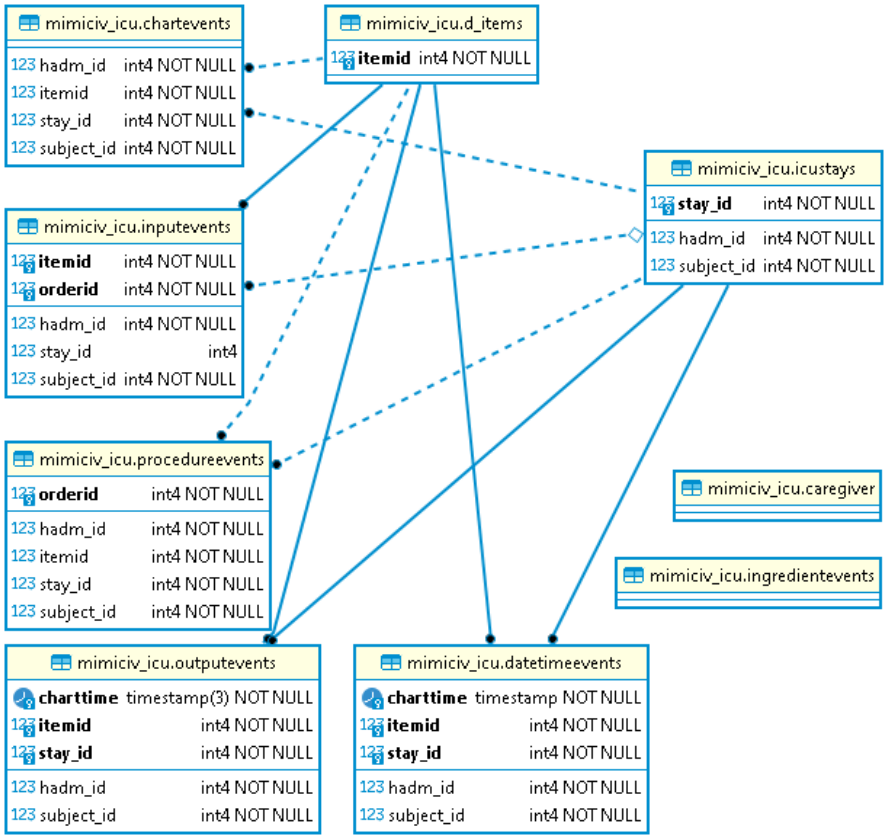



**Figura 2.9. Diagrama ER do módulo ICU do MIMIC-IV.**

É importante compreender que o módulo icu apresenta um volume de dados consideravelmente maior, pois representa medições contínuas – como frequência cardíaca, pressão arterial e saturação de oxigênio – coletadas a cada poucos minutos.

### 2.4.4. Relacionamentos e Identificadores

Uma característica essencial do MIMIC-IV é o uso de identificadores unificados que possibilitam a junção segura de dados entre as tabelas e os módulos. Os principais identificadores são:

- **subject_id**: identifica um paciente individual e é compartilhado entre todos os módulos.
- **hadm_id**: identifica uma internação hospitalar.
- **stay_id**: identifica uma internação na UTI.

Esses identificadores epseciais permitem, por exemplo, vincular exames laboratoriais ( *labevents*), com eventos de internação ( *admissions*), ou ainda cruzar dados de prescrição com registros de administração de medicamentos.

Além disso, o banco utiliza tabelas de apoio, como *d_items* e *d_lab_items*, que contém descrições dos itens presentes nas tabelas de eventos (*chartevents*, *labevents*, etc.), possibilitando a interpretação dos códigos numéricos.

### 2.4.5. Visões e Modelagem Relacional

A estrutura do MIMIC-IV é baseada nos princípios da modelagem relacional, seguindo normas clássicas de organização de dados em tabelas normalizadas. Esse modelo permite eliminar redundâncias e manter a consistência e integridade dos dados. No entanto, essa mesma estrutura exige que o usuário tenha familiaridade com conceitos fundamentais de bancos de dados relacionais para realizar consultas com precisão e eficiência. Neste ponto, é importante revisar alguns elementos centrais da modelagem relacional:

- **Tabelas**: Estruturas organizadas por colunas e registros (linhas), onde cada tabela representa uma entidade do modelo. Por exemplo, *patients* representa indivíduos, *admissions* representa episódios de internação, e *chartevents* contém observações clínicas contínuas.
- **Chaves primárias**: Um identificador único por registro, como *subject_id* em *patients*, que permite distinguir inequivocamente cada linha.
- **Chaves estrangeiras**: Atributos que criam uma ligação entre tabelas diferentes. Por exemplo, *subject_id* aparece em várias tabelas, estabelecendo relacionamentos entre elas.
- **Relacionamentos**: A modelagem relacional favorece uma estrutura onde as entidades estão interligadas por meio de identificadores desidentificados. Com isso, é possível navegar pelas informações do paciente entre diversas tabelas — da internação hospitalar até o nível dos sinais vitais.
- **Visões (Views)**: Embora o MIMIC-IV não forneça visões SQL pré-definidas, o uso de visões personalizadas é altamente recomendado durante as análises. Visões são consultas salvas que simulam tabelas e podem facilitar o reuso de lógicas complexas, principalmente em projetos com alto volume de *queries* recorrentes.

Esse modelo relacional torna o banco escalável, robusto e adequado para análises avançadas, mas impõe um desafio para iniciantes — especialmente aqueles com pouca familiaridade com SQL e com a lógica de bancos normalizados.

### 2.4.6. Granularidade, Temporalidade e Qualidade dos Dados

Um dos aspectos mais relevantes da estrutura do MIMIC-IV é sua granularidade. Ela varia consideravelmente entre as tabelas, impactando diretamente o tipo de análise que pode ser conduzida:

- **admissions**: Cada linha representa um episódio de internação hospitalar. A granularidade é uma entrada por internação.

- **chartevents**: Uma das tabelas mais volumosas. Cada linha representa uma observação clínica (sinais vitais, dados de monitoramento), com centenas ou milhares de entradas por paciente/dia.
- **noteevents**: Armazena notas clínicas em formato de texto livre. A granularidade depende do profissional e do tipo de nota registrada.

Essa granularidade exige atenção redobrada ao realizar junções (*joins*) entre tabelas, pois um relacionamento 1:n entre admissions e chartevents, por exemplo, pode gerar resultados inflados se a agregação não for realizada adequadamente.

Outro aspecto essencial é a temporalidade dos dados. A maioria das tabelas apresenta campos com marcações de tempo:

- **charttime**: representa o horário do evento registrado.
- **storetime**: horário em que o dado foi efetivamente armazenado no sistema.
- **starttime** e **endtime**: comuns em tabelas de prescrição ou procedimentos.

Esses campos permitem reconstruir a trajetória clínica do paciente ao longo do tempo, possibilitando análises longitudinais e temporais. No entanto, para preservar a privacidade dos indivíduos, todas as datas no MIMIC-IV foram alteradas com um deslocamento aleatório, diferente para cada paciente, mantendo a ordem e os intervalos entre os eventos [Johnson et al. 2023]. Esse deslocamento aleatório mantém a coerência temporal interna dos registros, mas inviabiliza análises agregadas por datas reais, como sazonalidade ou padrões semanais.

Além disso, há campos com valores ausentes ou registros generalizados (como códigos genéricos de procedimentos) devido a inconsistências no sistema original ou à política de privacidade.

### 2.4.7. Considerações sobre a estrutura do MIMIC-IV

A estrutura de dados do MIMIC-IV é, ao mesmo tempo, uma de suas maiores forças e um de seus maiores desafios. A modularidade entre os dados hospitalares (hosp) e os dados da UTI (icu) permite uma divisão lógica que reflete a origem dos registros, facilitando análises específicas por tipo de atendimento. Ao mesmo tempo, essa separação exige cuidado no planejamento das consultas, já que parte dos dados relevantes pode estar distribuída entre ambos os módulos.

A modelagem relacional com identificadores desidentificados aumenta a segurança e a privacidade dos dados, enquanto oferece uma estrutura robusta e flexível para o desenvolvimento de pesquisas. A granularidade variável e os aspectos temporais tornam o banco ideal para análises clínicas detalhadas, previsão de complicações, modelagem da permanência hospitalar, entre outros.

No entanto, o uso efetivo do MIMIC-IV requer um bom entendimento da arquitetura dos dados, da qualidade dos registros e das limitações impostas pela desidentificação.

Na próxima sessão, abordaremos como extrair e manipular esses dados de forma eficiente, utilizando comandos SQL, scripts em Python e práticas de modelagem de dados aplicadas à pesquisa clínica.

## 2.5. Recuperação de dados e análises exploratórias

### 2.5.1. MIMIC-IV Concepts

O MIMIC-IV é um banco de dados clínico complexo, composto por dezenas de tabelas brutas com estruturas distintas e, muitas vezes, difíceis de manipular diretamente. Para tornar a análise desses dados mais acessível e eficiente, a equipe mantenedora do projeto, em colaboração com a comunidade científica, desenvolveu o que se convencionou chamar de *Concepts* — consultas SQL que geram visões semânticas derivadas das tabelas originais.

As chamadas *Concept Tables* são visões organizadas que reúnem dados clínicos de interesse comum, como sinais vitais, exames laboratoriais, medicamentos administrados, entre outros. Esses conjuntos derivados são amplamente utilizados em pesquisas, pois oferecem padronização e facilitam a reprodutibilidade dos estudos.

Para auxiliar na criação dessas tabelas, o próprio repositório utilizado para importar o MIMIC-IV disponibiliza a pasta ./concepts-postgres, compatível tanto com PostgreSQL quanto com SQLite. Essa pasta contém scripts SQL prontos para gerar as *Concept Tables*, simplificando a manipulação dos dados e evitando erros frequentes associados à extração manual.

- O repositório oficial no GitHub pode ser acessado em: `https://github.com/MIT-LCP/mimic-code/tree/main/mimic-iv/concepts_postgres`
- As consultas estão organizadas por domínio clínico (ex: `demographics`, `firstday`, `measurement`, etc).
- O uso das *Concept Tables* é recomendado como ponto de partida para qualquer análise exploratória no MIMIC-IV.

### 2.5.2. Exemplo de utilização do *Concept*: recuperação de sinais vitais

Nesta subseção, mostraremos como construir uma sequência temporal com os principais sinais vitais registrados durante uma internação hospitalar no MIMIC-IV. Para isso, utilizaremos como exemplo uma internação selecionada aleatoriamente, da qual extrairemos frequência cardíaca, frequência respiratória, pressão arterial (sistólica, diastólica e média), temperatura corporal e saturação periférica de oxigênio (*SpO2*).

#### Etapa 1. Criação da *Concept Tables* vitalsign

Dentre os scripts de criação de *Concept Tables* disponibilizadas no repositório mencionado na seção anterior, a `Concept Table vitalsign` é a que consolida os principais sinais vitais medidos durante a internação dos pacientes. Essa tabela derivada facilita a

extração de informações como frequência cardíaca, pressão arterial, temperatura corporal, frequência respiratória e saturação de oxigênio (*SpO2*).

Para criar a `vitalsign`, basta executar o seguinte script SQL, adaptado do repositório oficial para criar uma nova tabela com essas informações:

```
CREATE TABLE vitalsign AS
SELECT
    ce.subject_id, ce.stay_id, ce.charttime
    , AVG(CASE WHEN itemid IN (220045) AND valuenum > 0 AND
       valuenum < 300 THEN valuenum END) AS heart_rate
    , AVG(CASE WHEN itemid IN (220179, 220050, 225309) AND
       valuenum > 0 AND valuenum < 400 THEN valuenum END) AS sbp
    , AVG(CASE WHEN itemid IN (220180, 220051, 225310) AND
       valuenum > 0 AND valuenum < 300 THEN valuenum END) AS dbp
    , AVG(CASE WHEN itemid IN (220052, 220181, 225312) AND
       valuenum > 0 AND valuenum < 300 THEN valuenum END) AS mbp
    , AVG(CASE WHEN itemid = 220179 AND valuenum > 0 AND
       valuenum < 400 THEN valuenum END) AS sbp_ni
    , AVG(CASE WHEN itemid = 220180 AND valuenum > 0 AND
       valuenum < 300 THEN valuenum END) AS dbp_ni
    , AVG(CASE WHEN itemid = 220181 AND valuenum > 0 AND
       valuenum < 300 THEN valuenum END) AS mbp_ni
    , AVG(CASE WHEN itemid IN (220210, 224690) AND valuenum > 0
       AND valuenum < 70 THEN valuenum END) AS resp_rate
    , ROUND(CAST(AVG(CASE
        WHEN itemid IN (223761) AND valuenum > 70 AND valuenum <
            120 THEN (valuenum - 32) / 1.8 -- converted to degC
           in valuenum call
        WHEN itemid IN (223762) AND valuenum > 10 AND valuenum <
            50 THEN valuenum END) -- already in degC, no
           conversion necessary
            AS NUMERIC), 2) AS temperature
    , MAX(CASE WHEN itemid = 224642 THEN value END) AS
       temperature_site
    , AVG(CASE WHEN itemid IN (220277) AND valuenum > 0 AND
       valuenum <= 100 THEN valuenum END) AS spo2
    , AVG(CASE WHEN itemid IN (225664, 220621, 226537) AND
       valuenum > 0 THEN valuenum END) AS glucose
FROM chartevents ce
WHERE ce.stay_id IS NOT NULL AND ce.itemid IN (
    220045 -- Heart Rate
    , 225309 -- ART BP Systolic
    , 225310 -- ART BP Diastolic
    , 225312 -- ART BP Mean
    , 220050 -- Arterial Blood Pressure systolic
    , 220051 -- Arterial Blood Pressure diastolic
    , 220052 -- Arterial Blood Pressure mean
    , 220179 -- Non Invasive Blood Pressure systolic
    , 220180 -- Non Invasive Blood Pressure diastolic
```

```
    , 220181 -- Non Invasive Blood Pressure mean
    , 220210 -- Respiratory Rate
    , 224690 -- Respiratory Rate (Total)
    , 220277 -- SPO2, peripheral
    -- GLUCOSE, both lab and fingerstick
    , 225664 -- Glucose finger stick
    , 220621 -- Glucose (serum)
    , 226537 -- Glucose (whole blood)
    -- TEMPERATURE
    -- 226329 -- Blood Temperature CCO (C)
    , 223762 -- "Temperature Celsius"
    , 223761  -- "Temperature Fahrenheit"
    , 224642 -- Temperature Site
    )
GROUP BY ce.subject_id, ce.stay_id, ce.charttime
```

### Etapa 2. Extração dos sinais vitais

Com a tabela `vitalsign` criada, podemos agora recuperar os sinais vitais registrados ao longo da internação de um paciente específico. O trecho de código abaixo exemplifica como obter uma sequência temporal dos dados fisiológicos a partir da data de admissão hospitalar, calculando o tempo decorrido (`offset`, em minutos) entre o momento da medição e o início da internação.

A variável `hadm_id` deve ser substituída pelo identificador da internação hospitalar desejada. Como resultado, teremos um conjunto de vetores dos sinais vitais definidos na consulta a seguir.

```
SELECT
    a.hadm_id
    , (julianday(v.charttime) - julianday(a.admittime)) * 24 *
       60 AS offset
    , v.heart_rate, v.sbp, v.dbp
    , v.sbp_ni, v.dbp_ni
    , v.resp_rate
    , v.temperature
    , v.spo2
    , v.glucose
FROM admissions a
INNER JOIN icustays i
    ON a.hadm_id = i.hadm_id
LEFT JOIN vitalsign v
    ON i.stay_id = v.stay_id
WHERE a.hadm_id = {hadm_id}
```

Com essa consulta, foi obtido um exemplo dos vetores de sinais vitais extraídos de uma única internação hospitalar (`hadm-id`). Como pode ser observado na Tabela 2.4, cada linha representa um registro temporal associado a um *offset* em minutos desde a

admissão, e cada coluna corresponde a um dos sinais vitais disponíveis. Os valores ausentes (NaN) indicam que o dado correspondente não estava disponível no momento da coleta. Esses vetores serão posteriormente organizados e interpolados para composição de janelas temporais com resolução horária padronizada.

**Tabela 2.4. Exemplo de amostra dos vetores de sinais vitais extraídos de uma internação hospitalar**

| hadm_id | offset | heart_rate | sbp | dbp | sbp_ni | dbp_ni | resp_rate | temperature | $SpO_2$ | glucose |
|---|---|---|---|---|---|---|---|---|---|---|
| x | 14310 | 80 | 115 | 66 | NaN | NaN | 15 | NaN | 100 | NaN |
| x | 14311 | 80 | 118 | 67 | NaN | NaN | 15 | 36.5 | 100 | NaN |
| x | 14321 | 80 | 104 | 60 | NaN | NaN | 17.50 | NaN | 100 | NaN |
| x | 14331 | 81 | 123 | 69 | NaN | NaN | 15 | NaN | 100 | NaN |
| x | 14336 | 70 | 113 | 62 | NaN | NaN | 9 | NaN | 100 | NaN |
| x | 14341 | 70 | 107 | 60 | NaN | NaN | 15 | NaN | 100 | NaN |
| x | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

### Etapa 3. Visualização dos sinais vitais

Com os dados de sinais vitais extraídos, o próximo passo é visualizá-los ao longo do tempo para facilitar a interpretação clínica e a análise exploratória. O código Python a seguir mostra como plotar essa série temporal utilizando a biblioteca `matplotlib`. A estratégia adotada permite representar a evolução dos parâmetros fisiológicos durante a internação hospitalar, com destaque para os principais sinais vitais.

O trecho de código é organizado em etapas:

- **Definição da Função `plot_numeric_data`**: Essa função recebe um eixo do matplotlib, um DataFrame com os dados de entrada, um dicionário de informações de plotagem e um tempo nulo opcional (usado para interromper linhas contínuas quando o paciente sai da UTI). Ela filtra dados ausentes, organiza por ordem cronológica (`offset`) e plota os dados com os parâmetros visuais definidos.

```
def plot_numeric_data(ax, df_input, plot_info, null_time=
    None):
     for i, (column, plot_args) in enumerate(plot_info.items
        ()):
         df = df_input[['offset', column]].copy().dropna()

         # insert a null when the patient leaves the ICU to
            break the line in the plot
         if null_time is not None:
             dff = pd.DataFrame([[null_time, None]], columns
                =['offset', column])
             df = pd.concat([df, dff], ignore_index=True)
         df.sort_values('offset', inplace=True)
         ax.plot(df['offset']/TIME_COEF, df[column], **
            plot_args)
```

- **Configurações Iniciais**: Define constantes como o `TIME_COEF`, usado para converter minutos em dias (`offset/1440`), facilitando a leitura dos eixos temporais, e configurações visuais da figura (tamanho, fontes, títulos etc.).

```
# === Initial Configurations ===
TIME_COEF = 24 * 60.0  # conversion from minutes to days
TIME_UNIT = 'days'
null_time = 8 * 24 * 60  # time used to separate distinct
    admissions

plt.rcParams.update({'font.size': 22})
fig = plt.figure(figsize=(20, 10))
gs = fig.add_gridspec(1, hspace=0.1)
ax0 = gs.subplots(sharex=True, sharey=False)

fig.suptitle(f'Vital Signs and Cardiac Markers - Patient ID:
    XXXXXXXX', fontsize=24, y=0.98)
ax0.set_xlim([0, 13])
```

- **Plotagem dos sinais vitais**:
  - **`Heart Rate`** e **`Respiratory Rate`** são plotados como linhas com marcadores circulares, utilizando a função `plot_numeric_data`.
  - **Pressão arterial (sistólica e diastólica)** é representada por uma área sombreada entre as curvas, indicando a faixa de variação da pressão.
  - **Temperatura corporal (°C)** é exibida com marcador em formato de losango, em duas fases (antes e depois do tempo nulo), permitindo distinguir diferentes internações, se aplicável.
  - **Saturação periférica de oxigênio ($SpO_2$)** é plotada com linha tracejada e marcadores quadrados.

```
# === Vital Signs ===
plot_info_vitals = OrderedDict([
    ['heart_rate', {'label': 'Heart Rate (bpm)', 'color':
       colors[0], 'marker': 'o', 'lw': 2.5, 'markersize':
       6}],
    ['resp_rate', {'label': 'Respiratory Rate (ipm)', 'color
       ': colors[1], 'marker': 'o', 'lw': 2.5, 'markersize':
        6}],
])
plot_numeric_data(ax0, vitals, plot_info_vitals, null_time=
    null_time)

# ===> Blood Pressure (shaded area)
bp = vitals[['offset', 'dbp', 'sbp']].dropna().copy()
plot_args = {'color': colors[3], 'alpha': 0.3}
bp_early = bp['offset'] < null_time
```

```
ax0.fill_between(bp.loc[bp_early, 'offset'] / TIME_COEF, bp.
   loc[bp_early, 'dbp'], bp.loc[bp_early, 'sbp'], label='
   Blood Pressure (mmHg)', **plot_args)
ax0.fill_between(bp.loc[~bp_early, 'offset'] / TIME_COEF, bp
   .loc[~bp_early, 'dbp'], bp.loc[~bp_early, 'sbp'], **
   plot_args)

# ===> Temperature
temp_early = temp['offset'] < null_time
ax0.plot(temp.loc[temp\_early, 'offset'] / TIME\_COEF, temp.
   loc[temp\_early, 'valuenum'], label='Temperature (C)',
   color=colors[2], marker='d', markersize=10, lw=2)
ax0.plot(temp.loc[~temp_early, 'offset'] / TIME_COEF, temp.
   loc[~temp_early, 'valuenum'], color=colors[2], marker='d'
   , markersize=10, lw=2)

# ===> SpO2
spo2_data = vitals[['offset', 'spo2']].dropna()
spo2_early = spo2_data['offset'] < null_time
ax0.plot(spo2\_data.loc[spo2_early, 'offset'] / TIME\_COEF,
   spo2_data.loc[spo2\_early, 'spo2'], label='SpO2 (%)',
         color=colors[7], marker='s', markersize=8, lw=2,
            linestyle='--')
ax0.plot(spo2_data.loc[~spo2_early, 'offset'] / TIME_COEF,
   spo2_data.loc[~spo2_early, 'spo2'],
         color=colors[7], marker='s', markersize=8, lw=2,
            linestyle='--')
```

- **Finalização**: Os elementos visuais do gráfico são refinados com legenda, rótulos, grades e ajustes de layout para melhor legibilidade.

```
ax0.legend(loc='upper right', bbox_to_anchor=(1.0, 1.0),
   fontsize=18)
ax0.set_ylabel('Values', fontsize=20)
ax0.set_xlabel('Days', fontsize=20)
ax0.grid(True, linestyle='--', alpha=0.7)

# === Final Layout ===
plt.tight_layout(rect=[0, 0, 1, 0.96])
```

A execução do código acima gera um gráfico consolidado que ilustra a evolução temporal dos sinais vitais do paciente ao longo do período de internação (Figura 2.10). Cada curva representa um parâmetro fisiológico distinto, com codificação de cores e marcadores que facilitam a identificação visual. A faixa sombreada referente à pressão arterial permite visualizar as variações entre os valores sistólico e diastólico. A divisão entre internações (ou fases distintas do atendimento) é indicada por grandes interrupções nas

linhas, controladas pelo parâmetro `null_time`. Essa visualização proporciona uma análise rápida do estado clínico do paciente ao longo do tempo e pode ser utilizada tanto para fins exploratórios quanto como base para métodos de modelagem preditiva.

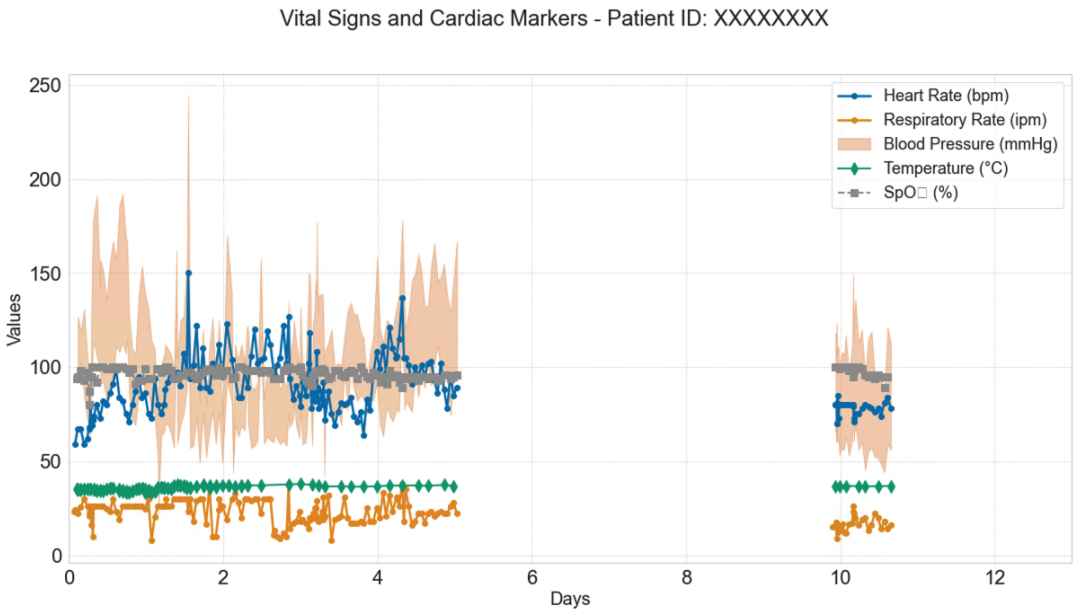



**Figura 2.10. Visualização temporal dos sinais vitais ao longo da internação hospitalar para um paciente selecionado.**

## 2.6. Aplicação prática com Python e SQL

Nesta sessão, você aprenderá a extrair um subconjunto específico de pacientes a partir do MIMIC-IV, utilizando Python com SQL. Este processo é útil para análises clínicas direcionadas, estudos estatísticos e desenvolvimento de modelos preditivos.

### Objetivo

Selecionar admissões de pacientes com os seguintes critérios:

- Sexo: Masculino
- Idade: entre 20 e 35 anos no momento da admissão
- Diagnóstico principal: Pneumonia

### Pré-requisitos

- Ter uma instância local do MIMIC-IV em um banco PostgreSQL ou arquivo local com o banco de dados em criado via SQLite
- Ter acesso autorizado ao MIMIC (via PhysioNet)
- Python instalado com as bibliotecas `pandas`, `sqlalchemy` ou `sqlite3`

### Etapa 1: Importação das bibliotecas

Vamos começar importando as bibliotecas necessárias:

- `pandas`: Para manipulação de dados em DataFrames
- `sqlalchemy`: Para conexão e execução de comandos SQL em bancos de dados relacionais, necessário para utilização do PostgreSQL
- `sqlite3`: Para conexão e execução de comandos SQL em bancos de dados relacionais, necessário para utilização do SQLite

Caso ainda não tenha as bibliotecas instaladas, use o seguinte comando:

```
pip install pandas sqlalchemy
```

```
import pandas as pd
import sqlalchemy
import sqlite3
```

### Etapa 2: Conexão com o Banco de Dados

Configure a conexão com o banco local que contém o MIMIC-IV.

**PostgreSQL**

```
def connect_postgres(user, password, host, port, db):
    url = f"postgresql://{user}:{password}@{host}:{port}/{db}"
    engine = sqlalchemy.create_engine(url)
    return engine
```

**SQLite**

```
engine = sqlite3.connect('mimic4.db') # Passando o caminho do
   arquivo do banco de dados como parâmetro
```

### Etapa 3: Consulta SQL - Extração de pacientes

Nesta etapa, construímos uma consulta SQL para buscar pacientes do sexo masculino.

```
query_pacientes = """
SELECT
    subject_id,
    gender,
    anchor_age,
    anchor_year,
    anchor_year_group,
    dod
FROM patients
WHERE gender = 'M'
"""
```

```
df_pacientes = pd.read_sql(query_pacientes, engine)
```

**Explicação das colunas utilizadas:**

- **subject_id** é o identificador único atribuído a cada paciente.
- **gender** indica o sexo biológico do paciente, sendo `M` para masculino e `F` para feminino.
- **anchor_age** representa a idade aproximada do paciente no ano de referência definido em **anchor_year**.
- **anchor_year** é o ano utilizado como ponto de ancoragem temporal para os dados clínicos do paciente, permitindo anonimização.
- **anchor_year_group** agrupa os pacientes com base no período de seu ano de referência (por exemplo, 2008–2010).
- **dod** (date of death) indica a data de óbito do paciente, quando disponível.

Agora, vamos transformar todos os subject_ids retornados em uma string para que possa ser utilizada posteriormente durante a busca de informações referentes a admissão desses pacientes e também dos diagnósticos deles.

O subject_id é o identificador único atribuído à cada paciente. Com ele, podemos recuperar as informações de cada um deles.

```
# Transforma a lista de IDs em uma string formatada para o SQL

subject_ids = df_pacientes['subject_id'].unique().tolist()
subject_ids_str = '(' + ', '.join(str(sid) for sid in
   subject_ids) + ')'
```

### Etapa 4: Consulta SQL - Extração de admissões

Agora, com auxílio dos subject_ids recuperados no passo anterior, vamos recuperar todas as admissões referente a estes pacientes.

```
query_admissoes = f"""
SELECT
    subject_id,
    hadm_id,
    admittime
FROM admissions
WHERE subject_id IN {subject_ids_str}
"""
df_admissoes = pd.read_sql(query_admissoes, engine)
```

**Explicação das colunas utilizadas:**

- **subject_id** é o identificador do paciente, como anteriormente.
- **hadm_id** é o identificador único de cada admissão hospitalar, permitindo a vinculação com outros eventos clínicos ocorridos durante a internação.
- **admittime** representa a data e hora da admissão do paciente no hospital.

Agora, vamos transformar todos os hadm_ids retornados em uma string para que possa ser utilizada posteriormente durante a busca de informações referentes aos diagnósticos.

O hadm_id é o identificador único atribuído à cada admissão. Com ele, podemos recuperar as informações de cada uma delas.

```
# Transforma a lista de IDs em uma string formatada para o SQL

hadm_ids = df_admissoes['hadm_id'].unique().tolist()
hadm_ids_str = '(' + ', '.join(str(hid) for hid in hadm_ids) + '
   )'
```

### Etapa 5: Consulta SQL - Extração dos Diagnósticos dos Pacientes

Agora, com o auxílio dos subject_ids e hadm_ids obtidos anteriormente, vamos buscar todos os diagnósticos realizados nas admissões destes pacientes.

```
query_diagnosticos = f"""
SELECT
    d.subject_id,
    d.hadm_id,
    d.seq_num,
    d.icd_code,
    d.icd_version,
    dx.long_title as diagnosis
FROM diagnoses_icd d
JOIN d_icd_diagnoses dx ON dx.icd_code = d.icd_code AND dx.
   icd_version = d.icd_version
WHERE d.subject_id in {subject_ids_str}
AND d.hadm_id in {hadm_ids_str}
"""

df_diag = pd.read_sql(query_diagnosticos, engine)
```

**Explicação das colunas utilizadas:**

- **subject_id** e **hadm_id** identificam o paciente e sua internação, respectivamente.
- **seq_num** indica a ordem do diagnóstico dentro da admissão, sendo 1 geralmente o diagnóstico principal.
- **icd_code** é o código da Classificação Internacional de Doenças (CID), que identifica a condição diagnosticada.

- **icd_version** informa a versão da classificação utilizada (por exemplo, ICD-9 ou ICD-10).
- **long_title** (renomeado como **diagnosis**) fornece a descrição textual completa da condição médica codificada.

### Etapa 6: Consulta SQL - Juntar Pacientes + Admissões + Diagnósticos

Agora que temos 3 dataframes contendo as informações sobre Pacientes, Admissões e Diagnósticos, precisamos unir tudo isso em um Dataframe único. Para isso, usaremos as colunas subject_id e hadm_id para realizarmos a união.

```
# Merge 1: pacientes + admissões
df_merged = pd.merge(df_pacientes, df_admissoes, on="subject_id"
    , how="inner")

# Merge 2: adicionar diagnósticos
df_merged = pd.merge(df_merged, df_diag, on=["subject_id", "
    hadm_id"], how="inner")
```

### Etapa 7: Aplicar Filtros Específicos

Agora que temos todos os dados em mãos, é hora de limpar o nosso subconjunto para que fiquem apenas os casos de interesse:

- Sexo: Masculino (já filtrado durante a obtenção de dados dos pacientes)
- Idade: Entre 40 e 50 anos (no momento da admissão)
- Diagnóstico: Pneumonia

Você deve ter percebido que não existe uma coluna de idade do paciente. Para calcularmos a idade de cada paciente, precisamos utilizar uma regra de negócio aplicada durante o processo de anonimização dos dados.

A idade é calculada desta forma:

- birth_year = anchor_year - anchor_age
- age_at_admission = admit_time - birth_year

```
# Para SQLite, descomente a linha a seguir para transformar a
    coluna admittime para o formato datetime
# df_merged['admittime'] = pd.to_datetime(df_merged['admittime
    '])

# Calcular a idade real na admissão
df_merged['age_at_admission'] = df_merged['admittime'].dt.year -
    (df_merged['anchor_year'] - df_merged['anchor_age'])
```

```
# Ajustar para verificar se o paciente já fez aniversário no ano
    da admissão
df_merged['age_at_admission'] -= ((df_merged['admittime'].dt.
   month < 1) |
                                  ((df_merged['admittime'].dt.
                                     month == 1) &
                                   (df_merged['admittime'].dt.
                                      day < 1))).astype(int)
```

Agora que anexamos a coluna com a idade no momento da admissão ao nosso conjunto de dados, podemos aplicar os filtros:

- Primeiro, manteremos apenas as linhas em que os pacientes com a idade dentro do intervalo de interesse: entre 40 e 50 anos.
- Por último, manteremos apenas as linhas em que o Diagnóstico contenha o termo "Pneumonia".

```
# Filtro por idade
df_filtrado = df_merged[(df_merged["age_at_admission"] >= 40) &
   (df_merged["age_at_admission"] <= 50)]

# Filtro por diagnóstico
df_final = df_filtrado[
    df_filtrado["diagnosis"].str.contains("pneumonia", case=
       False, na=False)
]
```

### Etapa 8: Exportar o Conjunto Final

Por fim, exportaremos o subconjunto obtido para um arquivo no formato .csv. Este arquivo pode ser usado para análises ou para auxiliar a extrair outras informações relevantes sobre os pacientes ou admissões nele contidos.

```
df_final.to_csv("subconjunto_mimiciv.csv", index=False)
print("Arquivo salvo como 'subconjunto_mimiciv.csv'")
```

### Revisando: O que aprendemos?

- Como se conectar ao MIMIC-IV com Python
- Como buscar e combinar diferentes tabelas (pacientes, admissões, diagnósticos)
- Como aplicar filtros específicos para extrair um subconjunto de dados
- Como exportar resultados para CSV

Com essa implementação, extraímos um subconjunto específico e relevante de dados, pronto para ser utilizado em análises clínicas, estudos estatísticos ou como base para investigações mais aprofundadas dentro do próprio MIMIC-IV, utilizando os identificadores de pacientes (subject_id) e de admissões hospitalares (hadm_id). Embora o exemplo apresentado tenha sido intencionalmente simples, ele demonstra, de forma prática, como o MIMIC-IV pode ser explorado com o uso combinado de Python e SQL.

O MIMIC-IV é um banco de dados robusto e abrangente, capaz de sustentar uma ampla gama de aplicações — desde pesquisas epidemiológicas e desenvolvimento de modelos preditivos até estudos longitudinais e análises de desfechos clínicos. A flexibilidade e a riqueza das informações disponíveis tornam este recurso extremamente valioso para pesquisadores e profissionais da saúde interessados em ciência de dados aplicada à área médica.

## 2.7. Aspectos éticos na utilização do MIMIC

O uso de dados clinicos reais, como os presentes no MIMIC, envolve uma série de responsabilidades éticas fundamentais. Embora o MIMIC tenha sido cuidadosamente projetado para proteger a privacidade dos indivíduos, o acesso ao banco completo está condicionado ao cumprimento de normas de ética em pesquisa, conforme exigido por regulamentos internacionais como o Health Insurance Portability and Accountability Act (HIPAA), além das diretrizes de proteção de dados sensíveis adotadas pelo MIT e pela plataforma PhysioNet.

### 2.7.1. Requisitos para Acesso ao MIMIC-IV

Para obter acesso aos dados do MIMIC-IV, o pesquisador deve cumprir os requisitos descritos na sessão 2.3.2. Esse processo garante que o usuário compreende os princípios éticos envolvidos, como o respeito à privacidade dos pacientes, a confidencialidade dos dados e a proibição explícita de qualquer tentativa de reidentificação dos indivíduos presentes no banco.

### 2.7.2. Anonimização dos Dados e Conformidade com o HIPAA

O MIMIC é uma base de dados desidentificada, em conformidade com a HIPAA. A anonimização é realizada por meio de diversas estratégias [Johnson et al. 2024]:

- Remoção de identificadores diretos, como nomes, números de documentos, endereços e contatos.
- Substituição de datas reais por datas deslocadas aleatoriamente, mantendo a coerência temporal interna para análises longitudinais, mas inviabilizando qualquer mapeamento com calendários reais.
- Codificação de identificadores, como subject_id, hadm_id e stay_id, que são únicos, porém artificiais.
- Generalização de valores sensíveis, como datas de nascimento e faixas etárias extremas (pacientes com mais de 89 anos, por exemplo, são representados como tendo "91 anos").

Além disso, os dados são revisados periodicamente pela equipe do MIT-LCP para garantir que continuam dentro dos padrões de desidentificação seguros mesmo após atualizações da base.

### 2.7.3. Termo de Uso de Dados e Responsabilidade dos Pesquisadores

Ao aceitar o termo de uso (DUA) [PhysioNet 2025], o pesquisador assume legalmente o compromisso de:

- Não tentar identificar qualquer indivíduo ou instituição referenciada nos dados restritos do PhysioNet;
- Exercer todo o cuidado razoável e prudente para evitar a divulgação da identidade de qualquer indivíduo ou instituição referenciada nos dados restritos do PhysioNet em qualquer publicação ou outra comunicação;
- Não compartilhar o acesso aos dados restritos do PhysioNet com ninguém;
- Exercer todo o cuidado razoável e prudente para manter a segurança física e eletrônica dos dados restritos do PhysioNet;
- Se encontrar informações dentro dos dados restritos do PhysioNet que possam permitir a identificação de qualquer indivíduo ou instituição, reportar prontamente a localização dessa informação por e-mail para PHI-report@physionet.org, citando a localização específica da informação em questão;
- Solicitar o acesso aos dados restritos do PhysioNet com o único propósito de uso legítimo em pesquisa científica, utilizando o privilégio de acesso, se concedido, exclusivamente para esse propósito;
- Completar um programa de treinamento em proteções de sujeitos de pesquisa humana e regulamentações HIPAA, fornecendo a comprovação de que o fez;
- Indicar o propósito geral para o qual pretende usar o banco de dados na solicitação de acesso;
- Se os resultados forem divulgados publicamente, contribuir também com o código utilizado para produzir esses resultados para um repositório aberto à comunidade de pesquisa;
- Reconhecer que o acordo pode ser rescindido por qualquer uma das partes a qualquer momento, mas que as obrigações com relação aos dados do PhysioNet continuarão após a rescisão.

O não cumprimento desses requisitos pode resultar em penalidades institucionais e legais, além de comprometer a integridade de toda a comunidade científica usuária do MIMIC.

### 2.7.4. Utilização Ética

Neste documento, algumas demonstrações e exemplos práticos foram elaborados com base na versão completa do MIMIC-IV, obtida mediante autorização formal, conforme os requisitos do PhysioNet. No entanto, todas as atividades apresentadas são inteiramente reproduzíveis utilizando a versão Demo do MIMIC-IV, que é amplamente divulgada pela plataforma PhysioNet e livre de restrições de acesso ou obrigações éticas adicionais.

A versão Demo contém um subconjunto representativo dos dados reais, suficientemente anonimizados e reduzidos para fins de ensino, testes e exploração de ferramentas. Isso garante que qualquer participante do minicurso poderá reproduzir os exemplos propostos sem infringir regulamentos éticos ou a política de uso da base de dados.

Reforçamos, também, que nenhuma parte deste material propõe ou orienta o uso da versão completa do MIMIC-IV sem a devida autorização. Ao adotar a versão Demo como referência para reprodução dos tutoriais, este documento mantém seu compromisso com as boas práticas em pesquisa e com os princípios éticos fundamentais na manipulação de dados clínicos.

## 2.8. Conclusões

O MIMIC-IV é amplamente reconhecido como uma importante fonte de dados clínicos públicos, reunindo informações detalhadas sobre pacientes em contextos hospitalares e de terapia intensiva. Sua estrutura relacional, modular e com granularidade temporal permite análises em múltiplas escalas, desde investigações populacionais até o monitoramento clínico individualizado.

Neste capítulo, foram abordadas as etapas essenciais para acessar e utilizar o MIMIC-IV, incluindo a configuração de ambientes locais com PostgreSQL e SQLite. Por meio de exemplos práticos com SQL e Python, mostramos como realizar análises exploratórias e extrair subconjuntos clinicamente relevantes, fornecendo uma base sólida para pesquisas mais acessíveis, reproduzíveis e alinhadas aos objetivos de diversos estudos.

Embora o MIMIC-IV seja tecnicamente complexo, essa complexidade pode ser superada com o uso de ferramentas adequadas e a disseminação de boas práticas. Ao reduzir a barreira técnica de entrada, ampliamos o acesso ao banco de dados, permitindo que mais pesquisadores e profissionais possam utilizá-lo, impulsionando o avanço da ciência de dados em saúde. Essa democratização do acesso fortalece uma comunidade multidisciplinar comprometida com a medicina baseada em evidências e com a evolução contínua da prática clínica por meio de análises informadas e inovadoras.

## Referências

[Fan et al. 2025] Fan, X., Xu, J., Ye, R., Zhang, Q., and Wang, Y. (2025). Retrospective cohort study based on the mimic-iv database: analysis of factors influencing all-cause mortality at 30 days, 90 days, 1 year, and 3 years in patients with different types of stroke. *Frontiers in Neurology*, 15.

[Johnson et al. 2024] Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., and Mark, R. (2024). Mimic-iv.

[Johnson et al. 2023] Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-W. H., Celi, L. A., and Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data*, 10(1):1.

[Johnson et al. 2016] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Sci. Data*, 3(1):160035.

[Jung et al. 2024] Jung, J., Kim, D., and Hwang, I. (2024). Exploring predictive factors for heart failure progression in hypertensive patients based on medical diagnosis data from the mimic-iv database. *Bioengineering*, 11(6):531.

[Lin et al. 2024] Lin, S., Lu, W., Wang, T., Wang, Y., Leng, X., Chi, L., Jin, P., and Bian, J. (2024). Predictive model of acute kidney injury in critically ill patients with acute pancreatitis: a machine learning approach using the mimic-iv database. *Renal Failure*, 46(1).

[MIT Laboratory for Computational Physiology 2023] MIT Laboratory for Computational Physiology (2023). MIMIC-IV Documentation. `https://mimic.mit.edu/docs/iv`. Acesso em: Abril de 2025.

[National Library of Medicine (US) 2025] National Library of Medicine (US) (2025). PubMed: Trending articles. `https://pubmed.ncbi.nlm.nih.gov/trending/`. Accessed: 2025-05-07.

[PhysioNet 2025] PhysioNet (2025). Physionet credentialed health data use agreement 1.5.0. `https://physionet.org/content/mimiciv/view-dua/3.1/`. Acesso em: 22 abr. 2025.

[Pérez-Tome et al. 2024] Pérez-Tome, J. C., Parrón-Carreño, T., Castaño-Fernández, A. B., Nievas-Soriano, B. J., and Castro-Luna, G. (2024). Sepsis mortality prediction with machine learning tecniques. *Medicina Intensiva (English Edition)*, 48(10):584–593.

[Sun et al. 2024] Sun, B., Man, Y.-l., Zhou, Q.-y., Wang, J.-d., Chen, Y.-m., Fu, Y., and Chen, Z.-h. (2024). Development of a nomogram to predict 30-day mortality of sepsis patients with gastrointestinal bleeding: An analysis of the mimic-iv database. *Heliyon*, 10(4):e26185.

# Capítulo 3

# Ciência de Dados em Saúde: Primeiros Passos na Preparação e Análise de Dados

Ivan Rodrigues de Moura, Francisco José da Silva e Silva, Luciano Reis Coutinho, Ariel Soares Teles, Nailton dos Reis Maciel e Danilo Gameleira Dias

***Abstract***

*Data preparation and exploratory analysis are essential stages in data science applied to healthcare, ensuring the integrity, reliability, and relevance of the information used in clinical, epidemiological, and other biomedical applications. This chapter presents a set of fundamental techniques for organizing, cleaning, and exploring health-related data, focusing on biomedical and epidemiological datasets. Topics covered include data reading, structuring, cleaning, handling of missing and duplicate values, data manipulation and aggregation, and data visualization methods. The tools discussed include widely used libraries in the Python ecosystem, such as Pandas, Seaborn, and Plotly, offering a practical and accessible foundation for exploratory data analysis in healthcare. The content is intended for students and professionals in technology and healthcare, including those with no prior experience in data science, aiming to support their introduction to the initial data analysis processes in this interdisciplinary domain.*

***Resumo***

*A preparação de dados e a análise exploratória constituem etapas essenciais na ciência de dados aplicada à saúde, pois asseguram a integridade, a confiabilidade e a relevância das informações utilizadas em estudos clínicos, epidemiológicos e outras aplicações biomédicas. Este capítulo apresenta um conjunto de técnicas fundamentais para a organização, limpeza e exploração de dados de saúde, com foco em conjuntos de dados biomédicos e epidemiológicos. São abordados tópicos como leitura, estruturação e higienização de dados, tratamento de valores ausentes e duplicados, manipulação e agregação de informações, bem como métodos de visualização de dados. As ferramentas discutidas incluem bibliotecas amplamente utilizadas no ecossistema Python, como Pandas, Seaborn e Plotly, proporcionando uma base prática e acessível para a análise exploratória de*

*dados em saúde. O conteúdo é direcionado a estudantes e profissionais das áreas de tecnologia e saúde, mesmo aqueles sem experiência prévia em ciência de dados, e tem como objetivo facilitar a introdução aos processos iniciais de análise de dados neste domínio interdisciplinar.*

## 3.1. Introdução

Atualmente, a grande quantidade de dados providos do setor de saúde oferece uma janela de oportunidades significativa para aprimorar diagnósticos, personalizar tratamentos e otimizar a gestão hospitalar [Chattu 2021]. No entanto, características como a complexidade e heterogeneidade dos dados impõem desafios significativos à análise e interpretação eficaz dessas informações, exigindo soluções avançadas como análise de dados e ciência de dados [Awrahman et al. 2022]. A ciência de dados, combinando métodos estatísticos, aprendizado de máquina e visualização de dados, emerge como uma ferramenta essencial para extrair conhecimento útil a partir dessas grandes bases de dados, com potencial para transformar desde diagnósticos médicos até a gestão hospitalar [Subrahmanya et al. 2022]. Por exemplo, modelos de aprendizado de máquina podem ser utilizados para predizer o desenvolvimento de doenças como diabetes, depressão e hipertensão [Moura et al. 2023]. Além disso, técnicas de análise de dados e visualização de dados podem ser utilizadas para identificar padrões úteis na tomada de decisões médicas [Subrahmanya et al. 2022].

Diante desse cenário, este capítulo de livro tem como objetivo fornecer uma introdução às principais ferramentas e técnicas da ciência de dados aplicadas à saúde, com ênfase na preparação e análise de dados. Primeiramente, serão apresentadas as principais técnicas de manipulação de dados utilizando a biblioteca Pandas, vindo a demonstrar na prática operações como seleção de dados, operações de agregamento, agrupamentos e filtros. Em seguida, serão apresentadas as principais bibliotecas de visualização de dados, em especial as bibliotecas Seaborn e Plotly. Serão criados gráficos de diversos tipos com essas bibliotecas, como barra, pizza, linha, boxplot, mapas de calor, dentre outros. Também serão apresentadas as principais técnicas de limpeza e pré-processamento de dados, isto é, trataremos valores nulos, duplicados e outlines. Finalmente, serão discutidos desafios éticos e regulatórios associados ao uso da ciência de dados e análise de dados na saúde.

É importante destacar que este capítulo de livro não aborda a criação de modelos de aprendizado de máquina. Em vez disso, seu foco está nas etapas iniciais da ciência de dados, como manipulação, limpeza e análise exploratória dos dados. Dessa forma, objetivamos desenvolver uma base sólida para futuras aplicações em modelagem preditiva e aprendizado de máquina. Ao explorar esses tópicos, objetivamos contribuir para a capacitação de estudantes, pesquisadores e profissionais interessados em aplicar técnicas de análise de dados para melhorar a eficiência e a qualidade dos serviços médicos. O avanço da ciência de dados na área da saúde tem o potencial de revolucionar a medicina moderna, promovendo diagnósticos mais precisos, tratamentos personalizados e uma gestão hospitalar mais eficiente, sempre com a preocupação de equilibrar inovação tecnológica e responsabilidade ética.

## 3.2. Ciência de Dados na Saúde

A ciência de dados é uma disciplina que tem causado uma profunda transformação em diversos setores da sociedade, e em especial, na área da saúde [Bao et al. 2019]. Esse impacto se deve à capacidade da ciência de dados de extrair conhecimento relevante a partir de grandes volumes de dados, muitas vezes heterogêneos e complexos. Na saúde, o crescimento exponencial na geração de dados biomédicos, clínicos e epidemiológicos nos últimos anos representa uma verdadeira janela de oportunidades. Com a popularização dos prontuários eletrônicos, o uso de dispositivos vestíveis, a digitalização de exames e a expansão dos sistemas públicos de informação em saúde, é possível acessar uma quantidade sem precedentes de informações sobre indivíduos e populações [Teles et al. 2025]. Essa abundância de dados permite a criação de soluções inovadoras baseadas em evidências que podem aprimorar significativamente o diagnóstico precoce de doenças, o monitoramento contínuo de pacientes, a personalização de tratamentos, a gestão eficiente dos recursos hospitalares e o desenho de políticas públicas mais eficazes e equitativas. Assim, a ciência de dados emerge como um eixo estratégico para a modernização do cuidado em saúde, promovendo maior precisão, agilidade e racionalidade nas decisões clínicas, administrativas e governamentais.

A ciência de dados é uma área interdisciplinar que combina conhecimentos oriundos da estatística, aprendizado de máquina, inteligência artificial, domínio específico da área de aplicação (por exemplo, saúde e indústria), dentre outros [Provost and Fawcett 2013]. Seu principal objetivo é extrair conhecimento significativo a partir de grandes volumes de dados, transformando informações brutas em insights úteis para a tomada de decisões estratégicas. Para atingir esse objetivo, a ciência de dados segue um fluxo de trabalho bem definido, composto por diferentes etapas que se complementam.

A Figura 3.1 apresenta as principais etapas de ciência de dados na saúde. Especificamente, o fluxo de trabalho de ciência de dados envolve as seguintes etapas: (i) coleta de dados, que pode envolver desde a extração de registros eletrônicos até o uso de sensores e APIs de sistemas de saúde; (ii) organização e limpeza, que visa padronizar, corrigir erros e tratar valores ausentes ou duplicados; (iii) análise exploratória, responsável por identificar padrões, tendências e possíveis relações entre variáveis; (iv) modelagem e predição, onde técnicas estatísticas e algoritmos são utilizados para construir modelos preditivos ou classificatórios; (v) por fim, a interpretação dos resultados e comunicação, etapa crucial para traduzir os achados técnicos em recomendações práticas.


**Figura 3.1. Fluxo de trabalho típico em ciência de dados na área da saúde.**

Cada uma das etapas apresentadas anteriormente é essencial e interdependente, de modo que a qualidade de uma etapa afeta diretamente o desempenho das demais. Um erro na preparação dos dados, por exemplo, pode comprometer toda a análise subsequente. Por essa razão, compreender a estrutura desse processo e dominar as ferramentas adequadas para cada etapa é fundamental para o sucesso de qualquer projeto de ciência de dados, especialmente em domínios sensíveis e complexos como a saúde.

### 3.2.1. Papel da Ciência de Dados na Melhoria da Saúde Pública

Em sua essência, a saúde é uma área fortemente orientada por dados. Por exemplo, tomadas de decisões referentes a serviços como vigilância epidemiológica e criação de políticas públicas dependem fortemente da coleta, análise e interpretação de grandes volumes de dados [Provost and Fawcett 2013]. Portanto, nesse contexto, a ciência de dados surge como uma solução poderosa para prover compreensões relevantes para apoiar ações preventivas e otimizar alocações de recursos.

A ciência de dados aplicada à saúde possibilita a criação de soluções para monitorar e detectar surtos epidemiológicos em tempo real, reconhecer padrões comportamentais, delinear estratégias para implementação de campanhas de vacinação, otimizar fluxo de trabalho dentro de hospitais, dentre outros serviços [O'connor 2018, Moura et al. 2023]. Um exemplo palpável da aplicabilidade da ciência de dados na saúde foram as tomadas de decisão contra a pandemia de COVID-19 [Latif et al. 2020]. Os gestores e autoridades competentes delinearam ações baseadas em evidências levantadas através de técnicas de ciência de dados, como monitoramento contínuo de dados sobre hospitalizações, óbitos,

vacinação, áreas mais afetadas, dentre outros dados. Especificamente, foram desenvolvidas soluções como painéis interativos contendo diversos tipos de gráficos e modelos de aprendizado de máquina, que representaram ferramentas fundamentais no processo de tomadas de decisão para conter a pandemia de COVID-19.

Além de cenários críticos como a pandemia de COVID-19, a análise de dados desempenha um papel fundamental na vigilância constante de enfermidades crônicas, como transtornos mentais, diabetes, pressão alta e distúrbios cardíacos [Moura et al. 2023], favorecendo a elaboração de estratégias voltadas à prevenção e ao fortalecimento da saúde pública. A integração de informações oriundas de registros clínicos digitais, bancos de dados populacionais, sistemas hospitalares e até tecnologias vestíveis amplia as oportunidades de identificar tendências comportamentais e prever ameaças à saúde coletiva.

Contudo, a aplicação da ciência de dados no âmbito da saúde pública demanda uma abordagem criteriosa, que leve em consideração tanto desafios técnicos quanto implicações éticas. A manipulação de grandes volumes de informações sensíveis exige o compromisso com a proteção da privacidade e da confidencialidade dos dados pessoais, respeitando os direitos individuais dos cidadãos [Arellano et al. 2018]. Portanto, a aplicabilidade da ciência de dados na saúde deve ser aliada à governança de dados e princípios éticos que garantam a equidade em saúde.

### 3.2.2. Aplicações em Saúde: Diagnósticos, Epidemiologia e Tratamentos

A ciência de dados pode ser aplicada em diversos segmentos da saúde, apoiando o processo de tomada de decisão médica em nível clínico e estratégico [Subrahmanya et al. 2022]. Sua aplicabilidade se estende a serviços como auxílio em diagnósticos médicos, monitoramento contínuo de comportamentos, análise epidemiológica, tratamentos personalizados, dentre outros serviços de saúde fundamentados em dados.

Um dos usos mais promissores da ciência de dados é apoiar o diagnóstico e tomada de decisão médica [Teles et al. 2025]. Especificamente, pesquisadores têm analisado e processado dados médicos para implementar modelos de aprendizado de máquina capazes de classificar e predizer estados de saúde. Esses modelos são projetados usando principalmente dados provenientes de fontes como prontuários eletrônicos, imagens médicas e dados de sensores de dispositivos móveis e vestíveis [Teles et al. 2025]. Por exemplo, algoritmos aplicados à análise de imagens podem auxiliar na detecção de câncer, lesões pulmonares e doenças cardiovasculares com alta acurácia [Talwar et al. 2023]. Além disso, sistemas inteligentes têm sido capazes de analisar grandes volumes de dados para detectar padrões comportamentais (por exemplo, padrões de mobilidade, sono e sociabilidade) processando dados de dispositivos pessoais como *smartphone* e *smartwatch* [Moura et al. 2022].

## 3.3. Introdução ao Python

Antes de explorarmos as aplicações do Python na análise de dados, é imprescindível compreender seus conceitos fundamentais. Portanto, nesta seção, apresentaremos os principais conceitos da programação em Python, a saber: (i) indentação em Python; (ii) declaração de variáveis e tipos de dados; (iii) operadores aritméticos, relacionais e booleanos; e (iv) estruturas condicionais. O objetivo desta etapa é fornecer ao leitor uma base sólida para o estudo da análise de dados com Python.

### 3.3.1. Indentação em Python

Uma das características mais marcantes do Python é a indentação obrigatória. Isso significa que os blocos de código são definidos por espaços horizontais (espaços ou tabulações), e não o uso de delimitadores como chaves ({}) ou ponto e vírgula (;), comuns em diversas outras linguagens de programação.

A seguir, apresenta-se um exemplo que ilustra a indentação na linguagem Python. Observe que o recuo do código define, de maneira clara, quais instruções pertencem ao bloco de uma estrutura condicional.

```
1 if True:
2     print("Este bloco está corretamente indentado.")
```

Caso a indentação seja ignorada, o Python gerará um erro. Portanto, respeitar a indentação é essencial para garantir o funcionamento correto do seu programa.

### 3.3.2. Declaração de Variáveis e Tipos de Dados

Variáveis são espaços na memória usados para armazenar dados temporariamente. Em Python, não é necessário declarar o tipo da variável antes de usá-la — basta atribuir um valor com o sinal =. No entanto, é fundamental observar algumas regras para a definição de nomes de variáveis:

- Devem começar com uma letra ou com um underline (_).
- Podem conter letras (a-z, A-Z), números (0-9) e underline.
- Não podem conter espaços, acentos ou caracteres especiais como ($, #, @, !, etc.).
- Não podem usar palavras reservadas do Python, como `print`, `if`, `else`, `True`, `False`, `min`, `max`, etc.

### 3.3.3. Tipos de Dados em Python

A Tabela 3.3.3 apresenta os principais tipos de dados utilizados na linguagem Python, juntamente com suas descrições e exemplos práticos. Cada tipo de dado possui características específicas que determinam como as informações são armazenadas e manipuladas no programa.

A tabela apresenta desde os tipos numéricos, como int (números inteiros) e float (números decimais), até tipos mais complexos, como list e dict. O tipo str é utilizado para armazenar textos, enquanto o tipo bool é empregado para valores lógicos (verdadeiro ou falso). Já as coleções list e tuple permitem armazenar múltiplos valores, com a diferença de que listas são mutáveis (podem ser alteradas após a criação) e tuplas são imutáveis. Por fim, o tipo dict representa um dicionário, estrutura que utiliza pares chave-valor para organizar os dados.

Compreender esses tipos é fundamental para trabalhar de forma eficiente com variáveis e estruturas em Python, além de permitir o desenvolvimento de soluções mais organizadas e eficazes no tratamento de dados.

| **Tipo** | **Descrição** | **Exemplo** |
|---|---|---|
| `int` | Representa números inteiros, positivos ou negativos, sem parte decimal. | `10, -3, 0, 42` |
| `float` | Representa números reais (decimais), permitindo valores fracionários. | `3.14, 0.5, -2.7` |
| `str` | Representa sequências de caracteres, usadas para armazenar textos. | `'Olá', "Python", "123"` |
| `bool` | Representa valores lógicos, indicativos de verdadeiro ou falso. | `True, False` |
| `list` | Coleção ordenada e mutável de elementos, que podem ser de tipos variados. | `[1, 2, 3], ["a", "b", "c"]` |
| `tuple` | Coleção ordenada e imutável de elementos, definida por parênteses. | `(1, 2, 3), ("a", "b")` |
| `dict` | Estrutura de dados composta por pares chave-valor, usada para mapeamentos. | `{"nome": "Ana", "idade": 25}` |

Entender a diferença entre cada um dos tipos de dados e seu funcionamento é essencial para trabalhar com diferentes tipos de datasets e realizar o tratamento correto dos dados.

### 3.3.4. Operadores Aritméticos, Relacionais e Booleanos

Em Python, os operadores matemáticos, relacionais e booleanos são amplamente utilizados ao longo do código. Conhecê-los é essencial para a manipulação de dados, análises e comparações.

Os operadores aritméticos são utilizados para realizar operações matemáticas básicas entre números. A Tabela 3.3.4 apresenta os principais operadores aritméticos utilizados na linguagem Python, acompanhados de suas respectivas descrições e exemplos. Esses operadores são fundamentais para a realização de cálculos e expressões matemáticas nos programas.

| **Operador** | **Descrição** | **Exemplo** |
|---|---|---|
| `+` | Realiza a adição entre dois valores. | `10 + 5 → 15` |
| `-` | Realiza a subtração entre dois valores. | `10 - 3 → 7` |
| `*` | Realiza a multiplicação entre dois valores. | `4 * 2 → 8` |
| `/` | Realiza a divisão entre dois valores. | `10 / 2 → 5.0` |
| `**` | Calcula a exponenciação (potência). | `2 ** 3 → 8` |
| `%` | Retorna o resto da divisão (módulo). | `10 % 3 → 1` |

A Tabela 3.3.4 apresenta de forma concisa algumas das funções matemáticas mais usadas em Python, mostrando o que fazem e como aplicá-las em exemplos práticos.

| **Função** | **O que Faz** | **Exemplo** |
|---|---|---|
| `abs(x)` | Converte qualquer número para sua magnitude positiva. | `abs(-12) → 12` |
| `pow(x, y)` | Eleva `x` à potência `y`; aceita opcionalmente um módulo. | `pow(3, 4) → 81` |
| `sqrt(x)` | Retorna a raiz quadrada de `x`; importe com `from math import sqrt`. | `sqrt(9) → 3.0` |
| `min(iterável)` | Percorre uma coleção e devolve o menor valor; também aceita múltiplos args. | `min([7, 2, 5]) → 2` |
| `max(iterável)` | Retorna o maior elemento de uma sequência ou vários argumentos. | `max([7, 2, 5]) → 7` |

Os operadores relacionais são usados para comparar dois valores e retornam um valor booleano (True ou False), indicando se a comparação é verdadeira ou falsa. A Tabela 3.3.4 apresenta os principais operadores relacionais em Python, suas descrições e exemplos de uso.

| **Operador** | **Descrição** | **Exemplo** |
|---|---|---|
| == | Igual a: verifica se dois valores são iguais. | `5 == 3 → False` |
| != | Diferente de: verifica se dois valores são diferentes. | `5 != 3 → True` |
| > | Maior que: testa se o valor à esquerda é maior que o da direita. | `7 > 2 → True` |
| >= | Maior ou igual a: verifica se um valor é maior ou igual ao outro. | `5 >= 5 → True` |
| < | Menor que: testa se o valor à esquerda é menor que o da direita. | `2 < 7 → True` |
| <= | Menor ou igual a: verifica se um valor é menor ou igual ao outro. | `3 <= 3 → True` |

Os operadores booleanos são usados para realizar operações lógicas entre expressões que retornam valores booleanos (True ou False). A Tabela 3.3.4 apresenta os principais operadores booleanos em Python, suas descrições e exemplos de uso.

| **Operador** | **Descrição** | **Exemplo** |
|---|---|---|
| `and` | Retorna `True` se ambas as condições forem verdadeiras. | `True and False → False` |
| `or` | Retorna `True` se pelo menos uma das condições for verdadeira. | `True or False → True` |
| `not` | Inverte o valor lógico: `True` vira `False` e vice-versa. | `not True → False` |

### 3.3.5. Estruturas Condicionais

Estruturas condicionais são utilizadas para alterar o fluxo do código com base em regras e parâmetros definidos pelo desenvolvedor. Um exemplo clássico de aplicação de estruturas condicionais é o desenvolvimento de um sistema de avaliação de desempenho acadêmico. Suponha que precisamos determinar a situação de um aluno com base na sua nota final, seguindo os critérios abaixo:

- Nota maior ou igual a 7 → Aprovado
- Nota maior ou igual a 4 e menor que 7 → Recuperação
- Nota abaixo de 4 → Reprovado

A implementação dessa lógica em Python pode ser feita utilizando uma estrutura condicional `if-elif-else`, que seleciona o resultado adequado com base na nota fornecida. Veja como isso ficaria no código:

```
notafinal = 5  # Valor de exemplo

# Estrutura de Decisão
if notafinal >= 7:  # SE notafinal for maior ou igual a 7,
   execute isso
    print("O aluno(a) está aprovado!")

elif notafinal >= 4:  # SENÃO, MAS SE notafinal for maior ou
   igual a 4, execute isso
    print("O aluno(a) está de recuperação!")

else:  # SENÃO, execute isso
    print("O aluno(a) está reprovado!")
```

## 3.4. Fundamentos do Python para Análise de Dados

Python é uma linguagem de programação de alto nível, amplamente reconhecida por sua versatilidade, clareza sintática e ampla aplicabilidade em diferentes áreas do conhecimento. Nesta seção, será apresentada a utilização do Python como uma ferramenta fundamental no contexto da análise de dados. Especificamente, iremos abordar os conceitos e fundamentos essenciais que servirão de base para o seu correto uso ao longo deste estudo.

### 3.4.1. Estruturas de Dados: DataFrame e Series no Pandas

O **Pandas** é uma biblioteca de código-aberto para Python que fornece ferramentas de alto desempenho e fáceis de usar para manipulação e análise de dados. Seu principal atrativo é a capacidade de trabalhar de forma eficiente com dados tabulares (planilhas, bases relacionais, arquivos CSV, etc.), permitindo filtrar, agrupar, transformar e resumir conjuntos de dados de forma muito mais simples do que usando apenas listas e dicionários.

**Series** são estruturas unidimensionais do Pandas: vetores rotulados capazes de armazenar qualquer tipo de dado (numérico, texto, data, etc.). Cada elemento de uma Series possui um índice (label), o que facilita:

- Seleção de faixas e alinhamento automático em operações aritméticas;
- Integração com outras bibliotecas (NumPy, Matplotlib etc.).

**DataFrame** é a estrutura bidimensional do Pandas, equivalente a uma tabela de banco de dados ou a uma planilha de Excel. Internamente, um DataFrame é composto por várias

Series compartilhando o mesmo índice de linhas, mas com colunas independentes que podem ter tipos diferentes. Isso o torna ideal para:

- *Leitura e escrita* de formatos diversos (CSV, Excel, JSON, SQL);
- *Seleção e filtragem* de linhas e colunas de forma intuitiva;
- *Transformações vetorizadas*, aplicando funções a colunas inteiras em uma única operação;
- *Agrupamentos e agregações* para resumos estatísticos (soma, média, contagem, etc.);
- *Junções e mesclagens* de tabelas distintas, como em bancos relacionais.

### 3.4.2. Preparando o Ambiente

Antes de iniciar a análise de dados, é fundamental garantir que o ambiente esteja devidamente configurado. Além do Pandas utilizaremos bibliotecas como, NumPy, Seaborn e Plotly. Recomendamos o uso do Google Colab (`https://colab.research.google.com/`) ou Jupyter Notebook (`https://jupyter.org/`), pois essas plataformas facilitam a visualização de resultados e gráficos interativos.

Se estiver utilizando um ambiente local, você pode instalar as bibliotecas com o seguinte comando:

```
!pip install pandas numpy seaborn plotly
```

No Google Colab, essas bibliotecas geralmente já estão instaladas por padrão. Neste trecho de código abaixo, realizamos a importação das bibliotecas que serão utilizadas ao longo de todo o conteúdo.

```
import pandas as pd
import numpy as np
import seaborn as sns
import plotly.express as px
import matplotlib.pyplot as plt
```

### 3.4.3. Importação de Dados para Análise

Nesta seção, utilizaremos um dataset sobre diabetes, que contém informações sobre pacientes e diversos fatores que podem influenciar o risco de desenvolver a doença. Para fazer o download do arquivo `.csv`, acesse o link para download do CSV.

O objetivo é aplicar técnicas de análise de dados para explorar, entender e extrair insights valiosos a partir dessas informações. O dataset inclui variáveis como: número de gestações, níveis de glicose, pressão arterial, espessura da pele, insulina, IMC, histórico familiar de diabetes, idade e o resultado do teste de diabetes.

Agora, vamos carregar o conjunto de dados que utilizaremos ao longo das próximas seções. Abaixo, apresentamos o código para ler diretamente o CSV do Google Drive usando o método `read_csv` do Pandas:

```
document_id = "1kH99hEJLsT1B2d5DWUUN2i1ATRdiy9Uj"
url = f"https://drive.google.com/uc?id={document_id}&export=
    download"

diabetes_df = pd.read_csv(url)
```

Se preferir, você também pode baixar o arquivo pelo link acima e enviá-lo manualmente ao Google Colab (aba "Files" → "Upload"). Em seguida, basta carregá-lo com:

```
import pandas as pd

diabetes_df = pd.read_csv("diabetes.csv")
```

### 3.4.4. Manipulação de Dados

A manipulação de dados é uma etapa essencial na análise. É nela que organizamos e preparamos os dados para que estejam prontos para serem analisados, incluindo ações como renomear colunas, criar novas variáveis e visualizar estatísticas básicas.

Para facilitar a leitura e interpretação do conjunto de dados, é recomendável renomear e padronizar os nomes das colunas, tornando-os mais claros e descritivos. No exemplo a seguir, realizamos essa padronização atribuindo novos nomes diretamente ao atributo columns do DataFrame.

```
diabetes_df.columns = [
    'Gravidez',
    'Glicotese',
    'Pressao',
    'Pele',
    'Insulina',
    'IMC',
    'Hereditariedade',
    'Idade',
    'Diabetes'
]
```

Dessa forma, os nomes originais são substituídos por termos em português, que representam de forma clara as informações contidas em cada coluna do dataset.

#### 3.4.4.1. Explorando Estatísticas Descritivas

Uma das primeiras etapas na análise exploratória de dados é obter uma visão geral das principais características numéricas do conjunto de dados. O Pandas oferece o método

describe(), que fornece estatísticas como média, desvio padrão, valores mínimos e máximos para cada coluna numérica. É essencial para obter uma visão geral dos dados e identificar possíveis valores extremos. O código abaixo apresenta seu uso no DataFrame diabetes_df e a Figura 3.2 apresenta o resultado da execução desse código.

```
diabetes_df.describe()
```

| | Gravidez | Glicotese | Pressao | Pele | Insulina | IMC | Hereditariedade | Idade | Diabetes |
|---|---|---|---|---|---|---|---|---|---|
| count | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 |
| mean | 3.845052 | 120.894531 | 69.105469 | 20.536458 | 79.799479 | 31.992578 | 0.471876 | 33.240885 | 0.348958 |
| std | 3.369578 | 31.972618 | 19.355807 | 15.952218 | 115.244002 | 7.884160 | 0.331329 | 11.760232 | 0.476951 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.078000 | 21.000000 | 0.000000 |
| 25% | 1.000000 | 99.000000 | 62.000000 | 0.000000 | 0.000000 | 27.300000 | 0.243750 | 24.000000 | 0.000000 |
| 50% | 3.000000 | 117.000000 | 72.000000 | 23.000000 | 30.500000 | 32.000000 | 0.372500 | 29.000000 | 0.000000 |
| 75% | 6.000000 | 140.250000 | 80.000000 | 32.000000 | 127.250000 | 36.600000 | 0.626250 | 41.000000 | 1.000000 |
| max | 17.000000 | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 | 2.420000 | 81.000000 | 1.000000 |

**Figura 3.2. Resultado da Função `Describe`.**

O método describe() gera um resumo estatístico automático das colunas numéricas que incluem as seguintes informações:

- count: quantidade de valores não nulos;
- mean: média aritmética;
- std: desvio padrão;
- min: valor mínimo;
- 25%, 50% e 75%: quartis (percentis que indicam a distribuição dos dados);
- max: valor máximo.

### 3.4.5. Selecionando Linhas e Colunas

Uma das tarefas fundamentais na análise de dados é a seleção de informações específicas dentro de um DataFrame. O Pandas oferece maneiras simples e flexíveis de acessar tanto colunas quanto linhas, permitindo que o analista foque apenas nos dados relevantes para cada etapa da análise.

Nas primeiras etapas, é comum realizar uma inspeção inicial para compreender sua estrutura e verificar se a leitura foi realizada corretamente. Para isso, é possível exibir rapidamente o início e o fim do DataFrame usando os métodos ‘head()‘ e ‘tail()‘, conforme ilustrado nas Figuras 3.3 e 3.4.

```
# Exibe as 5 primeiras linhas (padrão)
diabetes_df.head()
```

| | Gravidez | Glicotese | Pressao | Pele | Insulina | IMC | Hereditariedade | Idade | Diabetes |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 2 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 3 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 4 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

**Figura 3.3. Resultado da função `Head`.**

```
# Exibe as 10 últimas linhas (definido pelo parâmetro)
diabetes_df.tail(10)
```

| | Gravidez | Glicotese | Pressao | Pele | Insulina | IMC | Hereditariedade | Idade | Diabetes |
|---|---|---|---|---|---|---|---|---|---|
| 758 | 1 | 106 | 76 | 0 | 0 | 37.5 | 0.197 | 26 | 0 |
| 759 | 6 | 190 | 92 | 0 | 0 | 35.5 | 0.278 | 66 | 1 |
| 760 | 2 | 88 | 58 | 26 | 16 | 28.4 | 0.766 | 22 | 0 |
| 761 | 9 | 170 | 74 | 31 | 0 | 44.0 | 0.403 | 43 | 1 |
| 762 | 9 | 89 | 62 | 0 | 0 | 22.5 | 0.142 | 33 | 0 |
| 763 | 10 | 101 | 76 | 48 | 180 | 32.9 | 0.171 | 63 | 0 |
| 764 | 2 | 122 | 70 | 27 | 0 | 36.8 | 0.340 | 27 | 0 |
| 765 | 5 | 121 | 72 | 23 | 112 | 26.2 | 0.245 | 30 | 0 |
| 766 | 1 | 126 | 60 | 0 | 0 | 30.1 | 0.349 | 47 | 1 |
| 767 | 1 | 93 | 70 | 31 | 0 | 30.4 | 0.315 | 23 | 0 |

**Figura 3.4. Resultado da função `Tail`.**

Observação: Tanto o head() quanto o tail() permitem a definição da quantidade de linhas que se deseja visualizar, passando um número inteiro como argumento.

Além de exibir as primeiras e últimas linhas de um DataFrame, também é possível acessar linhas de qualquer intervalo, o que é bastante útil para examinar partes específicas dos dados sem precisar percorrer todo o conjunto. No exemplo a seguir, utilizamos a notação de fatiamento do Pandas para exibir as linhas que estão no intervalo de índices 20 a 24 (Figura 3.5).

```
diabetes_df[20:25]
```

| | Gravidez | Glicotese | Pressao | Pele | Insulina | IMC | Hereditariedade | Idade | Diabetes |
|---|---|---|---|---|---|---|---|---|---|
| **20** | 3 | 126 | 88 | 41 | 235 | 39.3 | 0.704 | 27 | 0 |
| **21** | 8 | 99 | 84 | 0 | 0 | 35.4 | 0.388 | 50 | 0 |
| **22** | 7 | 196 | 90 | 0 | 0 | 39.8 | 0.451 | 41 | 1 |
| **23** | 9 | 119 | 80 | 35 | 0 | 29.0 | 0.263 | 29 | 1 |
| **24** | 11 | 143 | 94 | 33 | 146 | 36.6 | 0.254 | 51 | 1 |

**Figura 3.5. Resultado do intervalo `[20:25]`.**

Essa notação remete a uma lista, onde o primeiro valor representa o índice inicial (inclusive) e o segundo valor indica o índice final (exclusive). Ou seja, no Python, os intervalos sempre param no número anterior ao final especificado. Portanto, o código acima mostrará as linhas de índice 20 a 24.

Após a inspeção inicial dos dados, é comum surgir a necessidade de realizar seleções mais específicas, como acessar colunas, linhas ou até mesmo valores individuais dentro do DataFrame. Para esse propósito, o Pandas disponibiliza dois métodos principais de seleção:

- .loc[]: permite acessar dados com base nos rótulos do índice.
- .iloc[]: realiza a seleção com base na posição numérica, onde tanto linhas quanto colunas são indexadas a partir de zero.

O exemplo abaixo ilustra o uso dos métodos .loc[] e .iloc[] para acessar linhas específicas de um DataFrame:

```
# Usando loc para acessar uma linha específica por índice
diabetes_df.loc[10]

# Usando iloc para acessar uma linha pela posição
diabetes_df.iloc[10]

# Intervalo de linhas com iloc (da linha 5 até a 9)
diabetes_df.iloc[5:10]

# Acessando um valor específico (linha 0, coluna 'Idade')
df.loc[0, 'Idade']
```

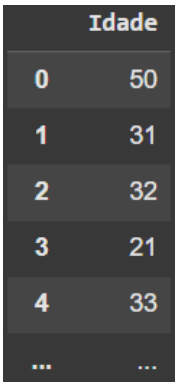
Essas ferramentas são essenciais quando queremos inspecionar, editar ou filtrar os dados de forma precisa.

#### 3.4.5.1. Selecionando Colunas

Além de selecionar linhas, também é possível acessar colunas de forma simples, utilizando os nomes definidos no DataFrame. A Figura 3.6 ilustra a seleção da coluna 'Idade' do DataFrame 'diabetes_df'.

```
diabetes_df['Idade']
```



| | Idade |
|---|---|
| 0 | 50 |
| 1 | 31 |
| 2 | 32 |
| 3 | 21 |
| 4 | 33 |
| ... | ... |

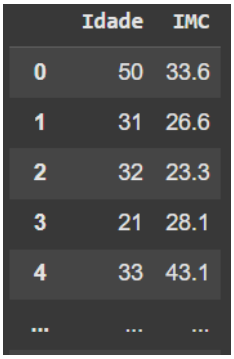
**Figura 3.6. Selecionando a coluna `idade`.**

Explicação: Basta informar o nome do DataFrame (diabetes_df) seguido por colchetes com o nome da coluna desejada entre aspas simples ou duplas.

Para selecionar duas ou mais colunas simultaneamente, basta passar uma lista com os nomes das colunas desejadas dentro dos colchetes:

```
diabetes_df[['Idade', 'IMC']]
```

Dessa forma, retornamos apenas as colunas especificadas, o que é útil para análises focadas em variáveis específicas, como mostrado na Figura 3.7.



| | Idade | IMC |
|---|---|---|
| 0 | 50 | 33.6 |
| 1 | 31 | 26.6 |
| 2 | 32 | 23.3 |
| 3 | 21 | 28.1 |
| 4 | 33 | 43.1 |
| ... | ... | ... |

**Figura 3.7. Selecionando as colunas `idade e IMC`.**

### 3.4.6. Operações de Agregamento e Agrupamento

Ao trabalhar com conjuntos de dados, muitas vezes é necessário resumir informações ou identificar padrões a partir de agrupamentos. O Pandas oferece recursos poderosos para realizar operações de agregação, como soma, média, contagem, valor máximo e mínimo, entre outras.

#### 3.4.6.1. Agregações Básicas

Utilizando funções de agregação do Pandas, é possível obter rapidamente dados relevantes sobre o DataFrame analisado. O Pandas oferece diversas funções de agregação que permitem calcular dados estatísticos essenciais sobre um conjunto de dados. Entre as funções mais comuns estão:

- `count()`: retorna o número de valores não nulos em cada coluna ou série.
- `min()`: retorna o valor mínimo observado em cada coluna numérica.
- `max()`: retorna o valor máximo observado em cada coluna numérica.
- `mean()`: calcula a média aritmética dos valores de cada coluna numérica.
- `std()`: calcula o desvio-padrão dos valores de cada coluna numérica.

O código a seguir ilustra o uso de funções de agregação para realizar cálculos sobre diferentes colunas do 'diabetes_df'. Além disso, a Figura 3.8 apresenta o resultados da execução desses códigos.

```
# Média da idade
diabetes_df['Idade'].mean()

# Maior valor de IMC
diabetes_df['IMC'].max()

# Menor valor de Glicotese
diabetes_df['Glicotese'].min()

# Contagem de entradas na coluna 'Diabetes'
diabetes_df['Diabetes'].count()

# Retorna o desvio padrão dos valores da coluna
diabetes_df['Insulina'].std()
```

| | |
|---|---|
| Média da idade | 33,240885416666664 |
| Maior valor de IMC | 67,1 |
| Menor valor de Glicotese | 0 |
| Contagem de ntradas na coluna 'Diabetes' | 768 |
| Desvio padrão da Insulina | 115,24408931535337 |

**Figura 3.8. Resultado das funções de agregação no DataFrame diabetes.**

### 3.4.6.2. Estatísticas Descritivas Agrupadas por Categorias

O Pandas fornece o método groupby() para segmentar dados em grupos com base em uma ou mais colunas. Esse método permite realizar operações específicas sobre cada grupo, facilitando a análise de dados de maneira estruturada e eficiente. Ao aplicar o groupby(), os dados são divididos em subconjuntos, e funções agregadoras podem ser aplicadas a cada grupo individualmente.

Abaixo, apresentamos um código que utiliza o método groupby() para realizar análises segmentadas com base na coluna 'Diabetes' do conjunto de dados 'diabetes_df'. A seguir, aplicamos a função de agregação 'mean()' para calcular o valor médio para cada grupo. Enfatizamos que poderíamos ter utilizado qualquer outra função de agregação em combinação com o método groupby(), como max(), min(), std(), dentre outros. A Figura 3.9 apresenta o resultado da execução desse código.

```
# Média de idade para cada grupo (diabéticos e não-diabéticos)
diabetes_df.groupby('Diabetes')['Idade'].mean()
```

| Diabetes | Idade |
|---|---|
| 0 | 31.190000 |
| 1 | 37.067164 |

dtype: float64

**Figura 3.9. Resultado do agrupamento pelas colunas Diabetes e Idade.**

No exemplo de código abaixo, utilizamos o método groupby() combinado com a função agg() para calcular várias métricas de agregação para diferentes colunas do DataFrame, agrupando os dados com base na coluna 'Diabetes' (diabéticos e não-diabéticos). A Figura 3.10 apresenta o resultado da execução do código abaixo.

```
# Várias métricas por grupo
diabetes_df.groupby('Diabetes').agg({
    'Idade': ['mean', 'max'],
    'IMC': ['mean', 'min']
}).reset_index() # A função reset_index serve para ajustar a
   visualização.
```

| Diabetes | Idade | | IMC | |
|---|---|---|---|---|
| | mean | max | mean | min |
| 0 | 31.190000 | 81 | 30.304200 | 0.0 |
| 1 | 37.067164 | 70 | 35.142537 | 0.0 |

**Figura 3.10. Resultado da função `Group`.**

#### 3.4.6.3. Contagem de Ocorrências Únicas

O pandas fornece o método value_counts() para contar a frequência de valores únicos em uma coluna de um DataFrame. Esse método realiza a contagem de cada valor único presente, ordenada de forma decrescente (do valor mais frequente para o menos frequente). Por exemplo, no código abaixo, esse método foi utilizado para contar a frequência de cada valor único na coluna 'Gravidez' do DataFrame 'diabetes_df'. A Figura 3.11 apresenta o retorno da execução desse código. Especificamente, é retornada uma Série com os valores únicos encontrados na coluna 'Gravidez' e suas respectivas contagens, ordenadas do valor mais frequente para o menos frequente.

```
diabetes_df['Gravidez'].value_counts()
```

| Gravidez | |
|---|---|
| 1 | 135 |
| 0 | 111 |
| 2 | 103 |
| 3 | 75 |
| 4 | 68 |
| 5 | 57 |

**Figura 3.11. Resultado da função `value_counts`.**

### 3.4.7. Filtros

Filtrar dados significa criar subconjuntos que atendam a uma condição lógica. Isso é útil, por exemplo, para estudar apenas pacientes com IMC elevado ou apenas aqueles acima

de certa idade. O código abaixo apresenta exemplos de filtros de dados. O primeiro filtra pacientes com idades superiores a 30 anos. O segundo código filtra pacientes que possuem mais de trinta anos e que apresentam IMC acima de 25. Por fim, o último código filtra pacientes com diabetes ou que apresentem IMC acima de 35.

```
# Pacientes com mais de 30 anos
diabetes_df[diabetes_df['Idade'] > 30]

# Pacientes com mais de 30 anos E IMC acima de 25
diabetes_df[(diabetes_df['Idade'] > 30) & (diabetes_df['IMC'] >
    25)]

# Pacientes com diabetes OU IMC acima de 35
diabetes_df[(diabetes_df['Diabetes'] == 1) | (diabetes_df['IMC']
     > 35)]
```

Lembre-se de sempre utilizar parênteses em torno das condições ao usar os operadores lógicos & (E) e | (OU).

### 3.4.8. Ordenação

Para ordenar os dados em ordem crescente ou decrescente, utilizamos o método sort_values(). Esse método permite organizar os dados com base em uma ou mais colunas, por meio do parâmetro by. Já o parâmetro ascending define se a ordenação será crescente (True) ou decrescente (False). No código abaixo, apresentamos alguns exemplos de ordenação de dados. O primeiro exemplo, diabetes_df.sort_values('Idade') ordena os dados pela coluna 'Idade' em ordem crescente. Já diabetes_df.sort_values('IMC', ascending=False) organiza os dados da coluna 'IMC' do maior para o menor valor. Também é possível ordenar por múltiplos critérios, como em diabetes_df.sort_values(by=['Diabetes', 'Idade'], ascending=[False, True]), que ordena primeiro pela coluna 'Diabetes' em ordem decrescente e, em seguida, pela coluna 'Idade' em ordem crescente.

```
# Ordenar por idade (crescente)
diabetes_df.sort_values('Idade')

# Ordenar por IMC (decrescente)
diabetes_df.sort_values('IMC', ascending=False)

# Ordenar por múltiplos critérios
diabetes_df.sort_values(by=['Diabetes', 'Idade'], ascending=[
    False, True])
```

Ordenar os dados é uma prática comum para facilitar análises, gerar gráficos ou destacar os extremos.

## 3.5. Limpeza e Tratamentos de Dados

A limpeza e tratamento de dados é uma das etapas cruciais no processo de análise de dados. Independentemente da fonte ou formato, raramente os dados chegam em condições

ideais para análise. Dados incompletos, inconsistentes, duplicados ou fora do padrão são desafios comuns enfrentados por cientistas de dados. Nessa fase, o objetivo é garantir que o conjunto de dados esteja confiável, coerente e pronto para ser explorado com técnicas analíticas e modelos preditivos. Portanto, nesta seção, objetivamos realizar a limpeza e o tratamento de dados utilizando as ferramentas do *python*, com foco na biblioteca Pandas.

### 3.5.1. Configuração do Ambiente

A execução das etapas de limpeza e tratamento de dados demonstradas nessa seção usarão as bibliotecas Pandas, NumPy e Seaborn. Como apresentado nas seções anteriores, essas ferramentas são amplamente exploradas na análise de dados e oferecem funcionalidades robustas para manipulação, inspeção e visualização de dados. Portanto, abaixo apresentamos o código de importação necessário para usar essas bibliotecas.

```
import pandas as pd
import numpy as np
import seaborn as sns
```

A seguir, apresentamos o código que cria um DataFrame simulado com dados de saúde, representando um conjunto de informações coletadas em um estudo clínico fictício com pacientes. O objetivo é reproduzir, de forma prática, situações comuns encontradas na preparação de dados reais, como: (i) presença de valores ausentes (por falhas na coleta ou entrada de dados); (ii) registros duplicados (por exemplo, quando um paciente é registrado mais de uma vez); (iii) valores discrepantes (outliers), que podem ser causados por erros de digitação ou por dados extremos fora da faixa esperada. A opção por dados simulados permite controlar intencionalmente a presença de valores ausentes, duplicados e discrepantes, garantindo que cada tipo de problema possa ser demonstrado e tratado de forma clara e objetiva.

```
dados_saude = pd.DataFrame({
    'id': [
        101, 102, 103, 104, 105, 106, 106, 107
    ],
    'nome': [
        'Ana', 'Bruno', 'Carlos', 'Diana', 'José',
        'Fernando', 'Fernando', 'Gabriela'
    ],
    'idade': [
        29, 35, 42, np.nan, 150, 38, 38, 27
    ],
    'peso_kg': [
        65.0, 80.5, np.nan, 70.0, 60.0,
        300.0, 300.0, 48.0
    ],
    'pressao_sistolica': [
        120, 130, 125, 118,
        119, 250, 250, np.nan
    ],
    'fumante': [
```

```
        'Não', 'Sim', 'Não', 'Não', 'Sim',
        np.nan, np.nan, 'Não'
    ],
    'data_coleta': [
        2024-01-10', '2024-01-12', '2024-01-15',
        '2024-01-17', '2024-01-19', '2024-01-21',
        '2024-01-21', '2024-01-23'
    ]
})
```

Ao executar o código anterior, resultará na criação do DataFrame denominado 'dados_saude' apresentado na Figura 3.12, que contém as seguintes colunas: 'id', 'nome', 'idade', 'peso_kg', 'pressao_sistolica', 'fumante' e 'data_coleta'. Esse DataFrame já é suficiente para trabalhar os seguintes pontos didáticos: valores ausentes (np.nan em idade, peso_kg, pressao_sistolica, fumante), tratamento de dados duplicados (id e nome duplicados no caso de Fernando); valores discrepantes (idade = 150, peso_kg = 300, pressao_sistolica = 250) e conversão de tipo de dados.

| | id | nome | idade | peso_kg | pressao_sistolica | fumante | data_coleta |
|---|---|---|---|---|---|---|---|
| 0 | 101 | Ana | 29.0 | 65.0 | 120.0 | Não | 2024-01-10 |
| 1 | 102 | Bruno | 35.0 | 80.5 | 130.0 | Sim | 2024-01-12 |
| 2 | 103 | Carlos | 42.0 | NaN | 125.0 | Não | 2024-01-15 |
| 3 | 104 | Diana | NaN | 70.0 | 118.0 | Não | 2024-01-17 |
| 4 | 105 | Eduarda | 150.0 | 60.0 | 119.0 | Sim | 2024-01-19 |
| 5 | 106 | Fernando | 38.0 | 300.0 | 250.0 | NaN | 2024-01-21 |
| 6 | 106 | Fernando | 38.0 | 300.0 | 250.0 | NaN | 2024-01-21 |
| 7 | 107 | Gabriela | 27.0 | 48.0 | NaN | Não | 2024-01-23 |

**Figura 3.12. Dataframe com dados simulados de saúde contendo valores ausentes, duplicados e discrepantes.**

### 3.5.2. Análise Inicial do DataFrame

Antes de iniciar qualquer processo de limpeza e tratamento de dados, é fundamental entender a estrutura básica do conjunto de dados que estamos manipulando. Uma das primeiras ferramentas que o pandas oferece para essa tarefa é o método **.info()**. A Figura 3.13 apresenta o resultado ao executar o comando 'dados_saude.info()'. A partir desse resultado, é possível identificar as seguintes informações: (i) o DataFrame possui 8 linhas e 7 colunas; (ii) existem valores ausentes nas colunas idade, peso_kg, pressao_sistolica, pois a quantidade de valores não nulos é menor que a quantidade total de linhas; e (iii) as colunas estão com tipos incorretos de dados (por exemplo, idade está float e data_coleta está object).

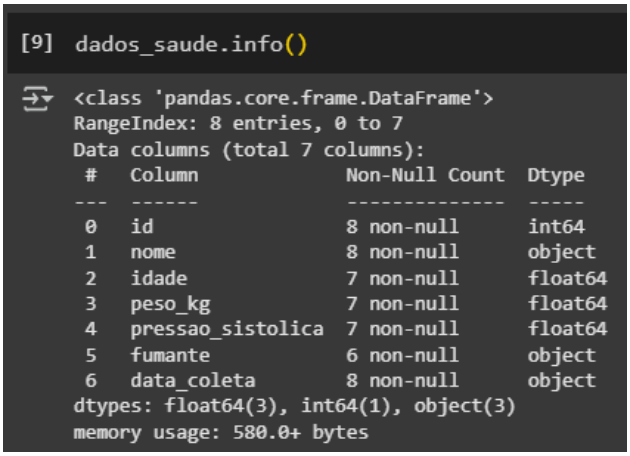



**Figura 3.13. Resultado da execução do comando .info() aplicado ao DataFrame dados_saude.**

Neste momento temos uma visão global de nosso conjunto de dados e, identificamos problemas que devem ser tratados, como valores ausentes, e colunas com tipos incorretos. Nas próximas seções, apresentaremos um conjunto de técnicas de higienização de dados que serão aplicadas a esse DataFrame.

### 3.5.3. Tratamento de Valores Ausentes

Em dados do mundo real, é comum encontrar informações ausentes em conjuntos de dados, que são conhecidas como valores nulos ou NaN (Not a Number). Tratar esses valores corretamente é uma tarefa fundamental de higienização de dados, já que eles podem atrapalhar análises, distorcer gráficos e comprometer resultados de modelos de aprendizado de máquina.

O primeiro passo para lidar com valores ausentes é identificar quais colunas contêm essas lacunas. Para isso, o Pandas oferece o método **.isna()**, que verifica cada posição do DataFrame e retorna uma tabela com valores booleanos: True onde há dados ausentes e False onde os dados estão preenchidos. Como o resultado mantém o mesmo formato do DataFrame original, fica fácil visualizar onde estão as falhas. Além disso, combinando esse método com **.sum()**, é possível contar o total de valores ausentes em cada coluna. Abaixo apresentamos o código que realiza a contagem de valores ausentes no DataFrame dados_saude. A Figura 3.14 apresenta o resultado do comando que realiza a contagem de valores ausentes. Ao analisar o resultado, observamos que há exatamente um valor ausente nas colunas nome, idade e peso_kg. Já a coluna fumante apresenta dois valores ausentes.

```
1 dados_saude.isna().sum()
```


**Figura 3.14. Resultado da contagem de valores ausentes no DataFrame dados_saude.**

Após identificar os valores ausentes, é necessário preenchê-los com um valor específico, como a média, mediana, zero ou uma string padrão. Com esse objetivo, o pandas fornece o método **.filna()**, que é capaz de preencher as lacunas vazias por dados de nosso interesse. O código abaixo trata os valores ausentes de dados_saude usando o método **.fillna()**. Primeiramente, preenchemos o valor ausente da coluna idade com a mediana ( método **.median()** ) das idades. Em seguida, preenchemos o valor faltante da coluna peso_kg com a média dos pesos ( **.mean()** ). Para a coluna pressao_sistolica usamos a estratégia de substituir pelo valor zero. Por fim, na coluna fumante, preenchemos os valores ausentes com a mensagem 'Não informado'.

```
idade_median = dados_saude['idade'].median()
dados_saude['idade'] = dados_saude['idade'].fillna(idade_median)

peso_mean = dados_saude['peso_kg'].mean()
dados_saude['peso_kg'] = dados_saude['peso_kg'].fillna(peso_mean
    )

dados_saude['pressao_sistolica'] = dados_saude['
    pressao_sistolica'].fillna(0)

dados_saude['fumante'] = dados_saude['fumante'].fillna('Não
    informado')
```

Em algumas situações, é necessário realizar a remoção das linhas que contêm valores ausentes, em vez de preenchê-los. Para este objetivo, o pandas fornece o método **.dropna()**, que elimina todas as linhas que contenham valores ausentes. Além disso, o método .dropna() permite o uso do parâmetro inplace=True. Quando esse parâmetro é definido como verdadeiro, a remoção das linhas ocorre diretamente no DataFrame original, sem a necessidade de criar uma nova variável ou sobrescrever manualmente. Caso o inplace não seja especificado ou seja definido como False, o método apenas retorna uma nova versão do DataFrame, deixando o original inalterado. Portanto, em nosso exemplo, caso a melhor decisão de higienização de dados fosse remover todas as linhas com valores ausentes, poderíamos executar o código abaixo.

```
dados_saude.dropna(inplace=True)
```

Neste momento, se executarmos novamente o comando **dados_saude.isna().sum()**, verificaremos que não existem mais valores ausentes, uma vez que tratamos usando o método **.fillna()** e o método **.dropna()**.

### 3.5.4. Remoção de Duplicatas

Outro problema comum que encontramos em conjuntos de dados é a duplicação de registros, que também pode distorcer as análises e visualizações de dados geradas e, portanto, precisa ser tratada cuidadosamente. O primeiro passo para tratar desse problema é identificar se o nosso DataFrame possui dados duplicados. Para isso, o pandas fornece o método .duplicated(), que é utilizado para identificar linhas repetidas dentro de um DataFrame. Ele retorna uma lista de valores booleanos, em que cada linha é marcada como True se for uma cópia exata de alguma linha anterior (considerando todas as colunas, por padrão). O código abaixo usa esse comando para verificar linhas duplicadas no DataFrame dados_saude. A Figura 3.15 apresenta o resultado da execução desse código, no qual é possível identificar que a linha 6 representa um registro duplicado, pois retornou um valor booleano verdadeiro.

```
dados_saude.duplicated()
```


**Figura 3.15. Resultado da identificação de linhas duplicadas no DataFrame dados_saude.**

O próximo passo que iremos executar é a remoção dessa linha duplicada. O pandas fornece o método **.drop_duplicates()** que é capaz de remover todos os registros duplicados de um DataFrame, mantendo apenas uma ocorrência (por padrão, a primeira). Além disso, esse método também permite o uso do parâmetro inplace=True, para que a remoção das duplicatas seja feita diretamente no DataFrame original, sem a necessidade de atribuir o resultado a uma nova variável. O código abaixo remove as linhas duplicadas do DataFrame dados_saude.

```
dados_saude.drop_duplicates()
```

Neste momento, se executarmos novamente o comando **dados_saude.duplicated()**, verificaremos que não existem mais registros duplicados, uma vez que tratamos esse problema usando o método **.drop_duplicates()**.

### 3.5.5. Conversão de tipos

Em tarefas de ciência de dados, é comum que os dados importados de diferentes fontes apresentem tipos inadequados para as análises. Por exemplo, no DataFrame dados_saude, a coluna idade e pressao_sistolica foram carregadas com o tipo float, quando o apropriado seria o tipo int. Esse tipo de inconsistência pode comprometer tanto as análises estatísticas quanto as visualizações de dados.

Para solucionar esse problema, o Pandas disponibiliza o método **.astype()**, que permite converter o tipo de dados de uma ou mais colunas de forma simples e eficiente. No exemplo a seguir, utilizamos esse método para converter as colunas idade e pressao_sistolica para o tipo inteiro (int), e as colunas nome e fumante para o tipo string.

```
dados_saude['idade'] = dados_saude['idade'].astype(int)
dados_saude['pressa_sistolica'] = dados_saude['pressao_sistolica
    '].astype(int)
dados_saude['nome'] = dados_saude['nome'].astype('string')
dados_saude['fumante'] = dados_saude['fumante'].astype('string')
```

Além das colunas tratadas anteriormente, também é possível identificar que a coluna 'data_coleta' está com o tipo incorreto. Essa inconsistência pode dificultar operações usuais em campos temporais, como calcular diferenças de tempo, ordenar cronologicamente e extrair componentes da data (como ano, mês e dia). Com o propósito de tratar esse problema, o pandas oferece o método .to_datetime(), que converte o tipo de colunas para representarem datas em objetos do tipo datetime, permitindo que essas informações sejam interpretadas corretamente pelo Python. Abaixo, apresentamos o código que utiliza esse método para converter o tipo da coluna 'data_coleta' para date.

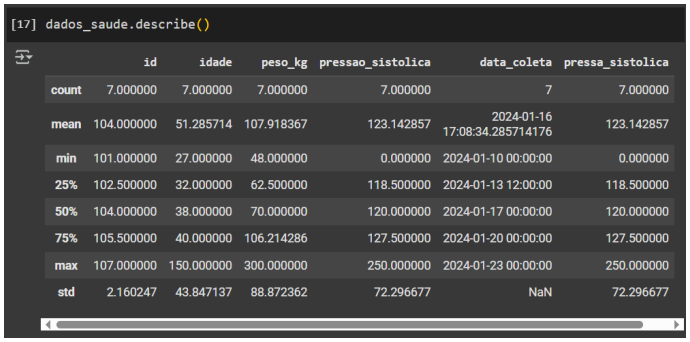
```
dados_saude['data_coleta'] = pd.to_datetime(dados_saude['
    data_coleta'])
```

### 3.5.6. Lidando com Outliers

Em ciência de dados, um passo importante é identificar e tratar valores discrepantes (isto é, outliers) presentes no conjunto de dados. Esses outliers podem ser resultado de diversos fatores, como erros de digitação e falhas de medição de sensores e equipamentos. No DataFrame 'dados_saude', é possível identificar alguns casos de outliers, a saber: (i) um paciente com idade de 150 anos; (ii) um peso corporal registrado como 300 kg; e (iii) uma pressão sistólica de 250, que está acima dos limites fisiológicos humanos comuns.

A primeira forma de identificar valores discrepantes é usar o método .describe() (apresentado na Seção 3.4.4.1), que resume as informações numéricas das colunas e permite perceber valores que não se ajustam ao intervalo esperado. A Figura 3.16 detalha as informações estatísticas do DataFrame 'dados_saude', no qual é possível identificar valores fora da faixa esperada para algumas colunas.



```
[17] dados_saude.describe()
```

| | id | idade | peso_kg | pressao_sistolica | data_coleta | pressa_sistolica |
|---|---|---|---|---|---|---|
| count | 7.000000 | 7.000000 | 7.000000 | 7.000000 | 7 | 7.000000 |
| mean | 104.000000 | 51.285714 | 107.918367 | 123.142857 | 2024-01-16 17:08:34.285714176 | 123.142857 |
| min | 101.000000 | 27.000000 | 48.000000 | 0.000000 | 2024-01-10 00:00:00 | 0.000000 |
| 25% | 102.500000 | 32.000000 | 62.500000 | 118.500000 | 2024-01-13 12:00:00 | 118.500000 |
| 50% | 104.000000 | 38.000000 | 70.000000 | 120.000000 | 2024-01-17 00:00:00 | 120.000000 |
| 75% | 105.500000 | 40.000000 | 106.214286 | 127.500000 | 2024-01-20 00:00:00 | 127.500000 |
| max | 107.000000 | 150.000000 | 300.000000 | 250.000000 | 2024-01-23 00:00:00 | 250.000000 |
| std | 2.160247 | 43.847137 | 88.872362 | 72.296677 | NaN | 72.296677 |

**Figura 3.16. Resultado da execução do método .describe() no DataFrame dados_saude.**

Outra maneira prática de detectar outliers é utilizando visualizações gráficas, como o boxplot. Essa ferramenta gráfica é bastante eficiente para identificar valores extremos, pois resume a distribuição dos dados de forma visual e intuitiva. Esse tipo de gráfico será detalhado na Seção 3.6.

Após identificar os outliers, é necessário aplicar alguma estratégia para tratá-los. A primeira estratégia que pode ser aplicada é remover todas as linhas que contêm valores discrepantes. Outra alternativa viável é substituir os outliers por valores plausíveis, usando estatísticas como média e mediana. Além disso, também é possível aplicar regras de negócio previamente estabelecidas, definindo limites máximos e mínimos aceitáveis com base no conhecimento especializado da área e, assim, filtrar os dados que estiverem fora desses parâmetros.

No caso do Dataframe dados_saude, trataremos outliers usando a estratégia de remoção de linhas com valores discrepantes. Especificamente, utilizaremos filtros (conforme apresentado na Seção 3.4.7) para remover outliers. Abaixo, apresentamos o código que trata os valores discrepantes em dados_saude.

```
# Removendo idade acima de 120
dados_saude = dados_saude[dados_saude['idade'] <= 120]

# Removendo peso acima de 200 kg
dados_saude = dados_saude[dados_saude['peso_kg'] <= 200]

# Removendo pressão sistólica acima de 200
dados_saude = dados_saude[dados_saude['pressao_sistolica'] <=
    200]
```

### 3.5.7. Salvando o Dataset Tratado

Após realizar todo o processo de limpeza de dados, é necessário salvar o dataset tratado para garantir que as alterações realizadas sejam preservadas e possam ser utilizadas em análises futuras. Essa prática evita a repetição desnecessária do processo de limpeza sempre que os dados forem utilizados novamente, além de garantir a consistência e a reprodutibilidade das análises. O pandas oferecem métodos simples que permitem salvar DatFrames em diversos formatos de arquivos, como JSON, CSV, Excel, dentre outros formatos. No código abaixo,

utilizamos o método '.to_csv()' para salvar o DataFrame dados_saude em arquivo com formato CSV. Além disso, atribuímos valor falso ao parâmetro index, para que os índices sejam ignorados ao criar o arquivo.

```
dados_saude.to_csv('dados_saude_tratados.csv', index=False)
```

Neste momento, finalizamos todas as etapas de limpeza e tratamento de dados em nosso DataFrame dados_saude. Utilize os conhecimentos adquiridos nessa seção para verificar se o conjunto de dados foi realmente higienizado (por exemplo, use os métodos .info e .describe). Em conclusão, dominar técnicas de limpeza de dados é essencial para qualquer profissional da área, já que dados bem preparados são a base para conclusões mais precisas e modelos mais robustos.

## 3.6. Explorando Dados com Gráficos na Saúde

A exploração visual de dados é uma das etapas mais importantes da ciência de dados, pois permite identificar padrões, tendências, distribuições e possíveis anomalias de maneira visual e intuitiva. Esta importância é ressaltada quando lidamos com dados da área da saúde, pois visualizações gráficas podem ser uma ferramenta poderosa em tarefas como tomadas de decisões clínicas, políticas públicas e apoio a diagnósticos médicos.

Nesta seção, exploraremos o uso das bibliotecas Seaborn e Plotly para criar gráficos a partir de um conjunto de dados sobre pacientes com e sem histórico de Acidente Vascular Cerebral (AVC). Os gráficos serão empregados para ilustrar relações entre variáveis, distribuições de valores, proporções e agrupamentos relevantes. Através dessa abordagem, seremos capazes de compreender melhor este conjunto de dados e extrair insights relevantes para o contexto médico.

### 3.6.1. Configuração do Ambiente

Antes de começarmos a explorar os dados por meio de visualizações, é necessário preparar o ambiente de trabalho com as bibliotecas que serão utilizadas ao longo desta seção. Abaixo, apresentamos o código que importa as bibliotecas pandas, matplotlib, seaborn e plotly.

```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import plotly.express as px
```

Com as bibliotecas devidamente configuradas, podemos seguir para o carregamento dos dados. O código abaixo faz a leitura do conjunto de dados sobre pacientes com e sem histórico de AVC diretamente do Google Drive. Caso opte por baixar os dados e carregar manualmente, o leitor pode baixar o conjunto de dados a partir deste link: `https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset`.

```
dataset_code = '1DP5u0SzZXtgb7mfs07cXjzrUwOE0rDEd'
url = f'https://drive.google.com/uc?id={dataset_code}'
```

```
avc_df = pd.read_csv(url)
```

Com os dados devidamente carregados, podemos visualizar as colunas e as cinco primeiras linhas do DataFrame 'avc_df' executando o código abaixo. A Figura 3.17 apresenta o resultado da execução desse código. Analisando esse resultado, é possível identificar que o DataFrame avc_df reúne informações clínicas e demográficas de pacientes para análise de fatores associados ao AVC. Esse conjunto de dados inclui colunas como id (identificador do paciente), gender (gênero), age (idade), hypertension e heart_disease (indicam presença dessas condições), ever_married (histórico de casamento), work_type (tipo de ocupação), Residence_type (zona de residência), avg_glucose_level (nível médio de glicose), bmi (índice de massa corporal), smoking_status (histórico de tabagismo) e stroke (indica se o paciente sofreu AVC).

```
avc_df.head()
```

| | id | gender | age | hypertension | heart_disease | ever_married | work_type | Residence_type | avg_glucose_level | bmi | smoking_status | stroke | faixa_idade |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 9046 | Male | 67.0 | 0 | 1 | Yes | Private | Urban | 228.69 | 36.6 | formerly smoked | 1 | 51–70 |
| 1 | 51676 | Female | 61.0 | 0 | 0 | Yes | Self-employed | Rural | 202.21 | NaN | never smoked | 1 | 51–70 |
| 2 | 31112 | Male | 80.0 | 0 | 1 | Yes | Private | Rural | 105.92 | 32.5 | never smoked | 1 | 71+ |
| 3 | 60182 | Female | 49.0 | 0 | 0 | Yes | Private | Urban | 171.23 | 34.4 | smokes | 1 | 31–50 |
| 4 | 1665 | Female | 79.0 | 1 | 0 | Yes | Self-employed | Rural | 174.12 | 24.0 | never smoked | 1 | 71+ |

**Figura 3.17. Detalhamento do DataFrame avc_df que contém dados sobre pacientes com e sem histórico de AVC.**

No decorrer desta seção, iremos utilizar as bibliotecas gráficas Seaborn e Plotly para criar diversos tipos de gráficos usando esse conjunto de dados. Enfatizamos também que utilizaremos os nomes de colunas originais dos dados (isto é, em inglês), mas, caso necessário, o leitor poderá mudar para português conforme ensinado na Seção 3.4.4.

### 3.6.2. Criação de Gráficos com a Biblioteca Seaborn

A biblioteca Seaborn é uma poderosa ferramenta de visualização de dados baseada no Matplotlib, projetada para tornar a criação de gráficos estatísticos mais simples e informativa. Nesta seção, exploraremos como utilizar o Seaborn para gerar visualizações no contexto da saúde, a partir do nosso DataFrame avc_df.

#### 3.6.2.1. Histograma: Distribuição de Idade dos Pacientes

Histograma é um tipo de gráfico de barras utilizado para representar a distribuição de uma variável numérica. Ele agrupa os dados em intervalos (chamados de bins) e mostra quantos valores caem em cada intervalo. No código a seguir, utilizamos a Seaborn para criar um histograma da variável age do conjunto de dados avc_df, que representará a distribuição de idades dos pacientes. O primeiro comando sns.histplot(avc_df['age'], bins=10, kde=True, color='red') gera o gráfico com 10 intervalos de idade (isto é, 10 bins), adiciona uma linha de densidade para suavizar a visualização da distribuição (kde) e define a cor das barras como vermelha. Em seguida, configuramos o título do gráfico e os rótulos dos eixos

utilizando plt.title, plt.xlabel e plt.ylabel. Por fim, exibimos o gráfico com o comando plt.show(). A Figura 3.18 apresenta o gráfico gerado, no qual é possível observar como as idades dos pacientes estão distribuídas.

```
#Plota o histograma da idade dos pacientes
sns.histplot(avc_df['age'], bins=10, kde=True, color='red')
#Insere um título para o gráfico
plt.title('Distribuição de Idade dos Pacientes')
#Adicionam rótulo ao eixo X
plt.xlabel('Idade')
#Adicionam rótulo ao eixo Y
plt.ylabel('Frequência')
#Exibe o gráfico
plt.show()
```

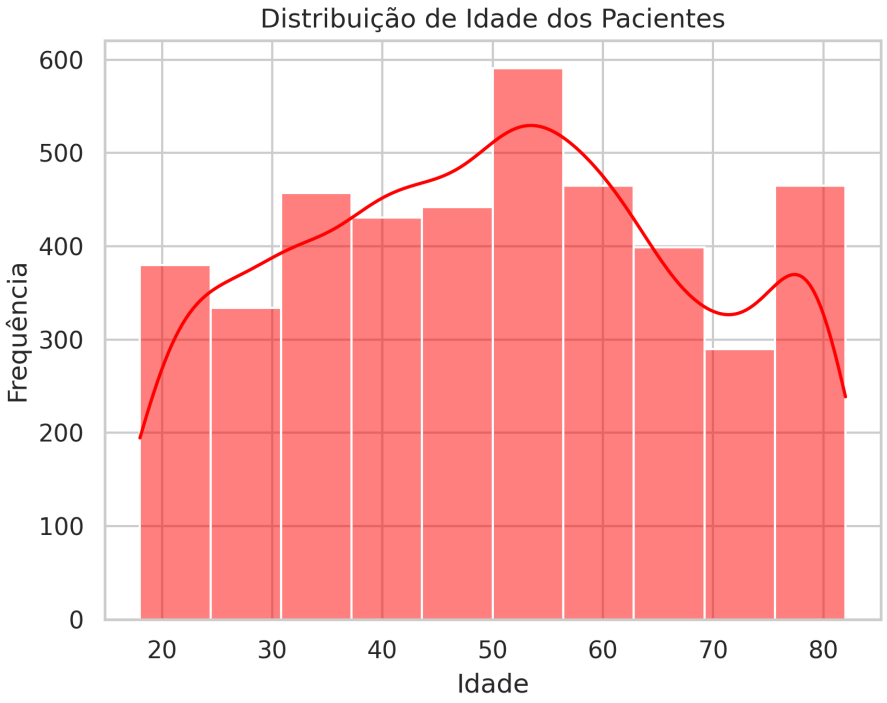



**Figura 3.18. Histograma que exibe a frequência de faixas etárias.**

### 3.6.2.2. Gráfico de Contagem: Relação Hipertensão x AVC

O gráfico de contagem (ou count plot) é utilizado para visualizar a frequência de categorias em uma variável categórica, permitindo comparações rápidas entre os grupos. No código abaixo, utilizamos a Seaborn para criar um gráfico de contagem relacionando a presença de hipertensão (hypertension) com a ocorrência de AVC (stroke). O primeiro comando sns.countplot() constrói o gráfico considerando hypertension no eixo x e separando as contagens pelo valor de stroke (usando cores distintas definidas pela paleta 'viridis').

Configuramos o título e os rótulos dos eixos com plt.title, plt.xlabel e plt.ylabel, e exibimos o gráfico com o comando plt.show().

```
sns.countplot(x='hypertension', hue='stroke', data=avc_df,
   palette='viridis')
plt.title('Relação entre Hipertensão e Ocorrência de AVC')
plt.xlabel('Hipertensão (0 = Não, 1 = Sim)')
plt.ylabel('Quantidade')
plt.show()
```

A Figura 3.19 apresenta o gráfico gerado, que possibilita observar se pacientes hipertensos tiveram uma frequência maior ou menor de AVC em comparação aos pacientes sem hipertensão.

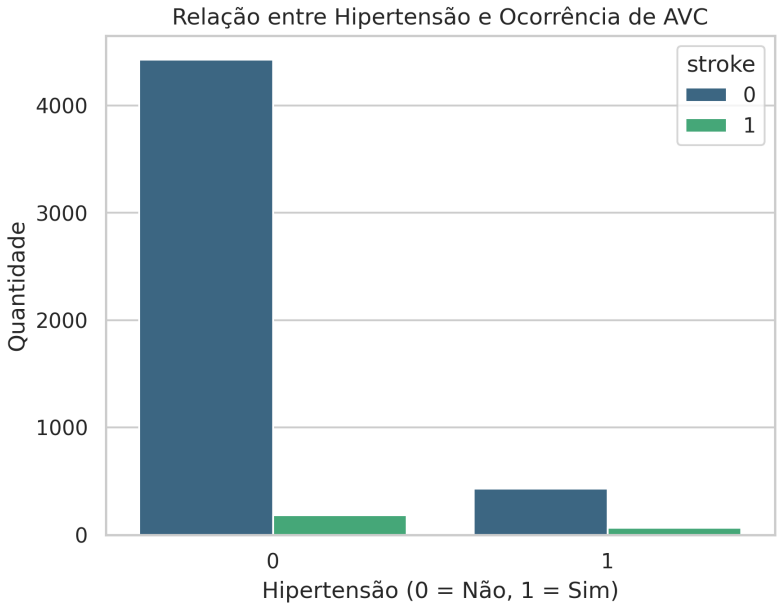



**Figura 3.19. Gráfico countplot que representa a relação entre hipertensão e AVC.**

### 3.6.2.3. Gráfico de Dispersão: Relação Idade x Glicose

O gráfico de dispersão (scatterplot) é utilizado para identificar a relação entre duas variáveis numéricas. Esse gráfico plota um ponto baseado nas coordenadas de cada valor, isto é, sua posição é determinada pelos valores das duas variáveis escolhidas, uma no eixo X e outra no eixo Y. No código abaixo, criamos um gráfico de dispersão usando o comando sns.scatterplot() para identificar a relação entre a idade dos pacientes e seu nível médio de glicose, distinguindo-os conforme a presença ou ausência de hipertensão. No eixo X, representamos a idade (age) e no eixo Y, o nível médio de glicose (avg_glucose_level). As cores dos pontos indicam se o paciente é hipertenso (1) ou não (0), utilizando a paleta de cores coolwarm. A legenda foi configurada para deixar claro o significado dos valores de hipertensão (isto é, 0 = não, 1 = sim).

```
sns.scatterplot(x='age', y='avg_glucose_level', data=avc_df, hue
    ='hypertension', palette='coolwarm')
plt.title('Idade vs Nível Médio de Glicose - Diferenciando
    Hipertensão')
plt.xlabel('Idade (anos)')
plt.ylabel('Nível Médio de Glicose')
plt.legend(title='Hipertensão (0 = Não, 1 = Sim)')
plt.show()
```

A Figura 3.20 apresenta o gráfico gerado, que possibilita identificar se há alguma tendência, como níveis mais elevados de glicose em pacientes hipertensos de determinadas faixas etárias. Especificamente, identificamos que pacientes com idades mais avançadas (isto é, acima de 50 anos) tendem a sofrer de hipertensão e apresentar altos níveis de glicose.

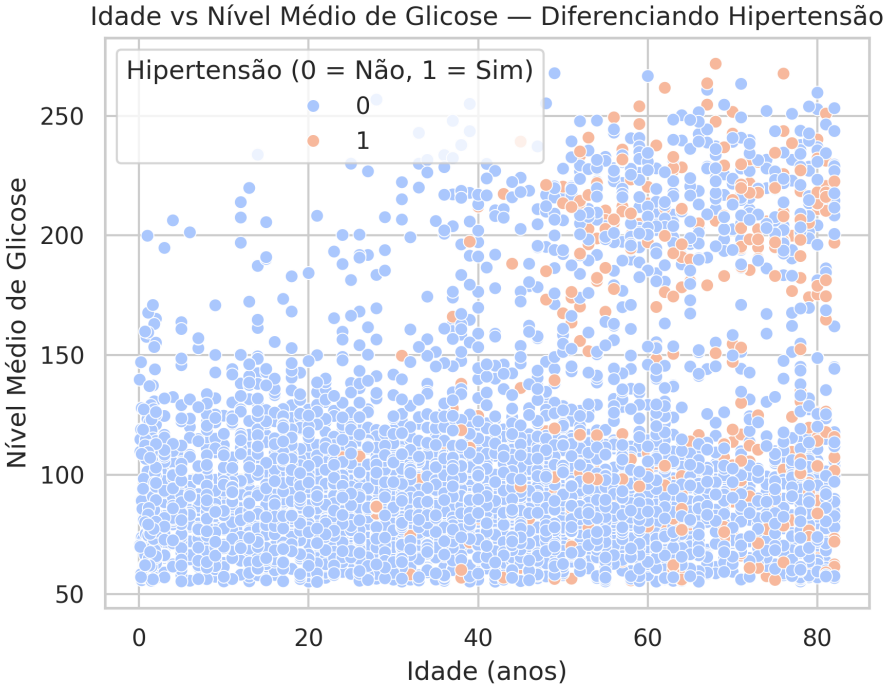



**Figura 3.20. Gráfico de dispersão que representa a relação entre idade e nível médio de glicose.**

### 3.6.2.4. Box Plot: BMI por Tipo de Trabalho

O boxplot é um gráfico utilizado para mostrar a distribuição de um conjunto de dados numéricos, destacando seus principais componentes estatísticos. Ele fornece uma maneira eficaz de visualizar a mediana, a variabilidade e os outliers de uma variável. Os principais elementos de um boxplot são:

- Caixas: representam o intervalo interquartil, que vai do primeiro quartil (Q1) ao

terceiro quartil (Q3). A caixa contém 50% dos dados, sendo que a linha dentro da caixa indica a mediana (Q2).

- Bigodes: as linhas que se estendem da caixa mostram os valores máximos e mínimos dentro de 1.5 vezes o intervalo interquartil a partir dos quartis. Valores fora desse intervalo são considerados outliers.

- Outliers: são os pontos que ficam fora da área definida pelos bigodes, podendo indicar dados incomuns ou discrepantes.

O código abaixo cria um boxplot que visualiza a distribuição do Índice de Massa Corporal (bmi) para diferentes tipos de trabalho (work_type). Para tanto, utilizamos a função sns.boxplot() do Seaborn, com a variável work_type no eixo X e bmi no eixo Y. O argumento hue='work_type' permite colorir as caixas de acordo com o tipo de trabalho, facilitando a distinção visual entre as categorias. A paleta de cores utilizada é pastel, proporcionando cores suaves. O comando plt.title() define o título do gráfico e os comandos plt.xlabel() e plt.ylabel() são usados para rotular os eixos X e Y, respectivamente. O comando plt.xticks(rotation=45) inclina os rótulos no eixo X, melhorando assim a legibilidade do gráfico.

```
sns.boxplot(x='work_type', y='bmi', data=avc_df, hue='work_type'
    , palette='pastel')
plt.title('Distribuição do IMC por Tipo de Trabalho')
plt.xlabel('Tipo de Trabalho')
plt.ylabel('Índice de Massa Corporal (IMC)')
plt.xticks(rotation=45)
plt.show()
```

A Figura 3.21 apresenta o gráfico gerado, que possibilita observar diferenças entre as categorias e identificar outliers. Especificamente, identificamos que o imc dos pacientes que nunca trabalharam (never_worked) é significativamente menor em comparação com os demais tipos de trabalho. Além disso, reconhecemos que existe uma maior quantidade de outliers nos dados dos pacientes que trabalham no setor privado.

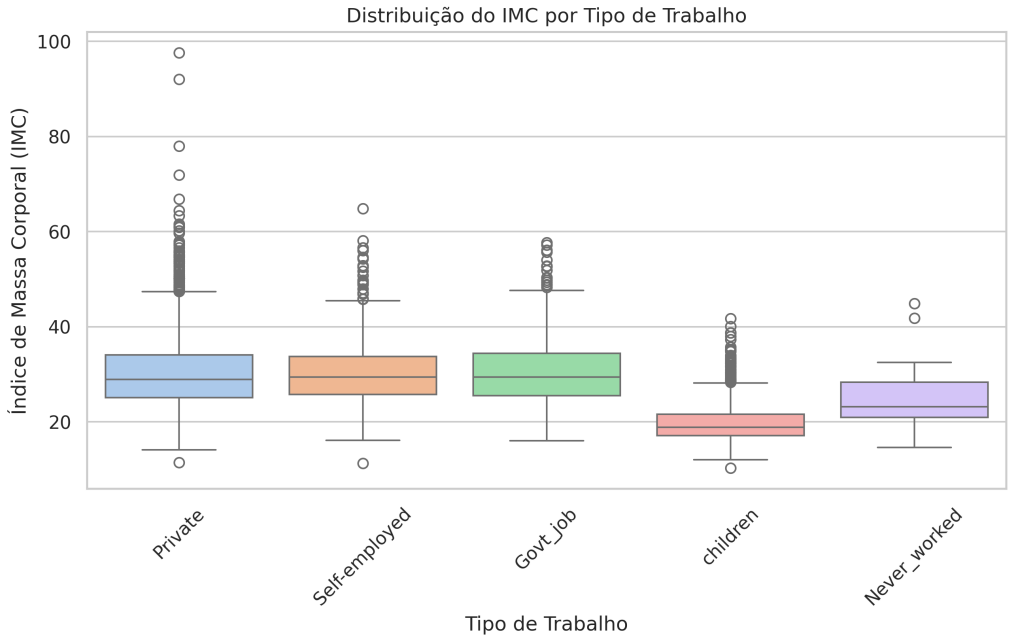



**Figura 3.21. Box plot que representa a distribuição dos dados e revela possíveis outliers de bmi por tipo de trabalho.**

### 3.6.2.5. Mapa de Calor: Correlações Entre Variáveis Numéricas

Mapa de calor (heatmap) é um gráfico que representa valores por meio de tonalidades de cores, em vez de apenas números. Esse tipo de gráfico possui uma escala de tonalidades que facilita a identificação de padrões e correlações entre variáveis. Seu uso mais comum é criar visualizações de matrizes de correlação, já que ele destaca rapidamente quais variáveis têm relações mais fortes (positivas ou negativas) entre si. O código abaixo cria um mapa de calor a partir das correlações das variáveis numéricas do DataFrame avc_df. Primeiramente, calculamos a matriz de correlação usando o comando avc_df.corr(numeric_only=True), que resulta em valores entre -1 (correlação fortemente negativa) e 1 (correlação fortemente positiva). Em seguida, usamos o comando sns.heatmap() para criar o mapa de calor com os seguintes parâmetros: annot=True que define que os valores numéricos sejam exibidos, cmap='coolwarm' define uma paleta de cores que varia do azul (correlação negativa) ao vermelho (correlação positiva), e fmt=".2f"que formata os números para duas casas decimais.

```
corr = avc_df.corr(numeric_only=True)
sns.heatmap(data=corr,annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlações')
plt.show()
```

A Figura 3.22 apresenta o mapa de calor gerado, que apresenta a força da relação entre as variáveis numéricas. Especificamente, as cores com tons vermelhos fortes indicam correlações fortes (positivas ou negativas), enquanto tons azuis fortes indicam pouca ou nenhuma correlação.

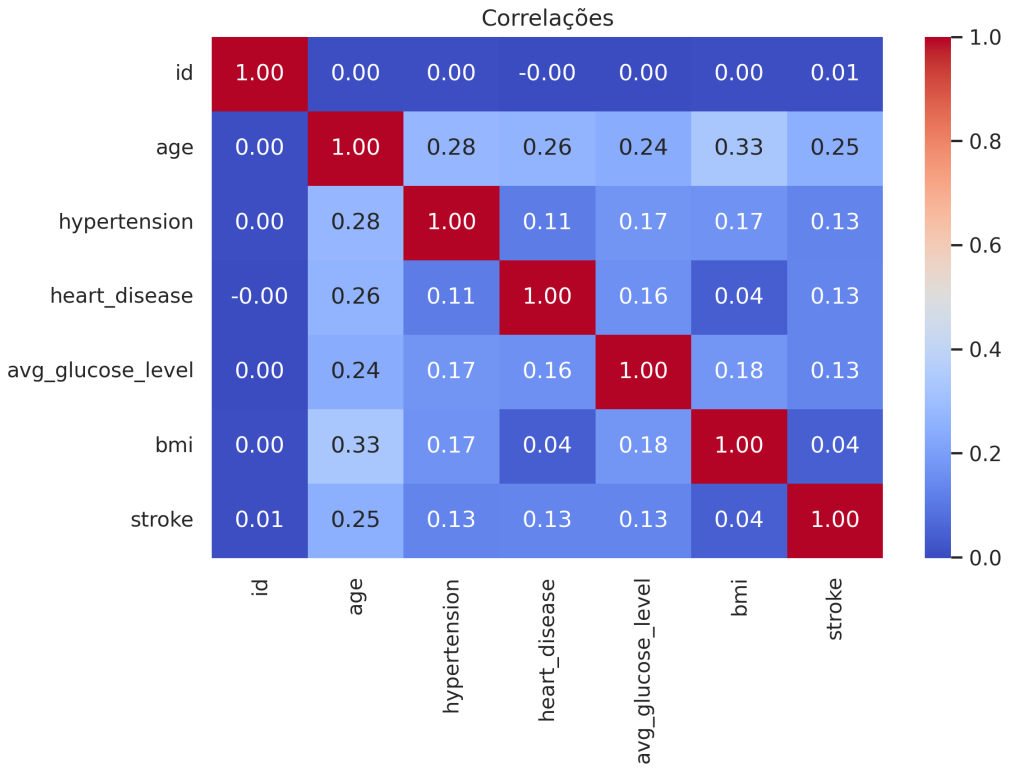



**Figura 3.22. Mapa de calor das correlações entre as variáveis do DataFrame avc_df.**

### 3.6.3. Criação de Gráficos com a Biblioteca Plotly

Nesta seção, aprenderemos a criar gráficos interativos usando a biblioteca Plotly. Essa biblioteca permite criar gráficos interativos e dinâmicos de dados, o que é especialmente útil em análises exploratórias e apresentações. Embora as imagens apresentadas nesta seção sejam gráficos estáticos, é importante ressaltar que todos os gráficos criados com Plotly são interativos. Isso significa que, ao visualizar o gráfico em um ambiente compatível (por exemplo, Google Colab), você pode passar o mouse sobre os elementos para ver informações detalhadas, fazer zoom e mover-se pelo gráfico. Portanto, aprenderemos a construir diversos tipos de gráficos com Plotly utilizando o conjunto de dados avc_df. Através de exemplos práticos, mostraremos como representar informações de forma clara e atrativa, favorecendo a análise e a tomada de decisões.

#### 3.6.3.1. Gráfico de Pizza Interativo: Proporção de Incidência de AVC por Tipo de Trabalho

Um gráfico de pizza é uma visualização circular usada para mostrar a proporção de diferentes categorias dentro de um todo. Cada fatia representa uma categoria, e seu tamanho é proporcional à quantidade ou frequência dessa categoria. O código abaixo cria um gráfico de pizza utilizando a biblioteca Plotly Express (px) para representar a proporção de casos de AVC conforme o tipo de trabalho dos pacientes. O comando px.pie() é utilizado para gerar o gráfico, onde apresenta os seguintes parâmetros: names='work_type' define que cada fatia representará uma categoria desta coluna; values='stroke' indica que o

tamanho das fatias será proporcional ao número de casos de AVC registrados; e title define o título do gráfico. Em seguida, fig.update_traces(textinfo='percent+label') ajusta as informações exibidas em cada fatia, mostrando tanto o nome da categoria quanto a porcentagem que ela representa em relação ao total. Por fim, o comando fig.show() exibe o gráfico interativo na tela.

```
fig = px.pie(avc_df,
             names='work_type',
             values='stroke',
             title='Proporção de AVCs por Tipo de Trabalho'
)

fig.update_traces(textinfo='percent+label')
fig.show()
```

A Figura 3.23 apresenta o gráfico de pizza gerado, em que cada fatia representa a proporção de pacientes com AVC dentro de uma categoria de tipo de trabalho.

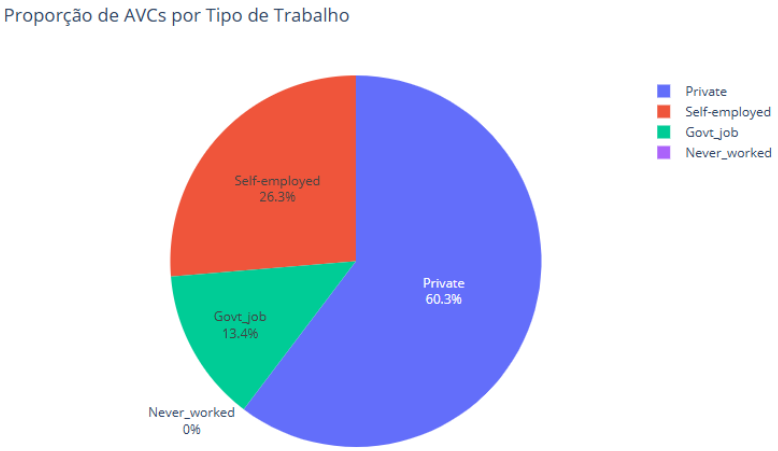



**Figura 3.23. Gráfico de pizza que apresenta a proporção de incidência de AVC por tipo de Trabalho.**

### 3.6.3.2. Gráfico de Violino: Distribuição do Nível Médio de Glicose por Estado Civil

Um gráfico de violino (violin plot) é uma visualização que combina o boxplot e a densidade de probabilidade. Especificamente, este gráfico adiciona a visualização da concentração de dados ao boxplot, em que a forma de 'violino' reflete a densidade dos dados. O código abaixo cria um gráfico de violino para visualizar a distribuição do nível médio de glicose entre diferentes grupos de estado civil. Utilizamos o comando px.violin() para criar esse gráfico, em que foram usados os seguintes parâmetros: x='ever_married' e y='avg_glucose_level' para definir o eixo x e y, respetivamente; color='ever_married' para diferenciar por cores os dois grupos de estado civil (casados e não casados); box=True para adicionar um boxplot dentro do gráfico de violino, fornecendo informações sobre a mediana e os quartis da distribuição; points='all' para exibir todos os pontos de dados no

gráfico; e definimos o parâmetro ‘title’ para plotar o título do gráfico. Por fim, definimos os rótulos dos eixos X e Y usando o comando fig.update_layout() para melhorar a legibilidade do gráfico.

```
fig = px.violin(avc_df,
                x='ever_married',  y='avg_glucose_level', color=
                   'ever_married', box=True,
                points='all',
                title='Distribuição do Nível Médio de Glicose
                   por Estado Civil')
fig.update_layout(
    xaxis_title='Estado Civil',
    yaxis_title='Nível Médio de Glicose'
)
fig.show()
```

A Figura 3.24 apresenta o gráfico de violino gerado, que permite observar como os níveis de glicose variam entre os grupos de estado civil, mostrando a distribuição, mediana e dispersão dos dados. No ambiente de execução, o leitor poderá interagir com o gráfico, executando ações como zoom e passar o mouse por cima para visualizar informações com mais detalhes.

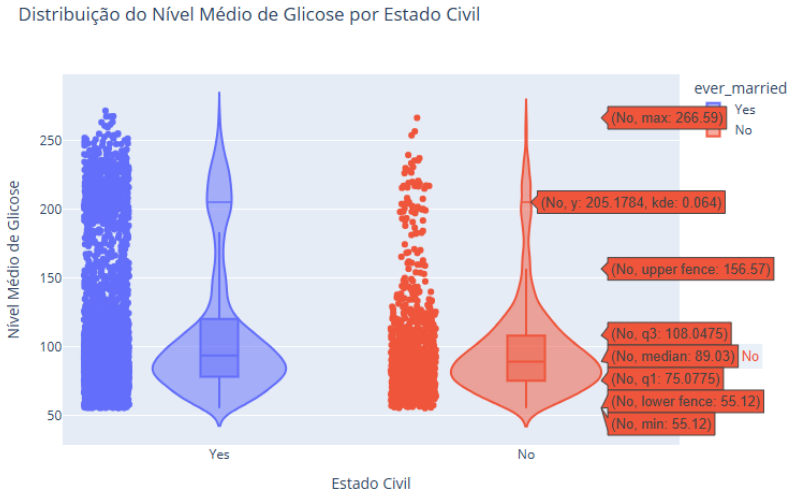



**Figura 3.24. Gráfico de violino da distribuição do nível médio de glicose por estado civil.**

### 3.6.3.3. Gráfico Sunburst: Distribuição Hierárquica entre Gênero Trabalho e Hipertensão

O sunburst é um tipo de gráfico hierárquico em formato circular, onde cada nível é representado por um anel. Ele apresenta como as categorias se subdividem e qual a proporção de cada segmento em relação ao todo. No código abaixo, utilizamos o método px.sunburst() para criar um gráfico do tipo sunburst, que representa hierarquicamente a relação entre gênero, tipo de trabalho e hipertensão. Primeiramente, definimos cada nível

da hierarquia usando o parâmetro path=['gender', 'work_type', 'hypertension'], isto é, o centro representa o gênero, o próximo nível o tipo de trabalho e, por fim, a presença ou não de hipertensão. A coloração dos segmentos é baseada na coluna stroke, que indica se o paciente teve ou não AVC. O método fig.update_traces(textinfo='label+percent entry') adiciona ao gráfico os rótulos das categorias e a porcentagem correspondente em relação ao total de entradas. Por fim, fig.show() exibe o gráfico.

```
fig = px.sunburst(
    avc_df,
    path=['gender', 'work_type', 'hypertension'],
     color='stroke',
    title='Distribuição Hierárquica por Gênero, Tipo de Trabalho
        e Hipertensão'
)
# Mostrar os rótulos nas seções
fig.update_traces(textinfo='label+percent entry')
fig.show()
```

A Figura 3.25 apresenta o gráfico sunburst gerado, que apresenta a distribuição hierárquica dos pacientes por gênero, tipo de trabalho e presença de hipertensão, com cores indicando a ocorrência de AVC. É possível visualizar como essas categorias se subdividem e contribuem proporcionalmente para o total de registros.

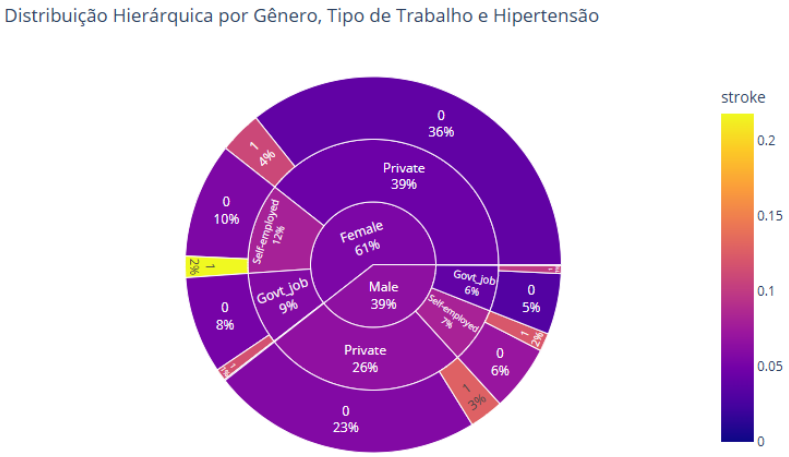



**Figura 3.25. Gráfico sunburst que apresenta de forma hierárquica a relação entre gênero, tipo de trabalho e hipertensão.**

### 3.6.3.4. Gráfico de Barras: Proporção de Hipertensão por Faixas de Idade

Nesta seção, iremos criar um gráfico de barras para apresentar a proporção de pacientes com hipertensão agrupados por faixas de idade. Antes da criação deste gráfico, vamos realizar um processamento inicial dos dados para permitir a análise por faixas etárias. O código abaixo apresenta esse processamento. Especificamente, realizamos as seguintes etapas: Primeiramente, criamos a coluna age_range utilizando o método pd.cut(), que segmenta a variável contínua age em quatro faixas: Até 30 anos, 31–50 anos, 51–70 anos e

71+ anos. Em seguida, os dados são agrupados com base nessas faixas etárias, e calculamos a média da variável hypertension em cada grupo. Como essa variável é binária (0 para ausência e 1 para presença de hipertensão), a média representa diretamente a proporção de indivíduos com hipertensão em cada faixa. Por fim, essa proporção é convertida para porcentagem, facilitando a interpretação no gráfico que será gerado.

```
# Criação da coluna faixa_idade com pd.cut()
avc_df['age_range'] = pd.cut(avc_df['age'],bins=[0, 30, 50, 70,
    100],
                             labels=['Até 30 anos', '31-50 anos',
                                     '51-70 anos', '71+ anos'])

# Agrupamento e cálculo da proporção de AVC por faixa etária
hipertencao_df = avc_df.groupby('age_range').agg(
    hypertension_proportion = ('hypertension', 'mean')
).reset_index()

# Convertendo a proporção para um formato percentual
hipertencao_df['hypertension_percent'] =
                    hipertencao_df['hypertension_proportion'] *
                       100
```

Após o processamento inicial, podemos utilizar o DataFrame hipertencao_df para criar o gráfico de barras. O código a seguir gera uma visualização da proporção de hipertensão em diferentes faixas etárias. A função px.bar() é utilizada para construir o gráfico, onde o eixo X representa as faixas etárias (age_range) e o eixo Y exibe a proporção percentual de pacientes com hipertensão (hypertension_percent). As cores das barras são definidas com base nas faixas etárias, facilitando a distinção visual entre os grupos. Em seguida, o método fig.update_layout() é empregado para configurar os rótulos dos eixos e o título da legenda.

```
fig = px.bar(hipertencao_df,
        x='age_range', y='hypertension_percent', color='
           age_range',
        title='Proporção de Hipertensão por Faixa Etária')

fig.update_layout(
    xaxis_title='Faixa Etária',
    yaxis_title='Proporção de Hipertensão',
    legend_title_text='Faixa Etária'
)
fig.show()
```

A Figura 3.26 apresenta o resultado: um gráfico de barras interativo que ilustra a proporção de pacientes com hipertensão em cada faixa etária, com base nos dados previamente agrupados e convertidos em percentuais.

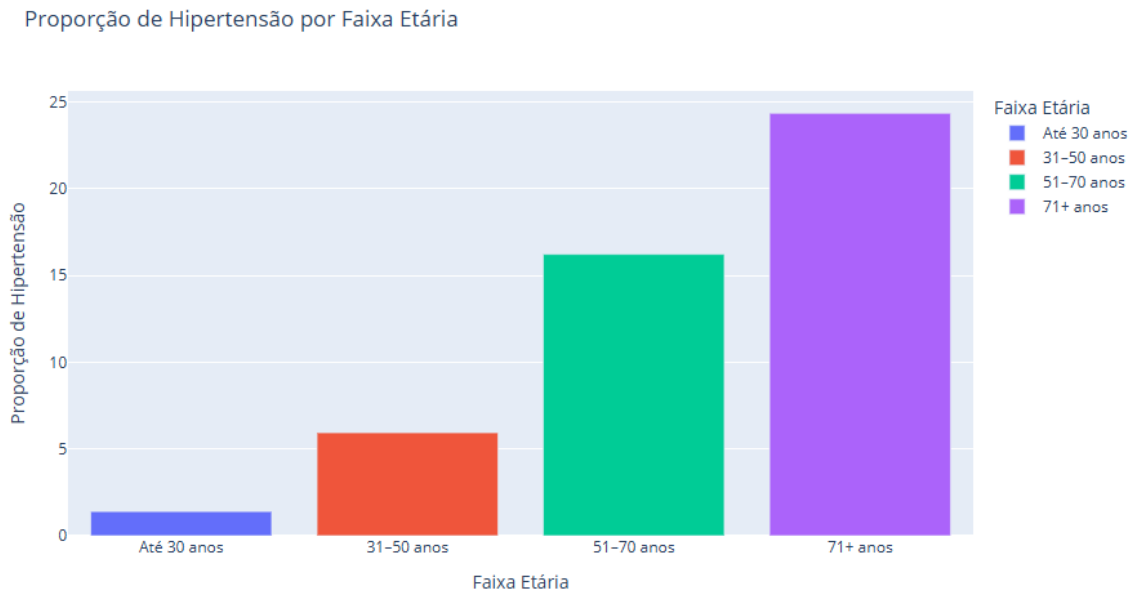



**Figura 3.26. Gráfico de barras que apresenta a proporção de hipertensão por faixas de idade.**

### 3.6.3.5. Gráfico de Linha: Glicose Média por Idade

Um gráfico de linhas (line plot) é um tipo de visualização que mostra a variação de um ou mais valores ao longo de um eixo contínuo, geralmente o tempo. Ele conecta os pontos de dados com linhas, facilitando a identificação de padrões e comparações entre séries. O código abaixo cria um gráfico de linhas que apresenta a variação da glicose média em função da idade dos pacientes. Antes de criar o gráfico efetivamente, realizamos dois processamentos iniciais, a saber: (i) aplicamos um filtro para considerar apenas pacientes com 18 anos ou mais e (ii) agrupamos os dados por idade e calculamos a média dos níveis de glicose para cada idade. O resultado é armazenado no DataFrame media_glicose_df. No próximo passo, utilizamos o método px.line() para criar um gráfico de linhas com marcadores (markers=True), em que o eixo X representa a idade e o eixo Y mostra a glicose média correspondente. Por fim, usamos o método fig.update_layout() para definir os rótulos dos eixos e o gráfico é exibido com a execução do fig.show().

```
#Filtro de idade
avc_df = avc_df[avc_df['age'] >= 18]

# Cálculo da glicose média por idade
media_glicose_df = avc_df.groupby('age').agg(
    media_glicose = ('avg_glucose_level', 'mean')
).reset_index()

fig = px.line(media_glicose_df,
              x='age',
              y='media_glicose',
              markers=True,
```

```
                title='Glicose Média por Idade')

fig.update_layout(
    xaxis_title='Idade',
    yaxis_title='Glicose Média',
)
fig.show()
```

A Figura 3.27 apresenta a variação da glicose média em função da idade dos pacientes. O eixo X representa as idades, enquanto o eixo Y exibe a glicose média correspondente a cada faixa etária. As linhas conectam os pontos de dados, facilitando a visualização das tendências. Cada ponto é marcado, destacando a glicose média para cada idade. Esse gráfico ajuda a entender como os níveis médios de glicose variam à medida que a idade dos pacientes aumenta.

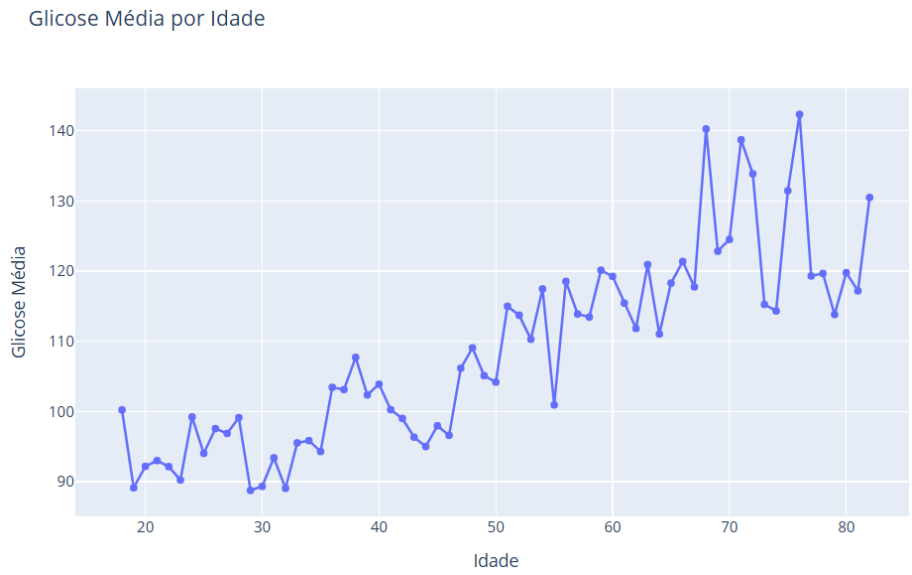



**Figura 3.27. Gráfico de linhas que apresenta a variação da glicose média em função da idade.**

## 3.7. Tendências em Ciência de Dados na Saúde

As ferramentas e técnicas exploradas ao longo deste material, como limpeza de dados, análise exploratória e visualização com Seaborn e Plotly, representam a base prática da ciência de dados aplicada à saúde. Essas habilidades estão no centro de diversas tendências emergentes que vêm transformando o setor. A seguir, apresentamos as principais tendências em ciência de dados aplicada à saúde.

O uso de Inteligência Artificial e Aprendizado de Máquina na saúde tem ganhado destaque por seu potencial em transformar o cuidado com o paciente [Provost and Fawcett 2013]. Algoritmos são capazes de analisar grandes volumes de dados clínicos, laboratoriais e de imagem para identificar padrões que escapam ao olhar humano. Técnicas de aprendizado supervisionado, não supervisionado e aprendizado profundo (deep learning) têm sido amplamente utilizadas em tarefas como detecção precoce de doenças, classificação de imagens

médicas (por exemplo, radiografias e ressonâncias), predição de readmissões hospitalares e estratificação de risco [Moura et al. 2023, Teles et al. 2025]. A capacidade preditiva dessas ferramentas contribui para diagnósticos mais rápidos, precisos e personalizados.

A popularização de dispositivos vestíveis (wearables), como smartwatches e sensores biomédicos, tem permitido a coleta contínua de dados fisiológicos em tempo real [Orphanidou 2019, Moura et al. 2023]. Esses dispositivos capturam informações como batimentos cardíacos, nível de atividade física, qualidade do sono e níveis de glicose. Integrados a plataformas digitais de monitoramento, esses dados podem ser analisados para identificar mudanças no estado de saúde, prevenindo complicações clínicas e permitindo intervenções precoces [Moura et al. 2022]. A saúde digital amplia o alcance do cuidado e promove um modelo mais proativo e centrado no paciente.

A ciência de dados na saúde enfrenta o desafio de lidar com dados oriundos de múltiplas fontes: prontuários eletrônicos, exames laboratoriais, sensores de dispositivos, bases genômicas e dados populacionais [Awrahman et al. 2022]. Esses dados são, muitas vezes, heterogêneos em formato, granularidade e qualidade. A integração dessas fontes permite uma visão mais completa e abrangente do paciente e da população, essencial para análises preditivas, avaliação de intervenções e formulação de estratégias de saúde. Ferramentas de big data e pipelines de processamento são fundamentais nesse contexto.

Modelos preditivos que integram dados clínicos, genômicos e ambientais têm sido amplamente empregados para antecipar a ocorrência de doenças, prever desfechos terapêuticos e estimar riscos individuais à saúde. [Shailaja et al. 2018]. Ao considerar variáveis específicas de cada paciente, é possível propor tratamentos personalizados que maximizem a eficácia e reduzam efeitos adversos. Essa abordagem, conhecida como medicina personalizada ou de precisão, é uma das mais promissoras na interseção entre ciência de dados e saúde [Collins and Varmus 2015]. A modelagem estatística, regressões e algoritmos de aprendizado de máquina são ferramentas centrais nesse processo.

Com o aumento do uso de dados pessoais na saúde, surgem questões críticas relacionadas à ética e à privacidade. O compartilhamento e a análise de informações sensíveis exigem o cumprimento de normativas como a LGPD no Brasil e o GDPR na Europa [Taddeo and Floridi 2018]. É fundamental garantir a anonimização dos dados, o consentimento informado dos pacientes e a transparência nos processos algorítmicos. Além disso, há preocupações sobre vieses em modelos de IA e o impacto ético de decisões automatizadas na saúde. A regulação eficaz é essencial para assegurar o uso responsável e justo da ciência de dados.

## 3.8. Conclusões

Este capítulo de livro forneceu uma introdução abrangente à aplicação de técnicas de ciência de dados na área da saúde, abordando desde a manipulação básica de dados até a criação de gráficos interativos para visualização de padrões e insights. Começamos com a fundamentação do Python, essencial para a construção de um raciocínio lógico sólido e para a criação de soluções eficazes na análise de dados. A compreensão de variáveis, operadores e estruturas condicionais proporcionou uma base essencial para o processamento de dados em saúde.

A manipulação de dados com o Pandas foi explorada de forma prática, permitindo ao leitor entender como selecionar, agregar e filtrar informações relevantes. A análise de dados de saúde, como no estudo de pacientes com risco de diabetes, foi facilitada pela utilização de dataframes, onde cada variável poderia ser manipulada de maneira eficiente e concisa. A introdução a funções de agregação, agrupamento e ordenação, por exemplo, proporcionou uma visão detalhada das técnicas necessárias para explorar grandes volumes de dados e extrair informações úteis.

A seção dedicada à limpeza e tratamento de dados destacou a importância de garantir a qualidade dos dados antes de prosseguir para a análise. O uso de técnicas como a remoção de duplicatas, o tratamento de valores ausentes e a gestão de outliers assegurou que os resultados finais fossem baseados em dados confiáveis e representativos.

Na parte de visualização de dados, a escolha de gráficos adequados, utilizando bibliotecas como Seaborn e Plotly, permitiu a criação de representações claras e informativas. Os gráficos de dispersão, histograma e mapa de calor mostraram como explorar a relação entre variáveis, enquanto os gráficos interativos de Plotly, como o gráfico de pizza e o gráfico Sunburst, possibilitaram uma exploração mais dinâmica dos dados. Esses recursos são essenciais para a comunicação de resultados, especialmente em estudos clínicos, onde a compreensão rápida e precisa dos dados pode ter implicações significativas para a saúde pública.

Ao longo do capítulo, o uso de datasets reais e simulados proporcionou uma compreensão prática das técnicas aplicadas à análise de dados em saúde. O dataset sobre diabetes, por exemplo, ilustrou como diferentes fatores podem impactar o risco de desenvolver doenças, enquanto o estudo de pacientes com histórico de AVC evidenciou a importância de se ter uma visão holística das condições de saúde ao realizar análises.

Por fim, discutimos sobre as tendências em ciência de dados na saúde, destacando como esta área tem impulsionado avanços importantes na saúde, por meio de tecnologias como inteligência artificial, dispositivos vestíveis e modelos preditivos personalizados. Esses recursos ampliam a precisão e a eficiência do cuidado, mas também exigem atenção a questões éticas e regulatórias, reforçando a importância de uma atuação crítica e responsável na área.

Em suma, este capítulo serve como um guia introdutório valioso para aqueles interessados em aplicar técnicas de ciência de dados no campo da saúde. Ao combinar teoria, ferramentas práticas e exemplos do mundo real, os leitores são equipados com as habilidades necessárias para começar a explorar, analisar e comunicar dados de saúde de maneira eficaz e eficiente.

## Referências

[Arellano et al. 2018] Arellano, A. M., Dai, W., Wang, S., Jiang, X., and Ohno-Machado, L. (2018). Privacy policy and technology in biomedical data science. *Annual review of biomedical data science*, 1(1):115–129.

[Awrahman et al. 2022] Awrahman, B. J., Aziz Fatah, C., and Hamaamin, M. Y. (2022). A review of the role and challenges of big data in healthcare informatics and analytics. *Computational intelligence and neuroscience*, 2022(1):5317760.

[Bao et al. 2019] Bao, Y., Chen, Z., Wei, S., Xu, Y., Tang, Z., and Li, H. (2019). The state of the art of data science and engineering in structural health monitoring. *Engineering*, 5(2):234–242.

[Chattu 2021] Chattu, V. K. (2021). A review of artificial intelligence, big data, and blockchain technology applications in medicine and global health. *Big Data and Cognitive Computing*, 5(3):41.

[Collins and Varmus 2015] Collins, F. S. and Varmus, H. (2015). A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795.

[Latif et al. 2020] Latif, S., Usman, M., Manzoor, S., Iqbal, W., Qadir, J., Tyson, G., Castro, I., Razi, A., Boulos, M. N. K., Weller, A., et al. (2020). Leveraging data science to combat covid-19: A comprehensive review. *IEEE Transactions on Artificial Intelligence*, 1(1):85–103.

[Moura et al. 2022] Moura, I., Teles, A., Coutinho, L., and Silva, F. (2022). Towards identifying context-enriched multimodal behavioral patterns for digital phenotyping of human behaviors. *Future Generation Computer Systems*, 131:227–239.

[Moura et al. 2023] Moura, I., Teles, A., Viana, D., Marques, J., Coutinho, L., and Silva, F. (2023). Digital phenotyping of mental health using multimodal sensing of multiple situations of interest: A systematic literature review. *Journal of Biomedical Informatics*, 138:104278.

[O'connor 2018] O'connor, S. (2018). Big data and data science in health care: What nurses and midwives need to know. *Journal of Clinical Nursing*.

[Orphanidou 2019] Orphanidou, C. (2019). A review of big data applications of physiological signal data. *Biophysical reviews*, 11(1):83–87.

[Provost and Fawcett 2013] Provost, F. and Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. "O'Reilly Media, Inc.".

[Shailaja et al. 2018] Shailaja, K., Seetharamulu, B., and Jabbar, M. (2018). Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*, pages 910–914. IEEE.

[Subrahmanya et al. 2022] Subrahmanya, S. V. G., Shetty, D. K., Patil, V., Hameed, B. Z., Paul, R., Smriti, K., Naik, N., and Somani, B. K. (2022). The role of data science in healthcare advancements: applications, benefits, and future prospects. *Irish Journal of Medical Science (1971-)*, 191(4):1473–1483.

[Taddeo and Floridi 2018] Taddeo, M. and Floridi, L. (2018). How ai can be a force for good. *Science*, 361(6404):751–752.

[Talwar et al. 2023] Talwar, R., Sharma, M., Singh, H., and Sagar, P. (2023). A comparative analysis on image processing-based algorithms and approaches in healthcare. In *Handbook of Research on Thrust Technologies' Effect on Image Processing*, pages 1–14. IGI Global.

[Teles et al. 2025] Teles, A. S., de Moura, I. R., Silva, F., Roberts, A., and Stahl, D. (2025). Ehr-based prediction modelling meets multimodal deep learning: A systematic review of structured and textual data fusion methods. *Information Fusion*, page 102981.

# Capítulo 4

# Construindo Modelos Justos: Fundamentos, Estratégias e Desafios para uma IA Ética e Equitativa na Saúde

Bianca Matos de Barros, Diego Dimer Rodrigues, Gabriela Bellardinelli Oliveira, Mariana Recamonde-Mendoza[1]

***Abstract***

*The use of artificial intelligence (AI) in healthcare raises concerns about biases that may perpetuate or exacerbate structural inequalities. This chapter provides an overview of how such biases can emerge throughout the machine learning (ML) pipeline – from data collection to model deployment – leading to unequal performance across different population groups. In addition, it discusses strategies for identifying and mitigating bias at each stage of this process. By integrating fundamental concepts, practical examples, and applicable tools, the chapter serves as a concise reference and emphasizes the importance of interdisciplinary approaches and continuous monitoring to ensure fairness in ML applications within healthcare contexts.*

***Resumo***

*O uso de inteligência artificial (IA) na área da saúde suscita preocupações em relação a vieses que podem perpetuar ou amplificar desigualdades estruturais. Este capítulo apresenta uma visão geral de como esses vieses podem surgir ao longo do pipeline de aprendizado de máquina (AM) – da coleta de dados à implantação do modelo –, gerando desempenhos desiguais entre diferentes grupos populacionais. Além disso, são discutidas estratégias para a identificação e mitigação de vieses em cada etapa desse processo. Ao integrar conceitos fundamentais, exemplos práticos e ferramentas aplicáveis, o capítulo configura-se como uma referência concisa e enfatiza a importância de abordagens interdisciplinares e do monitoramento contínuo para assegurar a equidade nas aplicações de AM em contextos de saúde.*

[1]Todos os autores são afiliados ao Instituto de Informática, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brasil. M. Recamonde-Mendoza também é afiliada ao Núcleo de Bioinformática, Hospital de Clínicas de Porto Alegre (HCPA), Porto Alegre, Brasil.

## 4.1. Introdução

A inteligência artificial (IA) pode ser definida como sistemas computacionais projetados para executar tarefas que, tradicionalmente, requerem inteligência humana, como reconhecimento de padrões, tomada de decisões e resolução de problemas [Haenlein and Kaplan 2019]. A adoção da IA tem provocado transformações profundas em diversas áreas do conhecimento, com potencial para reformular o método científico, os processos de descoberta de conhecimento e o desenvolvimento e operação de soluções. Isso se deve, sobretudo, à sua capacidade de automatizar e otimizar tarefas e decisões de maneira eficiente e escalável. Análises recentes evidenciam o engajamento crescente da comunidade científica com a IA, que já não se restringe a campos específicos, mas se estende a uma ampla gama de áreas do conhecimento [Hajkowicz et al. 2023, Duede et al. 2024].

Dentre essas áreas, a saúde se destaca como um campo em que o potencial transformador da IA é amplamente reconhecido. Aplicações inovadoras baseadas em IA vêm sendo desenvolvidas com foco em diagnósticos e tratamentos personalizados e de alta precisão, predição de pacientes com maior risco de desfechos desfavoráveis (como óbito ou reinternação) e otimização de protocolos. Tais soluções oferecem benefícios que transcendem a melhora direta dos resultados clínicos, incluindo redução de custos, economia de tempo e minimização de erros humanos [Alowais et al. 2023].

De acordo com Schwalbe e Wahl [Schwalbe and Wahl 2020], os usos atuais de IA em saúde podem ser agrupados em quatro eixos principais: (i) diagnóstico, (ii) avaliação do risco de morbidade ou mortalidade do paciente, (iii) previsão e vigilância de surtos de doenças e (iv) planejamento de políticas de saúde pública. Os mesmos autores afirmam que o potencial da IA é ainda mais evidente em países de baixa e média renda (LMICs, do inglês *Low and Middle-Income Countries*), onde a escassez de profissionais, a fragilidade dos sistemas de vigilância e a alta incidência de doenças infecciosas tornam a IA uma ferramenta promissora para superar desafios estruturais nos sistemas de saúde.

Grande parte das aplicações mencionadas é viabilizada por uma subárea da IA chamada de aprendizado de máquina (AM). O AM permite que algoritmos extraiam padrões a partir de dados e aprimorem seu desempenho progressivamente, com base na experiência, sem a necessidade de regras explicitamente programadas [Faceli et al. 2021]. Essa abordagem é particularmente relevante por possibilitar, além da automatização da tomada de decisões, a identificação de fatores associados aos desfechos de interesse que podem eventualmente escapar à análise humana, dada a complexidade ou sutileza das relações presentes nas evidências históricas [Barocas et al. 2023].

[Barocas et al. 2023] sumarizam o ciclo de AM em quatro etapas tipicamente empregadas no uso destes algoritmos, conforme mostrado na Figura 4.1. A primeira etapa é a **medição**, ou seja, o processo de transformar o estado do mundo real relacionado à tarefa que se deseja resolver em um conjunto de dados que possa ser processado pelos algoritmos, sejam dados estruturados (dispostos em tabelas com linhas, colunas e valores) ou não-estruturados (imagens, vídeos, textos, *etc.*). Embora o termo sugira neutralidade, essa etapa é cheia de decisões humanas subjetivas sobre como representar computacionalmente uma realidade complexa e frequentemente desorganizada.

A segunda etapa consiste no **aprendizado** (ou modelagem), momento em que o

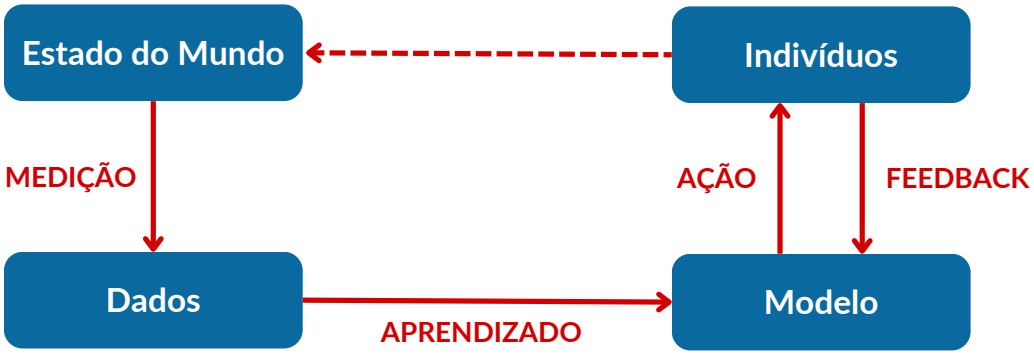



**Figura 4.1. O ciclo fundamental no uso de aprendizado de máquina. Adaptado de [Barocas et al. 2023].**

sistema transforma os dados em um modelo, sumarizando padrões presentes nos dados e fazendo generalizações. Existem diferentes abordagens computacionais que podem ser utilizadas, mas o modelo resultante é uma representação matemática dos padrões identificados, frequentemente expressa por meio de pesos, parâmetros ou, ainda, de regras. A terceira etapa é a de **predição** (ou ação), quando o modelo é aplicado a novos dados, gerando saídas que guiam ações humanas. As ações derivadas dessas previsões impactam os indivíduos e, coletivamente, modificam o estado do mundo, influenciando os padrões futuros. Por fim, em alguns sistemas, temos uma etapa final de **retroalimentação** (*feedback*), na qual as reações dos usuários às decisões são registradas e retroalimentam o modelo, podendo reforçar padrões iniciais.

Entender este ciclo básico do AM é fundamental para compreendermos a importância de fornecermos dados de qualidade para o desenvolvimento de modelos preditivos. Como requisitos mínimos, estes dados devem ser numerosos, diversos (em termos de características e comportamentos observados), representativos do domínio de interesse e coerentemente anotados [Mohammed et al. 2025]. No entanto, mesmo atendendo a estes critérios, não há garantias de que a generalização do modelo será capaz de produzir predições precisas, confiáveis, ou justas. Atualmente, um dos grandes desafios no desenvolvimento de modelos preditivos, especialmente em domínios sensíveis como a saúde, está no fato de que os dados históricos frequentemente carregam preconceitos sociais, estereótipos culturais e desigualdades demográficas, o que pode levar os modelos a aprender e reproduzir essas mesmas distorções.

No contexto da saúde, os dados usados no treinamento de modelos de AM costumam derivar de registros de atendimentos em sistemas de saúde, nos quais disparidades sociais e estruturais são amplamente documentadas [Leal et al. 2005, Goes and Nascimento 2013, Greenwood et al. 2020]. No Brasil, desigualdades regionais e socioeconômicas impactam o acesso a exames, tratamentos e acompanhamento médico, especialmente em populações negras, indígenas e em comunidades periféricas. Mundialmente, estudos demonstram que minorias raciais e étnicas recebem, em média, cuidados menos intensivos ou são subdiagnosticadas para diversas condições. Ao refletirem essas distorções, os dados históricos podem induzir algoritmos a reproduzir padrões discriminatórios, resultando em decisões automatizadas que desprivilegiam os mesmos grupos já vulnera-

bilizados [Silva 2022]. Ou seja, na etapa de medição, o mundo real é representado por dados que já carregam essas disparidades; ao serem utilizados para gerar um modelo, as decisões tomadas com auxílio destes modelos reforçam tais padrões, realimentando as desigualdades existentes.

Esse processo pode levar à criação de modelos com comportamentos enviesados. Em AM, o termo **viés** refere-se a distorções sistemáticas no desempenho do modelo, geralmente causadas pela sub-representação ou baixa qualidade dos dados referentes a determinados grupos. Essas distorções comprometem a capacidade de generalização do modelo, resultando em predições menos precisas ou menos justas para essas populações. Embora o conceito de viés seja discutido com mais profundidade nas seções seguintes, é importante antecipar que ele pode se manifestar de diferentes formas, ter diversas origens e gerar impactos distintos no desempenho e na equidade dos modelos.

Um exemplo ilustrativo de como disparidades sociais podem ser incorporadas não intencionalmente em modelos de AM é o caso analisado por [Obermeyer et al. 2019], que identificaram viés racial em um sistema amplamente utilizado nos Estados Unidos para prever quais pacientes necessitariam de cuidados médicos mais intensivos. O modelo utilizava o custo histórico com cuidados de saúde como variável substituta para a necessidade de cuidados futuros. No entanto, pacientes negros, devido a desigualdades sistêmicas no acesso e na qualidade do atendimento, historicamente geravam menores custos médicos, mesmo apresentando condições de saúde semelhantes às de pacientes brancos. Na etapa de medição, essa escolha de variável resultou em uma representação distorcida da real necessidade de cuidados. O modelo aprendeu, na etapa de aprendizado, que pacientes negros tendem a demandar menos atenção médica e, consequentemente, passou a priorizar pacientes brancos na etapa de predição. Como resultado, o sistema automatizado perpetuou desigualdades preexistentes, impactando negativamente o acesso de populações negras a intervenções preventivas e tratamentos especializados.

O estudo de Obermeyer *et al.* evidencia os riscos de se usar informações que podem carregar disparidades ou desigualdades sociais históricas durante o desenvolvimento dos modelos preditivos, e destaca a importância de incorporar a análise crítica de equidade desde as etapas iniciais deste processo. Conforme será apresentado ao longo deste capítulo, já são inúmeros os casos de vieses em modelos preditivos para a saúde registrados, como penalização em avaliações de desempenho de estabelecimentos de saúde que atendem populações menos favorecidas [Joynt Maddox et al. 2019], redução na acurácia dos diagnósticos [Burlina et al. 2021, Estiri et al. 2022] e na frequência de recomendação de intervenções assistivas [Borgese et al. 2022] para grupos minoritários como pessoas não brancas, mulheres e idosos. Estes casos mostram que graves prejuízos podem surgir quando tais aspectos são negligenciados.

A existência de vieses e a falta de transparência sobre como os modelos tomam decisões são fatores que comprometem a confiança de profissionais e usuários dos serviços de saúde nos modelos de IA. Como consequência, observa-se uma clara disparidade entre a crescente quantidade de pesquisas científicas desenvolvendo soluções em IA para a área da saúde e o pequeno conjunto destas soluções que de fato chegam à prática clínica [Rajpurkar et al. 2022]. A adoção da IA na saúde se mostra mais lenta do que em outros setores, devido a fatores como falta de validação dos modelos baseados em IA com dados

externos, insuficiência tecnológica nas organizações, necessidade de adaptação cultural e dificuldades com regulamentações e políticas, bem como na compreensão da própria tecnologia e das questões éticas envolvendo seu uso [Aldwean and Tenney 2023, Lin et al. 2024]. Abordar esses problemas e oferecer metodologias robustas que sejam capazes de mitigá-los é um passo indispensável na redução da barreira de adoção da IA na saúde.

Assim, este capítulo tem como objetivo apresentar uma abordagem abrangente sobre os vieses em modelos de AM, oferecendo uma base teórica consistente aliada a exemplos práticos que possam servir de referência para pesquisadores e profissionais que desejam iniciar ou aprofundar seus estudos sobre o tema. A discussão será estruturada com base nas etapas do ciclo de desenvolvimento de modelos preditivos, conforme delineado em trabalhos anteriores [Suresh and Guttag 2021], com o intuito de contextualizar de maneira sistemática as possíveis origens dos vieses e seus impactos ao longo de todo o processo de modelagem. O foco do capítulo está no desenvolvimento de modelos preditivos a partir de dados estruturados; embora existam manifestações relevantes de vieses em contextos como visão computacional e processamento de linguagem natural, esses serão apenas brevemente mencionados a título de ilustração.

Este capítulo está organizado da seguinte forma. A Seção 4.2 apresenta os fundamentos do AM necessário para contextualizar o processo de treinamento de modelos preditivos, enquanto a Seção 4.3 explora e esclarece o conceito de viés no escopo de AM. A Seção 4.4 aborda os tipos e origens de vieses em AM. Em seguida, a Seção 4.5 discute os impactos que esses vieses podem causar, especialmente em grupos vulneráveis, e como afetam os usuários dos sistemas automatizados. A Seção 4.6 descreve os principais métodos e métricas utilizados na detecção e quantificação de vieses. Na Seção 4.7, são apresentadas as principais estratégias de mitigação que podem ser adotadas ao longo do ciclo de desenvolvimento dos modelos. A Seção 4.8 aborda as implicações éticas e legais associadas ao uso de modelos enviesados. A Seção 4.9 reúne as principais ferramentas e pacotes de software disponíveis para apoiar a detecção e mitigação de vieses em dados e modelos. Por fim, a Seção 4.10 sintetiza as principais contribuições do capítulo, destacando limitações, desafios em aberto e perspectivas para pesquisas futuras.

## 4.2. Fundamentos de aprendizado de máquina

Esta seção apresenta uma revisão concisa dos principais conceitos de AM, cuja compreensão é essencial para o entendimento das origens, manifestações e impactos dos vieses em modelos preditivos. Trata-se de uma exposição introdutória, e não exaustiva, dos fundamentos da área. Para uma discussão mais aprofundada, sugerimos a consulta a referências complementares, como [Faceli et al. 2021].

### 4.2.1. Paradigmas de IA e o papel do aprendizado de máquina

Em suas décadas iniciais, entre os anos de 1950 e 1990, a pesquisa em IA foi amplamente guiada pelo paradigma simbólico-dedutivo, no qual o raciocínio é aplicado por meio de regras em um modelo de mundo bem definido, dentro de um espaço específico de cálculo, para resolver problemas e realizar inferências. A partir dos anos 1990, com o notável avanço na capacidade computacional, no desenvolvimento de algoritmos mais eficientes e na crescente disponibilidade de dados (muitas vezes, em grandes volumes), a pesquisa

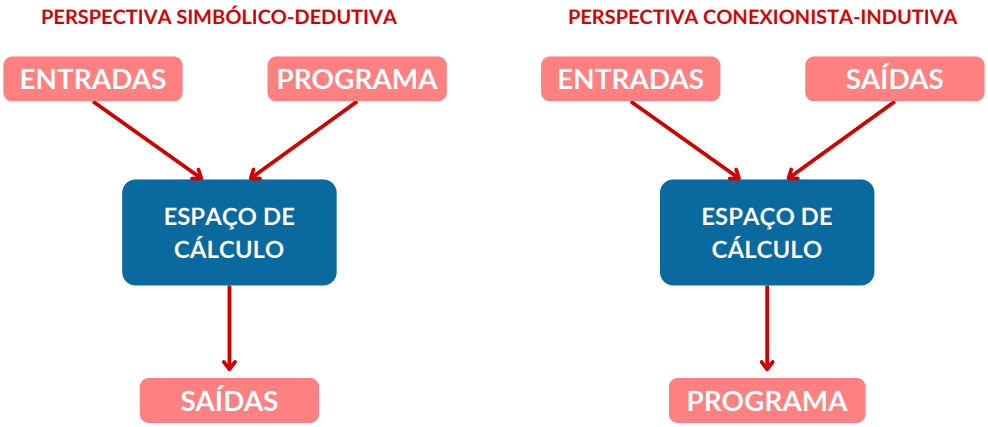



**Figura 4.2. Paradigmas simbólico-dedutivo × conexionista-indutivo na inteligência artificial. Adaptado de [Silva 2022].**

em IA passou a ser principalmente orientada pela perspectiva conexionista-indutiva [Silva 2022]. Nessa abordagem, os cálculos baseiam-se em aspectos correlacionais presentes nos dados utilizados para representar o mundo real, sem o uso de modelos simbólicos.

Os paradigmas simbólico-dedutivo e conexionista-indutivo são comparados na Figura 4.2. Essa mudança de paradigma possibilitou o surgimento e a consolidação do AM como eixo central das aplicações modernas de IA, especialmente após os avanços em dados clínicos digitais, poder computacional e algoritmos. Enquanto a abordagem simbólica proporciona maior transparência e interpretabilidade, a abordagem conexionista — embora mais eficiente em termos de resposta — tende a ser menos compreensível e mais suscetível a erros. Essa distinção pode ser associada ao modelo de cognição humana proposto por Daniel Kahneman, no qual o Sistema 1, intuitivo e rápido, se assemelha à IA baseada em aprendizado (isto é, paradigma conexionista), enquanto o Sistema 2, deliberativo e racional, remete à IA simbólica [Geffner 2018].

Assim, o AM desloca o foco do conhecimento programado para o conhecimento extraído dos dados. Essa transformação estabeleceu uma relação de retroalimentação entre IA e sociedade: os modelos aprendem com dados que refletem as condições sociais existentes e, ao serem aplicados, influenciam comportamentos, decisões e práticas coletivas. Isso torna os sistemas de IA não apenas ferramentas técnicas, mas também agentes sociais, cujas decisões podem reforçar padrões existentes ou gerar novos efeitos.

A hierarquia clássica dos algoritmos de AM os divide em dois grandes grupos de tarefas: as preditivas e as descritivas. As tarefas preditivas são abordadas por meio do aprendizado supervisionado, no qual os algoritmos são treinados com dados rotulados e aprendem a estimar os rótulos de novas instâncias com base em padrões observados nos dados de entrada. Já as tarefas descritivas são tratadas pelo aprendizado não supervisionado, que utiliza dados não rotulados para identificar estruturas ou regularidades, como agrupamentos de instâncias similares ou padrões de associação entre atributos. Há também outras abordagens que não se enquadram rigidamente nessa divisão. São os casos dos aprendizado semissupervisionado (usa dados rotulados e não rotulados no treinamento), ativo (usa dados rotulados e não rotulados junto a uma estrutura de oráculo) e por reforço

(baseia-se na maximização de recompensas acumuladas, aprendendo a partir da interação com um ambiente dinâmico) [Faceli et al. 2021].

Diversas revisões da literatura apontam que o aprendizado supervisionado é a abordagem mais frequentemente utilizada no AM aplicado à saúde [Rajpurkar et al. 2022], especialmente no contexto de estudos que buscam detectar, discutir ou mitigar vieses em nível individual [Caton and Haas 2024]. As demais abordagens de aprendizado aparecem com pouca representatividade, o que se justifica pelo fato de que as aplicações predominantes da IA em saúde — como diagnóstico, prognóstico, triagem, vigilância epidemiológica e apoio à tomada de decisão — são, majoritariamente, de natureza preditiva [Schwalbe and Wahl 2020].

### 4.2.2. Aprendizado de máquina indutivo e supervisionado

No aprendizado supervisionado, os modelos são programados de forma a aprender a partir de experiências passadas. Para isso, utilizam a indução, um princípio de inferência que permite obter conclusões genéricas a partir de um conjunto de dados. Os dados utilizados devem conter variáveis de entrada (também chamadas de atributos ou preditores), cujas relações são exploradas pelos algoritmos, bem como variáveis de saída (ou rótulos), cujos valores o modelo busca prever. Durante o treinamento, os algoritmos visam extrair padrões e relações entre as entradas e os rótulos. Ao final desse processo, espera-se que o modelo seja capaz de generalizar, isto é, aplicar o conhecimento adquirido para realizar predições precisas sobre dados não vistos anteriormente [Faceli et al. 2021].

A tarefa de predição pode assumir diferentes formas, dependendo da natureza dos rótulos. Quando os rótulos pertencem a um conjunto finito e discreto de categorias, a tarefa é denominada classificação. Por outro lado, quando os rótulos são valores numéricos contínuos, a tarefa é caracterizada como regressão. Diversos algoritmos são comumente empregados no aprendizado supervisionado, diferenciando-se entre si pelo tipo de viés indutivo que impõem no processo de aprendizagem. É importante distinguir este conceito técnico do viés no sentido ético ou social. O termo viés indutivo refere-se ao conjunto de suposições, implícitas ou explícitas, que um algoritmo adota para generalizar a partir de dados finitos, causando uma preferência por certas funções em detrimento de outras [Mitchell 1980, Hellström et al. 2020]. Por outro lado, o viés (no contexto ético e social) refere-se a distorções sistemáticas nos dados ou no processo de modelagem que resultam em decisões injustas ou discriminatórias.

Alguns algoritmos supervisionados têm se destacado na análise preditiva em saúde [Badawy et al. 2023]. Por exemplo, algoritmos baseados em instâncias, como o k-vizinhos mais próximos (kNN), assumem que exemplos próximos entre si no espaço de atributos tendem a compartilhar o mesmo rótulo. Árvores de decisão impõem uma estrutura hierárquica e interpretável, na qual as decisões são tomadas com base em divisões sucessivas dos atributos com foco na redução da heterogeneidade ou variância nos rótulos das instâncias. Modelos probabilísticos, como o Naïve Bayes, assumem independência condicional entre os atributos, o que simplifica o cálculo das probabilidades condicionais envolvidas. Redes neurais artificiais, por sua vez, são métodos conexionistas que simulam por meio de representações matemáticas o comportamento de redes de neurônios biológicos organizados em camadas interconectadas, permitindo a aprendizagem de

representações complexas a partir de dados, porém com menor interpretabilidade.

Também merece destaque a regressão logística, um modelo estatístico amplamente utilizado em tarefas de classificação binária, que estima a probabilidade de um evento (como a presença ou ausência de uma condição clínica) a partir de uma combinação linear dos atributos, transformada por uma função sigmoide. Por fim, temos os métodos ensemble, como *Random Forests* ou *Gradient Boosting Trees*, que combinam diferentes modelos (sejam de alta variância ou com diferentes vieses indutivos), buscando maior robustez e capacidade preditiva. A escolha do algoritmo mais apropriado depende, portanto, não apenas das características dos dados e da tarefa, mas também do tipo de suposições que se deseja (ou se pode) fazer sobre o problema.

Na última década, os métodos conexionistas ganharam destaque ao impulsionar avanços significativos na IA, especialmente com a popularização do aprendizado profundo. O aprendizado profundo é uma subcategoria dos métodos conexionistas caracterizada pela presença de múltiplas camadas ocultas entre as camadas de entrada e saída de uma rede neural. Essas camadas intermediárias permitem a modelagem de relações complexas e não lineares nos dados de entrada. Entre os exemplos mais relevantes estão as Redes Neurais Convolucionais (CNNs, do inglês Convolutional Neural Networks), amplamente utilizadas em tarefas de visão computacional pela sua capacidade de extração de características significativas de dados visuais, e os Transformers, que sustentam os mais avançados modelos de processamento de linguagem natural da atualidade, como o ChatGPT [Bazzan et al. 2023]. Embora muitos dos conceitos discutidos neste capítulo possam ser aplicáveis a cenários envolvendo aprendizado profundo, o foco da análise será voltado para modelos preditivos desenvolvidos a partir de dados estruturados.

### 4.2.3. Pipeline de desenvolvimento de modelos preditivos

Conforme demonstrado na Figura 4.3, o processo de aprendizado supervisionado pode ser entendido como um pipeline que inclui coleta de dados, pré-processamento, criação do modelo – compreendendo treinamento e avaliação – e pós-processamento. Além destas etapas centrais no desenvolvimento dos modelos preditivos, também são importantes a formulação do problema e a implantação e monitoramento do modelo. Nesta seção, revisaremos brevemente cada etapa, dada a importância do entendimento deste processo para a posterior compreensão das possíveis origens de vieses (discutidas na Seção 4.4). Cada uma destas etapas é explorada em mais detalhes em referências como [Faceli et al. 2021] e [Burkov 2020].

**Formulação do Problema** O ciclo de vida de um projeto de AM tem início com a formulação do problema. Nessa etapa, define-se qual fenômeno se deseja modelar, que tipo de tarefa será realizada (como classificação ou regressão) e quais são os objetivos e restrições do sistema. Essa definição orienta todas as demais decisões do projeto, como a coleta dos dados, a escolha das métricas de avaliação e os requisitos de interpretabilidade ou desempenho. Uma formulação clara e precisa é fundamental para o sucesso do modelo em contexto prático.

**Coleta de Dados** O ponto de partida para qualquer projeto de AM é a obtenção dos dados que serão usados para treinar e avaliar os modelos. Esse processo envolve definir uma população-alvo, selecionar variáveis relevantes (atributos e rótulos) e estabelecer

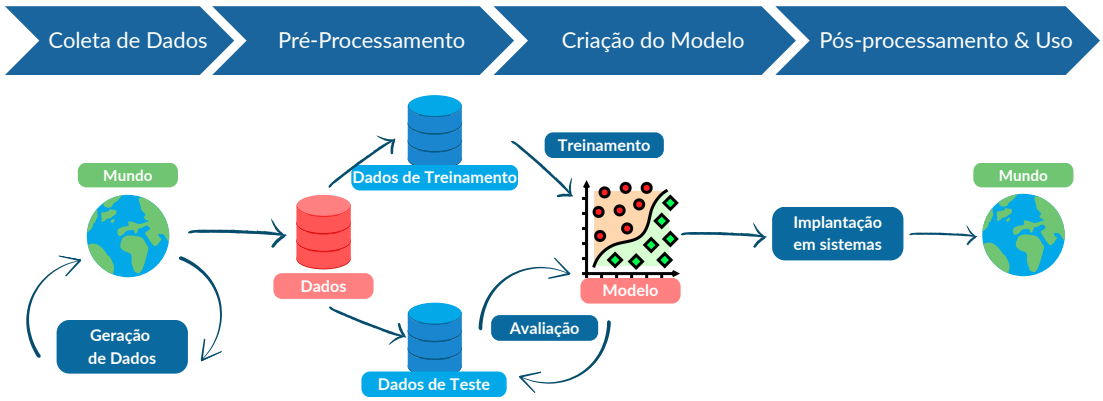



**Figura 4.3. Etapas do aprendizado de máquina supervisionado. Adaptado de [Ruback et al. 2022].**

como essas informações serão coletadas. Frequentemente, por limitações práticas ou financeiras, trabalha-se com uma amostra da população em vez de coletar dados de forma exaustiva, buscando-se, nesse caso, garantir que a amostra seja o mais representativa possível. Também é comum utilizar dados secundários — ou seja, informações que já foram coletadas anteriormente para outros fins. Nesses casos, o projeto já parte de uma base de dados existente. Independentemente da origem dos dados, é fundamental realizar uma análise crítica para verificar se eles contêm informações suficientes para representar adequadamente o problema que se deseja modelar [Burkov 2020]. É necessário avaliar, por exemplo, se os dados cobrem bem os padrões esperados de entrada, se refletem os tipos de situações que o modelo encontrará na prática, e se os rótulos (ou saídas) são consistentes. Além disso, é essencial considerar a qualidade dos dados. Dados coletados de forma retrospectiva ou prospectiva podem conter ruídos ou distorções que não representam fielmente o fenômeno ou domínio de interesse. Tais distorções podem introduzir vieses no modelo — conceito que será discutido com mais detalhes na Seção 4.3.

**Pré-processamento** O pré-processamento é uma etapa essencial que visa transformar os dados brutos em um formato adequado para o treinamento dos modelos. Além disso, é comum que os dados apresentem falhas – como ruídos, valores ausentes, atributos de naturezas distintas, dentre outros – que devem ser tratadas para não gerar dificuldades na etapa de criação do modelo. Os procedimentos mais comuns incluem a integração de dados (quando existem múltiplas fontes), a limpeza dos dados (como tratamento de valores ausentes ou ruidosos, e remoção de duplicatas), transformação de atributos (como normalização ou padronização de atributos numéricos, e codificação de atributos categóricos), e a redução de dimensionalidade (por seleção de atributos ou uso de técnicas como Análise de Componentes Principais). A engenharia de atributos também é uma parte fundamental dessa etapa, buscando criar novas variáveis mais informativas a partir das informações existentes (como transformar um texto com notas clínicas em um vetor estruturado a ser processado pelo algoritmo de AM).

Outro aspecto importante do pré-processamento é a divisão do conjunto de dados em subconjuntos distintos, de treino e teste. O conjunto de treino é usado para ajustar o modelo, enquanto o conjunto de teste é reservado para avaliar e reportar o desempenho do modelo. Em algumas situações, também é gerado um terceiro subconjunto nesta etapa,

denominado de validação, usado para escolher hiperparâmetros e realizar ajustes durante o desenvolvimento. Para evitar distorções na estimativa de desempenho, essas partições devem ser disjuntas e, preferencialmente, construídas por amostragem aleatória – simples ou estratificada, dependendo da necessidade de preservar a distribuição das classes. Não existe uma proporção ideal para cada subconjunto. Algumas fontes sugerem divisões de 70%/30% ou 80%/20% para treino e teste [Faceli et al. 2021], sendo o segundo conjunto normalmente subdividido em dois quando incluímos o subconjunto de validação. Os subconjuntos de validação e teste são utilizados apenas para calcular estatísticas que refletem o desempenho do modelo e, por isso, precisam ser suficientemente grandes para fornecer estimativas confiáveis – em geral, recomenda-se ao menos uma dúzia de exemplos por classe, sendo que algumas centenas por classe em cada conjunto garantem uma avaliação mais robusta [Burkov 2020].

Dentre as estratégias de amostragem, destacam-se: o holdout, que faz uma divisão única entre treino e teste; a validação cruzada, que particiona os dados em k subconjuntos (*i.e., folds)* e avalia o modelo repetidamente variando o subconjunto de teste; e o bootstrap, que utiliza amostragem com reposição para gerar diferentes conjuntos de treino e teste. A escolha da estratégia depende do tamanho do conjunto de dados e da robustez desejada na avaliação. Adicionalmente é importante evitar *data leakage* (contaminação de dados) na preparação dos dados, caracterizado pela introdução indevida de informações dos dados de teste no treinamento do modelo, resultando em estimativas de desempenho artificialmente infladas e em um modelo que não generaliza bem para novos dados. *Data leakage* pode ocorrer desde uma sobreposição indevida entre os subconjuntos de treino e teste (por exemplo, imagens de um mesmo paciente são distribuídas entre os dois subconjuntos, não respeitando uma independência entre eles), como pelo uso de exemplos de teste para transformar os dados (fazer imputação, normalização, *etc.*) ou pela introdução de atributos que não estariam disponíveis no momento da predição (como por exemplo, em predições de óbito em UTI, variáveis como o uso de antibióticos administrados apenas após o agravamento do quadro do paciente podem revelar indiretamente o desfecho antes da previsão). [Kapoor and Narayanan 2023] oferecem uma excelente revisão sobre o tema, discutindo o seu impacto na avaliação dos modelos e a sua relação com a crise de reprodutibilidade em modelos de AM.

**Criação do modelo** Na etapa de criação do modelo, parte-se da modelagem do problema, definindo como será feito o ajuste aos dados. É crucial avaliar se um único modelo é suficiente para representar adequadamente todos os grupos presentes ou se múltiplos modelos são necessários. Nesta etapa, também são selecionados algoritmos compatíveis com o tipo de tarefa e com requisitos específicos do problema, como interpretabilidade, uso de memória e tempo de treinamento. Quando há muitos algoritmos candidatos, pode-se realizar uma etapa inicial de spot-checking, na qual os modelos são treinados com poucas variações de hiperparâmetros para identificar os mais promissores para o conjunto de dados em questão [Burkov 2020]. Com base nesses resultados, um subconjunto de algoritmos é escolhido para a etapa de otimização de hiperparâmetros, buscando a melhor configuração para maximizar o desempenho do modelo. A avaliação é feita com base no desempenho preditivo sobre dados não vistos durante o treinamento, como um subconjunto de validação ou o processo de validação cruzada, utilizando métricas apropriadas. Para regressão, são comuns métricas como *mean absolute error* (MAE) e

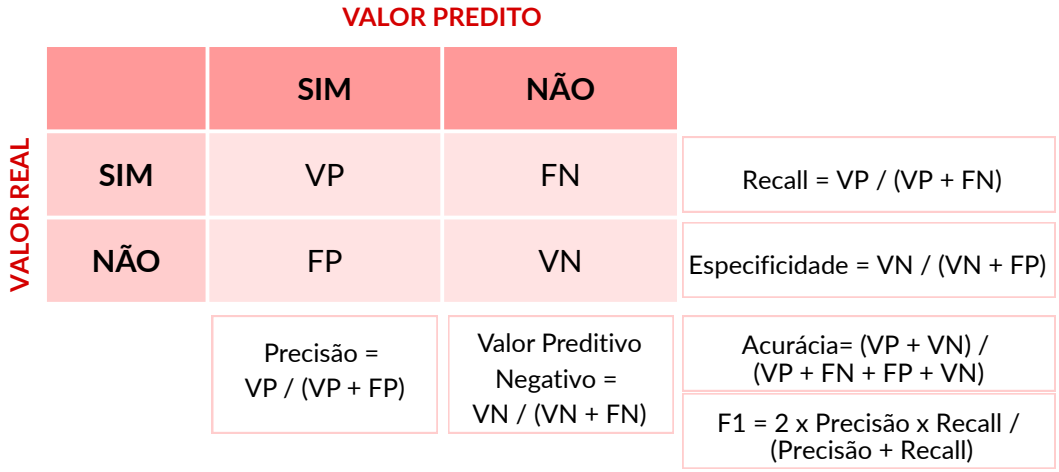



**Figura 4.4. Matriz de confusão e principais métricas de avaliação para classificadores.**

*mean square error* (MSE). Para classificação, utilizam-se métricas derivadas da matriz de confusão (Figura 4.4) e, em domínios como a saúde, métricas baseadas em variação de limiares, como a área sob a curva ROC (AUC-ROC) e a área sob a curva de Precisão-Recall (PR-AUC) são amplamente aplicadas. Após a otimização, o modelo final é avaliado com o conjunto de teste reservado no início do processo.

Um aspecto essencial da avaliação é verificar se o modelo generaliza bem, ou seja, se não sofre de sobreajuste (*overfitting*) ou subajuste (*underfitting*). O sobreajuste ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, obtendo alto desempenho neles, mas baixo desempenho nos dados independentes (validação ou teste). Isso indica que o modelo aprendeu padrões específicos do treinamento que não se repetem no mundo real. Já o subajuste ocorre quando o modelo apresenta baixo desempenho mesmo nos dados de treinamento, sugerindo que ele é incapaz de capturar os padrões relevantes, seja por simplicidade excessiva ou por dados pouco representativos.

Além da avaliação quantitativa do desempenho, é fundamental interpretar o modelo treinado para compreender como ele toma decisões e quais variáveis influenciam suas previsões. A interpretabilidade é especialmente importante em domínios sensíveis, como saúde, justiça e finanças, onde decisões automatizadas podem ter consequências significativas. Técnicas como análise de importância de atributos, uso de modelos interpretáveis por construção (como árvores de decisão), e métodos pós-hoc (como SHAP ou LIME) ajudam a garantir maior transparência e confiabilidade, além de auxiliar na detecção de vieses e na validação científica dos resultados [Molnar 2025].

**Pós-processamento** Após o treinamento do modelo, pode ser necessário realizar etapas de pós-processamento para adaptar as saídas do modelo ao contexto de uso ou às exigências práticas da aplicação. Por exemplo, em tarefas de classificação binária, os modelos frequentemente retornam uma probabilidade associada a uma das classes. No entanto, para a tomada de decisão ou exibição ao usuário, é preciso definir um limiar (*threshold*) para converter essa probabilidade em uma classe do problema. A escolha desse limiar pode ser ajustada de acordo com a sensibilidade desejada do modelo, priorizando, por exemplo, a redução de falsos positivos ou falsos negativos. Em outros casos, as saídas

podem ser agrupadas em categorias mais interpretáveis, como faixas de risco. Por exemplo, para prever risco de reinternação de paciente, o modelo pode gerar um score entre 0 e 1, o qual pode ser mapeado para rótulos como “baixo risco", “incerto", e “alto risco".

**Implantação e Monitoramento** Após a finalização do modelo, sua implantação em ambiente de produção requer cuidados adicionais. É importante garantir que os dados de entrada no sistema operacional estejam no mesmo formato e distribuição dos usados durante o treinamento. Uma vez implantado, o modelo deve ser monitorado continuamente quanto ao seu desempenho, para detectar degradações causadas por mudanças nos dados ou no contexto de aplicação (conhecidas como *data drift* e *concept drift*). Modelos em produção podem precisar de revalidação e retreinamento periódicos. Mecanismos de feedback fornecem dados atualizados e corrigidos ao modelo, constituindo ferramentas que promovem tanto a atualização da área de domínio quanto a correção de falhas, promovendo um ciclo de melhoria contínua [Ruback et al. 2022, Burkov 2020].

## 4.3. Definição de viés em aprendizado de máquina

O termo viés pode assumir diferentes significados, a depender do contexto em que é utilizado. No campo dos algoritmos de AM, **viés** é frequentemente definido como um erro sistemático ou uma tendência inesperada de favorecer certos resultados em detrimento de outros [Mehrabi et al. 2021]. Em aplicações de IA na área da saúde, viés pode ser entendido como qualquer diferença sistemática e indesejada na forma como previsões são geradas para diferentes populações de pacientes [Hasanzadeh et al. 2025].

Essas diferenças, muitas vezes, decorrem de uma dependência indesejada do modelo em relação a atributos sensíveis, também chamados de **atributos protegidos**, como raça, gênero, idade, ou condição socioeconômica [Caton and Haas 2024]. Esses atributos são considerados sensíveis porque se referem a características pessoais ou demográficas que, por razões éticas, legais ou históricas, não deveriam influenciar a tomada de decisão, principalmente aquela feita de forma automatizada. Quando um modelo se apoia nesses atributos, de maneira direta ou indireta, pode gerar impactos negativos desproporcionais sobre certos grupos desprivilegiados – ou seja, grupos que já enfrentam desvantagens estruturais na sociedade. Isso pode resultar em previsões injustas ou discriminatórias e, quando aplicadas na prática clínica, levar a uma prestação desigual de cuidados em saúde.

Por exemplo, se um modelo preditivo de risco cardiovascular subestima o risco em mulheres negras em comparação com homens brancos com o mesmo perfil clínico, esse erro pode atrasar o diagnóstico e o início do tratamento para o primeiro grupo. Nesses casos, o viés é considerado injusto, tanto do ponto de vista ético quanto do ponto de vista legal, pois reforça desigualdades preexistentes no sistema de saúde. Esse tipo de viés racial já foi documentado em inúmeros trabalhos, conforme discutiremos na Seção 4.5, e não é exclusivo de sistemas baseados em AM – muitos scores de risco aplicados na saúde refletem essa fragilidade [Coots et al. 2025].

Cabe salientar que mesmo quando atributos sensíveis não são explicitamente incluídos no conjunto de variáveis utilizadas pelo modelo, informações correlacionadas a esses atributos podem ser indiretamente incorporadas por meio de variáveis substitutas, conhecidas como *proxies*. Um ***proxy*** é uma variável que, embora não represente diretamente um atributo sensível, está altamente correlacionada a ele, permitindo que o modelo

aprenda padrões que refletem as mesmas desigualdades [Mehrabi et al. 2021]. A complexidade das relações de saúde e a eventual indisponibilidade de conjuntos de dados que ofereçam todas as informações relevantes em quantidade suficiente leva ao uso de correlações ou aproximações muitas vezes não óbvias, mas detectáveis pelos algoritmos.

Por exemplo, o código postal de um paciente pode funcionar como *proxy* para raça ou condição socioeconômica, pois áreas geográficas frequentemente refletem segregações históricas e desigualdades no acesso a recursos. Da mesma forma, o tipo de seguro de saúde ou a frequência de visitas médicas anteriores podem indiretamente representar fatores como nível de renda ou acesso prévio a cuidados adequados [O'Neil 2021, Obermeyer et al. 2019]. Ainda que variáveis demográficas sensíveis, como raça, idade ou gênero sejam removidas das bases de dados, as correlações entre elas e outros atributos demográficos ou comportamentais podem ser deduzidas pelos modelos com alta acurácia [Kosinski et al. 2013]. O uso de *proxies* dificulta a detecção explícita de viés, tornando ainda mais desafiador o desenvolvimento de modelos verdadeiramente justos e imparciais.

Para uma compreensão mais precisa do conceito de viés, também é importante distinguirmos viés estatístico de viés social. **Viés estatístico** refere-se a qualquer desvio sistemático entre os valores estimados por um modelo e os valores reais esperados [Parikh et al. 2019]. Esse tipo de viés compromete a validade das inferências estatísticas ao introduzir distorções na predição que não decorrem do acaso, mas de falhas na formulação do problema, no desenho experimental, na seleção da amostra, na modelagem ou na análise dos dados. Um exemplo clássico na área da saúde é o escore de risco de Framingham, amplamente usado para doenças cardiovasculares, o qual foi desenvolvido a partir de uma população majoritariamente branca e não hispânica [Coots et al. 2025]. Embora não use explicitamente a informação de raça na avaliação, quando aplicado a populações negras, o score tende a subestimar o risco de eventos cardiovasculares porque não captura adequadamente os fatores de risco prevalentes nesse grupo, por falta de representatividade na amostra utilizada para derivar o score [Parikh et al. 2019].

O **viés social**, por outro lado, refere-se à desigualdade na prestação de cuidados em saúde que sistematicamente leva a resultados abaixo do ideal para um determinado grupo [Parikh et al. 2019]. Em outras palavras, observa-se uma reprodução, amplificação ou naturalização de desigualdades sociais, culturais e estruturais existentes nos cuidados em saúde pelos processos algorítmicos de tomada de decisão. Ao contrário do viés estatístico, que é definido por desvios sistemáticos em termos de estimativas quantitativas, o viés social emerge das escolhas normativas e das estruturas sociais que permeiam a coleta de dados, a definição de variáveis, a formulação de objetivos e a própria lógica de funcionamento dos sistemas automatizados.

Enquanto o viés estatístico compromete a precisão e validade preditiva do modelo, o viés social compromete a justiça e equidade nas decisões automatizadas. Em muitos casos, esses dois tipos de viés coexistem e se reforçam mutuamente, gerando sistematicamente piores resultados para determinados grupos sociais. Por isso, independentemente de sua origem técnica ou sociocultural, é essencial o desenvolvimento de estratégias para identificar, mitigar e monitorar os vieses em modelos de AM aplicados à saúde.

## 4.4. Tipos de viés e suas origens

Vieses em modelos de AM podem surgir em diferentes etapas do processo de desenvolvimento descrito na Seção 4.2.3. Adicionalmente, alguns vieses já estão presentes no domínio modelado, ou surgem através da implantação e uso do modelo final. Nas próximas seções, apresentamos uma sistematização dos principais vieses discutidos na literatura, relacionando tipo (ou seja, como ele afeta o modelo) e origem do viés (isto é, quando ele emerge no processo). É importante destacar que os tipos e origens de vieses discutidos na literatura são diversos e nem sempre correspondem entre si. Assim, os conceitos aqui definidos resultam de uma análise agregada de diferentes trabalhos, com o objetivo de combinar perspectivas, padronizar a terminologia e facilitar o entendimento do tema [Suresh and Guttag 2021, Mehrabi et al. 2021, Rajkomar et al. 2018].

### 4.4.1. Coleta de dados

A Tabela 4.1 resume os principais vieses que podem surgir na coleta de dados para modelos preditivos em saúde. Muitos desses vieses já estão presentes na própria realidade que os dados buscam representar, refletindo desigualdades sociais e históricas anteriores ao processo de coleta. Um exemplo é o viés histórico, que decorre de desigualdades acumuladas no acesso, na qualidade e no tipo de atendimento em saúde, bem como da baixa inclusão de determinados grupos — como mulheres e minorias étnicas — em ensaios clínicos [Mccarthy 1994]. Os dados não são fruto apenas de escolhas individuais, mas também são moldados por estruturas ideológicas, desigualdades históricas e representações sociais. Como resultado, carregam influências de sistemas como o colonialismo, o patriarcado e o racismo científico, cujos efeitos continuam a impactar a forma como pensamos, atuamos e coletamos informações, reproduzindo formas de exclusão como o racismo, a misoginia e o capacitismo [Ruback et al. 2022].

**Tabela 4.1. Principais tipos de vieses introduzidos na coleta de dados.**

| **Tipo de Viés** | **Origem** |
|---|---|
| Viés Histórico | Dados refletem desigualdades ou práticas discriminatórias do passado. |
| Viés de Anotação | As anotações realizadas dos dados refletem suposições, crenças ou estereótipos dos anotadores ou observadores. |
| Viés de Amostragem | Subgrupos são amostrados de forma não aleatória, comprometendo a generalização dos modelos e achados. |
| Viés de Representação | Certos grupos populacionais são sub ou super-representados no conjunto de dados coletado. |
| Viés de Medição | Surge da forma como escolhemos, utilizamos e medimos determinadas variáveis (atributos e rótulos). |
| Viés de População | A demografia dos usuários incluídos no estudo difere da população-alvo original. |
| Viés de Auto-Seleção | Os próprios participantes escolhem participar de uma pesquisa (de forma voluntária). |

O **viés de anotação** também pode ocorrer antes ou durante a coleta de dados. Por exemplo, médicos ou enfermeiros podem deixar de registrar sintomas relatados por pacientes com transtornos mentais por considerarem esses relatos menos confiáveis, ou subestimar queixas de dor de mulheres com base em estereótipos inconscientes, distor-

cendo os registros clínicos utilizados para treinar modelos de predição. Esse viés está fortemente relacionado ao chamado **viés de observação**, que ocorre quando o observador influencia ou interpreta os dados com base em suas crenças ou expectativas prévias.

Além dos vieses de natureza sistêmica ou estrutural, existem diversas distorções associadas aos métodos de coleta de dados, muitas vezes referidas genericamente como vieses de dados. Um exemplo importante é o **viés de representação**, que surge quando a amostra utilizada no desenvolvimento do modelo sub-representa determinados segmentos da população, comprometendo sua capacidade de generalização. Esse viés está associado ao "viés de minoria" quando o grupo protegido tem uma quantidade insuficiente de exemplos para que o modelo aprenda padrões estatísticos adequados [Rajkomar et al. 2018]. De fato, vieses sóciodemográficos já foram identificados em diversos modelos clínicos baseados em AM, como através da sub-representação de participantes negros, hispânicos e asiáticos [Colacci et al. 2024].

O **viés de amostragem** decorre da seleção não aleatória de subgrupos, priorizando certos tipos de instâncias em detrimento de outros. Isso resulta em conjuntos de dados que não refletem adequadamente a diversidade da população real. Por exemplo, [Fatumo et al. 2022] relataram que cerca de 86% dos indivíduos incluídos em estudos genômicos eram de ascendência europeia, embora apenas aproximadamente 9,3% da população mundial estivesse concentrada no continente europeu no mesmo período.

O **viés de população**, também conhecido como viés de coorte, ocorre quando os dados são coletados a partir de uma população específica que não representa a população-alvo. Isso pode ocorrer devido a restrições geográficas, institucionais ou temporais, resultando em modelos que não se generalizam adequadamente. Por exemplo, ao treinar um modelo preditivo de risco de câncer com dados de pacientes atendidos em um hospital privado de grande porte – predominantemente composto por pessoas de classes socioeconômicas mais altas –, é possível que o modelo apresente desempenho inferior quando aplicado em hospitais públicos ou em comunidades rurais, devido a diferenças na prevalência da doença, no acesso ao diagnóstico e no histórico médico dos pacientes.

Os modelos também estão suscetíveis ao **viés de medição**, caracterizado pelo uso de proxies inadequados ou inconsistentes entre grupos. Como destacado por [Suresh and Guttag 2021], as variáveis utilizadas pelos modelos são frequentemente *proxies* (medidas concretas) escolhidas para representar fenômenos que não são diretamente observáveis ou quantificáveis. Esse viés pode ocorrer, por exemplo, devido a calibrações diferentes de equipamentos entre instituições, introduzindo variações artificiais nos dados. Além disso, a prescrição de medicamentos (como antidepressivos ou insulina) é muitas vezes utilizada como indicador da presença de uma condição (como depressão ou diabetes), embora haja variações significativas no acesso a medicamentos entre diferentes grupos populacionais.

Por fim, destaca-se o **viés de auto-seleção**, que ocorre quando as amostras disponíveis para modelagem são geradas a partir de decisões voluntárias ou de barreiras estruturais que afetam quem aparece nos dados. Isso pode distorcer tanto as estimativas de prevalência quanto as relações entre variáveis, reduzindo a capacidade de generalização dos modelos para a população real. Por exemplo, dados provenientes de aplicativos de saúde (como aplicativos de monitoramento de sintomas) tendem a refletir um público mais jovem, com maior escolaridade e acesso à tecnologia. De forma semelhante, pes-

soas que respondem a pesquisas voluntárias sobre saúde mental podem ser aquelas que já têm interesse no tema ou que estão em busca de ajuda, enquanto indivíduos com sintomas mais graves ou em situação de vulnerabilidade podem não participar dessas iniciativas.

### 4.4.2. Pré-processamento

Alguns vieses podem surgir na etapa de pré-processamento dos dados (Tabela 4.2). Um deles é o **viés de seleção**, que ocorre quando mecanismos de amostragem utilizados para dividir os dados em subconjuntos de treinamento e teste – ou durante procedimentos como a validação cruzada – introduzem distorções nas partições geradas. Essas divisões podem resultar em conjuntos com distribuições significativamente diferentes, comprometendo a capacidade de generalização do modelo. Por exemplo, se o subconjunto de treinamento tiver uma maior proporção de indivíduos jovens, o modelo pode aprender padrões mais eficazmente para essa faixa etária, em detrimento de indivíduos mais idosos.

Outro viés possível é o **viés de variável omitida**, que ocorre quando variáveis relevantes para a modelagem do problema são removidas acidentalmente durante o pré-processamento. Vale destacar que esse viés também pode estar presente desde a etapa de coleta de dados, caso determinadas variáveis jamais tenham sido registradas. Um exemplo seria o desenvolvimento de modelos para predição de risco cardiovascular sem considerar o nível socioeconômico dos pacientes – uma variável que influencia tanto os fatores de risco (como dieta, estresse e acesso a cuidados preventivos) quanto os desfechos. Ao omitir essa informação, o modelo pode superestimar o efeito de variáveis como o Índice de Massa Corporal (IMC), confundindo os efeitos da pobreza ou da baixa escolaridade com fatores fisiológicos.

O **viés de exclusão** ocorre quando certos grupos ou amostras são removidos, intencionalmente ou não, durante a limpeza ou preparação dos dados, comprometendo a representatividade do conjunto usado para treinar o modelo. Em saúde, um exemplo comum é a exclusão de pacientes com prontuários incompletos ou histórico médico fragmentado. Embora essa prática busque melhorar a qualidade dos dados, ela pode eliminar desproporcionalmente indivíduos em situação de vulnerabilidade social, como pessoas em situação de rua. Relacionado a este viés temos o **viés de dados ausentes**, que ocorre quando informações relevantes estão faltando de forma não aleatória, afetando especialmente determinados grupos populacionais. Como os dados ausentes muitas vezes refletem desigualdades no acesso a serviços, diagnósticos ou registros clínicos, o modelo treinado

**Tabela 4.2. Principais tipos de vieses introduzidos na etapa de pré-processamento.**

| Tipo de Viés | Origem |
|---|---|
| Viés de Seleção | A divisão aleatória de dados pode não manter o balanceamento real das amostras ou das características, gerando partições enviesadas. |
| Viés de Variável Omitida | Ocorre quando variáveis importantes são deixadas de fora do modelo. |
| Viés de Dados Ausentes | Os dados podem estar ausentes de forma não aleatória para grupos protegidos, dificultando a geração de previsões precisas. |
| Viés de Exclusão | Ocorre quando grupos ou amostras são removidos (intencional ou acidentalmente) durante a limpeza ou preparação dos dados. |

pode ter desempenho inferior justamente para os grupos mais vulneráveis. Por exemplo, ao treinar um modelo para prever insuficiência renal com base em exames laboratoriais, é comum que pacientes em situação de vulnerabilidade socioeconômica ou moradores de áreas rurais tenham menos registros de exames como creatinina ou ureia, devido a barreiras de acesso ao sistema de saúde. A falta destas variáveis pode comprometer o desempenho do modelo para estes grupos específicos.

### 4.4.3. Criação do modelo

Os vieses que surgem nesta etapa, sumarizados na Tabela 4.3, estão principalmente relacionados ao funcionamento interno do algoritmo ou à forma como ele é avaliado. O **viés de modelagem**, ou **viés algorítmico**, ocorre quando escolhas algorítmicas realizadas durante o processo de desenvolvimento do modelo – como a função de otimização utilizada, critérios de regularização, e decisões sobre aplicar o modelo de forma global ou separada por subgrupos – introduzem ou amplificam disparidades de desempenho. O **viés de avaliação** ocorre quando os dados utilizados como referência para medir o desempenho de um modelo (como *benchmarks*) não representam adequadamente a população-alvo do seu uso. Esse viés pode levar ao desenvolvimento de modelos que apresentam bons resultados apenas para os grupos presentes nesses dados de avaliação, mas que falham ao serem aplicados em contextos reais mais diversos. Além disso, a escolha de métricas agregadas, como acurácia global, pode mascarar disparidades de desempenho entre subgrupos, ocultando taxas elevadas de erros em populações minoritárias.

Por exemplo, ao treinar um classificador para prever risco de complicações em pacientes hospitalizados, otimizar apenas a acurácia geral pode favorecer o grupo majoritário, resultando em muitos falsos negativos para minorias étnicas ou faixas etárias menos representadas. Além disso, modelos treinados com dados de um único hospital podem não generalizar bem para outras instituições, devido a diferenças no perfil dos pacientes, fatores regionais ou práticas institucionais.

Por fim, citamos o **viés de agregação** que ocorre quando um único modelo é aplicado a dados que contêm subgrupos com características distintas, desconsiderando que a relação entre variáveis de entrada e saída pode variar entre eles. Isso compromete o uso do modelo por causar diferenças de desempenho do modelo para subgrupos envolvidos. Por exemplo, o uso de um único modelo de risco para doenças cardíacas que não diferencia padrões entre homens e mulheres pode falhar na estimativa de risco para mulheres. Como fatores de risco e sintomas podem se manifestar de maneira diferente entre os sexos, o modelo treinado sobre o grupo dominante (geralmente homens) pode apresen-

**Tabela 4.3. Principais tipos de vieses introduzidos na etapa de criação do modelo.**

| Tipo de Viés | Origem |
|---|---|
| Viés de Modelagem | Quando decisões sobre arquitetura, funções de custo, ou algoritmos favorecem certos padrões ou grupos. |
| Viés de Avaliação | Modelos são avaliados com métricas ou conjuntos de dados que não refletem adequadamente a diversidade real dos usuários ou casos. |
| Viés de Agregação | Conclusões sobre indivíduos são tiradas com base em dados agregados, desfavorecendo alguns grupos específicos ou minoritários. |

tar desempenho inferior na predição de risco para mulheres, resultando em diagnósticos menos precisos ou atrasados.

### 4.4.4. Pós-processamento e uso

Nesta etapa, surgem os vieses relacionados a interpretação humana, também chamados de vieses de implantação, que são não-computacionais. Alguns exemplos estão listados na Tabela 4.4. Uma das fontes deste tipo de viés é quando há uma incompatibilidade entre o problema que o modelo pretende resolver e a maneira como ele é realmente utilizado (**viés de uso**). Por exemplo, um modelo treinado para prever risco de complicações em pacientes hospitalizados de um centro urbano pode não ser adequado para hospitais rurais, onde o perfil dos pacientes e as condições locais são diferentes. Já o **viés de feedback** aparece quando profissionais de saúde seguem recomendações incorretas do modelo, reforçando e perpetuando erros em versões futuras. Um caso comum é a decisão clínica automatizada que, se errada, acaba influenciando futuras coletas de dados e treinamentos, criando um ciclo vicioso que mantém o modelo enviesado.

O **viés de automação** acontece quando os profissionais confiam cegamente no modelo, mesmo quando ele tem desempenho inferior para certos grupos, como minorias étnicas ou faixas etárias específicas. Isso pode levar a decisões médicas imprecisas, prejudicando esses pacientes. Outro viés relevante é o **viés de discrepância na alocação**, em que grupos protegidos recebem menos predições positivas, resultando em menor alocação de recursos essenciais, como atenção clínica ou suporte social. Por exemplo, pacientes de grupos socioeconômicos vulneráveis podem ter menos chances de receber intervenções preventivas por conta dessa discrepância nas recomendações do modelo.

## 4.5. Exemplos de vieses em modelos preditivos na saúde

Discutimos até aqui o conceito de viés, seus diversos tipos e as maneiras como podem ser introduzidos ao longo do ciclo de desenvolvimento de modelos preditivos baseados em AM. Para compreender melhor o impacto social e ético desses vieses e motivar a discussão sobre estratégias de mitigação, é importante examinarmos exemplos concretos de como eles têm sido caracterizados no contexto de IA aplicada à saúde. Vale destacar que os impactos negativos do viés em sistemas de apoio à decisão não se restringem apenas a este domínio. Há mais de uma década, diferentes áreas da sociedade vêm sendo

**Tabela 4.4. Principais tipos de vieses introduzidos na etapa de pós-processamento e uso do modelo.**

| Tipo de Viés | Origem |
|---|---|
| Viés de Uso | Decorre da diferença entre o contexto de uso real e o cenário para o qual o modelo foi treinado. |
| Viés de Feedback | Profissionais seguem a recomendação do modelo mesmo que esteja errada, perpetuando erros que serão incorporados em futuras versões do modelo. |
| Viés de Automação | Profissionais não sabem que o modelo erra mais para certos grupos e, por isso, confiam demais em previsões incorretas. |
| Viés de Discrepância na Alocação | Grupos protegidos recebem menos previsões positivas, levando a menor alocação de recursos (como atenção clínica ou serviços sociais). |

afetadas por esses problemas. Um exemplo é o mapeamento realizado por Tarcízio Silva sobre Danos e Discriminação Algorítmica[2] – anteriormente conhecido como Linha do Tempo do Racismo Algorítmico –, que evidencia a frequência e a gravidade desses casos em diferentes contextos.

Esta breve revisão de alguns exemplos recentes de vieses preditivos na saúde será orientada pelas variáveis sensíveis envolvidas nestes casos. Já mencionamos o exemplo emblemático do estudo de Obermeyer *et al.*, no qual os autores identificaram que um algoritmo utilizava o custo com saúde como *proxy* para necessidades de saúde, o que levou a uma subestimação sistemática das necessidades de pacientes negros em relação a pacientes brancos, já que, historicamente, a população negra tem menos acesso e recebe ou faz menos investimentos em saúde [Obermeyer et al. 2019]. Trata-se de um viés racial, perpetuando as disparidades já existentes no acesso aos cuidados em saúde. [Huang et al. 2022] fizeram uma revisão focada no viés racial, observando que este tipo de viés foi reportado em diagnósticos por imagem de retinopatia diabética, predição de depressão pós-parto e uso indevido de opioides. Em todos estes casos, observou-se menor poder preditivo para indivíduos negros em relação aos indivíduos brancos.

Quanto ao viés de gênero, [Solans Noguero et al. 2023] investigaram disparidades de desempenho entre homens e mulheres em algoritmos de detecção de anorexia nervosa em postagens de redes sociais, identificando taxas de falsos negativos significativamente maiores para grupos sub-representados (normalmente mulheres). [Larrazabal et al. 2020] demonstraram que modelos de classificação de imagens médicas treinados majoritariamente com dados de homens apresentavam desempenho inferior ao diagnosticar imagens de pacientes do sexo feminino. Os autores ressaltam que este é um desafio, pois muitos conjuntos de imagens disponibilizados publicamente e usados para treinar (ou pré-treinar) modelos, não contém informação de gênero para cada indivíduo contido na amostra, impossibilitando a detecção ou mitigação deste tipo de viés.

O problema de etarismo em modelos de IA foi revisado e discutido por [Stypinska 2023], sumarizando evidências de que a análise de sentimentos e o reconhecimento facial com algoritmos de IA possuem um significativo viés de idade. Por exemplo, um estudo mostrou que frases contendo adjetivos mais “joviais" tinham 66% mais probabilidade de serem pontuadas positivamente na análise de sentimentos do que frases idênticas com adjetivos mais “velhos"[Díaz et al. 2018]. Outra análise mostrou que modelos de reconhecimento facial para prever idade e gênero a partir de fotografias tinham um desempenho ruim em faixas etárias mais velhas (com 60 anos ou mais) [Meade et al. 2021]. Além disso, observou-se que o pior desempenho era obtido em mulheres mais velhas e negras, um exemplo de intersecção entre três tipos de vieses: de raça, de gênero e etário. No geral, estes resultados negativos estão atrelados à sub-representação de idosos (ou de suas características ou hábitos) no conjunto de treinamento.

Viés relacionado à ancestralidade (*i.e.,* viés genético) também tem sido fonte de preocupação na área da saúde, tendo em vista que trabalhos anteriores já apontaram falhas em modelos genômicos ao generalizar para populações não-europeias [Martin et al. 2019]. Estima-se que indivíduos de origem africana, asiática ou hispânica representem menos de 10% dos dados disponíveis em bases genômicas públicas, o que prejudica dire-

[2] `https://desvelar.org/casos-de-discriminacao-algoritmica/`

tamente o desempenho de modelos preditivos nessas populações [Guerrero et al. 2018]. [Hatoum et al. 2021] investigaram como modelos de AM treinados para prever o transtorno por uso de opioides podem ser enviesados pela ancestralidade. Eles descobriram que os modelos treinados com variantes genéticas candidatas apresentaram desempenho elevado quando havia confusão entre casos e ancestralidade, mas esse desempenho caiu para o nível do acaso quando os dados foram balanceados por ancestralidade.

Além dos casos relatados, destacamos que no campo do processamento de linguagem natural (PLN), amplamente usado para análise de prontuários médicos e sistemas de apoio à decisão clínica, há outros exemplos de modelos que já demonstraram replicar preconceitos relacionados a racismo, misoginia, homofobia e xenofobia [Papakyriakopoulos et al. 2020]. Além disso, diferenças têm sido observadas nas recomendações clínicas geradas para pacientes pertencentes a grupos minoritários [Borgese et al. 2022]. Por fim, em visão computacional, onde algoritmos de IA são utilizados para tarefas como detecção e segmentação de lesões, análise de imagens biomédicas e geração de representações tridimensionais, diversos relatos apontam que esses sistemas podem apresentar desempenho inferior em pessoas de diferentes gêneros ou tons de pele, reforçando desigualdades no diagnóstico e tratamento [Daneshjou et al. 2022, Buolamwini and Gebru 2018].

Diante desse cenário, diversos estudos têm destacado a importância de princípios como justiça (*fairness*), igualdade e equidade na prestação de cuidados em saúde [Hasanzadeh et al. 2025]. É fundamental distinguir esses conceitos, pois eles se relacionam diretamente à presença – e à mitigação – de vieses em sistemas de IA. Justiça em saúde envolve tanto aspectos distributivos quanto socio-relacionais, exigindo uma abordagem holística que leve em conta os contextos sociais, culturais e ambientais dos indivíduos. A igualdade busca garantir o mesmo nível de acesso e de resultados para todos, enquanto a equidade reconhece que diferentes grupos podem demandar apoios específicos para alcançar os mesmos benefícios. Nesse sentido, estratégias uniformes – mesmo bem-intencionadas – podem acentuar ainda mais disparidades preexistentes.

## 4.6. Métricas e métodos para detecção de viés

Quantificar o viés nos dados é o primeiro passo para corrigir as disparidades e evitar modelos injustos. A partir da definição do que são os vieses e como eles podem surgir nos dados, torna-se imperativo o entendimento do contexto e onde as diferentes medidas de análise de vieses podem ser aplicadas [Hardt et al. 2021]. Por exemplo, considerando o atributo **sexo** e, para simplificar, assumindo dois grupos demográficos (ou duas *classes*): homens e mulheres. Em um modelo de AM aplicado a um processo seletivo, a justiça pode ser pensada de diferentes maneiras. Primeiramente, um número igual de candidatos de cada grupo demográfico é aceito/rejeitado. Essa suposição pode considerar que os dois grupos têm tamanhos iguais entre os candidatos ou que ambos apresentam a mesma distribuição quanto à qualificação dos aplicantes – ou ainda, pode desconsiderar essas informações. Em segundo lugar, que a porcentagem de aprovados e rejeitados é igual entre os grupos, levando em conta (ou não) a distribuição de qualificação entre os grupos.

Os conjuntos de dados utilizados podem apresentar diferentes distribuições para os atributos de interesse, e essas diferenças podem indicar vieses – sejam eles inerentes aos dados ou não. Aplicar métricas para detecção de vieses começa pela definição de

quais métricas são apropriadas para o contexto e fazem sentido no domínio de aplicação. Bases distorcidas ou desbalanceadas representam apenas parte do desafio, e podem indicar problemas que frequentemente não são tratados no pipeline de desenvolvimento de modelos, como desigualdades estruturais na área da saúde, diferenças no atendimento entre pacientes, inclusão indevida de atributos sensíveis ou de *proxys* desses atributos, entre outros [Chen et al. 2021].

As métricas para detecção de viés apresentadas nesta seção, resumidas na Tabela 4.5, são divididas em dois grupos: no primeiro, **métricas pré-treino** (ou pré-modelo), que podem ser calculadas sem intervenção humana e requerem apenas o conjunto de dados a ser usado na modelagem (Seção 4.6.1); no segundo, **métricas pós-treino** (ou pós-modelo), que levam em conta os resultados do modelo e como as predições se distribuem entre as diferentes classes do atributo sensível considerado (Seção 4.6.2). As definições foram extraídas de [Hardt et al. 2021] e são focadas em problemas de classificação binária, nos quais o atributo alvo do modelo assume sempre um valor positivo (1) ou negativo (0). A classe privilegiada (ou favorecida) pelo viés será denotada pela letra $p$, e a classe desprivilegiada (ou desfavorecida), pela letra $d$. Os exemplos utilizados na definição das métricas foram adaptados de [Rodrigues 2023] e utilizam a base de dados *Heart Disease*, composta por registros médicos de pacientes coletados na *Cleveland Clinic Foundation*, com o objetivo de prever a presença de doença cardíaca [Janosi and Detrano 1989]. O conjunto possui 14 atributos, sendo que, nos exemplos, o atributo protegido analisado é **sex**, no qual feminino (valor 0) e masculino (valor 1) são considerados, respectivamente, como classes desfavorecida e favorecida pelo viés.

**Tabela 4.5. Resumo de métricas para identificação de viés em aprendizado de máquina.**

| Categoria | Métrica | Descrição |
|---|---|---|
| Pré-treino | Class Imbalance (CI) | Mede o desbalanceamento na distribuição de classes ou atributos protegidos no conjunto de dados. |
| | Kullback-Leibler Divergence (KL) | Avalia a divergência entre distribuições de saída para diferentes grupos do atributo protegido. |
| | Kolmogorov-Smirnov (KS) | Mede a diferença máxima entre distribuições de probabilidade entre grupos protegidos. |
| | Conditional Demographic Disparity (CDDL) | Quantifica a disparidade condicional nos rótulos, considerando um atributo adicional para estratificação. |
| Pós-treino | Difference in Positive Proportions (DPPL) | Compara a proporção de saídas positivas entre classes do atributo protegido. |
| | Disparate Impact (DI) | Razão entre as proporções de saídas favoráveis entre os grupos; usada amplamente em auditorias. |
| | Difference in Conditional Outcome (DCO) | Compara as saídas preditas com as observadas, avaliando equilíbrio entre os grupos. |
| | Difference in Conditional Acceptance (DCA) | Mede a diferença condicional na taxa de aceitação entre grupos protegidos. |
| | Difference in Conditional Rejection (DCR) | Mede a diferença condicional na taxa de rejeição entre grupos protegidos. |
| | Recall Difference (RD) | Compara a taxa de verdadeiros positivos entre os grupos; relevante em diagnósticos. |
| | Difference in Acceptance Rates (DAR) | Avalia a diferença nas taxas de predições positivas corretas entre os grupos. |
| | Difference in Rejection Rates (DRR) | Avalia a diferença nas taxas de predições negativas corretas entre os grupos. |

### 4.6.1. Métricas pré-treino para análise de viés

**Class Imbalance (CI)**, ou desbalanceamento de classes, pode aparecer quando um atributo tem pouca representação para uma classe ou categoria específica. Estendendo o conceito clássico de balanceamento de classes, esta métrica também é aplicada no contexto de atributos protegidos. A Equação 1 apresenta o cálculo da métrica, em que $n_p$ e $n_d$ correspondem, respectivamente, ao número de amostras da classe $p$ (privilegiada) e da classe $d$ (desprivilegiada). O valor da métrica varia entre $-1$ e 1: valores positivos indicam predominância da classe privilegiada, enquanto valores negativos indicam predominância da classe desprivilegiada. Um valor igual a 1 ($-1$) indica que todas as amostras pertencem exclusivamente à classe privilegiada (desprivilegiada). O valor ideal é próximo de 0, refletindo uma distribuição equilibrada entre os grupos.

$$CI = (n_p - n_d)/(n_p + n_d) \tag{1}$$

Por exemplo, considerando uma base de dados com 100 instâncias, das quais 80 correspondem a pacientes do sexo masculino e 20 ao sexo feminino, um modelo treinado a partir desses dados pode atribuir importância ao atributo sexo em seu processo decisório e tornar-se mais propenso a cometer erros para a classe feminina, devido à menor exposição a exemplos dessa categoria. Aplicando a fórmula de CI, obtemos o valor de 0.6. Esse resultado indica um possível desbalanceamento em favor da classe favorecida (neste caso, o sexo masculino).

**Kullback-Leibler Divergence (KL Divergence)**, ou divergência de Kullback-Leibler, também conhecida na literatura como entropia relativa, é uma medida assimétrica que quantifica a divergência entre duas distribuições de probabilidade. É importante destacar que a divergência KL não deve ser interpretada estritamente como uma métrica de distância, pois não é simétrica – em geral, $KL(P_p \parallel P_d) \neq KL(P_d \parallel P_p)$. A fórmula da divergência KL é apresentada na Equação 2, onde $P_x(y)$ representa a distribuição de probabilidade observada na faceta $x$, dado o valor $y$ do atributo $Y$. Para problemas de classificação binária, essa distribuição é calculada como a proporção de amostras na classe $x$ com saída $z$, em relação ao total de amostras da classe $x$, considerando todos os possíveis valores de $Y$ (atributo alvo da predição). O valor dessa métrica varia de 0 a infinito, com valores próximos de 0 representando que as saídas são distribuídas de maneira similar, e valores positivos significando uma divergência entre os atributos de saída – quanto maior o valor, maior a divergência. Vale ressaltar que o cálculo da divergência KL não está restrito a saídas binárias, podendo abranger múltiplos valores de $y$, o que amplia o número de termos na equação (no caso binário, são apenas dois).

$$KL(P_p||P_d) = \sum_{Y} P_p(y) * \log[P_p(y)/P_d(y)] \tag{2}$$

Considerando o exemplo do conjunto de dados para predição de doença cardíaca, em um cenário onde as instâncias com saída positiva para a doença, respectivamente para homens e mulheres, são 20 e 70, e as instâncias com saída negativa são 80 e 30, calculamos as probabilidades conforme a Equação 2 e obtemos $KL = 0.8 * \log(0.8/0.3) + 0.2 * \log(0.2/0.7)$. O valor da métrica resulta em 0.53, indicando a divergência entre as

distribuições do atributo predito, e um viés positivo para a classe privilegiada.

**Kolmogorov-Smirnov (KS)** é um teste estatístico não paramétrico utilizado para avaliar a compatibilidade entre duas amostras. Para a análise de vieses, flexibilizamos a definição da métrica, aplicando-a na tarefa de identificar o rótulo *mais desbalanceado* em um conjunto de dados. Utilizando a fórmula na Equação 3, calculamos a divergência máxima das probabilidades para todas as possíveis saídas do modelo (no caso de uma classificação binária, calculamos $P_x(0)$ e $P_x(1)$ para as diferentes classes $x$ do atributo protegido). Essa métrica, assim como a divergência KL, pode ser aplicada a cenários onde o atributo predito não é binário, aumentando o número de termos na equação. O valor da métrica varia entre 0 e 1, sendo que 0 indica distribuição igualitária entre as classes, valores positivos indicam desbalanceamento em alguma classe (não indicando qual delas seria "privilegiada") e o valor 1 indica que todas as amostras pertencem a uma única classe.

$$KS = max(|P_p(y) - P_d(y)|) \tag{3}$$

Considerando o conjunto de dados para doença cardíaca, repetindo o exemplo anterior, onde as instâncias com saída positiva para a doença são, respectivamente, 20 para homens e 70 para mulheres, e as instâncias com saída negativa são 80 para homens e 30 para mulheres, temos: $KS = max(|0.2 - 0.7|, |0.8 - 0.3|)$. O valor final da métrica é 0.5, indicando um possível desbalanceamento para alguma das classes.

**Conditional Demographic Disparity in Labels (CDDL)**, ou disparidade demográfica condicional nos rótulos, mede a disparidade nas saídas entre duas classes (as classes do atributo protegido), considerando também a disparidade em subgrupos por meio de um atributo adicional do conjunto de dados utilizado como variável *correlacionada* para estratificação. Nessa métrica, introduzimos o cálculo da disparidade demográfica (*DD*), que representa a taxa com que uma classe específica apresenta determinado resultado (positivo ou negativo); dizemos que existe disparidade quando há diferença entre essas taxas. A fórmula para o cálculo da métrica está na Equação 4, onde $DD_i$ é definido conforme a Equação 5. Nas fórmulas, $n$ representa o número total de amostras, e $i$ são as diferentes saídas para os atributos correlacionados. O valor dessa métrica varia entre -1 e 1, sendo que 1 indica ausência de saídas negativas na classe privilegiada ou subgrupo e ausência de saídas positivas na classe desprivilegiada ou subgrupo; valores positivos indicam disparidade demográfica, pois a classe desprivilegiada, ou subgrupo, apresenta mais saídas desfavoráveis do que a classe privilegiada. Valores negativos indicam que a classe desprivilegiada, ou subgrupo, possui mais saídas favoráveis do que a classe privilegiada (o que, dependendo do contexto, pode ser o objetivo da aplicação de *fairness*); já -1 indica ausência de instâncias com saída desfavorável na classe desprivilegiada ou subgrupo e ausência de instâncias com saída favorável na classe privilegiada ou subgrupo. A definição de saídas favoráveis e desfavoráveis é sensível ao contexto do problema. A interpretação mais simples desta métrica é que valores diferentes de 0 indicam viés, enquanto valores próximos de 0 indicam ausência dele.

$$CDDL = \frac{1}{n} * \sum_i n_i * DD_i \tag{4}$$

$$DD_i = \frac{n_d^{(0)}}{n^{(0)}} - \frac{n_d^{(1)}}{n^{(1)}} \tag{5}$$

Para exemplificar essa métrica, considere um conjunto de dados com 20 instâncias, igualmente divididas entre 10 mulheres e 10 homens. Cinco instâncias de cada grupo têm mais de 20 anos (sendo a idade o atributo correlacionado utilizado para estratificação). Para mulheres acima de 20 anos, 4 possuem diagnóstico de doença cardíaca e 1 não possui; para mulheres abaixo de 20 anos, os valores são, respectivamente, 2 e 3. Para homens acima de 20 anos, são 3 instâncias com diagnóstico de doença cardíaca e 2 sem, e para homens abaixo de 20 anos, respectivamente, 1 e 4. Substituindo esses valores na fórmula, obtemos a Equação 6 (note que, para o problema de detecção de doença cardíaca, a saída "favorável"é a ausência de doença, representada por 0). O valor final da métrica, $0,23$, indica a existência de disparidade, pois a classe desprivilegiada (mulheres) apresenta mais saídas desfavoráveis (diagnóstico positivo) que a classe privilegiada, em ambos os subgrupos.

$$CCDL = \frac{1}{20} * \left[ 10 * \left( \frac{4}{(4+3)} - \frac{1}{(1+2)} \right) + 10 * \left( \frac{2}{(2+1)} - \frac{3}{(3+4)} \right) \right] \tag{6}$$

### 4.6.2. Métricas pós-treino para análise de viés

As três primeiras métricas de pós-treino apresentadas nesta seção focam em avaliar e quantificar diferentes taxas a partir das predições do modelo, considerando a existência de uma saída "favorável" (ou positiva) e "desfavorável" (ou negativa). Em contextos gerais, como em um modelo que auxilia na decisão de conceder ou não um empréstimo financeiro, a definição dessas saídas é geralmente direta. Entretanto, no contexto de modelos aplicados à área da saúde, onde o objetivo pode ser o prognóstico de uma doença, a definição do que constitui uma "saída favorável" nem sempre é clara. Para fins de ilustração, nestas métricas, consideraremos as saídas negativas para a doença como favoráveis, e as saídas positivas (presença da doença) como desfavoráveis. Ressalta-se, contudo, que a aplicação dessas métricas requer uma interpretação cuidadosa do contexto específico, e que os resultados podem variar conforme a definição adotada para saídas favoráveis ou desfavoráveis. Mesmo assim, essas métricas permanecem úteis para identificar e quantificar vieses nos dados.

**Difference in positive proportions in predicted labels (DPPL)**, ou diferença na proporção de positivos para os rótulos preditos, é definida como a diferença direta entre as predições positivas (com a "saída favorável", geralmente 1) entre as diferentes classes do atributo protegido. A fórmula é apresentada na Equação 7, onde calculamos a distribuição dos rótulos a partir da divisão do número de entradas com saída favorável para uma determinada classe, representado por $\hat{n}^{(1)}$, pelo número de registros naquela classe, $n$. Essa métrica produz valores entre -1 e 1 no caso de classificação binária, em que valores positivos indicam que a classe privilegiada apresenta maior proporção de saídas positivas, enquanto valores negativos indicam que a classe desprivilegiada possui maior proporção. Valores próximos a zero indicam proporções similares entre as classes.

Embora seja usualmente aplicada em problemas de classificação binária, essa métrica também pode ser estendida para saídas contínuas ou multiclasse.

$$\hat{q}_p = \frac{\hat{n}_p^{(1)}}{n_p} \qquad \hat{q}_d = \frac{\hat{n}_d^{(1)}}{n_d} \qquad DPPL = \hat{q}_p - \hat{q}_d \tag{7}$$

Supondo a tarefa de predição de doença cardíaca em um cenário onde homens e mulheres têm, respectivamente, 30% e 50% de saídas positivas indicando doença cardíaca (0.7 e 0.5 de $\hat{q}_p$ e $\hat{q}_d$, respectivamente), teríamos um valor de 0.2 para DPPL, indicando um leve desbalanceamento favorecendo a classe privilegiada.

**Disparate Impact (DI)**, ou impacto díspar (também referido como impacto discrepante), mede a razão entre as proporções das previsões do modelo para as diferentes classes do atributo protegido. A fórmula utilizada para o cálculo está apresentada na Equação 8, onde os valores de $\hat{q}_p$ e $\hat{q}_d$ são calculados conforme a fórmula da Equação 7. Essa métrica varia de 0 a $\infty$, sendo que valores menores que 1 indicam que a classe privilegiada possui maior proporção de saídas favoráveis em relação à classe desprivilegiada, enquanto valores maiores que 1 indicam que a classe desprivilegiada tem maior proporção de saídas favoráveis.

$$DI = \frac{\hat{q}_d}{\hat{q}_p} \tag{8}$$

Aplicando no exemplo de predição de doença cardíaca, supondo que as taxas de doença para homens e mulheres são, respectivamente, 30% e 50% (ou seja $\hat{q}_p = 0.7$ e $\hat{q}_d = 0.5$), obtemos um valor de DI aproximado de 0.7, indicando que a classe privilegiada, homens, apresenta uma maior proporção de saídas favoráveis em comparação à classe desprivilegiada, mulheres.

**Difference in Conditional Outcome (DCO)**, denominada diferença na saída condicional, compara os rótulos observados nos dados com os rótulos preditos pelo modelo, e compara se o balanceamento da variável alvo da predição é a mesma nas diferentes classes do atributo protegido. A motivação dessa métrica é que a simples análise da proporção de saídas favoráveis para cada classe pode não capturar nuances importantes para a interpretação dos resultados. Por exemplo, considere um conjunto de dados para concessão de empréstimos, com 100 instâncias do sexo masculino e 50 do sexo feminino. Suponha que o modelo tenha aprovado empréstimos para 60 homens e 30 mulheres, ou seja, 60% de aprovação em ambas as classes. Segundo a métrica DPPL, o modelo seria considerado "livre de viés". Contudo, o número absoluto de aprovações para homens é 33% maior, indicando um desbalanceamento favorável à classe privilegiada. A partir da definição da diferença condicional no rótulo, derivam-se duas métricas distintas: ***Difference in Conditional Acceptance (DCA)*** e ***Difference in Conditional Rejection (DCR)***, que diferem na definição de saída favorável e desfavorável. As fórmulas para DCA e DCR estão nas Equações 9 e 10, respectivamente, onde $\hat{n}_a^{(x)}$ representa o número de instâncias da classe $a$ preditas com valor $x$, e $n_a^{(x)}$ é o número de instâncias da classe $a$ com o valor alvo da predição $x$. Essa métrica varia de $-\infty$ a $+\infty$, onde valores positivos indicam desbalanceamento para a classe privilegiada, valores próximos de zero indicam proporções similares

entre os diferentes grupos, e valores negativos indicam desbalanceamento para a classe desprivilegiada.

$$c_p = \frac{n_p^{(1)}}{\hat{n}_p^{(1)}} \qquad c_d = \frac{n_d^{(1)}}{\hat{n}_d^{(1)}} \qquad DCA = c_p - c_d \tag{9}$$

$$r_p = \frac{n_p^{(0)}}{\hat{n}_p^{(0)}} \qquad r_d = \frac{n_d^{(0)}}{\hat{n}_d^{(0)}} \qquad DCR = r_d - r_p \tag{10}$$

Considerando o exemplo anterior, a taxa predita de doença cardíaca pelo modelo para homens e mulheres é, respectivamente, 30% e 50% (representadas por $\hat{n}_p^{(0)}$ e $\hat{n}_d^{(0)}$), o que implica que os valores de $\hat{n}_p^{(1)}$ e $\hat{n}_d^{(1)}$ são, respectivamente, 70% e 50%. Supondo que os dados estejam igualmente divididos, os valores de $n_p^{(0)}$, $n_d^{(0)}$, $n_p^{(1)}$ e $n_d^{(1)}$ são todos iguais a 100. Aplicando as fórmulas das Equações 9 e 10, obtemos a Equação 11, na qual o valor final da métrica DCA é $-0.57$ e da métrica DCR é $-1.3$. Esses resultados indicam um desbalanceamento em favor da classe desprivilegiada, pois os dados mostram que a saída favorável (diagnóstico negativo para doença cardíaca) ocorre com maior frequência na classe privilegiada.

$$c_p = \frac{100}{70} \qquad c_d = \frac{100}{50} \qquad r_p = \frac{100}{30} \qquad r_d = \frac{100}{50} \tag{11}$$

**Recall Difference (RD)**, ou diferença na revocação, é especialmente importante quando o objetivo do modelo é auxiliar em um diagnóstico, pois avalia com que frequência o modelo prevê uma saída positiva para casos em que essa saída é esperada (verdadeiros positivos). A revocação ideal ocorre quando o modelo é capaz de identificar corretamente todos os casos positivos. Calculamos a diferença de revocação pela fórmula apresentada na Equação 12, onde $VP_x$ e $FN_x$ representam, respectivamente, o número de verdadeiros positivos e falsos negativos para a classe $x$. O valor dessa métrica varia entre -1 e 1, sendo que valores positivos indicam maior taxa de revocação para a classe privilegiada, valores próximos a zero indicam taxas semelhantes entre as classes, e valores negativos indicam maior revocação para a classe desprivilegiada. Nota-se que tanto valores positivos quanto negativos podem ser considerados formas de viés, já que o modelo está mais propenso a encontrar os verdadeiros positivos para uma classe específica do conjunto de dados.

$$RD = \frac{VP_a}{VP_p + FN_p} - \frac{VP_d}{VP_d + FN_d d} \tag{12}$$

Para ilustrar a diferença de revocação, suponha que um modelo tenha taxas de verdadeiros positivos de 33 e 26 para homens e mulheres, respectivamente, e taxas de falsos negativos de 10 e 1. A fórmula aplicada a esses valores está na Equação 13. O valor calculado para RD é $-0.19$, indicando que a classe desprivilegiada apresenta uma taxa de revocação maior que a classe privilegiada. Isso sugere que o modelo identifica melhor os verdadeiros positivos para a classe $d$.

$$RD = \frac{33}{33+10} - \frac{26}{26+1} \tag{13}$$

**Difference in Label Rates (DLR)**, ou diferença nas taxas dos rótulos, parte da ideia de que rótulos podem ser positivos ou negativos, e que a diferença nas taxas das predições positivas ou negativas pode evidenciar viés. A partir dessa métrica, derivamos duas fórmulas: ***Difference in Acceptance Rates (DAR)***, que mede se as predições do modelo que deveriam ser verdadeiras são preditas corretamente e ***Difference in Rejection Rates (DRR)***, que mede se as taxas de diagnósticos negativos são preditas corretamente. Em contextos de saúde, onde falsos negativos podem ser mais prejudiciais que falsos positivos, essas métricas são fundamentais para analisar se o modelo "aceita" e "rejeita" de forma balanceada entre as classes do atributo protegido. A fórmula para DAR e DRR está na Equação 14, onde $VP$ significa verdadeiros positivos, $VN$ verdadeiros negativos, $FP$ falsos positivos e $FN$ falsos negativos. O valor dessas métricas varia entre -1 e 1, sendo que valores ideais estão próximos a zero. Para DAR, valores positivos indicam possível viés contra a classe desprivilegiada, indicando mais falsos positivos nessa classe; valores negativos indicam viés favorecendo a classe desprivilegiada, com mais falsos positivos na classe privilegiada. Para DRR, valores positivos indicam viés favorecendo a classe desprivilegiada, devido a mais falsos negativos na classe privilegiada, enquanto valores negativos indicam viés favorecendo a classe privilegiada, com menos falsos positivos.

$$DAR = \frac{VP_p}{VP_p + FP_p} - \frac{VP_d}{VP_d + FP_d} \qquad DRR = \frac{VN_d}{VN_d + FN_d} - \frac{VN_p}{TN_p + FN_p} \tag{14}$$

Para ilustrar essas métricas, considerando-se o conjunto de dados para predição de doença cardíaca, o resultado do modelo gera as matrizes da Figura 4.5, onde temos as predições para a classe privilegiada e para a classe desprivilegiada. Considerando a fórmula para cálculo de DAR e DRR, temos a Equação 15. O valor final de DAR, de $-0.21$ indica um possível viés favorecendo a classe privilegiada, observável pelo número maior de falsos positivos para essa classe, concordando com o valor final de DRR, de 0.09, indicando a presença de viés pelo número maior de falsos negativos para a classe privilegiada.

$$DAR = \frac{33}{33+9} - \frac{26}{26+0} \qquad DRR = \frac{7}{7+1} - \frac{36}{36+10} \tag{15}$$

## 4.7. Estratégias para mitigação de viés

Nesta seção, apresentaremos as principais estratégias para mitigação de viés conhecidas na literatura, agrupadas pela etapa de desenvolvimento de modelos em que são aplicadas, conforme sumarizado na Tabela 4.6. Em suma, três classes de estratégias serão abordadas [Mehrabi et al. 2021]: (i) estratégias de *pre-processing* (também denominadas de estratégias baseadas em dados), que incluem técnicas de amostragem de dados e otimização de hiperparâmetros anteriores ao treinamento dos modelos; (ii) estratégias *in-processing* (também denominadas de estratégias baseadas em algoritmos), que envolvem formas de considerar a justiça algorítmica durante o treinamento dos modelos; e (iii) estratégias

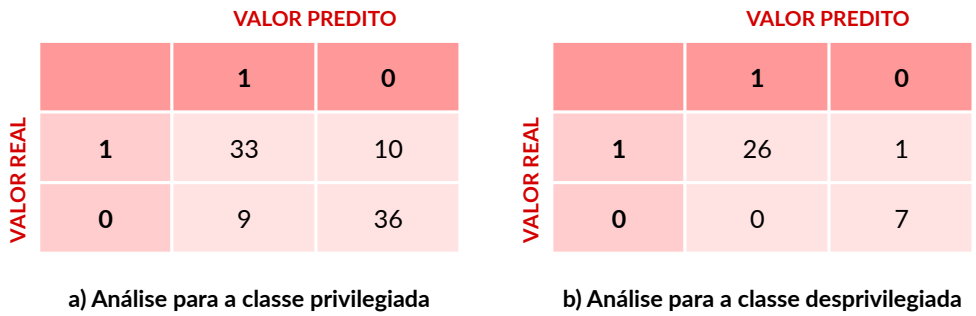



**Figura 4.5. Matrizes de confusão para o problema de detecção de doença cardíaca, com análise para a classe (a) privilegiada e (b) desprivilegiada.**

*post-processing* (também denominadas de estratégias baseadas em pós-treinamento), englobando métodos para adicionar critérios de justiça aos modelos após o treinamento.

### 4.7.1. Métodos de pré-processamento

Os métodos de *pre-processing* (ou pré-processamento) são técnicas aplicadas antes da etapa de treinamento de modelos de aprendizado de máquina (AM) [Wan et al. 2022]. A motivação para seu uso decorre do fato de que vieses frequentemente estão presentes nos próprios dados utilizados no treinamento, seja por problemas de distribuição, desbalanceamento de classes ou falta de representatividade de determinados grupos. A principal

**Tabela 4.6. Resumo de métodos de mitigação de viés em aprendizado de máquina**

| Categoria | Método | Descrição |
|---|---|---|
| Pré-processamento | Reweighting | Pesos são atribuídos a subgrupos para priorizar instâncias e reduzir o viés. |
| | Data Massaging | Altera a variável resposta no conjunto de dados. |
| | Disparate Impact Remover | Realiza uma perturbação nos dados, alterando a distribuição das variáveis. |
| Em processamento | Discrimination-Aware Tree Construction | Alteração de critérios de divisão de nós para contemplar penalização por ganho de informação do atributo sensível. |
| | GridSearch Reduction | Realiza uma busca em grade de penalizações de justiça. |
| | Exponentiated Gradient Reduction | Atualiza iterativamente os pesos das restrições de justiça com base em quão fortemente elas são violadas nas previsões do modelo. |
| | Prejudice Remover | Utiliza um termo de regularização em modelos probabilísticos. |
| | Adversarial Debiasing | Usa adversário para remover informação do atributo sensível. |
| Pós-processamento | Equalized Odds Postprocessing | Equilibra as taxas de verdadeiro e falso positivo entre os grupos sensíveis. |
| | Reject Option Classification | Reclassifica instâncias com baixa confiança favorecendo o grupo sensível desfavorecido. |
| | Leaf Relabeling | Altera o rótulo de folhas de árvore de decisão. |

vantagem desses métodos é a facilidade de implementação em pipelines de dados, pois funcionam como etapas adicionais que podem ser inseridas antes do treinamento, sem necessidade de alterar o modelo propriamente dito. Entretanto, uma limitação é que a modificação dos dados originais pode acarretar perda de informações relevantes. Além disso, o pré-processamento por si só não garante a eliminação completa dos vieses, já que o modelo ainda pode ser afetado por outras fontes de viés durante o treinamento e avaliação. Os métodos descritos a seguir baseiam-se em referências da área, como [Tawakuli and Engel 2024, Wan et al. 2022].

**Reweighting** [Calders et al. 2009] é um método que atribui pesos às instâncias do conjunto de dados, utilizados durante o treinamento para mitigar vieses presentes na distribuição. Os pesos são definidos com base na combinação do atributo sensível e da variável resposta. Dessa forma, todas as instâncias pertencentes ao mesmo grupo – isto é, que compartilham o mesmo valor de atributo sensível e rótulo – recebem o mesmo peso. Uma variação proposta em [Li and Liu 2022] individualiza os pesos para cada instância, permitindo um ajuste mais fino e sensível ao contexto local dos dados.

**Disparate Impact Remover**[Feldman et al. 2015] propõe uma perturbação controlada nas variáveis preditoras do conjunto de dados, com o objetivo de reduzir a possibilidade de inferir a variável sensível a partir desses atributos. Essa modificação preserva a distribuição dos dados dentro de cada grupo sensível, mantendo a ordem relativa entre as instâncias, e busca preservar a capacidade preditiva em relação à variável resposta. O ajuste é realizado com base na posição de cada instância dentro da distribuição do seu grupo sensível. Para cada quantil, calcula-se a mediana (ou média) dos valores correspondentes entre os grupos, e o novo valor da instância é interpolado entre o valor original e o valor reparado. Na prática, o método promove uma aproximação entre as distribuições dos grupos protegido e não protegido para cada atributo considerado. Uma limitação importante é que a técnica é restrita a variáveis ordenáveis, sendo inadequada para atributos categóricos não ordenados.

**Data Massaging** [Kamiran and Calders 2009] é um método de pré-processamento que altera a variável resposta (*label*) de forma controlada, visando tornar as predições do modelo mais justas para grupos sensíveis. A ideia central é corrigir vieses nos dados de treinamento ajustando rótulos de algumas instâncias pontuais, especialmente aquelas próximas à fronteira de decisão. Para isso, utiliza-se um classificador auxiliar – originalmente um Naive Bayes – treinado sobre o conjunto de dados original, que estima a probabilidade de cada instância pertencer à classe positiva. Com base nessas probabilidades, instâncias negativas do grupo desfavorecido são ordenadas de forma decrescente (da maior para a menor probabilidade de serem positivas), enquanto instâncias positivas do grupo favorecido são ordenadas de forma crescente (da menor para a maior probabilidade). O algoritmo seleciona pares de instâncias (uma de cada grupo) e troca seus rótulos, promovendo a instância desfavorecida à classe positiva e rebaixando a favorecida para a classe negativa. Essa troca ocorre até que uma medida de justiça desejada seja atingida, como uma taxa de aprovação equilibrada entre os grupos.

Por alterar apenas exemplos próximos à fronteira, o *Data Massaging* reduz o viés nos dados com impacto mínimo na acurácia geral, desde que o número de alterações seja rigorosamente controlado. Do ponto de vista ético, a alteração de dados reais pode ser

controversa, tornando a aplicação do método delicada em certos contextos. Outro aspecto relevante é a dependência da performance do modelo auxiliar, cuja baixa qualidade pode comprometer os resultados. Além disso, é necessário que o atributo sensível esteja explicitamente presente no conjunto de dados, já que o método depende dessa informação para identificar grupos favorecidos e desfavorecidos.

Uma variação chamada *Local Massaging* [Zliobaite et al. 2011] utiliza métodos auxiliares para identificar regiões do conjunto de dados com maior evidência de viés. Apenas essas regiões selecionadas passam por alterações nos rótulos, evitando mudanças generalizadas e preservando melhor a estrutura global dos dados. De forma similar, [Luong et al. 2011] propuseram uma metodologia de alteração da variável resposta utilizando o algoritmo *k-Nearest Neighbors* (kNN).

### 4.7.2. Métodos em processamento

Os métodos *in-processing*, ou “em processamento", atuam diretamente na forma como o modelo de AM é treinado, buscando impedir que o processo de aprendizado incorpore vieses discriminatórios. Em vez de modificar os dados, essas técnicas interferem no ajuste dos parâmetros do modelo durante o treinamento. As principais estratégias incluem alterações na função de custo, imposição de restrições e modificações nas funções de otimização, visando reduzir a dependência do modelo em relação aos atributos sensíveis.

Uma das vantagens centrais desses métodos é que não requerem alterações nos dados originais, evitando potenciais perdas de informação associadas a técnicas de pré-processamento. Além disso, por dispensarem uma etapa adicional de preparação dos dados, são especialmente vantajosos em cenários com grandes volumes de dados, nos quais o pré-processamento pode ser oneroso. Outra vantagem é a maior capacidade de lidar com padrões discriminatórios complexos, como relações não lineares e interações entre múltiplos atributos sensíveis, que podem ser difíceis de capturar por abordagens mais simples. Esses métodos também oferecem maior flexibilidade para controlar o *trade-off* entre desempenho preditivo e mitigação de vieses, permitindo ajustar o modelo para equilibrar precisão e equidade entre grupos sensíveis.

Por outro lado, os métodos *in-processing* frequentemente demandam modificações na implementação do algoritmo de AM, o que pode limitar sua aplicabilidade a certos tipos de modelos — por exemplo, técnicas que alteram a função de perda costumam ser restritas a algoritmos probabilísticos, como regressão logística e máquinas de vetores de suporte (SVM). Consequentemente, não são facilmente integráveis a qualquer pipeline de aprendizado, ao contrário dos métodos de pré-processamento, que são independentes do modelo. Além disso, embora algumas bibliotecas, como o AIF360 [Bellamy et al. 2018], já ofereçam técnicas *in-processing* prontas, o conjunto de opções disponíveis ainda é limitado. Isso dificulta a adoção desses métodos por usuários que dependem de ferramentas populares como o Scikit-learn [Pedregosa et al. 2011], que não incorporam essas modificações de forma nativa.

De forma geral, os métodos *in-processing* podem ser classificados em abordagens explícitas e implícitas [Wan et al. 2022]. As explícitas atuam diretamente sobre a função objetivo do treinamento, incorporando métricas de equidade via restrições ou termos de regularização. Já as abordagens implícitas visam produzir representações latentes me-

nos enviesadas, reduzindo a correlação entre atributos sensíveis e resultados do modelo. Existem também abordagens híbridas que combinam esses mecanismos. É importante destacar que não há consenso na literatura sobre a categorização dessas estratégias, e diversos autores propõem subdivisões adicionais ou critérios alternativos [Hort et al. 2023], refletindo a complexidade e diversidade do campo.

**Prejudice Remover** [Kamishima et al. 2012] é um exemplo representativo, que integra um componente de penalização chamado *prejudice index* (índice de preconceito) na função de custo do modelo, com o objetivo de reduzir a correlação entre atributos sensíveis e a variável resposta. Dessa forma, o modelo produz resultados mais justos, minimizando vieses implícitos presentes nos dados ou nas previsões. A função de perda ajustada fica conforme a equação (16).

$$\text{Loss} = \text{Loss}_{\text{model}} + \eta \cdot \text{Loss}_{\text{prejudice}} \tag{16}$$

Onde $Loss_{\text{model}}$ é a função de custo tradicional, que mede o erro do modelo, $Loss_{\text{prejudice}}$ é o termo de penalização que mede o índice de preconceito, $\eta$ é o parâmetro de regularização que controla o grau de penalização aplicado, permitindo ajustar o *trade-off* entre a precisão do modelo e a mitigação do preconceito. Essa técnica é aplicável a modelos probabilísticos. Atualmente, no AIF360, é implementada em Regressão Logística, utilizando os parâmetros padrão da biblioteca Scikit-learn [Pedregosa et al. 2011].

A principal vantagem do *Prejudice Remover* está na facilidade de integração em pipelines existentes, especialmente quando se utiliza modelos simples como a regressão logística. Além disso, o parâmetro $\eta$ permite controlar de maneira intuitiva o *trade-off* entre a precisão preditiva e a mitigação de vieses. Por outro lado, a simplicidade do algoritmo, embora vantajosa em termos de facilidade de implementação, pode limitar sua eficácia em cenários onde padrões mais complexos de vieses. Outra desvantagem é a necessidade de identificar explicitamente a variável sensível, assumindo-se que o desenvolvedor tem conhecimento dessa variável e que ela está presente no conjunto de dados, o que pode não ser sempre o caso em cenários mais complexos. Além disso, essa técnica não lida adequadamente com intersecções de viés, como no caso de grupos que combinam múltiplos atributos sensíveis (por exemplo, o grupo de gênero “mulher"e raça “preta").

**Grid Search Reduction** e **Exponentiated Gradient Reduction** são métodos que buscam incorporar restrições de *fairness* diretamente no processo de treinamento, otimizando classificadores para equilibrar a precisão preditiva e a equidade [Agarwal et al. 2018]. O *Grid Search Reduction* é um método que seleciona um classificador mais justo ao otimizar a penalização associada às restrições de *fairness* impostas durante o treinamento do modelo. Isso é alcançado por meio da introdução de multiplicadores de Lagrange, que ajustam o peso dado a cada restrição de justiça, transformando o problema original em uma sequência de problemas de classificação custo-sensível, que pode ser definido pela equação (17).

$$L(Q,\lambda) = \widehat{\text{err}}(Q) + \lambda^{\top}\left(M\mu(Q) - \hat{c}\right) \tag{17}$$

Para isso, é realizada uma busca em *grid* sobre possíveis valores dos multiplicado-

res, que controlam a severidade com que cada restrição de justiça é tratada no treinamento. Ou seja, na prática, o método permite analisar diferentes pontos do *trade-off* entre equidade e desempenho preditivo, escolhendo o modelo que melhor se ajusta aos critérios definidos. O *Grid Search Reduction* é aplicável a uma ampla variedade de algoritmos de aprendizado de máquina, incluindo modelos *black-box*, desde que aceitem pesos de entrada para problemas custo-sensíveis.

Esse método retorna um classificador determinístico – ou seja, uma única função preditiva final – o que facilita sua aplicação prática. Por outro lado, pode ser computacionalmente custoso, especialmente quando a grade de valores de $\lambda$ é extensa. Além disso, devido à natureza discreta e heurística da busca em grade, não há garantia de que o ponto ótimo global será encontrado, já que o espaço contínuo de soluções é explorado parcialmente. O método requer o conhecimento e a presença explícita da variável sensível, mas permite a inclusão simultânea de múltiplas restrições de justiça.

O *Exponentiated Gradient Reduction* também realiza a otimização dos multiplicadores $\lambda$ em um problema custo-sensível, mas em vez de uma busca exaustiva, utiliza o algoritmo de gradiente exponenciado (*Exponentiated Gradient*) para atualizar iterativamente os valores de $\lambda$ com base nas violações das restrições de *fairness*. Essa abordagem constrói uma distribuição sobre classificadores, treinando cada classificador com um vetor diferente de custos derivados de $\lambda$. A solução final é obtida como uma média ponderada (ou amostragem) desses classificadores.

A principal vantagem do *Exponentiated Gradient Reduction* é a existência de garantias teóricas de convergência para uma solução próxima do ótimo, tanto em termos de erro quanto de equidade. Embora mais complexo, esse método é especialmente indicado quando se deseja controlar múltiplas métricas de justiça simultaneamente ou quando a busca determinística se torna inviável. Por fim, considerando que o objetivo é desenvolver um classificador para sistemas sensíveis a atributos protegidos, em alguns domínios, a adoção de um classificador aleatorizado pode não ser apropriada, devido a exigências de interpretabilidade, auditabilidade ou reprodutibilidade.

**Discrimination-Aware Tree Construction** [Kamiran et al. 2010] é uma das diversas abordagens desenvolvidas para incorporar critérios de *fairness* em algoritmos baseados em árvores de decisão (*Fair Decision Trees*). O método aplica o conceito de ganho de informação não apenas em relação à variável de resposta, mas também a um atributo sensível previamente definido. O ganho de informação com respeito ao atributo sensível, denotado como IGS, é calculado pela equação.

$$IGS = H_B - \sum_{i=1}^{k} \frac{|D_i|}{|D|} H_B(D_i) \tag{18}$$

Onde $H_B$ representa a entropia associada ao atributo sensível e $D_1, \ldots, D_k$: correspondem às partições do conjunto de dados geradas pela divisão em um nó da árvore.

A partir do cálculo do IGS, existem três estratégias para integrá-lo ao IGC (ganho de informação relacionado à variável de resposta):

- IGC – IGS: penaliza os ganhos de acurácia, ao subtrair o ganho de informação do

atributo sensível do ganho da variável de resposta.

- IGC/IGS: expressa o *trade-off* entre acurácia e discriminação, por meio da razão entre ambos os ganhos.

- IGC+IGS: soma direta dos ganhos. Isoladamente, tende a aumentar a discriminação; entretanto, pode ser combinada com a técnica de Leaf Relabeling (também proposta por [Kamiran et al. 2010], e discutida na seção de post-processing) para aprimorar a justiça do modelo.

Resultados experimentais indicam que, isoladamente, o método *Discrimination-Aware Tree Construction* não melhora significativamente métricas de *fairness* e pode acarretar perdas na acurácia. Contudo, a combinação do critério IGC + IGS com *Leaf Relabeling* (um método de pós-processamento) apresenta desempenho mais equilibrado, mitigando discriminação com impacto moderado na performance do modelo. Apesar da necessidade de acesso ao atributo sensível, trata-se de um método de implementação simples, com baixa complexidade computacional e compatível com algoritmos de árvore – o que garante interpretabilidade e transparência, características desejáveis em contextos sensíveis. Como limitação, destaca-se o fato de se basear em um algoritmo *greedy*, o que impede a obtenção de soluções ótimas globais.

**Adversarial Debiasing** [Zhang et al. 2018] é um método baseado no treinamento de redes neurais adversariais [Goodfellow et al. 2014]. Nesse método, existe um modelo preditor, cujo objetivo é prever o rótulo ($\hat{y}$, ou variável resposta), enquanto o modelo adversário tenta prever a variável sensível $Z$. Assim o treinamento fica baseado em um problema "min-max", podendo ser definido pela equação (19).

$$\min_{\theta} \max_{\phi} E_{(X,Y,Z)} \left[ \text{Loss}(f_\theta(X), Y) - \lambda \cdot \text{Loss}(g_\phi(f_\theta(X)), Z) \right] \tag{19}$$

Onde $\theta$ são os parâmetros do modelo preditor, $\phi$ são os parâmetros do modelo adversário, e $\lambda$ é um parâmetro de regularização que controla o *trade-off* entre minimizar a perda preditiva e maximizar a capacidade do adversário de prever $Z$. Durante o treinamento do modelo adversário, os parâmetros $\theta$ e $\phi$ são ajustados por meio de gradientes. O gradiente do adversário é usado para atualizar os parâmetros do preditor, de modo a minimizar a quantidade de informação transmitida sobre $Z$.

A principal vantagem do *Adversarial Debiasing* é sua aplicabilidade a uma ampla variedade de modelos baseados em gradientes, como regressão logística, redes neurais e máquinas de vetores de suporte (SVMs). Além disso, o método permite a utilização de atributos sensíveis contínuos (como idade), além dos discretos. Ele também é flexível quanto às definições de justiça, podendo ser adaptado para *Demographic Parity*, *Equalized Odds* e *Equal Opportunity*.

Entretanto, como outros métodos adversariais, o *Adversarial Debiasing* pode apresentar instabilidade no treinamento, podendo convergir para um ótimo local. Por isso, é necessário ajustar cuidadosamente os hiperparâmetros para balancear adequadamente as contribuições do adversário e do modelo principal. Além disso, assim como em outras

abordagens, o conhecimento prévio da variável sensível e sua presença explícita no conjunto de dados continuam sendo limitações importantes para a aplicação desse método.

### 4.7.3. Métodos de pós-processamento

Métodos de mitigação de vieses do tipo *post-processing*, ou pós-processamento, em português, têm como objetivo atuar após o treinamento do modelo, buscando mitigar vieses presentes nas predições. Esse tipo de abordagem pode tratar o classificador como uma caixa-preta, o que facilita sua integração em diferentes pipelines de tomada de decisão. De forma geral, esses métodos englobam qualquer tipo de intervenção que não envolva alterações no processo de treinamento, incluindo: modificações nos dados de teste (*input correction*), ajustes nas saídas do modelo (*output correction*), e correções aplicadas diretamente ao classificador já treinado (*classifier correction*), como a adaptação de limiares de decisão, estrutura e regras internas.

O método **Equalized Odds Postprocessing** [Hardt et al. 2016] tem como objetivo alterar a saída do modelo, se baseando na métrica *Equalized Odds*, que determina *fairness* como mesma probabilidade entre grupo protegido e desprotegido de obter um resultado positivo do modelo. Isso é feito derivando um segundo classificador derivado do primeiro, sendo na prática uma função probabilística do resultado original e atributo sensível.

Para cada grupo de atributo sensível, existe uma faixa de valores possíveis de taxas de verdadeiros positivos (*true positive rate*, TPR) e taxa de falsos positivos (*false positive rate*, FPR) que podem ser alcançados a partir das previsões originais do modelo, utilizando transformações como: manter a saída original, inverter a saída, prever sempre positivo ou sempre negativo. Essas combinações definem uma região viável (um quadrilátero) para cada grupo. A função de pós-processamento final é encontrada por meio de programação linear, buscando um par de TPR e FPR comum entre os grupos (atendendo ao critério de *Equalized Odds*) e que minimize a perda – isto é, a diferença entre os rótulos verdadeiros e as previsões transformadas. Assim, a saída do modelo é uma função aleatorizada com a probabilidade do resultado ser positivo ou negativo, dado o grupo sensível pertencente e o rótulo original.

Existe ainda uma variação de método, *Calibrated Equalized Odds*, que tem a mesma premissa básica de alterar a saída do modelo utilizando uma função probabilística. Entretanto, também tem como premissa não alterar a calibragem do modelo. Uma vantagem desse método está na simplicidade de implementação e na garantia formal de *fairness*, conforme definido pelo critério de *Equalized Odds*. Por outro lado, ele está limitado a classificadores binários e levanta questões éticas e de interpretabilidade, por envolver modificações nas previsões originais do modelo.

O método de **Leaf Relabeling** [Kamiran et al. 2010] parte de uma árvore de decisão previamente treinada e modifica o rótulo de algumas folhas com o objetivo de aumentar *fairness*, minimizando o impacto negativo sobre a acurácia. Para cada folha, são calculados o impacto na discriminação ($\Delta disc$) e a variação na acurácia ($\Delta acc$) resultantes da troca do rótulo. O *trade-off* entre esses dois critérios é expresso pela razão da equação (20).

$$\frac{\Delta disc}{|\Delta acc|} \tag{20}$$

O algoritmo opera de forma *greedy*, ordenando as folhas com base nesse índice, da maior para a menor razão. A substituição dos rótulos é feita sequencialmente, respeitando um limite pré-definido de perda de acurácia global. Considerando uma folha inicialmente rotulada como positiva, o impacto da troca para o rótulo negativo pode ser expresso pela equação (21).

$$\Delta acc_l = n - p \tag{21}$$

Em que $n$ denota o número de instâncias na folha com classe real negativa e $p$ o número de instâncias na folha com classe real positiva. Nesse cenário, a troca de rótulo melhora a acurácia se houver mais instâncias negativas do que positivas. Para folhas inicialmente negativas, a equação se adapta para $p - n$, seguindo o mesmo raciocínio. O impacto na discriminação é orientado pelo critério de *demographic parity* (conforme equação (22)), segundo o qual todos os grupos definidos por um atributo sensível $B$ devem ter igual probabilidade de receber predição positiva.

$$P(\hat{Y} = 1 \mid B = 1) - P(\hat{Y} = 1 \mid B = 0) \tag{22}$$

Dessa forma, o impacto na discriminação causado por uma folha que deixa de ser positiva é dado pela equação (23).

$$\Delta disc_l = \frac{s_l}{n_s} - \frac{r_l}{n_r} \tag{23}$$

Onde $n_s$ é o número de instâncias no conjunto de dados com $B = 1$; $n_r$ é o número de instâncias no conjunto de dados com $B = 0$; $s_l$ é o número de instâncias na folha com $B = 1$; $s_r$ é o número de instâncias na folha com $B = 0$. Essa equação mede a diferença entre as taxas de predição positiva para os dois grupos. Se $\Delta disc_l > 0$, o grupo $B = 1$ está sendo favorecido na folha; se $\Delta disc_l < 0$, o grupo $B = 0$ recebe mais classificações positivas. Quando o *relabeling* altera o rótulo de uma folha de negativo para positivo, a equação assume os mesmos termos, porém com sinal invertido, uma vez que os exemplos passam a contribuir para a taxa de predição positiva em cada grupo.

A combinação do método de *Leaf Relabeling* com o critério de divisão IGC + IGS, proposto na abordagem de *Discrimination-Aware Tree Construction*, demonstrou desempenho superior. Embora a soma direta dos ganhos de informação tenda, isoladamente, a aumentar a discriminação, sua interação com o *Leaf Relabeling* favorece a formação de folhas mais homogêneas tanto em termos de acurácia quanto de *fairness*. Isso ocorre porque as divisões tendem a agrupar instâncias mais similares, reduzindo a quantidade de folhas ambíguas que necessitam de correção posterior – o que beneficia algoritmos gananciosos ao reduzir o espaço de decisão e tornar o processo de *relabeling* mais eficiente.

Uma das principais vantagens do *relabeling* é sua aplicabilidade em modelos legados, permitindo a mitigação de viés sem a necessidade de reestruturar o modelo original. Por outro lado, seu principal risco está na limitação de generalização: como as decisões de troca de rótulo são feitas com base apenas nos dados de treinamento da própria folha, o método pode superajustar-se, especialmente na ausência de validação externa.

**Reject Option Classification** (ROC) [Kamiran et al. 2012] é um método que atua no estágio de pós-processamento, utilizando as probabilidades preditas por um classificador para reajustar as decisões em casos de incerteza. Considerando que o resultado desejado é o rótulo positivo, define-se uma região crítica composta por instâncias cuja confiança na classificação está abaixo de um limiar $\theta$.

$$max(P(C^+|X), 1 - P(C^+|X)) < \theta \tag{24}$$

Onde $P(C^+|X$ é a probabilidade da instância X ser classificada como pertencente à classe positiva e $\theta$ é um parâmetro que define o limiar de confiança. Ou seja, a zona de incerteza é composta por instâncias cuja probabilidade de pertencer a qualquer das classes (positiva ou negativa) é próxima de 0,5, indicando baixa confiança na predição. Para essas instâncias, a decisão original do classificador é substituída com base no grupo ao qual a instância pertence, de acordo com o atributo sensível considerado. Se $X$ pertence ao grupo desfavorecido, a instância é rotulada como positiva; se $X$ pertence ao grupo favorecido, é rotulada como negativa. Uma das principais vantagens do ROC é sua facilidade de integração a modelos já treinados, sem a necessidade de modificar o classificador ou os dados. Além disso, oferece ainda um mecanismo simples de controle do *trade-off* entre acurácia e justiça por meio da escolha do parâmetro $\theta$.

## 4.8. Aspectos éticos e legais

A crescente adoção de tecnologias baseadas em inteligência artificial na área da saúde tem trazido oportunidades significativas para aprimorar diagnósticos, tratamentos, gestão de sistemas e equidade no acesso aos cuidados. Contudo, esses avanços também impõem desafios éticos, legais e sociais que devem ser enfrentados com responsabilidade e visão crítica [Rajkomar et al. 2018]. Como alerta Tedros Adhanom Ghebreyesus, diretor-geral da Organização Mundial da Saúde (OMS), "*Como toda nova tecnologia, a inteligência artificial possui um enorme potencial para melhorar a saúde de milhões de pessoas em todo o mundo, mas, como toda tecnologia, também pode ser mal utilizada e causar danos*" [World Health Organization 2021]. A governança ética da IA em saúde busca justamente garantir que essas tecnologias sejam desenvolvidas e aplicadas com responsabilidade, promovendo o bem-estar coletivo e respeitando os direitos humanos.

A OMS estabeleceu seis princípios éticos fundamentais para orientar o uso responsável da IA na saúde: (1) proteção da autonomia humana; (2) promoção do bem-estar, da segurança humana e do interesse público; (3) transparência, explicabilidade e inteligibilidade; (4) responsabilidade e prestação de contas; (5) inclusão e equidade; e (6) promoção de uma IA responsiva e sustentável [World Health Organization 2021]. Esses princípios devem nortear o desenvolvimento de tecnologias que respondam às reais necessidades dos sistemas de saúde e das populações atendidas, com atenção especial aos grupos historicamente marginalizados.

A construção ética de sistemas de IA requer o envolvimento ativo de múltiplos *stakeholders* — incluindo pesquisadores, desenvolvedores, profissionais da saúde, pacientes, gestores, reguladores e representantes da sociedade civil. Esse envolvimento é essencial para assegurar legitimidade, justiça procedimental e melhores resultados técnicos e sociais [Floridi et al. 2018]. Além disso, papéis e responsabilidades devem estar claramente

definidos ao longo de todo o ciclo de vida da tecnologia, desde a concepção até sua aplicação clínica. A ausência de diretrizes claras pode levar à diluição de responsabilidades e dificultar a responsabilização em casos de dano. Outro ponto central é o desenvolvimento responsável da IA, que deve considerar não apenas eficácia e eficiência, mas também os impactos éticos e institucionais sobre os sistemas de saúde e os grupos vulneráveis. A avaliação prévia de riscos éticos, legais e sociais deve ser uma etapa obrigatória nos processos de inovação tecnológica, conforme recomendam frameworks como o "AI Ethics Impact Assessment" e as "Ethics Guidelines for Trustworthy AI" da Comissão Europeia [Jobin et al. 2019].

Vale destacar que mesmo sistemas construídos com boas práticas podem apresentar desvios comportamentais ao serem expostos a novos contextos, dados corrompidos ou populações distintas daquelas utilizadas no treinamento. Por isso, a governança da IA deve incluir mecanismos robustos de monitoramento contínuo, capazes de identificar falhas, vieses ou efeitos adversos ao longo do tempo. A OMS recomenda a realização de auditorias técnicas e avaliações regulares de impacto social após a implementação dos sistemas, promovendo a atualização constante dos modelos com base em dados reais e evidências científicas emergentes [World Health Organization 2021]. Esse acompanhamento deve incluir indicadores de desempenho ético, como equidade no acesso, ausência de discriminação, explicabilidade das decisões e nível de confiança dos usuários – especialmente profissionais da saúde e pacientes [Chen et al. 2021]. Tais mecanismos de retroalimentação são essenciais para assegurar que a IA permaneça sensível às necessidades dinâmicas dos sistemas de saúde e contribua para a redução – e não para o agravamento – das desigualdades em saúde.

Embora um arcabouço legal para regulamentar o uso ético e responsável da IA (de forma geral e na saúde) ainda esteja em desenvolvimento em muitos países, organizações internacionais têm incentivado a criação de estruturas legais e institucionais capazes de acompanhar a complexidade e a velocidade dos avanços tecnológicos [Jobin et al. 2019]. O Brasil tem avançado nesse sentido, com destaque para a aprovação do Projeto de Lei nº 2338/2023 pelo Senado Federal em dezembro de 2024. Esse projeto estabelece princípios fundamentais como a centralidade da pessoa humana, a proteção de direitos fundamentais e a promoção da transparência e da responsabilidade no desenvolvimento e aplicação de sistemas de IA. O texto também classifica os sistemas de IA de acordo com o nível de risco que representam, prevendo restrições para usos considerados de risco excessivo, como aqueles que exploram vulnerabilidades humanas ou disseminam conteúdos prejudiciais. A construção de soluções éticas, justas e sustentáveis dependerá da colaboração ativa entre governos, instituições de pesquisa, empresas privadas e sociedade civil organizada, de forma a garantir que os benefícios da inteligência artificial sejam distribuídos de maneira equitativa e socialmente responsável.

## 4.9. Ferramentas

Esta seção apresenta ferramentas que implementam as métricas, métodos e demais técnicas previamente discutidas. Para ilustrar seu uso, são fornecidos trechos de código aplicados ao conjunto de dados *Heart Disease* [Janosi and Detrano 1989]. Esse conjunto de dados contém 14 atributos, incluindo variáveis sensíveis como *age* e *sex*, sendo esta última selecionada para análise nos exemplos. A variável alvo é multiclasse, corres-

pondendo a diferentes diagnósticos de doença cardíaca, mas pode ser mapeada para um domínio binário, indicando se o paciente é saudável ou se possui a presença de alguma doença cardíaca. O conjunto de dados é composto por 303 instâncias, apresentando um desequilíbrio significativo em relação à variável *sex*, enquanto a variável *target* apresenta uma distribuição relativamente balanceada. No entanto, ao considerar a segmentação por sexo, observa-se que a variável *target* é extremamente desbalanceada no grupo feminino, o que evidencia uma potencial fonte de viés. Essa distribuição está ilustrada na Figura 4.6.

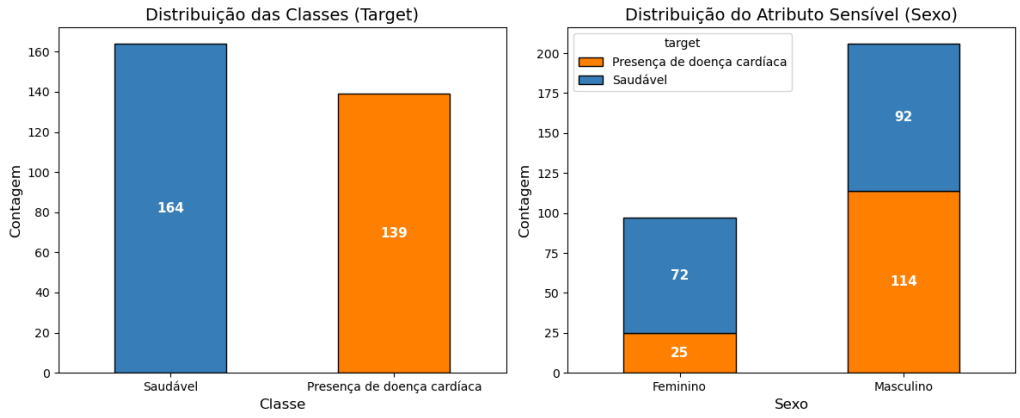



**Figura 4.6. Balanceamento da variável *target* (binária) e do atributo sensível *sex* (Fonte: Os autores)**

### 4.9.1. FairML

A ferramenta FairML[3], apresentada em [Adebayo 2016], tem como objetivo analisar modelos preditivos por meio da quantificação da dependência do modelo em relação às suas variáveis de entrada. Para isso, utiliza quatro métodos de ranqueamento de atributos: o algoritmo de projeção ortogonal iterativa (*Iterative Orthogonal Feature Projection* - IOFP), o critério de mínima redundância e máxima relevância (*minimum Redundancy, Maximum Relevance* - mRMR), o algoritmo de regressão LASSO (*Least Absolute Shrinkage and Selection Operator*) e o método de florestas aleatórias (*Random Forest*) para seleção de atributos. O resultado gerado pela ferramenta indica a significância relativa de cada variável no processo de decisão do modelo, permitindo identificar quais atributos são mais influentes ou importantes no processo de tomada de decisão. Caso um atributo sensível (ou protegido) figure entre os mais relevantes, isso pode indicar que o modelo está tomando decisões potencialmente injustas. FairML está disponível como uma biblioteca em Python. Um exemplo de sua aplicação no conjunto de dados para predição de doença cardíaca é apresentado no Algoritmo 4.1, e a execução do código gera o gráfico exibido na Figura 4.7.

```
1 import matplotlib.pyplot as plt
2 import pandas as pd
3 from fairml import audit_model, plot_dependencies
4 from sklearn.linear_model import LogisticRegression
5
```

[3]https://github.com/adebayoj/fairml

```
data = pd.read_csv("Heart_Disease_Prediction.csv")

y = data.target.values
data = data.drop("target", axis=1)
x = data.values

clf = LogisticRegression(penalty='l2', C=0.01)
clf.fit(x, y)

importancies, _ = audit_model(clf.predict, data)

fig = plot_dependencies(
    importancies.median(),
    reverse_values=False,
    title="FairML feature dependence logistic regression model")
```

**Algoritmo 4.1. Exemplo de uso da ferramenta FairML**

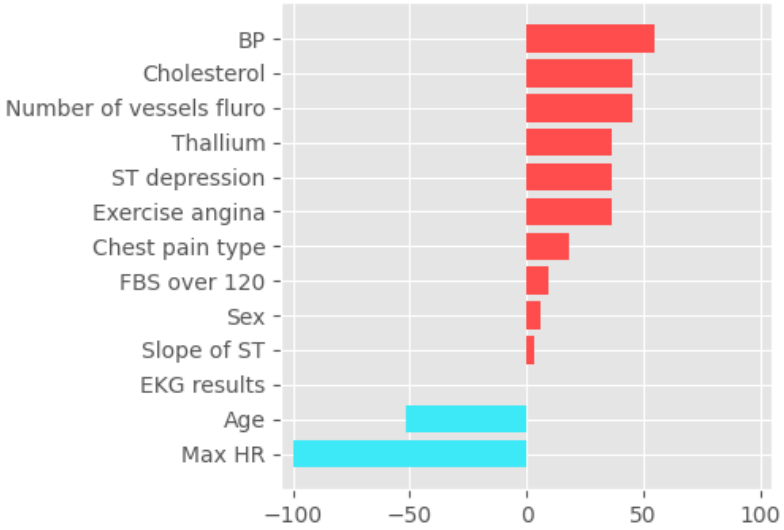



**Figura 4.7. Importância dos atributos para conjunto de dados de predição de doença cardíaca utilizando uma regressão logística e a ferramenta FairML. (Fonte: Os autores)**

### 4.9.2. FairnessMeasures

A ferramenta FairnessMeasures[4] implementa diversas métricas de *fairness*, conforme descritas em [Zehlike et al. 2017]. Trata-se de uma ferramenta open-source desenvolvida em Python, embora não esteja disponível como uma biblioteca tradicional. Para utilizá-la com novos conjuntos de dados, é necessário importar manualmente o código e adaptar os dados a um formato específico exigido pela aplicação. Um trecho de código utilizando essa ferramenta está disponível no Algoritmo 4.2. As métricas calculadas são, respectivamente: **Paridade Estatística**, que mede a distribuição de atributos protegidos e não protegidos nos resultados positivos, **Diferença Média**, que avalia a diferença média nos resultados entre os grupos dos atributos protegidos, **Diferença Normalizada**, que analisa

[4]https://fairnessmeasures.github.io/

a disparidade nos resultados, e normaliza em relação a um grupo, e **Proporção de Impacto** que avalia a razão de resultados positivos entre os grupos do atributo protegido, e **Razão de Chances**, que avalia as chances relativas de resultados positivos para os diferentes grupos.

```
from fairnessmeasures.src.measures.absolute_measures import *
from fairnessmeasures.src.measures.statistical_tests import *
from fairnessmeasures.src.data_structure.dataset import Dataset

data = pd.read_csv("Heart_Disease_Prediction.csv")
data.target = data.target.map({'Presence': 1, 'Absence': 0})
data.rename(columns={
    'Sex': 'protected_sex'}, inplace=True)

dataset = Dataset(data)

print(f"difference of means test: {t_test_ind(dataset, 'target', '
    protected_sex')}")
print(f"mean differences: {mean_difference(dataset, 'target', '
    protected_sex').T}")
print(f"normalized differences: {normalized_difference(dataset, 'target
    ', 'protected_sex')}")
print(f"impact ratio: {impact_ratio(dataset, 'target', 'protected_sex')
    }")
print(f"odds ratio: {fisher_exact_two_groups(dataset, 'target', '
    protected_sex')}")
```

**Algoritmo 4.2. Exemplo de uso da ferramenta FairnessMeasures**

### 4.9.3. Amazon Sagemaker

O *Amazon SageMaker*[5] é uma plataforma gerenciada da AWS que oferece suporte ao desenvolvimento, treinamento e implantação de modelos de AM. A análise de vieses nos dados é realizada por meio da ferramenta *SageMaker Clarify*, que disponibiliza diversas métricas para a preparação, validação e avaliação dos dados utilizados no treinamento dos modelos. As métricas oferecidas são organizadas em três categorias: pré-treinamento, durante o treinamento e pós-treinamento. Elas integram o pipeline de desenvolvimento da plataforma, que contempla todo o ciclo de vida dos modelos.

### 4.9.4. AIF360

O AIF360 (AI Fairness 360) [Bellamy et al. 2018] é uma biblioteca open-source desenvolvida pela IBM que oferece um conjunto robusto de ferramentas para mensuração e mitigação de vieses algorítmicos. Um de seus principais diferenciais é a integração facilitada com o Scikit-Learn [Pedregosa et al. 2011], permitindo a aplicação de seus recursos diretamente em pipelines de aprendizado de máquina com diferentes estimadores (ou seja, algoritmos). Apresentamos a seguir uma demonstração prática utilizando dados relacionados a doenças cardiovasculares, abordando desde o pré-processamento até a avaliação de métricas de fairness, incluindo também a aplicação de uma técnica de mitigação.

[5]https://aws.amazon.com/pt/sagemaker/

- **Pré-processamento**: Inicialmente, o conjunto de dados é carregado. Em seguida, são removidas entradas com valores nulos, e o rótulo de saída (originalmente multiclasse) é transformado em binário, indicando a presença ou ausência de doença.

```
    heart_disease = fetch_ucirepo(id=45)

    df = pd.concat([heart_disease.data.features, heart_disease.
data.targets], axis=1).dropna()
    df['target'] = (df.iloc[:, -1] > 0).astype(int)
```

**Algoritmo 4.3. Carregamento e pré-processamento do dataset**

- **Dataset AIF360**: O AIF360 utiliza uma estrutura de dados própria para classificação chamada BinaryLabelDataset, que organiza variáveis e rótulos de forma compatível com os métodos da biblioteca. Essa classe também oferece recursos como o split, facilitando a divisão do conjunto de dados para treino e teste.

```
    dataset = BinaryLabelDataset(
        df=df,
        label_names=["target"],
        protected_attribute_names=["sex"],
        favorable_label=1,
        unfavorable_label=0
    )
    dataset_train, dataset_test = dataset.split([0.7], shuffle=
True, seed=42)
```

**Algoritmo 4.4. Criação do dataset da estrutura do AIF360**

- **Modelo baseline**: Com variáveis preditoras normalizadas, é treinado um modelo de regressão logística simples como baseline. As predições geradas são então utilizadas para calcular métricas de *fairness*.

```
    model = LogisticRegression(random_state=42)
    model.fit(dataset_train.features, dataset_train.labels.ravel()
)
    y_pred = model.predict(dataset_test.features)

    dataset_test_pred = dataset_test.copy()
    dataset_test_pred.labels = y_pred
```

**Algoritmo 4.5. Criação do modelo baseline**

- **Cáculo de métricas**: A classe "*ClassificationMetric*"permite calcular de forma prática métricas de justiça algorítmica, como *Disparate Impact* e *Equal Opportunity Difference*, além de métricas tradicionais como acurácia.

```
    metric_test = ClassificationMetric(
    dataset_test, dataset_test_pred,
    unprivileged_groups=[{"sex": 0}],
    privileged_groups=[{"sex": 1}]
    )
    print(f"Disparate Impact: {metric_test.disparate_impact():.3f}
")
```

```
    print(f"Equal Opportunity Difference: {metric_test.
equal_opportunity_difference():.3f}")
    print(f"Accuracy: {metric_test.accuracy():.3f}")
```

**Algoritmo 4.6. Cálculo de métricas**

- **Mitigação de vieses**: Como exemplo de mitigação, será utilizado o algoritmo Prejudice Remover, disponível no AIF360. Após o ajuste, as mesmas métricas podem ser utilizadas para comparar os resultados com o modelo baseline.

```
    pr_model = PrejudiceRemover(sensitive_attr="sex", class_attr="
target", eta=25.0)
    pr_model.fit(dataset_train)
    dataset_test_pred_pr = pr_model.predict(dataset_test)
```

**Algoritmo 4.7. Utilização do Prejudice Remover para mitigar o viés**

A Tabela 4.7 apresenta os resultados das métricas de *fairness* e acurácia, comparando o modelo *baseline* com o modelo após mitigação de vieses. Observa-se uma melhora no *Disparate Impact*, embora com uma redução significativa na acurácia.

**Tabela 4.7. Métricas de Fairness e Acurácia: Baseline vs. Prejudice Remover**

| Métrica | Baseline | Prejudice Remover |
|---|---|---|
| Disparate Impact | 0.214 | 0.266 |
| Equal Opportunity Difference | -0.333 | -0.433 |
| Acurácia | 0.922 | 0.900 |

## 4.10. Considerações finais, desafios e perspectivas

Neste capítulo, discutimos a presença persistente de vieses em modelos preditivos aplicados à área da saúde, bem como as principais estratégias conhecidas para sua detecção e mitigação. Embora haja avanços significativos, a remoção completa dos efeitos históricos, políticos e culturais presentes nos dados ainda constitui um desafio considerável, pois esses vieses são profundamente enraizados nas estruturas sociais e institucionais. A relevância da análise de vieses em AM é especialmente crítica em domínios sensíveis como a saúde, onde decisões automatizadas impactam diretamente vidas humanas. Espera-se que a leitura deste capítulo tenha esclarecido a importância do tema e estimulado reflexões sobre a construção de sistemas mais justos, inclusivos e socialmente responsáveis.

Como perspectivas, destaca-se a necessidade de desenvolvimento de métodos capazes de lidar com a complexidade e interseccionalidade dos vieses presentes nos dados reais. A integração de técnicas avançadas de interpretabilidade e explicabilidade é fundamental para aumentar a transparência dos modelos, especialmente os do tipo “caixa-preta", permitindo aos usuários compreender não apenas os resultados, mas também os processos decisórios subjacentes, o que é crucial para mitigar vieses de interpretação humana. Além disso, o estabelecimento e a adoção de normas regulatórias rigorosas, aliadas a debates éticos consistentes, são essenciais para garantir a responsabilização dos desenvolvedores e orientar o uso adequado dessas tecnologias. A colaboração interdisciplinar

entre pesquisadores, profissionais da saúde, reguladores e a sociedade civil será decisiva para superar as lacunas atuais e fomentar uma IA que atenda aos princípios de justiça e equidade. Este campo de pesquisa permanece dinâmico e desafiador, mas seu impacto no uso ético da IA na saúde promete transformar positivamente práticas, políticas e resultados para populações diversas e vulneráveis.

## Agradecimentos

Agradecemos o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, e da Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), por meio dos projetos 21/2551-0002052-0 (Projeto MARCS) e 22/2551-0000390-7 (Projeto CIARS), às pesquisas do grupo que fundamentaram a construção do conhecimento consolidado neste capítulo. M. Recamonde-Mendoza é parcialmente financiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), por meio de Bolsa de Produtividade em Pesquisa – PQ2 [308075/2021-8].

## Referências

[Adebayo 2016] Adebayo, J. (2016). Fairml: Toolbox for diagnosing bias in predictive modeling.

[Agarwal et al. 2018] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, pages 60–69. PMLR.

[Aldwean and Tenney 2023] Aldwean, A. and Tenney, D. (2023). Artificial intelligence in healthcare sector: a literature review of the adoption challenges. *Open Journal of Business and Management*, 12(1):129–147.

[Alowais et al. 2023] Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., Yami, M. S. A., Harbi, S. A., and Albekairy, A. M. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*, 23(1):689.

[Badawy et al. 2023] Badawy, M., Ramadan, N., and Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: a survey. *Journal of Electrical Systems and Information Technology*, 10(1):40.

[Barocas et al. 2023] Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT press.

[Bazzan et al. 2023] Bazzan, A. L., Tavares, A. R., Pereira, A. G., Jung, C. R., Scharcanski, J., Carbonera, J. L., Lamb, L. C., Recamonde-Mendoza, M., da Silveira, T. L., and Moreira, V. (2023). “A nova eletricidade": Aplicações, riscos e tendências da IA moderna – “The new electricity": Applications, risks, and trends in current AI. *arXiv preprint arXiv:2310.18324*.

[Bellamy et al. 2018] Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.

[Borgese et al. 2022] Borgese, M., Joyce, C., Anderson, E. E., Churpek, M. M., and Afshar, M. (2022). Bias assessment and correction in machine learning algorithms: a use-case in a natural language processing algorithm to identify hospitalized patients with unhealthy alcohol use. In *AMIA Annual Symposium Proceedings*, volume 2021, page 247.

[Buolamwini and Gebru 2018] Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

[Burkov 2020] Burkov, A. (2020). *Machine Learning Engineering*. True Positive Inc.

[Burlina et al. 2021] Burlina, P., Joshi, N., Paul, W., Pacheco, K. D., and Bressler, N. M. (2021). Addressing artificial intelligence bias in retinal diagnostics. *Translational Vision Science & Technology*, 10(2):13–13.

[Calders et al. 2009] Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, page 13–18, USA. IEEE Computer Society.

[Caton and Haas 2024] Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38.

[Chen et al. 2021] Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., and Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science Annu. Rev. Biomed. Data Sci*, 2021:123–144.

[Colacci et al. 2024] Colacci, M., Huang, Y. Q., Postill, G., Zhelnov, P., Fennelly, O., Verma, A., Straus, S., and Tricco, A. C. (2024). Sociodemographic bias in clinical machine learning models: a scoping review of algorithmic bias instances and mechanisms. *Journal of Clinical Epidemiology*, page 111606.

[Coots et al. 2025] Coots, M., Linn, K. A., Goel, S., Navathe, A. S., and Parikh, R. B. (2025). Racial bias in clinical and population health algorithms: a critical review of current debates. *Annual Review of Public Health*, 46.

[Daneshjou et al. 2022] Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S. M., Bailey, E. E., Gevaert, O., et al. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147.

[Díaz et al. 2018] Díaz, M., Johnson, I., Lazar, A., Piper, A. M., and Gergle, D. (2018). Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

[Duede et al. 2024] Duede, E., Dolan, W., Bauer, A., Foster, I., and Lakhani, K. (2024). Oil & water? Diffusion of AI within and across scientific fields. *arXiv preprint arXiv:2405.15828*.

[Estiri et al. 2022] Estiri, H., Strasser, Z. H., Rashidian, S., Klann, J. G., Wagholikar, K. B., McCoy Jr, T. H., and Murphy, S. N. (2022). An objective framework for evaluating unrecognized bias in medical AI models predicting COVID-19 outcomes. *Journal of the American Medical Informatics Association*, 29(8):1334–1341.

[Faceli et al. 2021] Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. d., and Carvalho, A. C. P. d. L. F. d. (2021). *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC.

[Fatumo et al. 2022] Fatumo, S., Chikowore, T., Choudhury, A., Ayub, M., Martin, A. R., and Kuchenbaecker, K. (2022). A roadmap to increase diversity in genomic studies. *Nature Medicine*, 28(2):243–250.

[Feldman et al. 2015] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. KDD '15, page 259–268, New York, NY, USA. Association for Computing Machinery.

[Floridi et al. 2018] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Craglia, P., Dignum, M., Dignum, V., Lütge, C., Pagallo, R., Pasquale, F., et al. (2018). AI4People – an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707.

[Geffner 2018] Geffner, H. (2018). Model-free, model-based, and general intelligence. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI2018*.

[Goes and Nascimento 2013] Goes, E. F. and Nascimento, E. R. d. (2013). Mulheres negras e brancas e os níveis de acesso aos serviços preventivos de saúde: uma análise sobre as desigualdades. *Saúde em Debate*, 37:571–579.

[Goodfellow et al. 2014] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.

[Greenwood et al. 2020] Greenwood, B. N., Hardeman, R. R., Huang, L., and Sojourner, A. (2020). Physician–patient racial concordance and disparities in birthing mortality for newborns. *Proceedings of the National Academy of Sciences*, 117(35):21194–21200.

[Guerrero et al. 2018] Guerrero, S., López-Cortés, A., Indacochea, A., García-Cárdenas, J. M., Zambrano, A. K., Cabrera-Andrade, A., Guevara-Ramírez, P., González, D. A.,

Leone, P. E., and Paz-y Miño, C. (2018). Analysis of racial/ethnic representation in select basic and applied cancer research studies. *Scientific Reports*, 8(1):1–8.

[Haenlein and Kaplan 2019] Haenlein, M. and Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, 61(4):5–14.

[Hajkowicz et al. 2023] Hajkowicz, S., Sanderson, C., Karimi, S., Bratanova, A., and Naughtin, C. (2023). Artificial intelligence adoption in the physical sciences, natural sciences, life sciences, social sciences and the arts and humanities: A bibliometric analysis of research publications from 1960-2021. *Technology in Society*, 74:102260.

[Hardt et al. 2021] Hardt, M., Chen, X., Cheng, X., Donini, M., Gelman, J., Gollaprolu, S., He, J., Larroy, P., Liu, X., McCarthy, N., Rathi, A., Rees, S., Siva, A., Tsai, E., Vasist, K., Yilmaz, P., Zafar, M. B., Das, S., Haas, K., Hill, T., and Kenthapadi, K. (2021). Amazon sagemaker clarify: Machine learning bias detection and explainability in the cloud. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 2974–2983, New York, NY, USA. Association for Computing Machinery.

[Hardt et al. 2016] Hardt, M., Price, E., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

[Hasanzadeh et al. 2025] Hasanzadeh, F., Josephson, C. B., Waters, G., Adedinsewo, D., Azizi, Z., and White, J. A. (2025). Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *npj Digital Medicine*, 8(1):154.

[Hatoum et al. 2021] Hatoum, A. S., Wendt, F. R., Galimberti, M., Polimanti, R., Neale, B., Kranzler, H. R., Gelernter, J., Edenberg, H. J., and Agrawal, A. (2021). Ancestry may confound genetic machine learning: Candidate-gene prediction of opioid use disorder as an example. *Drug and Alcohol Dependence*, 229:109115.

[Hellström et al. 2020] Hellström, T., Dignum, V., and Bensch, S. (2020). Bias in machine learning – what is it good for? *arXiv preprint arXiv:2004.00686*.

[Hort et al. 2023] Hort, M., Chen, Z., Zhang, J. M., Harman, M., and Sarro, F. (2023). Bias mitigation for machine learning classifiers: A comprehensive survey.

[Huang et al. 2022] Huang, J., Galal, G., Etemadi, M., and Vaidyanathan, M. (2022). Evaluation and mitigation of racial bias in clinical machine learning models: scoping review. *JMIR Medical Informatics*, 10(5):e36388.

[Janosi and Detrano 1989] Janosi, Andras, S. W. P. M. and Detrano, R. (1989). Heart Disease. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C52P4X.

[Jobin et al. 2019] Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

[Joynt Maddox et al. 2019] Joynt Maddox, K. E., Reidhead, M., Hu, J., Kind, A. J., Zaslavsky, A. M., Nagasako, E. M., and Nerenz, D. R. (2019). Adjusting for social risk factors impacts performance and penalties in the hospital readmissions reduction program. *Health Services Research*, 54(2):327–336.

[Kamiran and Calders 2009] Kamiran, F. and Calders, T. (2009). Classifying without discriminating. In *Proceedings 2nd IEEE International Conference on Computer, Control and Communication (IC4 2009, Karachi, Pakistan, February 17-18, 2009)*, pages 1–6, United States. Institute of Electrical and Electronics Engineers.

[Kamiran et al. 2010] Kamiran, F., Calders, T., and Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pages 869–874.

[Kamiran et al. 2012] Kamiran, F., Karim, A., and Zhang, X. (2012). Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929.

[Kamishima et al. 2012] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Proceedings of the 2012th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, ECMLPKDD'12, page 35–50, Berlin, Heidelberg. Springer-Verlag.

[Kapoor and Narayanan 2023] Kapoor, S. and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9).

[Kosinski et al. 2013] Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

[Larrazabal et al. 2020] Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., and Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594.

[Leal et al. 2005] Leal, M. d. C., Gama, S. G. N. d., and Cunha, C. B. d. (2005). Desigualdades raciais, sociodemográficas e na assistência ao pré-natal e ao parto, 1999-2001. *Revista de saude publica*, 39:100–107.

[Li and Liu 2022] Li, P. and Liu, H. (2022). Achieving fairness at no utility cost via data reweighing with influence. In *International Conference on Machine Learning*, pages 12917–12930. PMLR.

[Lin et al. 2024] Lin, N., Paul, R., Guerra, S., Liu, Y., Doulgeris, J., Shi, M., Lin, M., Engeberg, E. D., Hashemi, J., and Vrionis, F. D. (2024). The frontiers of smart healthcare systems. In *Healthcare*, volume 12, page 2330. MDPI.

[Luong et al. 2011] Luong, B. T., Ruggieri, S., and Turini, F. (2011). k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 502–510, New York, NY, USA. Association for Computing Machinery.

[Martin et al. 2019] Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591.

[Mccarthy 1994] Mccarthy, C. R. (1994). Historical background of clinical trials involving women and minorities. *Academic Medicine*, 69(9):695–8.

[Meade et al. 2021] Meade, R., Camilleri, A., Geoghegan, R., Osorio, S., and Zou, Q. (2021). Bias in machine learning: how facial recognition models show signs of racism, sexism and ageism.

[Mehrabi et al. 2021] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

[Mitchell 1980] Mitchell, T. M. (1980). The need for biases in learning generalizations.

[Mohammed et al. 2025] Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., and Harmouch, H. (2025). The effects of data quality on machine learning performance on tabular data. *Information Systems*, 132:102549.

[Molnar 2025] Molnar, C. (2025). *Interpretable Machine Learning*. 3 edition.

[Obermeyer et al. 2019] Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

[O'Neil 2021] O'Neil, C. (2021). *Algoritmos de destruição em massa*. Editora Rua do Sabão.

[Papakyriakopoulos et al. 2020] Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., and Marco, F. (2020). Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 446–457.

[Parikh et al. 2019] Parikh, R. B., Teeple, S., and Navathe, A. S. (2019). Addressing bias in artificial intelligence in health care. *Jama*, 322(24):2377–2378.

[Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Rajkomar et al. 2018] Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., and Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12):866–872.

[Rajpurkar et al. 2022] Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. (2022). Ai in health and medicine. *Nature medicine*, 28(1):31–38.

[Rodrigues 2023] Rodrigues, D. D. (2023). Assessing pre-training bias in health data and estimating its impact on machine learning algorithms.

[Ruback et al. 2022] Ruback, L., Carvalho, D., and Avila, S. (2022). Mitigando vieses no aprendizado de máquina: Uma análise sociotécnica. *iSys-Brazilian Journal of Information Systems*, 15(1):23–1.

[Schwalbe and Wahl 2020] Schwalbe, N. and Wahl, B. (2020). Artificial intelligence and the future of global health. *The Lancet*, 395(10236):1579–1586.

[Silva 2022] Silva, T. (2022). *Racismo algorítmico: inteligência artificial e discriminação nas redes digitais*. Edições Sesc SP.

[Solans Noguero et al. 2023] Solans Noguero, D., Ramírez-Cifuentes, D., Ríssola, E. A., and Freire, A. (2023). Gender bias when using artificial intelligence to assess anorexia nervosa on social media: data-driven study. *Journal of Medical Internet Research*, 25:e45184.

[Stypinska 2023] Stypinska, J. (2023). AI ageism: a critical roadmap for studying age discrimination and exclusion in digitalized societies. *AI & society*, 38(2):665–677.

[Suresh and Guttag 2021] Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9.

[Tawakuli and Engel 2024] Tawakuli, A. and Engel, T. (2024). Make your data fair: A survey of data preprocessing techniques that address biases in data towards fair AI. *Journal of Engineering Research*.

[Wan et al. 2022] Wan, M., Zha, D., Liu, N., and Zou, N. (2022). In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 17.

[World Health Organization 2021] World Health Organization (2021). Ethics and governance of artificial intelligence for health. Accessado em 2025-05-07.

[Zehlike et al. 2017] Zehlike, M., Castillo, C., Bonchi, F., Baeza-Yates, R., Hajian, S., and Megahed", M. (2017). Fairness measures: A platform for data collection and benchmarking in discrimination-aware ml. `https://fairnessmeasures.github.io`.

[Zhang et al. 2018] Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

[Zliobaite et al. 2011] Zliobaite, I., Kamiran, F., and Calders, T. (2011). Handling conditional discrimination. pages 992–1001.

# Capítulo 5

# Biofeedback na avaliação da experiência do usuário em ambientes imersivos

Ingrid Winkler, Paulo E. Ambrósio, Regina M. C. Leite, André M. Cordeiro, Lucas G. G. Almeida, Yasmim Thasla, Alexandre G. Siqueira, Marcio F. Catapan, Luciana O. Berretta

***Abstract***

*This short course explores how the integration of biofeedback data can revolutionize the evaluation of User Experience in Virtual Reality environments. Essential concepts of VR, eye tracking and heart rate monitoring are explored, demonstrating how this information can be applied to gain valuable insights into the behavior and reactions of users in immersive experiences. Concepts and applications related to the health of participants in immersive experiences with a focus on professional training are also covered.*

***Resumo***

*Neste minicurso é abordado como a integração de dados de biofeedback pode revolucionar a avaliação da Experiência do Usuário em ambientes de Realidade Virtual. São explorados conceitos essenciais de RV, rastreamento ocular e monitoramento de frequência cardíaca, demonstrando como essas informações podem ser aplicadas para obter insights valiosos sobre o comportamento e as reações dos usuários em experiências imersivas. Também são abordados conceitos e aplicações relacionados à saúde dos participantes em experiências imersivas com foco em treinamento profissional.*

## 5.1. Introdução: Ambientes Imersivos e *Biofeedback*: Conceitos, Tecnologias e Aplicações

A Realidade Virtual (RV) representa uma das mais expressivas inovações tecnológicas das últimas décadas, permitindo a criação de ambientes tridimensionais imersivos que simulam situações reais ou imaginárias. Por meio de dispositivos como óculos, sensores e controladores, os usuários podem interagir com cenários digitais em tempo real,

explorando espaços e realizando tarefas com liberdade de movimento e sensação de presença.

Paralelamente, o *biofeedback* surge como uma abordagem que permite o monitoramento e a devolutiva de sinais fisiológicos à pessoa em tempo real, proporcionando maior consciência corporal e emocional. A integração entre RV e *biofeedback* tem potencial revolucionário em diversas áreas: educação, saúde, treinamentos industriais, entretenimento, reabilitação e pesquisa em interação humano-computador.

Este capítulo introduz os conceitos fundamentais de ambientes imersivos em Realidade Virtual e *Biofeedback*, com ênfase em três sensores amplamente utilizados: Rastreamento Ocular, Variabilidade da frequência cardíaca e Eletroencefalografia (EEG). Apresentam-se aplicações práticas, implicações técnicas e éticas, e exemplos de como essas tecnologias podem ser combinadas para criar experiências imersivas responsivas e adaptativas.

## 5.2. Fundamentos de ambientes imersivos e Realidade Virtual

Ambientes Virtuais Imersivos (AVI) são espaços simulados digitalmente que buscam replicar ou expandir a realidade percebida pelo usuário, promovendo sensações de presença, interatividade e engajamento por meio de estímulos sensoriais diversos.

Os AVIs são ambientes persistentes, podendo ser permanentes, gerados por computador. Neles, múltiplos usuários em diferentes locais físicos e remotos podem interagir em tempo real, seja para fins de trabalho, diversão ou entretenimento. Também conhecidos como mundos virtuais imersivos, mundos virtuais 3D ou metaverso, os AVIs fazem parte de um subconjunto de aplicações de realidade virtual.

A Realidade Virtual é definida como uma representação digital tridimensional gerada por computador, que permite ao usuário sentir-se presente e interagir com o ambiente simulado. Por meio de óculos de RV, sensores de movimento e fones de ouvido, cria-se uma imersão sensorial que simula presença física e engajamento cognitivo. Conhecida também como realidade ficcional, a RV não se limita à mera reprodução do real, mas oferece experiências imersivas em mundos possíveis, mesmo que completamente fictícios.

Para os seres humanos, a percepção da realidade combina informação sensorial com os mecanismos cerebrais que dão sentido a essa informação. Assim, ainda que um ambiente imersivo em Realidade Virtual seja artificial, ele pode ser interpretado como real, despertando emoções, prazer, aprendizado e entretenimento, respondendo às ações do usuário [Tori; Hounsell; Kirner, 2018].

Esse efeito de realismo é potencializado pela união de periféricos especializados, softwares avançados, computadores de alto desempenho e gráficos tridimensionais. Juntos, esses elementos permitem que objetos virtuais sejam sentidos e manipulados de maneira intuitiva [Cardoso; Lamounier, 2006].

A interação com o mundo virtual acontece quando o usuário explora, manipula e altera os elementos desse ambiente, utilizando seus sentidos e movimentos naturais do corpo, como gestos, olhares e comandos de voz [Tori; Kirner, 2006]. Essa familiaridade torna a experiência mais fluida e envolvente, sem a necessidade de aprender comandos complexos.

Os ambientes em Realidade Virtual podem ser utilizados para simular espaços reais (como salas de aula, usinas ou hospitais) ou imaginários (como mundos artísticos ou lúdicos). A Realidade Virtual é caracterizada pela interatividade, imersão e envolvimento ativo do usuário, sendo especialmente poderosa em situações de aprendizado prático, treinamento sob risco e terapias baseadas em exposição.

A tecnologia de RV que conhecemos hoje foi construída ao longo de décadas, impulsionada por pioneiros que desempenharam um papel crucial no desenvolvimento dessas inovações. Graças aos avanços contínuos, a Realidade Virtual segue evoluindo e se consolidando como uma das mais fascinantes ferramentas digitais da atualidade.

A Realidade Aumentada (RA) e a Realidade Mista (MR) são tecnologias que complementam o mundo real com elementos virtuais, criando experiências em que objetos digitais parecem coexistir e interagir com o ambiente físico. A RA é caracterizada por combinar elementos reais e virtuais, permitindo a interatividade em tempo real e interação em três dimensões [Azuma, 1997]. Enquanto a RA insere camadas de informação ou objetos sobre o mundo real, a MR vai além, permitindo uma fusão mais avançada entre o real e o virtual, com interações dinâmicas e respostas em tempo real. Já o termo Realidade Estendida (XR) atua como um guarda-chuva conceitual que abrange tanto a RV quanto a RA e a MR, sendo utilizado para descrever qualquer tecnologia imersiva que expanda ou modifique a percepção do usuário em relação ao mundo real ou virtual.

As experiências oferecidas por AVIs são potencializadas pelo uso de interfaces multisensoriais e dispositivos avançados de interação. Interfaces hápticas, por exemplo, proporcionam ao usuário a sensação de toque e textura, aumentando o realismo da experiência. Além disso, sistemas de rastreamento ocular, reconhecimento de voz e sensores corporais permitem que os ambientes virtuais respondam de forma natural aos movimentos e comandos do usuário, ampliando a imersão e a interatividade. Esses recursos transformam a navegação e manipulação em ações intuitivas, aproximando ainda mais a experiência digital da vivência no mundo físico.

Nos últimos anos, a convergência entre essas tecnologias (Realidade Virtual, Realidade Aumentada, Realidade Mista e Ambientes Virtuais Imersivos) tem pavimentado o caminho para o desenvolvimento do metaverso. Esse conceito se refere a uma rede interconectada e em larga escala de mundos virtuais tridimensionais, persistentes e renderizados em tempo real, nos quais os usuários podem interagir de forma síncrona com outras pessoas e com o ambiente digital.

O metaverso é definido como uma rede massivamente dimensionada e interoperável de mundos virtuais 3D renderizados em tempo real que podem ser experimentados de forma síncrona e persistente por um número efetivamente ilimitado de usuários com um senso de presença individual e com continuidade de dados, como identidade, histórico, direitos, objetos, comunicações e pagamentos [Ball, 2022]. Nesse

cenário, os Ambientes Virtuais Imersivos não apenas compõem o metaverso, mas são fundamentais para sua construção, oferecendo a base para experiências digitais complexas, contínuas e integradas entre o real e o virtual.

### 5.2.1. Aplicações Práticas de Tecnologias Imersivas: Educação, Indústria e Entretenimento

As tecnologias imersivas estão deixando de ser tendências experimentais para se tornarem ferramentas práticas em diversos setores. Esta seção exemplifica como RV, RA e MR estão sendo utilizadas na educação, indústria e no entretenimento e lazer, evidenciando seus impactos, benefícios e desafios. Ao longo dos casos, será possível observar como essas tecnologias ampliam possibilidades de aprendizagem, otimizam processos produtivos e reinventam experiências culturais.

#### 5.2.1.1. Educação

A Realidade Virtual tem se consolidado como uma ferramenta inovadora no ensino, expandindo suas aplicações e proporcionando experiências de aprendizagem imersivas que antes eram inimagináveis [Dalgarno; Lee, 2010]. Nos últimos anos, sua integração aos currículos escolares e universitários tornou-se mais estruturada, à medida que educadores e desenvolvedores colaboram para criar soluções tecnológicas que complementam e enriquecem o ensino tradicional.

Estudos mostram que a RV pode melhorar significativamente a retenção do conhecimento e a compreensão de conceitos abstratos. Ambientes de aprendizagem baseados em RV proporcionam resultados superiores em comparação com métodos convencionais, destacando seu potencial para transformar a maneira como os alunos interagem com conteúdos educacionais [Merchant, 2014].

Um dos aspectos mais promissores da Realidade Virtual na educação é sua capacidade de tornar o aprendizado mais inclusivo e acessível. Essa tecnologia pode ser adaptada para atender às necessidades de alunos com deficiências, criando ambientes personalizados que ajudam a superar barreiras físicas e cognitivas [Smith; Hamilton, 2015]. Dessa forma, a VR democratiza a educação, garantindo que todos os estudantes tenham experiências de aprendizagem significativas.

A personalização do conteúdo educacional em Realidade Virtual pode aumentar a eficiência e a satisfação dos alunos, tornando o aprendizado mais dinâmico e envolvente [Radianti et al., 2020]. Além disso, a RV viabiliza o acesso a ambientes antes inacessíveis, como visitas virtuais a locais históricos, exploração de ecossistemas remotos e até simulações de viagens espaciais, ampliando o horizonte dos estudantes e tornando o ensino mais enriquecedor [Dalgarno; Lee, 2010].

Além do aspecto educativo, o turismo virtual revolucionou a forma como as pessoas experienciam as viagens a partir do conforto das suas casas, oferecendo uma experiência imersiva e simulando a sensação de estar fisicamente presente em um local turístico, proporcionando também um vislumbre das viagens futuras [Ouerghemmi et al., 2023]. Como exemplo, a implementação da realidade virtual no contexto do turismo de cacau na Mata Atlântica não só oferece uma solução inovadora, mas também promove a sustentabilidade, a educação e o desenvolvimento econômico, alinhando-se

com as metas globais de preservação ambiental e desenvolvimento sustentável (Figura 5.1).


**Figura 5.1. Turismo virtual na Cabruca (Alves 2025, MTILab/UESC)**

A Realidade Virtual tem sido cada vez mais utilizada no ensino de Artes e Humanidades, oferecendo experiências imersivas que aprofundam a compreensão da cultura, história e arte. A RV também revoluciona o ensino de Humanidades, tornando-o mais experiencial e contextualizado. Por exemplo, uma aula de literatura pode ser enriquecida com uma visita virtual ao cenário descrito em um romance, promovendo uma absorção mais aplicada dos conceitos [Luo et al., 2021]. No ensino de História, visitas a museus virtuais possibilitam a interação do estudante com objetos e documentos de difícil acesso (Figura 5.2).

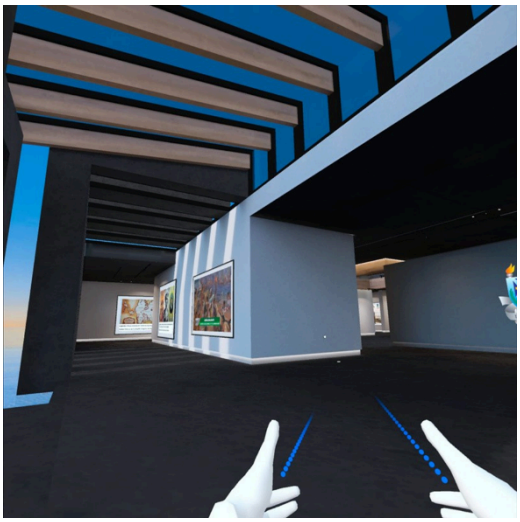


**Figura 5.2. Imersão na História da Independência da Bahia (MTILab/UESC)**

Além de suas aplicações educacionais, a RV democratiza o acesso à arte e cultura, permitindo que indivíduos de diversas regiões e com diferentes capacidades participem plenamente dessas experiências. Instituições culturais podem usar a RV para criar exposições virtuais acessíveis a todos, reduzindo barreiras geográficas e físicas [Smith; Hamilton, 2015].

Jogos desenvolvidos com técnicas de Realidade Virtual, podem não apenas entreter, mas também educar e promover a compreensão intercultural, contribuindo assim para uma maior valorização e preservação das diferentes culturas [Titus e Ng'ambi, 2023]. O jogo IsoPuzzle (Figura 5.3), apresenta-se como uma ferramenta intrincada de aprendizado, é um jogo educativo interdisciplinar que combina quebra-cabeças, geometria isométrica e simbologia Adinkra.

**Figura 5.3. IsoPuzzle (Moura 2024)**

#### 5.2.1.1.1. Laboratórios imersivos de ciências

Em diversas universidades, como a Arizona State University (EUA), a realidade virtual tem sido usada para simular experimentos químicos, biológicos e físicos em ambientes virtuais controlados. O ensino de ciências tem sido fortemente impactado pela RV, permitindo visualizações tridimensionais de moléculas, observação detalhada de reações químicas e até viagens pelo corpo humano para entender processos biológicos [Merchant et al., 2014]. Os estudantes podem explorar moléculas em 3D, manipular reagentes e observar reações em escala molecular sem riscos reais. Isso democratiza o acesso ao conhecimento prático, especialmente em instituições que não dispõem de laboratórios físicos.

Além disso, laboratórios virtuais oferecem um espaço seguro para experimentações, permitindo que os alunos pratiquem habilidades investigativas sem

riscos e aprendam com erros através da repetição ilimitada [Makransky et al., 2020]. A Realidade Virtual também possibilita a exploração de ambientes inacessíveis no mundo real, como o espaço sideral, o fundo do oceano ou ecossistemas distantes, aumentando o engajamento dos alunos e tornando a aprendizagem mais memorável [Dalgarno; Lee, 2010].

Outro benefício significativo da Realidade Virtual é sua capacidade de promover interação e colaboração entre alunos. Ambientes virtuais podem ser configurados para suportar múltiplos usuários, permitindo que estudantes trabalhem juntos em projetos, discutam conceitos e resolvam problemas de maneira cooperativa [Roussou, 2004]. Essa abordagem fortalece habilidades comunicativas e de trabalho em equipe, essenciais para a aprendizagem moderna.

#### 5.2.1.1.2. Realidade Aumentada em livros e materiais didáticos

A Realidade Aumentada tem revolucionado a forma como o conteúdo educacional é apresentado, oferecendo novas possibilidades para o engajamento dos estudantes e a compreensão de conceitos abstratos. Em vez de limitar-se a imagens estáticas e textos, livros e materiais didáticos com RA permitem que conteúdos ganhem vida por meio de elementos digitais tridimensionais sobrepostos ao mundo físico. Essa integração torna o aprendizado mais interativo, visual e dinâmico, beneficiando especialmente alunos do ensino básico, que frequentemente aprendem melhor por meio da exploração e da experimentação.

Plataformas como o Merge Cube permitem que alunos do ensino fundamental visualizem estruturas complexas, como o sistema solar, órgãos do corpo humano ( Figura 5.4) ou dinossauros em 3D, diretamente sobre um cubo físico, usando apenas um celular. Essa abordagem permite que os alunos interajam com os objetos digitais, girando-os, ampliando-os e observando-os de diferentes ângulos, promovendo uma experiência mais envolvente e significativa. Professores de biologia, por exemplo, podem transformar aulas expositivas em experiências interativas de exploração anatômica.

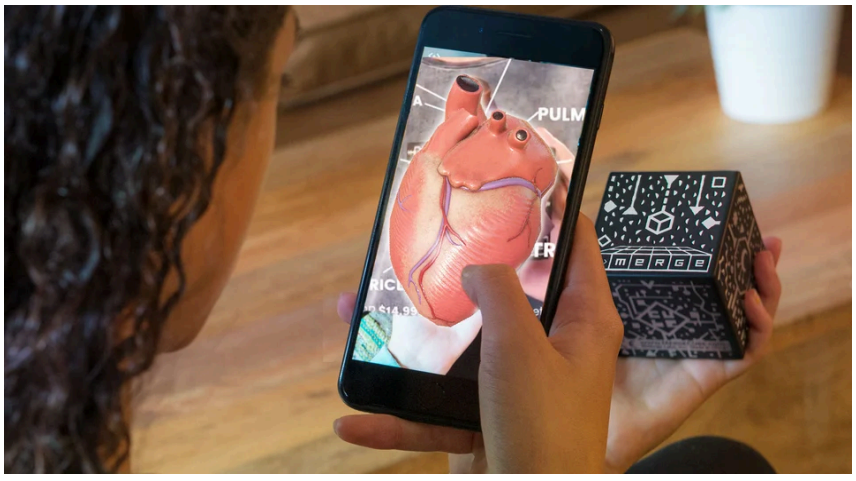



**Figura 5.4: Uso do Merge Cube em aulas remotas (SQUIRRELS LLC 2020)**

Além de conteúdos científicos, a RA também tem sido aplicada em livros infantis e obras literárias, onde personagens saltam das páginas e interagem com o leitor. Isso cria uma conexão emocional mais forte com o conteúdo e pode ajudar na alfabetização e no estímulo à leitura. Em contextos bilíngues ou multilíngues, recursos de RA podem incluir traduções instantâneas, pronúncias guiadas e contextos culturais visuais, promovendo uma aprendizagem mais rica.

Muitos recursos de RA são compatíveis com celulares comuns, o que permite sua adoção mesmo em ambientes escolares com infraestrutura tecnológica limitada. A integração da Realidade Aumentada ao ambiente escolar como ferramenta pedagógica contribui para a formação de competências digitais e prepara os estudantes para um futuro no qual tecnologias imersivas estarão cada vez mais presentes em diversos campos profissionais.

#### 5.2.1.1.3. Realidade Mista para treinamentos técnicos

A Realidade Mista tem se destacado em aplicações para o treinamento técnico em ambientes industriais e profissionais. Diferente da Realidade Virtual, que imerge completamente o usuário em um ambiente digital, a MR combina o mundo real com elementos virtuais interativos, possibilitando que os usuários trabalhem em cenários que mesclam objetos físicos e digitais de forma integrada. Isso cria um ambiente seguro e controlado para o desenvolvimento de habilidades práticas, especialmente em áreas que envolvem máquinas complexas ou riscos operacionais.

No Brasil, instituições como o SENAI CIMATEC têm adotado a MR para capacitação de operadores de máquinas industriais. Os alunos podem visualizar simulações virtuais de equipamentos reais sobrepostos ao ambiente físico da sala de aula. Dessa forma, é possível praticar a manipulação de botões, válvulas ou paineis com feedback visual em tempo real. Essa interação realista permite que os alunos experimentem diferentes procedimentos, compreendam a sequência correta de operações e aprendam a responder a falhas ou situações de emergência sem riscos reais.

Esse método de ensino promove não apenas a segurança, eliminando a exposição direta a máquinas em funcionamento, mas também aumenta a retenção do conteúdo. O aprendizado ativo, por meio da prática simulada, facilita a fixação das informações e prepara o aluno para a realidade do ambiente industrial com maior confiança.

A aplicação da Realidade Mista em treinamentos técnicos também contribui para a modernização dos processos educacionais e produtivos, alinhando a formação profissional às demandas de Indústrias 4.0. O uso de tecnologias imersivas potencializa a capacitação de trabalhadores, melhora a produtividade e reduz custos relacionados a erros operacionais e acidentes.

### 5.2.1.2. Indústria

Os novos métodos de produção que incorporam Sistemas Ciberfísicos nas áreas de manufatura, logística e serviços estão impulsionando uma transformação significativa na

estrutura industrial, rumo ao modelo da Indústria 4.0, reconhecido por seu expressivo potencial econômico [Lee; Bagheri; Kao, 2015].

Nesse contexto de inovação, as tecnologias imersivas estão desempenhando um papel fundamental na otimização dos processos industriais. Essas ferramentas possibilitam simulações avançadas, treinamento interativo para colaboradores e uma visualização aprimorada de dados e operações, aumentando a eficiência e reduzindo custos operacionais.

A evolução tecnológica não ocorre isoladamente, mas traz consigo uma série de mudanças organizacionais, promovendo o surgimento de novos modelos de negócios e incentivando uma maior participação dos colaboradores na gestão e operação dos sistemas. Com o apoio das tecnologias imersivas, os trabalhadores podem interagir com interfaces digitais intuitivas, acessar informações em tempo real e solucionar problemas de forma mais ágil e precisa. Isso demanda um entendimento mais profundo sobre o funcionamento dessas tecnologias e o fluxo de informações, a fim de identificar soluções eficazes para desafios específicos [Erol; Schumacher; Sihn, 2016; Kagermann; Wahlster; Helbig, 2013].

#### 5.2.1.2.1. Simulações de segurança e operação

Em indústrias marcadas por riscos operacionais, acidentes podem resultar em perda de vidas, elevados prejuízos financeiros e graves consequências sociais ou ambientais. Assim, torna-se fundamental treinar e avaliar a eficácia dos treinamentos, especialmente no que diz respeito à sua capacidade de se preparar adequadamente para situações do mundo real.

Aplicação de tecnologias de RV têm o potencial de proporcionar altos níveis de fidelidade física e psicológica, imergindo os treinandos em cenários realistas e interativos. Em indústrias de alto risco, o treinamento em saúde e segurança baseado em RV oferece vantagens significativas, ao permitir que os participantes enfrentam situações complexas e inovadoras sem se exporem a perigos ou colocarem em risco equipamentos industriais de alto valor. Esse ambiente de treinamento imersivo possibilita a simulação de diversas falhas, incluindo emergências raras que geralmente não ocorrem nas operações cotidianas. Essa abordagem não apenas reduz riscos, protege a saúde e preserva os ativos da organização, como também permite uma avaliação comportamental abrangente e eficaz.

Empresas como a Shell e a Boeing utilizam ambientes virtuais para treinar seus trabalhadores em situações críticas, como vazamentos de gás, incêndios ou falhas operacionais. O uso de RV nesses casos reduz custos, evita acidentes e permite múltiplas repetições sem consequências reais. Um operador de plataforma de petróleo pode simular o abandono de área em um ambiente 100% virtual, reproduzido com fidelidade.

No contexto da construção civil, o trabalho em altura segue requisitos explícitos nas Normas Regulamentadoras NR18 e NR35 (Brasil, 2023), mesmo assim, a prevenção de acidentes nessas atividades ocorre com dificuldade devido ao treinamento deficitário, à falta de fiscalização e monitoramento e a negligência, especialmente na guarda e conservação dos equipamentos de segurança.

Esta constatação reforça a necessidade de medidas preventivas e treinamentos específicos para garantir a proteção dos trabalhadores. O descumprimento das normas de segurança e as más condições laborais permanecem como as principais causas de acidentes na construção civil. As quedas de altura, por exemplo, representam 36% do total de acidentes registrados, o que evidencia a relevância da Realidade Virtual como ferramenta para treinamento em altura. O uso dessa tecnologia auxilia na identificação e prevenção de riscos ocupacionais, além de reduzir incidentes e aumentar a produtividade [Getuli et al., 2021].

A realidade virtual imersiva permite a simulação realista dos canteiros de obras, viabilizando estratégias ativas de treinamento [Eiris et al., 2020] Quando aplicada na indústria da construção, facilita a estimativa dos riscos e contribui para sua prevenção. Segundo Jeelani et al. (2020), ambientes imersivos proporcionam uma melhoria de 39% no reconhecimento de perigos e de 44% na gestão de riscos. Mesmo trabalhadores experientes e previamente treinados nem sempre conseguem identificar e gerenciar riscos com eficácia, pois essas habilidades cognitivas dependem de atenção, exame visual e tomada de decisão.

Existem pesquisas que apresentam métodos que auxiliam no desenvolvimento e aplicação de treinamentos imersivos de forma colaborativa e multiplataforma, ampliando sua adoção no setor industrial [Almeida et al., 2023]. Com base nesses procedimentos, foi desenvolvido um treinamento virtual para operação segura de prensa hidráulica que conduz o operador pelas etapas de posicionamento bi-manual, tudo em ambiente livre de riscos, demonstrado na Figura 5.5. O cenário admite sessões multiusuário, nas quais instrutor e colegas podem acompanhar a execução em tempo real e orientar correções, recurso alinhado à tendência de AVIs industriais voltados à interação social [Almeida et al., 2023].

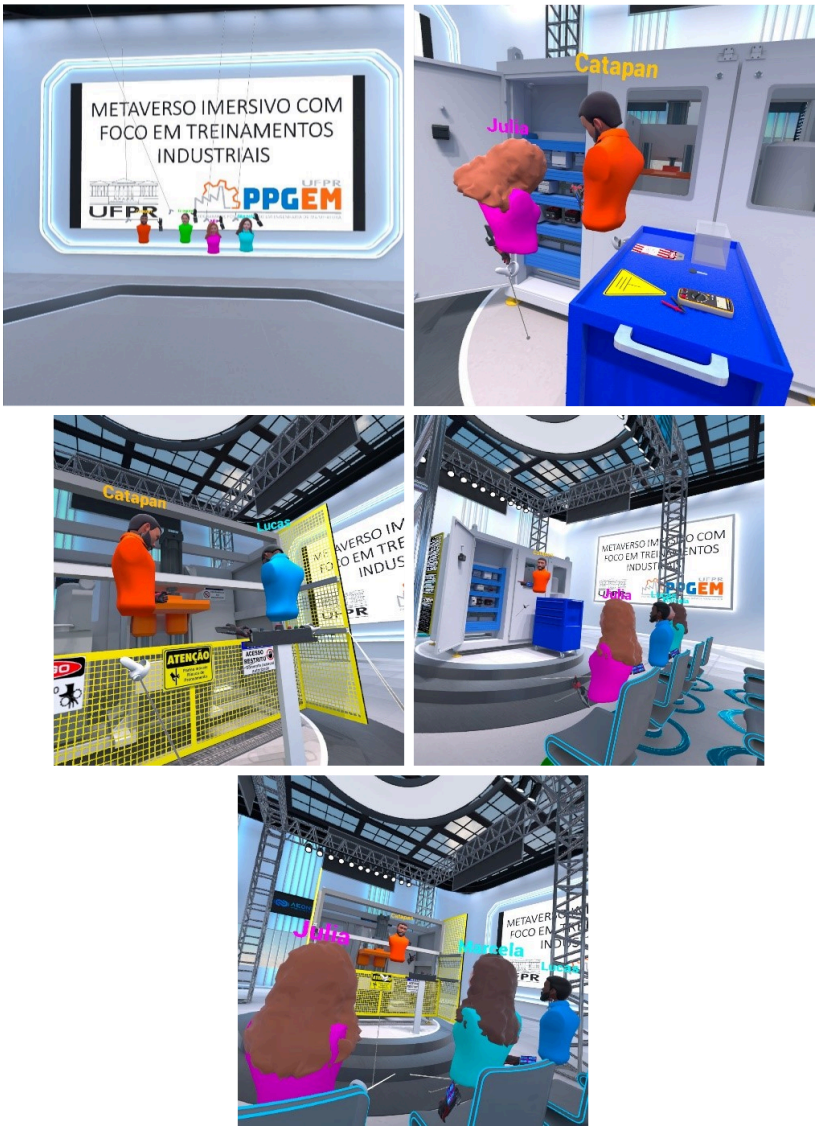



**Figura 5.5. Treinamento imersivo multiusuário com prensa hidráulica (Almeida et al., 2023)**

#### 5.2.1.2.2. Suporte remoto com Realidade Aumentada

Na manutenção de equipamentos industriais, a RA tem sido uma ferramenta para o suporte remoto. Por meio do uso de dispositivos vestíveis, como o óculos de realidade aumentada RealWear, que pode ser visto na Figura 5.6, técnicos em campo podem receber instruções visuais sobrepostas diretamente em seu campo de visão enquanto executam tarefas. Isso permite que eles tenham as mãos livres para trabalhar, mantendo acesso contínuo a informações críticas e orientação detalhada.

**Figura 5.6. Uso do óculos RealWear para suporte remoto na indústria (REALWEAR 2020)**

Uma das principais vantagens da aplicação dessa tecnologia para esse propósito é a possibilidade de interação em tempo real com especialistas localizados remotamente. Um engenheiro mais experiente pode ver exatamente o que o profissional em campo está visualizando, oferecer suporte personalizado e guiar a execução de procedimentos complexos, reduzindo erros e aumentando a eficiência. Essa colaboração remota minimiza a necessidade de deslocamentos físicos, acelerando o tempo de resposta e reduzindo os custos operacionais.

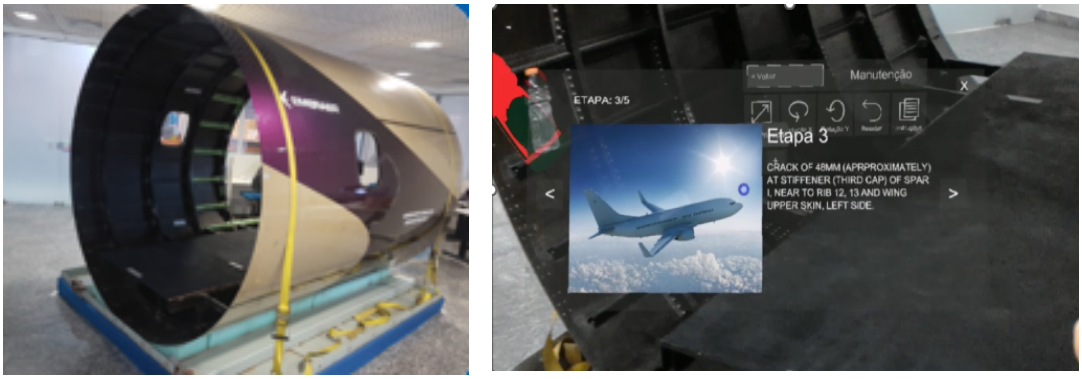



**Figura 5.7. Seção transversal de uma aeronave e aplicação de RA para manutenção (SENAI CIMATEC 2025)**

Além disso, a RA para suporte remoto contribui para diminuir o tempo de parada de máquinas, uma vez que o atendimento pode ser realizado com agilidade e precisão, evitando atrasos na produção. A sobreposição de diagramas e alertas instrutivos no ambiente real aumenta a clareza das instruções e facilita o entendimento do técnico, mesmo em situações de alta pressão ou complexidade. Como exemplo, a Figura 5.7 demonstra o projeto de assistência remota da EMBRAER que aplicou RA para indicar etapas da manutenção em aeronaves.

#### 5.2.1.2.3. Prototipagem e co-design com Realidade Virtual e Realidade Mista

A prototipagem e o co-design baseados em Realidade Virtual e Realidade Mista estão a ser cada vez mais aplicados, oferecendo vantagens significativas ao design colaborativo e ao desenvolvimento de produto [Kent et al., 2021; Wang et al., 2020]. A tecnologia permite sobrepor e ajustar componentes virtuais diretamente sobre protótipos físicos, acelerando ciclos de desenvolvimento e apoiando decisões mediante análises e visualizações aprimoradas [Kent et al., 2021]. Em sessões de co-design em tempo real (Figura 5.8), diversos participantes-chave podem manipular simultaneamente o mesmo modelo 3D integrado ao ambiente físico, inserindo anotações e validando requisitos em tempo real, prática que tem reduzido retrabalho e melhorado a convergência de ideias [Wang et al., 2020].

**Figura 5.8. Sessão de co-design com Realidade Mista. (Chamusca 2025)**

Especificamente na manufatura, onde o design de produto é crucial, os ambientes imersivos possibilitam diversas aplicações [Wang et al., 2020]. Um exemplo destacado é a indústria automobilística, onde sistemas exploratórios de RM permitem aos designers sobrepor modelos virtuais em ambientes físicos reais [Wang et al., 2020]. Isso potencializa atividades de design colaborativo e revisão de projetos, aprimorando significativamente a comunicação e a visualização entre participantes [Wang et al., 2020]. Além disso, os mesmos recursos que fortalecem a interação in loco podem ser estendidos a equipes distribuídas, apontando para um escopo de colaboração que transcende os limites físicos da fábrica.

Sob a ótica da prototipagem virtual, a MR (em conjunto com RV e RA) desponta como ferramenta versátil, capaz de substituir ou complementar protótipos físicos, permitir testes e análises digitais, e viabilizar a visualização interativa de múltiplas alternativas de design a menor custo [Kent et al., 2021; Wang et al., 2020]. Esses

protótipos podem ser réplicas fieis dos produtos ou incorporar atributos adicionais que enriquecem a experiência do usuário; a inserção direta de informações digitais no ambiente real torna-se decisiva para compreender o uso e aumentar a eficiência do processo de design [Kent et al., 2021]. Exemplos de protótipos em desenvolvidos em RV e MR são demonstrados nas Figuras 5.9 e 5.10.

**Figura 5.9. Protótipo de eletrodoméstico em Realidade Virtual (SENAI CIMATEC 2025)**

**Figura 5.10. Protótipo de veículo em Realidade Virtual (SENAI CIMATEC 2025)**

Expandindo além das fases de design e prototipagem, a RM também se destaca como ferramenta para colaboração remota [Wang et al., 2021]. Nesses contextos, a tecnologia possibilita interação em tempo real entre usuários geograficamente distantes, facilitando a comunicação e o compartilhamento não verbal de informações. Essa capacidade é útil em domínios como telemedicina, educação, treinamento, manutenção e engenharia [Wang et al., 2021].

Depois de examinar as aplicações e benefícios da MR, é importante considerar os obstáculos que ainda limitam sua adoção em escala. Apesar das vantagens, a implementação ampla da Realidade Mista enfrenta desafios significativos. Entre eles

estão a escassez de diretrizes claras de design, problemas de compatibilidade técnica, dificuldades na concepção de interfaces espaciais e integração aos fluxos de trabalho existentes [Krauß et al., 2021].

A percepção da irrelevância ou inadequação de recomendações para design em MR, fator que impacta diretamente as práticas de co-design, ao limitar orientações claras sobre como diferentes participantes podem colaborar e validar soluções em ambientes híbridos pode ser atribuída à falta de clareza sobre o uso pretendido e sobre o público-alvo, à abstração excessiva relativa a dispositivos específicos e dificuldades em encontrar conteúdos relevantes para profissionais em canais acadêmicos [Krauß et al., 2021]. Contudo, diante do potencial transformador que a MR apresenta para o design, colaboração remota e prototipagem, superar esses desafios é essencial para promover inovação e crescimento contínuos [Wang et al., 2021; Krauß et al., 2021].

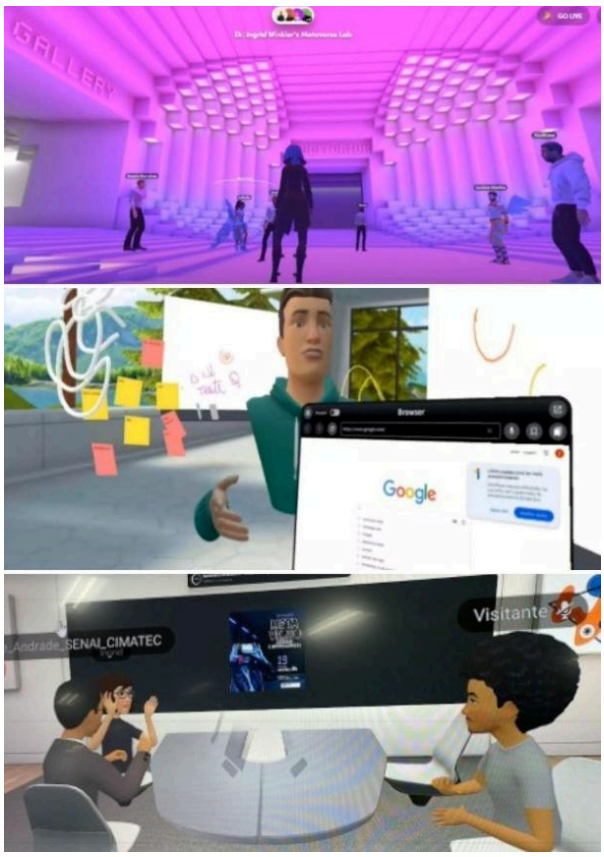



**Figura 5.11. Exemplos de colaboração remota nas plataformas de Realidade Virtual SPATIAL, Horizon Workrooms e Glue (SENAI CIMATEC 2025)**

### 5.2.1.3. Entretenimento e Lazer

A transformação digital está redefinindo a maneira como a sociedade se comunica, consome informações e interage com o mundo ao seu redor. Impulsionada por avanços tecnológicos cada vez mais sofisticados, essa mudança afeta desde a forma como

trabalhamos até como nos divertimos e socializamos. No setor de entretenimento e lazer, um dos aspectos mais inovadores dessa revolução é a ascensão das tecnologias imersivas, que oferecem experiências cada vez mais interativas e envolventes.

Essas tecnologias têm expandido os limites do entretenimento, permitindo que os usuários se transportem para mundos virtuais, interajam com conteúdos de maneira inédita e vivenciem novas formas de engajamento em jogos, filmes, atrações culturais e até eventos esportivos. Seja na indústria dos videogames, na exibição de filmes imersivos ou na criação de museus interativos, as tecnologias imersivas não apenas aprimoram a diversão, mas também criam novas possibilidades de aprendizado e conexão social.

#### 5.2.1.3.1. Jogos em Realidade Virtual

Os jogos em RV se destacam no entretenimento digital, oferecendo experiências imersivas além dos jogos tradicionais. Títulos populares como *Beat Saber, Half-Life: Alyx* e *The Walking Dead: Saints & Sinners* demonstram o potencial da RV para transportar os jogadores para mundos virtuais detalhados, onde a interação não é apenas visual, mas física e sensorial.

O PlayStation VR2 (Figura 5.12), lançado pela Sony como sucessor do PSVR, representa um avanço significativo nos jogos em realidade virtual para consoles. Compatível com o PlayStation 5, o dispositivo oferece gráficos em 4K HDR, rastreamento ocular, resposta tátil no headset e controles *Sense* com feedback preciso, elevando a sensação de presença e imersão. Equipamentos como esse demonstram como a tecnologia proporciona experiências envolventes, com jogabilidade intuitiva e ambientes virtuais ricos em detalhes, integrando performance técnica e conforto ao entretenimento doméstico.

**Figura 5.12. PlayStation VR2 (SONY INTERACTIVE ENTERTAINMENT LLC 2023)**

Por meio de controles de movimento e sensores posicionados no ambiente, os usuários podem se mover livremente, agachar, pular, pegar e manipular objetos dentro do jogo, ampliando a sensação de presença e controle. Essa integração entre o corpo e o

ambiente virtual redefine a forma como o jogador se relaciona com o conteúdo, tornando a experiência muito mais ativa e envolvente.

Além do entretenimento, os jogos em RV promovem benefícios cognitivos e físicos, estimulando a coordenação motora, o raciocínio espacial e o tempo de reação. Muitos jogos também incentivam a socialização virtual, com ambientes multiplayer que possibilitam a interação e cooperação entre jogadores em diferentes localidades. Os jogos em RV representam uma convergência entre tecnologia, arte e interação humana, ampliando os limites do entretenimento digital e experiências sensoriais

#### 5.2.1.3.2. Museus e exposições com Realidade Aumentada

A RA pode ampliar a forma como o público interage com museus e exposições ao proporcionar camadas adicionais de informação e experiências imersivas que enriquecem a visitação. Instituições como o Museu do Louvre, em Paris, e o MASP, em São Paulo, têm adotado aplicativos com RA para enriquecer exposições, transformando a observação passiva em uma jornada interativa para os visitantes.

Por meio de dispositivos móveis, como smartphones, o visitante pode apontar a câmera para obras de arte ou artefatos e visualizar informações contextuais, textos explicativos e reconstruções históricas em camadas digitais sobre o objeto real. O uso dessa tecnologia permite contextualizar as peças, ampliar detalhes ou recriar eventos relacionados ao objeto exposto.

Museus virtuais, exposições interativas e recriações históricas permitem que os usuários explorem conteúdos de maneira dinâmica, proporcionando um aprendizado envolvente e memorável [Roussou, 2004]. Isso aumenta o engajamento e a compreensão do público e também oferece acessibilidade ampliada, auxiliando pessoas com diferentes níveis de conhecimento ou necessidades especiais a explorar os acervos de forma personalizada e interativa. A mediação digital também abre espaço para novas formas de curadoria e narrativa, integrando elementos educativos e culturais.

#### 5.2.1.3.3. Espetáculos híbridos com Realidade Mista

Espetáculos híbridos com Realidade Mista estão transformando a forma como o público vivencia performances ao vivo, ao fundir o mundo físico com elementos digitais interativos em tempo real. Em eventos como o festival SXSW (South by Southwest), nos Estados Unidos, artistas têm explorado a MR para criar apresentações em que o espectador, equipado com óculos como o HTC Vive ou o Magic Leap, presencia simultaneamente a performance física do artista e camadas digitais imersivas, como cenários dinâmicos, dançarinos virtuais, avatares e efeitos visuais tridimensionais que reagem à música ou à movimentação do público.

**Figura 5.13. Apresentação da HTC com RV no SXSW (LANG; VARIETY 2022)**

Essa convergência entre arte, tecnologia e interatividade redefine o conceito tradicional de espetáculo. O palco deixa de ser um espaço limitado e passa a ser expandido virtualmente, permitindo que cada participante viva uma experiência única, adaptada à sua posição e interação com o ambiente. A narrativa pode ser fragmentada, envolvente e não linear, abrindo novas possibilidades para a expressão artística.

Além de concertos, essa tecnologia tem sido aplicada em peças teatrais, óperas e instalações imersivas, onde a presença física e a presença digital coexistem de forma contínua. A Realidade Mista, ao permitir a sobreposição de conteúdo digital ao ambiente real com liberdade de movimento e percepção tridimensional, proporciona não apenas um espetáculo visual, mas uma sensação de participação ativa do público.

#### 5.2.1.4. Saúde

O setor de saúde adotou a realidade virtual pela primeira vez na década de 1990, utilizada como uma ferramenta de planejamento para procedimentos cirúrgicos complexos. A adoção crescente da Realidade Virtual no setor de saúde é impulsionada por sua capacidade exclusiva de oferecer simulações imersivas, interativas e altamente realistas, abrindo caminho para avanços em treinamento médico, estratégias de reabilitação e intervenções psicológicas.

No treinamento médico os simuladores cirúrgicos permitem que estudantes e cirurgiões pratiquem procedimentos complexos em ambientes virtuais, sem riscos a pacientes reais.

As tecnologias de RV, utilizam dispositivos computacionais interativos, que têm evoluído para simular de forma cada vez mais fiel a interação humana em ambientes virtuais. Esse avanço tem possibilitado a criação de soluções realistas e seguras. Em contextos de reabilitação com uso de aparelhos, algumas abordagens baseadas em Realidade Virtual têm se destacado como alternativas promissoras aos métodos

tradicionais. Entre elas, as soluções que utilizam Jogos Sérios Virtuais ganham relevância por promoverem maior engajamento e motivação dos pacientes, ao introduzirem uma dimensão de entretenimento em atividades que, de outro modo, poderiam ser cansativas e repetitivas [Silva, 2017]. O CicloExergame, Figura 5.14, é um exemplo de um jogo sério para apoiar a realização de sessões de telerreabilitação que envolvem o cicloergômetro (bicicleta de cabeceira) na reabilitação de pacientes com disfunções motoras.

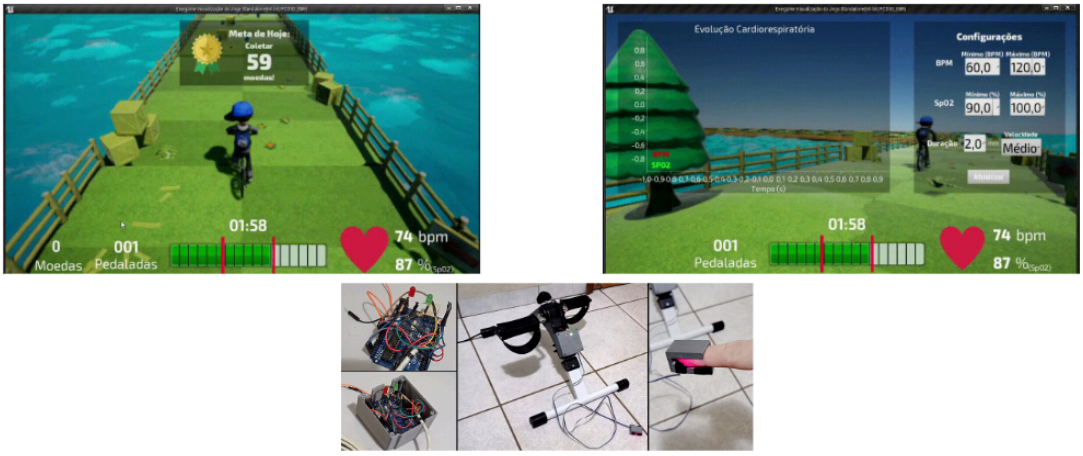



**Figura 5.14. CicloExergame (Souza 2021, LabJIS/UFG)**

Os jogos sérios têm se mostrado ferramentas promissoras no contexto dos cuidados com a saúde, ao aliarem elementos lúdicos com objetivos educativos e/ou terapêuticos. Eles podem contribuir para o engajamento de pacientes em seus tratamentos, facilitar o aprendizado de práticas saudáveis e melhorar a adesão a rotinas médicas, especialmente entre crianças, adolescentes e idosos. Além disso, permitem a simulação de cenários clínicos para a capacitação de profissionais da saúde, promovendo a aprendizagem prática em ambientes controlados e seguros. Combinando motivação, interatividade e feedback em tempo real, os jogos sérios fortalecem a relação entre tecnologia e bem-estar, ampliando o alcance e a efetividade das estratégias de promoção da saúde. O Lenda Daara e o Saúde Bucal, Figuras 5.15 e 5.16 respectivamente, apresentam exemplos de jogos sérios voltados para o cuidado com a saúde. O Lenda Daara é um jogo que utiliza a casa como forma de coletar dados do paciente diabético e/ou hipertenso e atua nesse ambiente de forma lúdica e que inspira o autocuidado. O Saúde Bucal é um jogo voltado para o público infantil que apoia na escolha de melhores alimentos e cuidados na escovação.

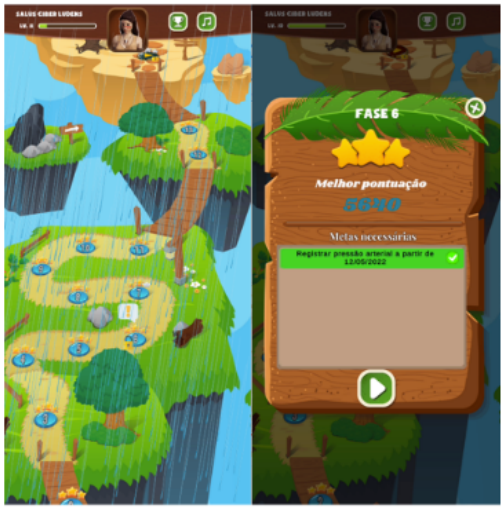



**Figura 5.15. Lenda Daara (Wanderley 2021)**

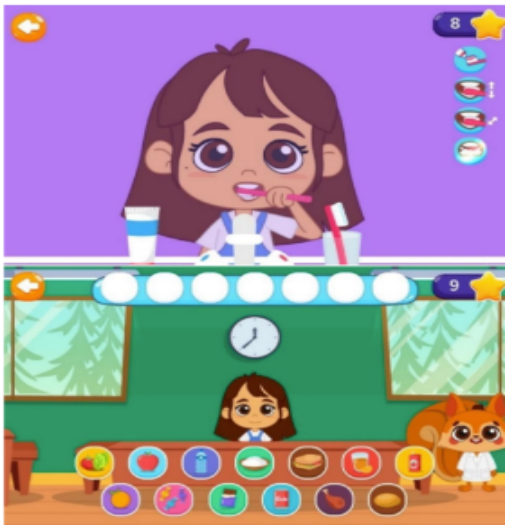


**Figura 5.16. Saúde Bucal (Mendonça 2022)**

Os jogos sérios têm ganhado destaque também como ferramentas complementares em intervenções psicológicas, oferecendo abordagens interativas e envolventes para o tratamento de diversas condições, como ansiedade, depressão, transtornos do espectro autista (TEA) e fobias. Esses jogos são projetados com fundamentos em teorias psicológicas e técnicas terapêuticas, como a terapia cognitivo-comportamental, promovendo o autoconhecimento, a regulação emocional e o desenvolvimento de habilidades sociais. Ao proporcionar um ambiente controlado e seguro para a experimentação de comportamentos e enfrentamento de situações desafiadoras, os jogos sérios facilitam o engajamento dos pacientes no processo terapêutico, especialmente entre crianças e adolescentes. Além disso, permitem o monitoramento do progresso de forma dinâmica e personalizada, contribuindo para intervenções mais eficazes e adaptadas às necessidades individuais. O DiagnosTEA, Figura 5.17, é um jogo digital implementado para auxiliar o diagnóstico e tratamento do TEA.


**Figura 5.17. DiagnosTEA (Barbosa 2024, LabJIS/UFG)**

Soluções imersivas têm potencial para apoiar o desenvolvimento de habilidades de orientação e mobilidade em pessoas cegas, ao oferecer ambientes virtuais seguros e controlados onde os usuários podem explorar, treinar e experimentar diferentes estratégias de locomoção. Utilizando recursos como áudio 3D, feedback tátil e

simulações realistas, essas soluções permitem que os participantes pratiquem a percepção espacial, a identificação de obstáculos e a tomada de decisões em trajetos urbanos ou internos, promovendo maior confiança e autonomia na vida real. Esse tipo de abordagem integra saúde, tecnologia e inclusão, ampliando as possibilidades de reabilitação e participação ativa no cotidiano. O Escape_INF-VR é um jogo imersivo do tipo escape que pode ser jogado por pessoas cegas e videntes (Figura 5.18). Além do entretenimento, para as pessoas cegas, o jogo com respostas sonoras e táteis pode apoiar o desenvolvimento da cognição espacial e das habilidades de orientação e mobilidade.

**Figura 5.18. Escape-INF-VR (Coronado 2024, LabJIS/UFG)**

À medida que a medicina avança em direção a uma maior complexidade, a demanda por metodologias inovadoras se intensifica, posicionando a realidade virtual como um caminho promissor para a área da saúde, especialmente considerando as restrições econômicas associadas às simulações presenciais tradicionais. Diferentemente de muitos outros domínios de alto risco, a área da saúde notavelmente carece de um protocolo padronizado para ensaio e simulação preventivos antes que os trainees se envolvam em cenários clínicos de alto risco ou realizem procedimentos complexos em pacientes reais.

Os exemplos expostos mostram que os ambientes imersivos já estão sendo adotados de maneira estratégica e escalável. Algumas tendências emergentes incluem a customização adaptativa, na qual os ambientes se ajustam ao perfil do usuário, a integração com inteligência artificial, permitindo agentes virtuais mais responsivos e personalizados e o aumento da portabilidade, com dispositivos cada vez mais leves, acessíveis e conectados.

Porém, ainda há desafios importantes: a infraestrutura tecnológica, a capacitação de professores e profissionais, as questões éticas envolvidas na coleta de dados e a necessidade de inclusão digital. Para que a adoção dessas tecnologias seja efetiva e equitativa, é fundamental desenvolver políticas públicas, promover pesquisa aplicada e investir em formação multidisciplinar.

## 5.3. Senso de presença em Realidade Virtual

O senso de presença em RV refere-se à sensação subjetiva de "estar presente" no ambiente digital. É o fenômeno psicológico que ocorre quando o usuário começa a reagir ao mundo virtual como se ele fosse real, mesmo sabendo que está em uma simulação. A presença não depende apenas da qualidade gráfica ou dos dispositivos utilizados, mas da integração entre aspectos tecnológicos e fatores humanos.

Embora os termos imersão e presença muitas vezes sejam usados como sinônimos, eles têm significados distintos. Imersão é uma característica objetiva do sistema de RV, refere-se à capacidade técnica do ambiente de simular um mundo convincente por meio de elementos como som tridimensional, imagens em 3D estereoscópico, rastreamento corporal e interfaces responsivas. Já a presença é o estado subjetivo em que o usuário se sente realmente dentro desse ambiente virtual. Em suma, a imersão é proporcionada pela tecnologia, enquanto a presença é uma experiência psicológica do usuário.

Tradicionalmente, a presença em ambientes virtuais é avaliada por meio de questionários subjetivos. No entanto, métodos objetivos como medidas comportamentais e monitoramento fisiológico são muito eficazes por captarem reações involuntárias do usuário, como o tempo de reação aos estímulos e o comportamento da sua visão na interação com o ambiente. No campo do monitoramento fisiológico, sinais corporais são usados para inferir o nível de presença, a seção a seguir aborda a utilização do *biofeedback* na Realidade Virtual.

## 5.4. *Biofeedback*: Conceitos e Sensores

O *biofeedback* em Realidade Virtual é a coleta, monitoramento e interpretação em tempo real de sinais fisiológicos do usuário durante a interação com o ambiente virtual. Esses dados permitem avaliar a experiência do usuário, adaptar dinamicamente o conteúdo da simulação e promover autorregulação e autoconsciência.

Em ambientes de RV, as medições de *biofeedback* apoiam a compreensão do estado físico, emocional e cognitivo do usuário. Entre as métricas mais utilizadas estão a frequência cardíaca (HR) e a variabilidade da frequência cardíaca (HRV), que indicam níveis de excitação, estresse ou relaxamento. A condutância da pele (GSR), por sua vez, mede alterações na sudorese, refletindo diretamente a intensidade da excitação emocional diante de estímulos virtuais.

Outra métrica importante é a eletromiografia (EMG), que registra a tensão muscular, especialmente em regiões como o rosto ou os ombros. A tensão muscular elevada pode sinalizar estresse, esforço físico ou reações intensas a eventos no ambiente virtual. A dilatação pupilar e os padrões de respiração também são utilizados para avaliar carga cognitiva e estados emocionais, pupilas mais dilatadas e respiração mais rápida geralmente indicam maior engajamento ou ansiedade. O eletroencefalograma (EEG), permite uma leitura direta da atividade cerebral, revelando níveis de atenção, concentração e sobrecarga cognitiva.

Essas medições fornecem dados para prever aspectos do comportamento do usuário. Por exemplo, uma queda na variabilidade da frequência cardíaca combinada com aumento da GSR pode indicar níveis elevados de estresse, o que é útil para ajustar a dificuldade de uma tarefa em tempo real. O EEG pode identificar momentos de sobrecarga cognitiva, permitindo que sistemas adaptativos reduzam a complexidade do ambiente ou ofereçam pausas estratégicas. Já padrões de EMG associados à tensão mandibular podem sinalizar frustração ou esforço excessivo, útil em treinamentos que envolvam precisão motora, como simulações cirúrgicas.

Combinadas, essas métricas possibilitam uma personalização profunda da experiência em RV, promovendo maior imersão e engajamento. Além disso, o uso do *biofeedback* permite tornar a realidade virtual uma ferramenta não apenas de entretenimento, mas também de avaliação emocional, treinamento adaptativo e suporte terapêutico. Ao compreender melhor os estados internos do usuário por meio de suas respostas fisiológicas, é possível criar experiências virtuais mais responsivas, seguras e eficazes.

Os sensores mais utilizados em contextos imersivos para captura de sinais que compõem o *biofeedback* incluem rastreamento ocular, variabilidade da frequência cardíaca e eletroencefalografia.

### 5.4.1. Rastreamento Ocular

O rastreamento ocular é uma ferramenta importante para medir e influenciar o senso de presença do usuário no ambiente virtual. Ele mede a direção, duração e sequência dos olhares do usuário. É amplamente utilizado para avaliar a atenção visual, o interesse e a carga cognitiva. Em RV, o rastreamento ocular permite que o ambiente responda ao foco visual do participante, ou que se meça seu engajamento durante tarefas específicas [Holmqvist et al., 2011].

Os sensores de rastreamento ocular, ou *eyetrackers*, são normalmente integrados aos óculos de RV. Esses sensores monitoram em tempo real para onde o usuário está olhando, permitindo captar o foco visual, os padrões de atenção e a movimentação dos olhos dentro do ambiente virtual. Essa tecnologia acrescenta uma nova camada de interatividade e personalização às experiências imersivas, tornando possível adaptar o conteúdo de acordo com o comportamento visual do usuário.

Uma das principais aplicações do rastreamento ocular em RV é a otimização do desempenho gráfico por meio da técnica chamada *foveated rendering*. Com ela, o sistema renderiza com alta resolução apenas a região para onde o usuário está olhando diretamente, enquanto reduz a resolução periférica. Isso melhora a performance dos sistemas e reduz a carga computacional sem comprometer a qualidade percebida da imagem. O rastreamento ocular também permite criar interações mais naturais com o ambiente, como selecionar objetos apenas com o olhar ou gerar respostas adaptativas a partir da interação que o usuário está tendo com o ambiente apenas pelos olhos.

Do ponto de vista da pesquisa e avaliação do usuário, o rastreamento ocular fornece dados objetivos sobre o nível de presença e engajamento. Por exemplo, quando os movimentos oculares seguem os eventos virtuais de natural, é um indicativo de que o usuário está imerso e emocionalmente envolvido com a experiência. Por outro lado,

padrões desorganizados ou foco visual difuso podem sinalizar distração, desconforto ou quebra de presença.

Além disso, o rastreamento ocular também pode ser utilizado em conjunto com outras métricas de *biofeedback* para enriquecer a análise do estado mental e emocional do usuário. Por exemplo, a combinação entre rastreamento ocular, dilatação pupilar e atividade cerebral (EEG) pode indicar momentos de alta carga cognitiva ou tomada de decisão.

### 5.4.2. Variabilidade da Frequência Cardíaca (HRV)

A variabilidade da frequência cardíaca mede as variações entre os intervalos de batimentos cardíacos consecutivos, refletindo o equilíbrio entre os sistemas simpático e parassimpático. É um indicador importante de estresse, recuperação e regulação emocional [Shaffer e Ginsberg, 2017]. Em ambientes de RV, pode-se utilizar a Variabilidade da Frequência Cardíaca para detectar sobrecarga emocional ou adaptar a experiência ao nível de ansiedade do usuário.

No caso do *biofeedback* de frequência cardíaca, sensores conectados ao usuário monitoram suas respostas fisiológicas e transmitem os dados para um computador, onde são processados. A visualização dessas informações em tempo real possibilita ao indivíduo modificar suas próprias reações, promovendo um processo de autorregulação [Tori; Hounsell; Kirner, 2018].

Essa técnica atua como uma intervenção cardiorrespiratória, utilizando eletrocardiograma (ECG) para fornecer dados da frequência cardíaca, batimento a batimento, enquanto o usuário realiza manobras de respiração lenta. Isso permite ajustes conscientes nos processos corporais, favorecendo reduções na ansiedade e estados de relaxamento [Zeier, 1984].

Além de seu impacto na saúde emocional, a respiração influencia funções autonômicas, regulando a atividade parassimpática do coração. Assim, sistemas de RV com *biofeedback* podem integrar práticas de Mindfulness para o alívio do estresse e ansiedade no ambiente de trabalho.

O treinamento com *biofeedback* de frequência cardíaca se assemelha a práticas meditativas, pois envolve atenção aos movimentos respiratórios e o desvio do foco de preocupações. Como resultado, o usuário aprende a suprimir ou estimular atividades cerebrais específicas, aprimorando seu funcionamento cognitivo e emocional [Lehrer; Gevirtz, 2014].

Além de contribuir para o bem-estar e saúde física, o *biofeedback* é reconhecido como um método terapêutico valioso [Lantyer; Viana; Padovani, 2013]. Em ambientes virtuais projetados para *biofeedback*, técnicas de gamificação podem ser usadas para engajar o usuário no treinamento de habilidades emocionais. Dessa forma, o sistema responde às alterações fisiológicas e recompensa os avanços positivos na assimetria cortical.

De acordo com Lehrer e Gervitz (2014), o *biofeedback* de frequência cardíaca é útil em diversas condições fisiológicas e psicológicas, como dor, ansiedade, depressão, controle da pressão arterial e desempenho atlético. Treinar o foco na respiração e na

atenção plena promove benefícios físicos e mentais, como redução do estresse, queda dos níveis de cortisol e melhoria na tomada de decisões e criatividade.

Embora tradicionalmente aplicado na área clínica [Lehrer e Gervitz , 2014], o *biofeedback* vem sendo investigado em contextos como esportes, gravidez e ambientes de trabalho, com impactos positivos na redução do estresse e no aumento do desempenho cognitivo [Deschodt-Arsac et al., 2018; Van Der Zwan et al., 2019; Sutarto; Wahab; Zin, 2010].

Na educação e treinamento, o *biofeedback* pode melhorar atenção, cognição, gestão do estresse e eficiência na tomada de decisões, tornando-se um recurso valioso para situações desafiadoras e avaliação de desempenho [França; Pereira Neto; Soares, 2017]. Estudiosos pioneiros vêm explorando suas aplicações em RV para treinamento e cyberterapia, ampliando suas possibilidades terapêuticas e educativas.

### 5.4.3. Eletroencefalografia (EEG)

A EEG capta a atividade elétrica cerebral e permite inferir estados como atenção, relaxamento ou excitação mental. Pode ser utilizada em Realidade Virtual para controlar objetos (neurofeedback simbólico), adaptar a narrativa conforme o engajamento cerebral ou monitorar processos de aprendizagem em tempo real [Gruzelier, 2014].

O eletroencefalograma (EEG) humano é uma técnica consolidada e amplamente utilizada para avaliar a atividade elétrica cerebral [Rios-Pohl; Yacubian, 2016]. Sua aplicação envolve a colocação de eletrodos no escalpo do usuário para detectar pequenas correntes elétricas geradas pela atividade neural. Cada neurônio conduz cargas elétricas, e a soma dessas cargas resulta na formação de um campo elétrico flutuante ao redor do couro cabeludo, cuja diferença de potencial pode ser medida por sensores específicos [Crepaldi e De Faria, 2013].

O EEG possui alta resolução temporal, permitindo a observação da atividade cerebral em uma escala de milissegundos [Vallabhaneni et al., 2005]. Seus sinais apresentam amplitudes entre 10 e 100 μV, com frequências variando de 1 a 100 Hz, sendo classificadas nas seguintes bandas:

a) Delta (1-4 Hz): Predominante em adultos durante o sono profundo, podendo indicar distúrbios neurológicos presentes em estados de vigília [Nicolas-Alonso e Gomez-Gil, 2012; Kubler et al., 2001].

b) Teta (4-7 Hz): Comum em crianças e adultos em estados de sonolência ou meditação. A atividade excessiva pode estar associada a condições neurológicas [Nicolas-Alonso e Gomez-Gil, 2012].

c) Alfa (8-12 Hz): Surge em momentos de relaxamento e quando os olhos estão fechados. Relaciona-se ao processamento visual e à memória, podendo ser suprimida por um aumento do esforço mental [Klimesch, 1997; Venables e Fairclough, 2009].

d) Beta (12-30 Hz): Associada às atividades motoras, sendo simétrica em repouso e alterada durante movimentos ativos [Pfurtscheller e Neuper, 2001].

e) Gama (30-100 Hz): Vinculada à percepção motora e sensorial. Seu papel na cognição é relevante para pesquisas em sistemas BCI (Brain-Computer Interface), pois melhora a transferência de informações e a especificidade espacial [Darvas et al., 2010; Miller et al., 2007].

O posicionamento dos eletrodos no couro cabeludo segue o sistema internacional 10-20, que define a disposição dos sensores em distâncias de 10% a 20% das referências anatômicas do crânio [Zetehaku et al., 2016]. Esse sistema conta com 21 eletrodos, dos quais 19 estão no couro cabeludo e 2 na região auricular (A1 e A2). Cada eletrodo é identificado por uma letra, representando a região cerebral correspondente, e por um número. Números ímpares indicam o hemisfério esquerdo e pares, o direito. Os eletrodos da linha média possuem apenas letras e são identificados pela segunda letra "z".

Além da sua aplicação tradicional na medicina, o EEG é uma ferramenta essencial na pesquisa da neuroergonomia. Por ser possível sua utilização em ambientes reais e simulados, ele contribui para investigações sobre desempenho humano em cenários exigentes e estressantes. Segundo Gevins e Smith (2007), os padrões do EEG variam conforme a carga mental da tarefa, esforço cognitivo e níveis de fadiga.

A interpretação dos dados do EEG representa um desafio metodológico, pois requer a correlação entre os sinais elétricos cerebrais e características emocionais e funcionais. Estudos sugerem que a assimetria entre os hemisférios cerebrais pode servir como um indicador do estado emocional, embora os mecanismos neurais ainda não sejam plenamente compreendidos [Cacioppo, 2004; Coan; Allen, 2004]. Esse parâmetro é usado como marcador de neuroergonomia, auxiliando no monitoramento das emoções e na análise da neuroplasticidade cerebral, promovendo o desenvolvimento de habilidades emocionais em usuários de realidade virtual.

## 5.5. Integração entre Realidade Virtual e *Biofeedback*

A integração entre Realidade Virtual e *biofeedback* representa um avanço na criação de experiências imersivas mais inteligentes e centradas no usuário. Existem três modalidades principais para essa integração: passiva, ativa e interativa.

Na modalidade passiva, os dados fisiológicos do usuário são registrados e analisados, mas não influenciam o ambiente virtual. Esse modelo é amplamente utilizado em contextos de pesquisa e avaliação, permitindo estudar reações emocionais, padrões de atenção ou respostas de estresse sem modificar o cenário virtual. É ideal para diagnósticos, validação de interfaces e estudos comportamentais.

Na modalidade ativa, o ambiente virtual responde automaticamente ao estado fisiológico do usuário. Por exemplo, se for detectado um aumento na frequência cardíaca ou condutância da pele, indicando estresse, o cenário pode suavizar a iluminação, desacelerar a ação ou introduzir estímulos relaxantes. Esse tipo de adaptação torna as experiências mais empáticas e personalizadas, sendo especialmente útil em aplicações terapêuticas, como no tratamento de ansiedade ou reabilitação psicológica.

Já a modalidade interativa transforma os sinais fisiológicos em comandos diretos dentro do ambiente virtual. O rastreamento ocular captado por sensores pode ser usado para mover ou selecionar objetos, enquanto a respiração pode controlar a intensidade da luz ou a velocidade de uma simulação. Esse nível de integração cria experiências verdadeiramente imersivas e centradas no corpo do usuário.

Em conjunto, essas três modalidades tornam possível o desenvolvimento de sistemas de RV adaptativos, que reconhecem e respondem ao estado interno do usuário.

## 5.6. Aplicações Práticas do *Biofeedback* em ambientes virtuais

Em diversas áreas, o uso de sistemas de RV com *biofeedback* viabilizou novas formas de adaptação em ciclo fechado (onde o ambiente ou cenário se ajusta continuamente com base no estado fisiológico do usuário). Essa responsividade em tempo real tem demonstrado grande potencial para personalizar intervenções, aumentar a validade ecológica e melhorar tanto os resultados psicológicos quanto a experiência do usuário em aplicações de RV. Vamos explorar alguns campos de aplicação do *biofeedback* em ambientes virtuais: educação e treinamento, saúde e reabilitação, pesquisa e experiência do usuário.

### 5.6.1. Educação e Treinamento

Aplicações educacionais e de treinamento devem ser submetidas a uma avaliação sistemática da experiência do usuário e ser rigorosamente avaliadas quanto à sua eficácia na promoção da aprendizagem ou aquisição de habilidades, assegurando, assim, a transferência bem-sucedida de conhecimentos e competências.

Integrando processos de treinamento com tecnologias imersivas e de monitoramento de dados biométricos, os instrutores podem melhor compreender as condutas dos treinandos, melhor conhecendo como estes reagem aos estímulos externos.

O Eye Tracking permite rastrear onde o treinando está focando sua atenção, se ele observou - ou não - algum risco operacional ou mesmo se deixou de verificar itens e condições de trabalho.


**Figura 5.19. Avaliação em Realidade Virtual de Trabalho em Altura (Cordeiro 2024)**

O monitoramento cardíaco e eletroencefalograma permitem detectar níveis de atenção, estresse, distrações, fadiga e outros dados que, combinados, permitem compreender melhor as reações e, em boa medida, as situações de trabalho simuladas, os acidentes simulados e suas causas [Cordeiro et al., 2024, 2025].

Assim sendo, ambientes de Realidade Virtual com *biofeedback* podem identificar queda de atenção e propor reforços didáticos, adaptações dinâmicas e uma melhor avaliação do processo de treinamento em geral. Em treinamentos técnicos, pode-se avaliar se o profissional está sob estresse e adaptar a tarefa. Isso permite jornadas de aprendizado personalizadas, mais eficazes e humanizadas.

### 5.6.2. Saúde e Reabilitação

Sistemas de Realidade Virtual com Variabilidade da Frequência Cardíaca e EEG têm sido aplicados em tratamentos de ansiedade, fobias, TDAH e reabilitação motora [Friedman, 2022]. A imersão promove engajamento, enquanto o *biofeedback* permite ao paciente visualizar e controlar seus estados internos, promovendo maior autonomia e bem-estar.

Além da integração de sinais fisiológicos como variabilidade da frequência cardíaca e EEG, experiências de Realidade Virtual Olfativa (RVO) têm demonstrado potencial terapêutico para o tratamento e prevenção do Transtorno de Estresse Pós-Traumático (TEPT). Ao incorporar odores específicos a ambientes virtuais imersivos, é possível acessar memórias autobiográficas e emoções com maior precisão, promovendo um envolvimento emocional controlado durante as sessões [Herz, 2021]. Quando integradas com *biofeedback*, essas experiências permitem que pacientes regulem suas respostas fisiológicas em tempo real, favorecendo o reaprendizado emocional em contextos seguros e graduais.

A percepção em tempo real de seus próprios estados emocionais e fisiológicos, por meio de métricas como batimentos cardíacos, condutância da pele e padrões de EEG, pode ser utilizada para ajustar os estímulos sensoriais e a complexidade das tarefas propostas nos cenários de RV. Essa adaptabilidade contribui para uma

abordagem mais personalizada e ajuda a modular o nível de excitação ou ansiedade durante terapias de exposição em RV, otimizando os resultados terapêuticos em distúrbios relacionados ao trauma.

### 5.6.3. Pesquisa e Experiência do Usuário

O uso de *biofeedback* em ambientes de RV permite acessar indicadores fisiológicos em tempo real, ampliando a compreensão sobre o comportamento e os estados internos dos usuários durante a interação. Esses dados oferecem uma visão objetiva sobre aspectos subjetivos da experiência, como engajamento, desconforto, fadiga mental ou prazer, e servem como base para decisões de design mais precisas e centradas no usuário. Além disso, permitem o desenvolvimento de experiências adaptativas, que ajustam o conteúdo e o ritmo com base nas respostas do usuário, tornando os testes mais eficientes e contribuindo para o avanço da Experiência do Usuário (UX) em ambientes imersivos.

Pesquisadores utilizam as tecnologias de *biofeedback* em Realidade Virtual para medir a presença, a sobrecarga cognitiva e a resposta emocional em experiências imersivas [Slater & Sanchez-Vives, 2016]. Esses dados são valiosos para projetar sistemas mais responsivos, confortáveis e inclusivos. A Figura 5.20 ilustra a integração de VR e rastreamento ocular em testes de usabilidade e a experiência do usuário na indústria automotiva.

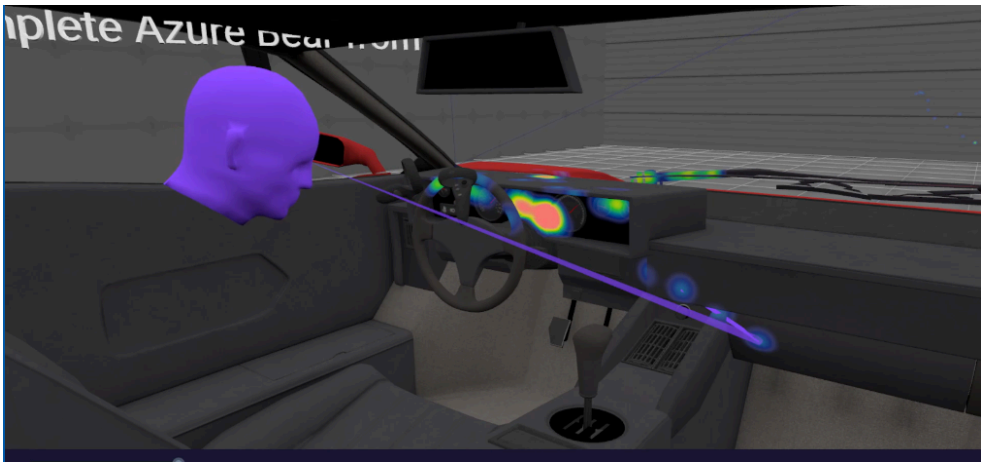


**Figura 5.20. Avaliação de UX por RV e UX na Automotiva (Autoria própria)**

## 5.7. Considerações Éticas e Finais

A integração entre Realidade Virtual e *biofeedback* exige cuidado com aspectos éticos e legais, sobretudo no que tange à privacidade dos dados fisiológicos. É fundamental garantir consentimento informado, segurança dos dados e uso transparente das informações coletadas.

Adicionalmente, é recomendado o cuidado para evitar incidentes e acidentes durante a aplicação de tecnologias de RV em combinação com elementos hápticos e físicos, mitigando ou eliminando riscos de tontura, quedas ou excessiva estimulação.

A perspectiva é que sistemas imersivos com sensores fisiológicos tornem-se cada vez mais comuns em contextos de formação, saúde e entretenimento. Essa convergência inaugura uma nova geração de tecnologias centradas no humano, que consideram não apenas a interação mecânica com a máquina, mas também os estados subjetivos que moldam a experiência humana.

Outro aspecto ético relevante é o impacto psicológico da exposição a ambientes imersivos que utilizam *biofeedback*. Como essas tecnologias podem induzir estados emocionais intensos, como estresse, euforia ou ansiedade, é essencial que profissionais capacitados estejam presentes para monitorar a experiência e intervir, se necessário. A ausência de supervisão ou o uso indiscriminado desses sistemas pode gerar consequências adversas, como gatilhos emocionais inesperados ou reforço de padrões cognitivos disfuncionais, especialmente em contextos terapêuticos e de reabilitação.

Deve-se considerar também a autonomia e a equidade no acesso às tecnologias. À medida que sistemas com *biofeedback* em RV se tornam mais sofisticados, corre-se o risco de concentrar tais soluções em contextos privilegiados, ampliando disparidades no acesso à saúde mental e ao bem-estar. A ética na aplicação dessas tecnologias também passa pela inclusão, pelo design acessível e pela atenção a populações vulneráveis que podem se beneficiar significativamente desses recursos, desde que adaptados às suas necessidades.

## References

Almeida, L. G. G. et al. (2023) “Innovating Industrial Training with Immersive Metaverses: A Method for Developing Cross-Platform Virtual Reality Environments” Applied Sciences, v. 13, n. 15, p. 8915, 2 ago. 2023.

Alves, A. O. (2025) Turismo imersivo em fazendas de cacau: diretrizes para a criação e desenvolvimento de conteúdos em 360º. Dissertação (Mestrado em Modelagem Computacional em Ciência e Tecnologia). Universidade Estadual de Santa Cruz, Ilhéus, 2025.

Azuma, R. T. A. (1997) Survey of augmented reality. Presence: Teleoperators and Virtual Environments, v. 6, n. 4, p. 355–385, 1997.

Barbosa, B. M., Ribeiro, M. W., Berretta, L, O,, Carvalho, S. T. (2024) “DiagnosTEA: a digital game as a tool for the diagnosis/therapy of Autism Spectrum Disorder” In: Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGAMES) - Trilha Saúde, 2024, Manaus. Anais do SBGAMES 2024. Porto Alegre: SBC, 2024. p. 1707-1718.

Marques, B., Moreira, D., Neves, M., Brás, S., Fernandes, J. M. (2025) “Battle Against Your Fears: Virtual Reality Serious Games and Physiological Analysis for Phobia Treatment” IEEE Computer Graphics and Applications, vol. 45, no. 1, pp. 67-75, Jan.-Feb. 2025

Cardoso, A.; Lamounier Jr, E. (2006) “A Realidade Virtual na Educação e Treinamento” In: Tori, R.; Kirner, C.; Siscoutto, R. Fundamentos e Tecnologia de Realidade Virtual

e Aumentada. Livro do Pré-simpósio, VIII Symposium on Virtual Reality. Porto Alegre: Editora SBC, 2006.

Cordeiro, A. et al. (2025) "Immersive Technologies to Height Safety: Training Evaluation and Insights" 2025 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR), Lisbon, Portugal, 2025, pp. 435-440, doi: 10.1109/AIxVR63409.2025.00082.

Cordeiro, A., Santos, F., Winkler, I. (2023) "Effectiveness of industrial training using virtual reality to mitigate risks associated with the work environment: a literature review" International Symposium on Innovation and Technology, Engineering and the Future of the Industry, 2023. ISSN: 2357-759.

Cordeiro, A., Almeida, L., Neves, C., Leite, R., Catapan, M., Siqueira, A., Silva, T., Winkler, I. (2024) IX International Symposium on Innovation and Technology, Innovation and Global Transformations for a Sustainable World - 2024. ISSN: 2357-759.

Coronado, A. ; Carvalho, S.; Berretta, L. O. (2024) Escape-INF-VR: An Accessible VR Escape Game Proposal for Blind Individuals. In: IFIP International Conference on Entertainment Computing (ICEC), 2024, Manaus. Proceedings of the Entertainment Computing - ICEC 2024: 23rd IFIP TC 14 International Conference , ICEC 2024, Lecture Notes in Computer Science. Cham: Springer, 2024. v. 15192. p. 337-341.

Dalgarno, B.; Lee, M. J. W. (2010) What are the learning affordances of 3-D virtual environments? British Journal of Educational Technology, 41(1), 10-32. 2010.

Deschodt-Arsac, V.; Lalanne, R.; Spiluttini, B.; Bertin, C.; Arsac, L.M. (2018) Effects of Heart Rate Variability Biofeedback Training in Athletes Exposed to Stress of University Examinations. PLoS ONE, 13(7): e0201388. 2018.

Dey, A.; Chatburn, A.; Billinghurst, M. (2019) Exploration of an EEG-Based Cognitively Adaptive Training System in Virtual Reality. 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, 2019, pp. 220-226.

França, A. C. P.; Soares, M. (2017) Review of Virtual Reality Technology: An Ergonomic Approach and Current Challenges. 8th International Conference on Applied Human Factors and Ergonomics (AHFE) and the Affiliated Conferences, 2017.

Friedman, D. (2022) Virtual Reality Therapy for Anxiety, Depression and Stress. Springer, 2022.

Gomes, PV et al. (2023) The use of artificial intelligence in interactive virtual reality adaptive environments with real-time biofeedback applied to phobias psychotherapy. In: VI Congreso Xove TIC: impulsando el talento científico. Outubro, 2023, A Coruña. Universidade da Coruña, Servizo de Publicacións, 2023. p. 275-279.

Gruzelier, J. H. (2014) EEG-neurofeedback for optimising performance. Progress in Brain Research, 215, 193–213. 2014.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011) Eye tracking: A comprehensive guide to methods and measures. Oxford University Press, 2011.

Jerald, J. (2015) The VR Book: Human-Centered Design for Virtual Reality. Association for Computing Machinery and Morgan & Claypool, 2015.

Kent, Lee et al. (2021) Mixed reality in design prototyping: A systematic review. Design Studies, v. 77, p. 101046, nov. 2021.

Kritikos, Jacob; ALEVIZOPOULOS, Georgios; KOUTSOURIS, Dimitris. (2021) Personalized virtual reality human-computer interaction for psychiatric and neurological illnesses. Frontiers in Human Neuroscience, v. 15, p. 596980, 2021.

Krauß, Veronika et al. (2021) Research and Practice Recommendations for Mixed Reality Design – Different Perspectives from the Community. In: VRST '21: 27TH ACM Symposium on Virtual Reality Software and Technology, Osaka Japan: ACM, 8 dez. 2021.

Lantyer, A. S.; Viana, M. B.; Padovani, R. C. (2013) Biofeedback no Tratamento de Transtornos Relacionados ao Estresse e à Ansiedade: Uma Revisão Crítica. Psico-USF, Bragança Paulista, v. 18, n. 1, p. 131 – 140, jan/abr 2013.

Lehrer, P. M.; Gevirtz, R. (2014) Heart Rate Variability Biofeedback: How and Why Does it Work? Frontiers in Psychology, Vol 5, Article 756, July 2014.

Luo, H., Li, G., Feng, Q., Yang, Y.; Zuo, M. (2021) Virtual reality in K-12 and higher education: A systematic review of the literature from 2000 to 2019. Journal of Computer Assisted Learning, 887–901. 2021.

Makransky, G., Terkildsen, T. S., Mayer, R. E. (2020) Adding immersive virtual reality to a science lab simulation causes more presence but less learning. Learning and Instruction, 60, 225-236. 2020.

Mark Billinghurst. Short course on the Psychology of XR at the University of South Australia.

Mendonça, T. S., (2022) Desenvolvimento de um serious game de educação em saúde bucal com participação de especialistas e usuários. 2022. Dissertação - Universidade Federal de Goiás.

Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., Davis, T. J. (2014) Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. Computers & Education, 70, 29-40. 2014.

Nasri, Mahsa. (2025) Towards Intelligent VR Training: A Physiological Adaptation Framework for Cognitive Load and Stress Detection. arXiv preprint, arXiv:2504.06461, 2025.

Norris, K.; Spicer, T.; Byrd. (2019) Virtual reality: the new pathway for effective safety training. Professional Safety, 64(6):36–39, 2019.

Ouerghemmi, C., Ertz, M., Bouslama, N., Tandon, U. (2023) “The impacto f virtual reality (vr) tour experience on tourists’ intention to visit”, Information, v. 14, n. 10, p. 546.

Radianti, J., Majchrzak, T. A., Fromm, J., Wohlgenannt, I. (2020) A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. Computers & Education, 147, 103778. 2020.

Roussou, M. (2004) Learning by doing and learning through play: An exploration of interactivity in virtual environments for children. Computers in Entertainment (CIE), 2(1), 1-23. 2004.

SENAI CIMATEC. (2025) Immersive Technologies for a Sustainable and Human-centric Industry 5.0 Lab. [Apresentação institucional]. Salvador, 10 abr. 2025.

Shaffer, F., & Ginsberg, J. P. (2017) An Overview of Heart Rate Variability Metrics and Norms. Frontiers in Public Health, 5, 258. 2017.

Silva, A. D. S. D.; Valenciano, P. J.; Fujisawa, D. S. (2017) Atividade lúdica na fisioterapia em pediatria: Revisão de literatura. Revista Brasileira de Educação Especial, 23:623 – 636, 2017.

Slater, M., & Sanchez-Vives, M. V. (2016) Enhancing Our Lives with Immersive Virtual Reality. Frontiers in Robotics and AI, 3, 74. 2016.

Slater, Mel, et al. (2022) A separate reality: An update on place illusion and plausibility in virtual reality. Frontiers in Virtual Reality, 3, 914392. 2022.

Slater, Mel et al. (2022) A separate reality: an update on place illusion and plausibility in virtual reality. Frontiers in Virtual Reality, v. 3, 2022.

Smith, S. R., & Hamilton, M. (2015) The efficacy of virtual reality technologies for educational purposes: A case study of special needs education. Journal of Educational Technology, 12(3), 15-25. 2015.

Souza, CHR, Oliveira, DM, Berretta, LO, Carvalho, ST. (2021) Jogos Sérios e Elementos de Jogos na Promoção de Engajamento em Contextos de Telerreabilitação de Pacientes. In: Simpósio Brasileiro de Jogos e Entretenimento Digital, 2021, Brasil. Anais Estendidos do XX Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames Estendido 2021). p. 896.

Sutarto, A. P.; Wahab, M. N. A.; Zin, N. M. (2010) Heart Rate Variability (HRV) Biofeedback: A New Training Approach for Operator’s Performance Enhancement. Journal of Industrial Engineering and Management, v3n1, p176-198, 2010.

Tori, R.; Hounsell, M. S.; Kirner, C. (2018) Realidade Virtual. In: Tori, R.; Hounsell, M. S. (Org.) Introdução à Realidade Virtual e Aumentada. Porto Alegre: SBC, 2018.

Tori, R.; Kirner, C. (2006) Fundamentos de Realidade Virtual. In: Tori, R.; Kirner, C.; Siscoutto, R. Fundamentos e Tecnologia de Realidade Virtual e Aumentada. Livro do Pré-simpósio, VIII Symposium on Virtual Reality. Porto Alegre: Editora SBC, 2006.

Van Der Zwan, J. E. et al. (2019) The Effect of Heart Rate Variability Biofeedback Training on Mental Health of Pregnant and Non-Pregnant Women: A Randomized Controlled Trial. International Journal of Environmental Research and Public Health, 16, 1051, 2019.

Wang, Peng et al. (2020) A comprehensive survey of AR/MR-based co-design in manufacturing. Engineering with Computers, v. 36, n. 4, p. 1715–1738, 2020.

Wang, Peng et al. (2021) AR/MR Remote Collaboration on Physical Tasks: A Review. Robotics and Computer-Integrated Manufacturing, v. 72, p. 102071, 2021.

Wanderley, L, Soares, S, Santos, SLV, Carvalho, ST. (2021) Desenvolvimento de um jogo para hipertensão utilizando a metodologia Design Science Research: equilibrando a Ciência e a Arte. In: Simpósio Brasileiro de Jogos e Entretenimento Digital, 2021, Brasil. Anais Estendidos do XX Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames Estendido 2021). p. 857.

Weber, R, Dash, A, Wriessnegger, SC. (2024) Design of a Virtual Reality-Based Neuroadaptive System for Treatment of Arachnophobia. 2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), St Albans, United Kingdom, 2024, pp. 255-259.

Zeier, H. (1984) Biofeedback and Self-Regulation, 9: 497, 1984

Zelis, H, Kubanka, T, Williams, L. (2025) Adaptive Virtual Reality Exposure Therapy with Biofeedback. SoutheastCon 2025, Concord, NC, USA, 2025, pp. 1059-1060, doi: 10.1109/SoutheastCon56624.2025.10971438.

# Capítulo 6

# Respostas a perguntas de clínica médica utilizando a Geração Aumentada por Recuperação (RAG)

Luciana Bencke

***Abstract***

*Large Language Models (LLMs) have excelled in clinical tasks, but they still face important challenges, such as generating factually incorrect responses (hallucinations) and the limitation of not reflecting recent updates in medical guidelines. These risks become critical in high-impact contexts, such as answering clinical questions. Retrieval Augmented Generation (RAG) seeks to mitigate these flaws by combining the capabilities of LLMs with information retrieved from external sources, public or private, enabling more accurate, up-to-date responses that are aligned with current clinical knowledge. Applications in healthcare include virtual assistants, decision support, educational systems, and other intelligent tools. This chapter discusses the motivation for using RAG in healthcare, its components, pipeline stages, and research opportunities in the area.*

***Resumo***

*Grandes Modelos de Linguagem (LLMs) têm se destacado em tarefas clínicas, mas ainda enfrentam desafios importantes, como a geração de respostas factualmente incorretas (alucinações) e a limitação de não refletirem atualizações recentes em diretrizes médicas. Esses riscos tornam-se críticos em contextos de alto impacto, como respostas a perguntas de clínica médica. A Geração Aumentada por Recuperação (RAG) busca mitigar essas falhas ao combinar a capacidade dos LLMs com informações recuperadas de fontes externas, públicas ou privadas, possibilitando respostas mais precisas, atualizadas e alinhadas com o conhecimento clínico vigente. Aplicações em saúde incluem assistentes virtuais, suporte à decisão, sistemas educacionais e outras ferramentas inteligentes. Esse capítulo aborda a motivação para o uso do RAG na saúde, seus componentes, etapas do pipeline e oportunidades de pesquisa na área.*

## 6.1. Introdução

Sistemas de resposta a perguntas são criados com o objetivo de suprir as demandas por informação dos seres humanos. Como grande parte do conhecimento está registrada em textos na internet, em e-mails ou em livros, esses sistemas têm forte relação com os buscadores. Atualmente, a distinção entre sistemas de perguntas e respostas e buscadores está cada vez mais tênue, à medida que os mecanismos de busca modernos incorporam grandes modelos de linguagem de grande especificamente treinados para responder a esse tipo de solicitação [Jurafsky and Martin 2025].

O escopo de um sistema de perguntas e respostas (*Question Answering*, QA) na área da saúde pode ser amplo, cobrindo uma diversidade de temas médicos e biomédicos, ou restrito, focando em subdomínios específicos, como cardiologia, oncologia ou saúde pública. A base de conhecimento utilizada por esses sistemas pode assumir múltiplas formas: desde documentos não estruturados, como prontuários, artigos científicos e diretrizes clínicas, até fontes estruturadas, como bancos de dados de saúde, ou ainda conhecimento embutido nos próprios parâmetros de modelos de linguagem, a chamada memória paramétrica.

As perguntas feitas a sistemas de QA também variam em complexidade e finalidade, sendo classificadas em factuais e não factuais [Cortes et al. 2024]. Perguntas factuais requerem respostas diretas e objetivas baseadas em evidências específicas, por exemplo "*Qual é a dosagem recomendada de paracetamol para adultos?*"ou "*Quais são os sintomas iniciais da COVID-19?*". Já perguntas não factuais exigem respostas mais extensas e contextualizadas, que envolvem a síntese de múltiplas fontes e maior elaboração interpretativa, por exemplo "*Explique as diferenças entre os protocolos de tratamento para hipertensão em idosos e adultos jovens*". Em ambientes clínicos, responder corretamente a ambos os tipos de perguntas é fundamental para garantir segurança, precisão e suporte à tomada de decisão baseada em evidências.

Aplicações baseadas em grandes modelos de linguagem (*Large Language Models*, LLMs), como o ChatGPT, exibem grande potencial em responder perguntas, mas apresentam vulnerabilidades como alucinações, vieses e imprecisões factuais — limitações especialmente críticas em ambientes clínicos, onde a exatidão das informações é essencial [Zakka et al. 2024]. LLMs têm o potencial de transformar a distribuição de informações médicas, destacando-se na geração de conteúdo e na comunicação interativa. No entanto, enfrentam limitações como desatualização, alucinações factuais e dependência de dados públicos, o que restringe seu uso na saúde [Ng et al. 2025].

Embora LLMs tenham alcançado excelente desempenho em várias tarefas que exigem resposta a perguntas de clínica médica, eles ainda enfrentam desafios com alucinações e conhecimento desatualizado [Xiong et al. 2024]. Os LLMs podem gerar respostas que parecem plausíveis, mas factualmente incorretas, conhecidas como alucinação [Chowdhury et al. 2025]. Além disso, os corpora de treinamento dos LLMs podem não incluir o conhecimento mais recente, como atualizações de diretrizes clínicas. Essas questões podem ser perigosas em domínios de alto risco, como assistência médica [Xiong et al. 2024]. [Singhal et al. 2023] avaliaram LLMs em respostas abertas sobre conhecimento clínico ao longo de eixos como factualidade, compreensão, raciocínio, possível dano e viés. Os autores constatam limitações dos modelos da época, reforçando a

importância das estruturas de avaliação e do desenvolvimento de métodos na criação de LLMs seguros e úteis para aplicações clínicas.

A Recuperação Aumentada por Geração (*Retrieval-Augmented Generation*, RAG) surge como uma solução para minimizar esses riscos na geração de respostas a perguntas clínicas, conectando LLMs a fontes externas — como artigos científicos, compêndios médicos e políticas institucionais — e permitindo acesso a informações além dos dados de treinamento. Com isso, ferramentas de Inteligência Artificial (IA) generativa podem integrar dados públicos e privados, ampliando sua aplicabilidade e precisão em contextos clínicos [Ng et al. 2025].

RAG combina técnicas de recuperação de informações com modelos generativos. A ideia central é recuperar informações relevantes de uma base de conhecimento externa ao modelo generativo antes de gerar uma resposta à pergunta, melhorando a precisão e a atualidade das respostas fornecidas por modelos de linguagem. Várias aplicações no domínio da saúde podem se beneficiar da utilização de RAG, como chatbots, assistentes virtuais, sistemas de educação na saúde, suporte à tomada de decisão clínica, entre outros.

Dois macrocomponentes do RAG são cruciais para bons resultados: o *Retriever*, responsável por recuperar fragmentos de texto que contenham informações relevantes para responder à pergunta, e o *Generator*, que se trata de um LLM que recebe essas informações e elabora a resposta, gerando o texto de forma coerente e fortemente fundamentado nos fragmentos recebidos. O *Generator* se baseia na capacidade de aprendizado em contexto (*In-Context Learning* [Brown et al. 2020]) dos LLMs, permitindo que o modelo produza respostas sobre informações que não foram explicitamente incluídas em seu treinamento. Assim, o modelo generativo elabora a resposta com base em um contexto recuperado de uma fonte externa, muitas vezes pertencente a um sistema privado. Esse processo reduz as limitações dos LLMs, garantindo maior precisão e alinhamento com diretrizes clínicas atualizadas.

A implementação de sistemas RAG pode seguir diferentes estratégias como proposto por [Gao et al. 2023]. Os autores destacam *Naïve RAG*, *Advanced RAG* e *Modular RAG* apresentadas na Figura 6.1. Cada estratégia se caracteriza por métodos específicos, com níveis distintos de complexidade, vantagens e limitações [Gao et al. 2023]. Os autores chamam essas estratégias de paradigmas RAG. O *Naïve RAG* é a forma mais básica (indexação, recuperação e geração), sendo simples de implementar, porém com piores resultados e mais suscetível a alucinações e incoerência. O paradigma *Advanced RAG* introduz melhorias em todas as etapas, como otimizações na indexação, reranking, compressão de contexto e busca híbrida, aumentando assim a qualidade das respostas, embora com maior complexidade e custo computacional. Já o *Modular RAG* adota uma arquitetura flexível com módulos especializados permitindo customização para diferentes cenários e domínios, mas exigindo maior esforço técnico. Segundo os autores, cada paradigma apresenta um equilíbrio diferente entre simplicidade, desempenho e adaptabilidade, sendo aplicável conforme o caso de uso.

As estratégias apresentadas na Figura 6.1 trazem uma ideia do nível de complexidade que sistemas RAG podem ter e enfatiza a natureza híbrida dos mesmos, onde resultados são afetados por contribuições de diversos componentes específicos. Assim, apesar dos benefícios significativos do RAG, especialmente em domínios sensíveis como

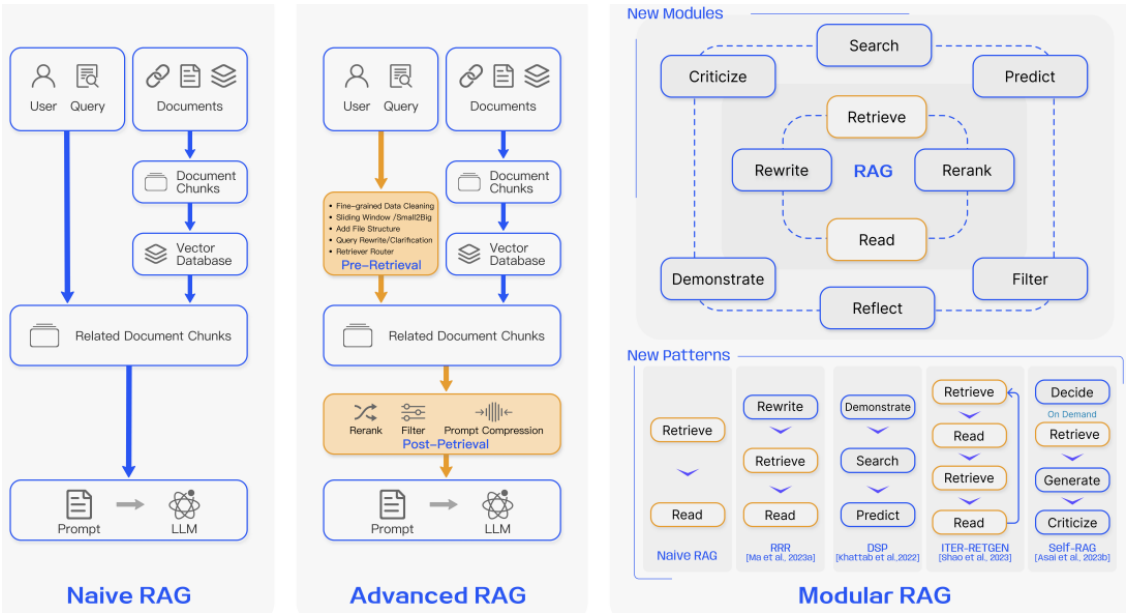



**Figura 6.1. Estratégias RAG. Extraído de [Gao et al. 2023]**

a saúde, existem diversos desafios na sua implementação e manutenção.

Um dos principais obstáculos está na **qualidade da recuperação**: o sistema pode retornar documentos irrelevantes ou genéricos, deixando de contemplar aspectos essenciais de casos específicos, por exemplo, ao buscar terapias para insuficiência cardíaca em idosos, pode recuperar textos sobre o tratamento em adultos, ignorando particularidades geriátricas. Além disso, o modelo enfrenta dificuldades na **seleção e fusão do contexto**, por exemplo quando precisa combinar informações contraditórias como no caso em que recuperam-se dois documentos sobre COVID-19 que mencionam tratamentos distintos: um sugere corticoterapia, o outro contraindica em determinadas fases. O modelo pode gerar uma resposta que mistura as informações, causando confusão ou erro clínico. Outro desafio técnico é o **limite da janela de contexto** dos modelos de linguagem, que os impede de lidar com documentos extensos, como diretrizes clínicas completas, podendo levar à omissão de dados cruciais.

A **atualização da base de conhecimento** também é um aspecto fundamental, pois o sistema pode se apoiar em evidências desatualizadas, como ocorre em casos que ainda seguem recomendações ultrapassadas sobre o uso de determinados medicamentos. A qualidade final da resposta está diretamente relacionada à qualidade dos documentos disponíveis. Caso o repositório contenha informações defasadas, isso será inevitavelmente refletido nas respostas geradas.

Além disso, há o **custo computacional elevado**, que pode comprometer a usabilidade em contextos de tempo crítico. No RAG, este custo se divide nas duas etapas principais: a recuperação de documentos e a geração de texto. Na primeira, o sistema busca os documentos mais relevantes para uma consulta, o que pode ser feito com métodos como BM25, menos custosos, ou com busca densa baseada em embeddings vetoriais, que exige mais recursos computacionais, especialmente em bases grandes. Na segunda

etapa, os documentos recuperados são usados como contexto para um modelo de linguagem gerar uma resposta, sendo essa fase impactada pelo tamanho do modelo, pela janela de contexto e pela estratégia de decodificação utilizada. Modelos maiores, mais precisos, tendem a ser mais caros em termos de processamento. Adicionalmente, o número de documentos usados como entrada afeta diretamente o tempo e a memória consumidos na geração.

Outro ponto de atenção é a **avaliação da qualidade das respostas**, que não pode se limitar às métricas automáticas; é necessário considerar completude, relevância clínica e impacto potencial da informação gerada. Há o risco da propagação de vieses ou desinformação presentes na base de documentos, o que pode reproduzir práticas ultrapassadas ou discriminatórias, com sérias implicações clínicas e éticas.

Diante desses desafios, torna-se essencial compreender o papel de cada componente do pipeline RAG, a fim de orientar a adoção de estratégias que mitiguem ou eliminem suas limitações. Essa compreensão é particularmente crítica em aplicações de perguntas e respostas na área clínica. Com o objetivo de oferecer uma visão abrangente, este capítulo está estruturado da seguinte forma:

Na Seção 6.2, são apresentados casos de uso de LLMs aplicados à área da saúde, utilizando o conceito de RAG para responder a perguntas clínicas. A Seção 6.3 detalha os diversos componentes do RAG, iniciando pela segmentação de texto (Seção 6.3.1). Na sequência, a Seção 6.3.2 aborda os modelos de embeddings fundamentais para a recuperação densa. A Seção 6.3.3 discute as estratégias de recuperação empregadas pelo *Retriever*, incluindo abordagens tradicionais, densas e híbridas. O *Generator*, responsável pela geração da resposta com base no contexto recuperado, é explorado na Seção 6.3.4, que também apresenta técnicas de geração de texto, engenharia de *prompts* e ajustes de modelos (*fine-tuning*) voltados a domínios sensíveis como o da saúde. A avaliação das respostas e da contribuição de cada componente para o desempenho do sistema é tratada na Seção 6.4. Por fim, a Seção 6.5 conclui o capítulo e destaca oportunidades de pesquisa futuras no contexto da saúde.

## 6.2. RAG no domínio da saúde

Os LLMs têm o potencial de transformar significativamente a forma como informações médicas são disseminadas. Entretanto, limitações como a dificuldade em acessar dados atualizados e privados e a possibilidade de gerar informações incorretas (alucinações) reduzem sua utilização em contextos clínicos, onde precisão e confiabilidade são cruciais. A abordagem RAG surge como uma alternativa promissora, ao integrar aos LLMs fontes externas de conhecimento. Com isso, os modelos podem acessar informações relevantes e atualizadas fora de seu treinamento original, ampliando sua aplicabilidade em cenários médicos. Embora o uso do RAG na área da saúde ainda esteja em estágio inicial, ele oferece grande potencial para transformar a comunicação e o acesso a informações clínicas [Ng et al. 2025].

Nesta são apresentados trabalhos que aplicam RAG dentro do domínio da saúde. Conceitos relacionados aos componentes RAG que estes trabalhos exploram serão tecnicamente abordados posteriormente na Seção 6.3.

### 6.2.1. ChatDoctor

[Li et al. 2023] desenvolveram o *ChatDoctor*, um chatbot capaz de responder a perguntas que um paciente faria ao seu médico. A Figura 6.2 representa os principais processos envolvidos na construção do ChatDoctor: (1) coleta de um conjunto de dados contendo conversas entre pacientes e médicos; (2) treinamento do modelo por *fine-tuning*; (3) uso de bases externas para enriquecer as respostas, abordagem que os autores denominam *Autonomous Knowledge Retrieval*; e (4) avaliação da performance do modelo.

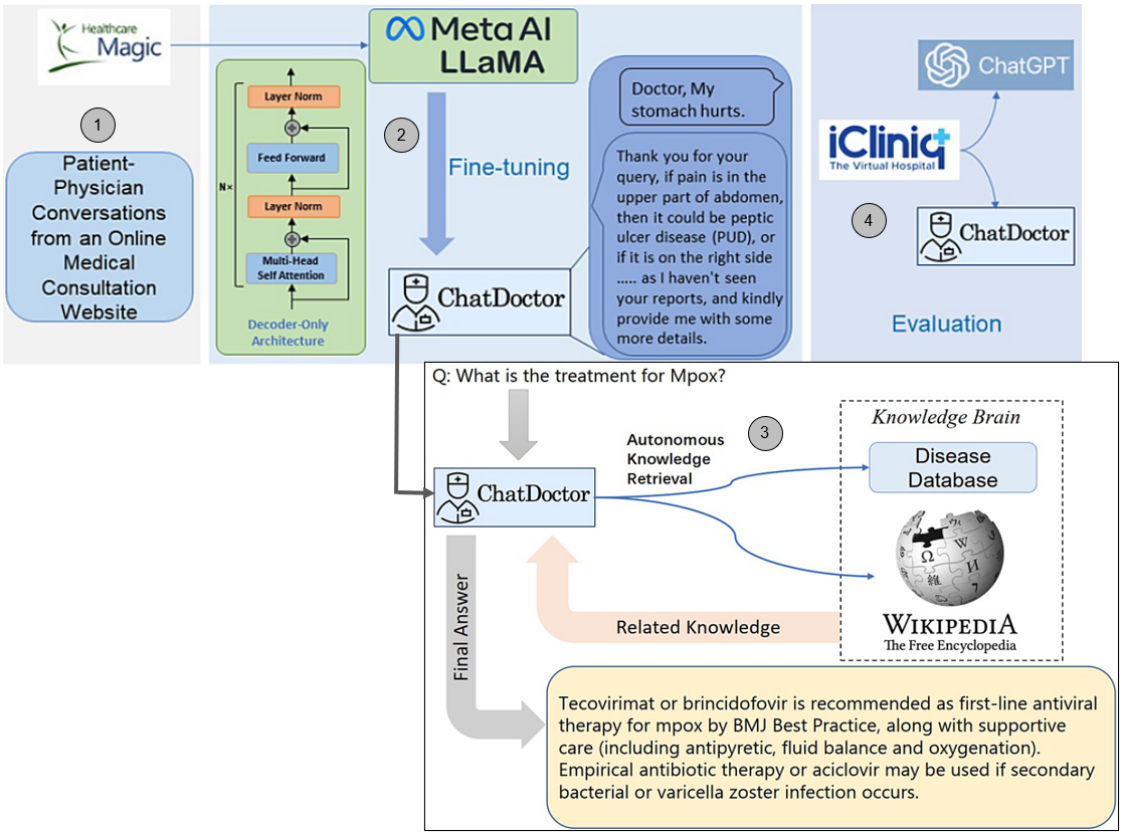



**Figura 6.2. Resumo dos processos do Chatdoctor. Adaptado de [Li et al. 2023]**

Na etapa (1) referente a coleta de dados, os autores utilizaram um grande conjunto de dados de 100.000 diálogos entre pacientes e médicos, provenientes de uma plataforma de consulta médica online[1] amplamente utilizada. As conversas foram limpas e anonimizadas para respeitar as questões de privacidade. Estes dados foram utilizados na etapa (2) para ajuste do modelo público LLaMA-7B disponibilizado pela Meta.

Além do refinamento do modelo, foi incorporado um mecanismo de recuperação de informações representado na etapa (3) da Figura 6.2. Ao utilizar o *ChatDoctor*, informações de fontes online como a Wikipédia, e dados de bancos de dados médicos offline previamente selecionados, são acessados em tempo real. Essas informações são incorporadas ao que é enviado ao LLM. Para recuperar as informações das bases externas os autores inicialmente usaram o LLM para identificar palavras chaves da pergunta

[1] www.healthcaremagic.com

como pode ser observado na Figura 6.3 no *prompt A*. Os termos identificados são usados então na busca por documentos que contenham os mesmos. Os textos foram divididos em seções iguais num tamanho que o LLM conseguiria administrar. As seções contendo as palavras-chaves são ordenadas pelo número de ocorrências destes termos, e as $N$ primeiras seções são adicionadas ao *prompt*. O modelo *ChatDoctor* usa as cinco primeiras ($N = 5$), usando o *prompt B* da Figura 6.2 para selecionar as informações pertinentes. Por fim, a pergunta e as informações selecionadas são adicionadas ao *prompt C*, que solicita a geração da resposta final.

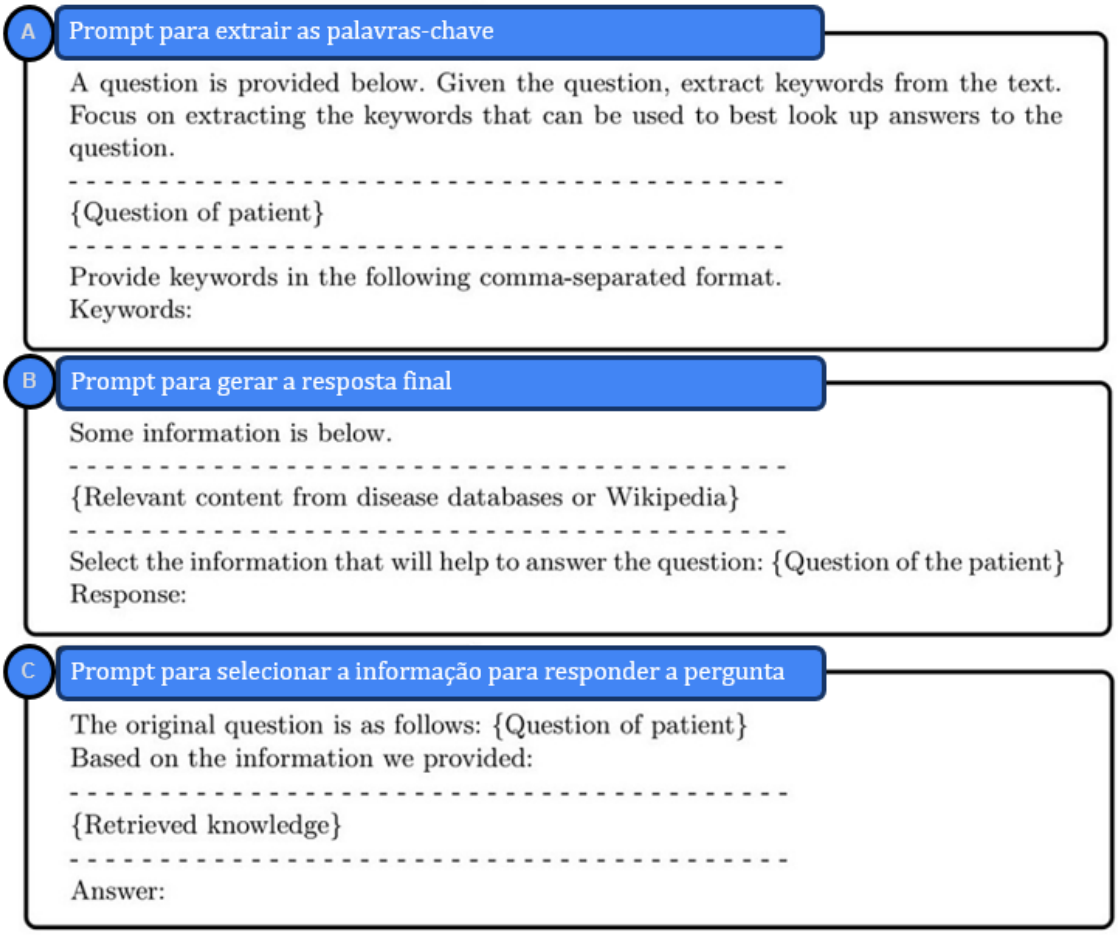



**Figura 6.3. *Prompts* usados no *Chatdoctor*. Adaptado de [Li et al. 2023]**

Para uma avaliação quantitativa do desempenho do *ChatDoctor*, representado na etapa (4) da Figura 6.2, foram utilizadas aproximadamente 10 mil conversas adicionais de um site independente de consulta médica online, o iCliniq[2]. Tratam-se de perguntas efetuadas por pacientes e as respectivas respostas de médicos que servem como referência para avaliação. Foram geradas as respostas usando o *ChatDoctor* e o ChatGPT (usando o modelo gpt-3.5-turbo). Para avaliar a similaridade semântica entre as respostas dos modelos e as respostas de referência, a métrica BERTScore [Zhang et al. 2020] foi utilizada. O *ChatDoctor* apresentou desempenho significativamente superior ao ChatGPT.

[2]https://www.icliniq.com/

Além disso, os autores destacam sua capacidade de lidar com doenças emergentes. Para demonstrar esse aspecto, o modelo foi testado com uma variedade de consultas médicas sobre temas recentes. Um exemplo é uma pergunta relacionada à "Varíola dos Macacos", termo recém designado pela Organização Mundial da Saúde (OMS) em 28 de novembro de 2022. Por se tratar de uma designação recente, a versão do ChatGPT utilizada à época não foi capaz de reconhecer o termo, ao contrário do *ChatDoctor*.

### 6.2.2. Almanac

Outro framework para RAG aplicado a dados clínicos é o *Almanac* [Zakka et al. 2024], que combina a capacidade dos LLMs com recuperação aumentada, sendo projetado para apoiar a tomada de decisões clínicas com segurança e precisão. O *Almanac* é composto por diversos componentes, conforme ilustrado na Figura 6.4. A pergunta feita no início do fluxo (1) é simplificada com o auxílio de um LLM (GPT-4), que gera termos ou palavras-chave para pesquisa. Esses termos são inseridos em um navegador especializado (2), que acessa exclusivamente fontes verificadas, como PubMed, UpToDate e BMJ Best Practice. Apenas fontes previamente registradas no *Almanac* como confiáveis são consultadas durante o processo.

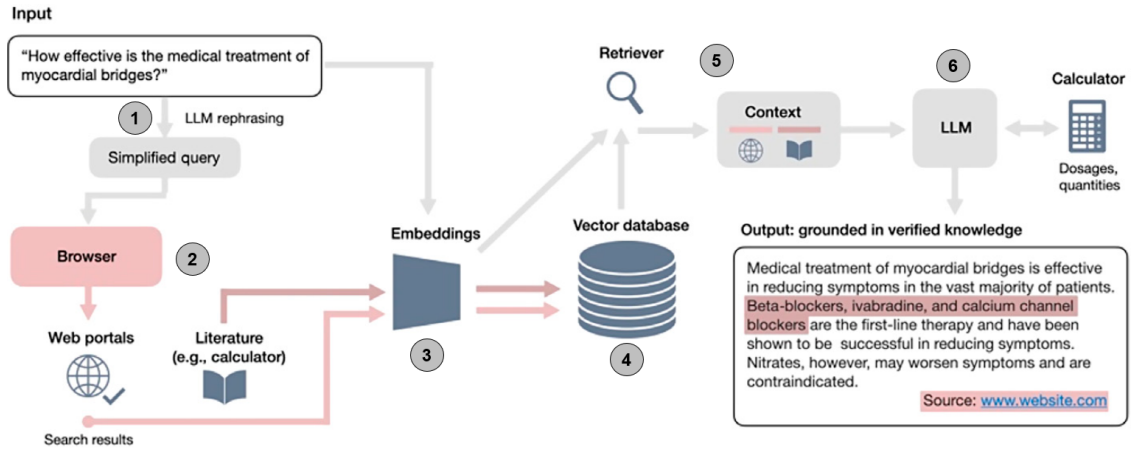



**Figura 6.4. Visão geral do *Almanac*. Adaptado de [Zakka et al. 2024]**

Os documentos recuperados pelo *Almanac* são divididos em fragmentos de 1000 tokens. As representações numéricas (embeddings) dos fragmentos são geradas, correspondente à etapa (3) na Figura 6.4. Cada fragmento tem os embeddings gerados separadamente usando o modelo text-embedding-ada-002 da OpenAI e armazenados no banco de dados vetorial Qdrant[3] (4). O Qdrant é então utilizado como *Retriever* para recuperar conteúdo relevante à pergunta na base de dados usando a similaridade de cosseno (5). O Qdrant também é inicializado com mais de 500 calculadoras clínicas. Essas calculadoras são obtidas diretamente do MDCalc[4] e suas indicações e instruções clínicas são usadas como metadados para recuperação. O LLM é usado em vários momentos com diferentes *prompts*. Alguns exemplos estão apresentados na Tabela 6.1. Inicialmente o LLM é usado para identificar os termos-chave na consulta (1) e depois para sintetizar as respostas finais

[3]https://qdrant.tech/
[4]https://www.mdcalc.com/

com citações no texto e para utilizar as calculadoras relevantes dependendo dos termos da consulta (6). Os autores não esclarecem quanto à persistência dos documentos indexados no Qdrant, ou seja, por quanto tempo os fragmentos indexados são mantidos na base.

**Tabela 6.1. Exemplos de *prompts* usados nas diferentes etapas do *Almanac*.**

| Etapa | Prompt |
|---|---|
| Simplificação da consulta | `Given question (Q) convert it to a simple Google search term.` |
| Geração da Resposta completa | `Generate a thorough and concise answer for a given question (Q) on the basis of the provided context (C). If you are asked to calculate a value, output the final equation in Python code. Use an unbiased and journalistic tone. If you cannot find a relevant answer, write "I apologize but there doesn't seem to be a clear answer to your question based on my sources ... Answer the question based on your own knowledge." Cite sources as [1] or [2] or [3] after each sentence to back up your answer (Ex: Correct: [1], Correct: [2][3], Incorrect: [1, 2]).` |
| Sumarização da Resposta | `Given the question (Q), context (C), and output (O), generate a thorough and concise answer for the given question (Q).` |

Além do framework, os autores compilaram um novo benchmark, o ClinicalQA, com 314 perguntas clínicas abertas desenvolvidas por 10 especialistas. Diversas especialidades médicas estão contempladas no dataset, com tópicos que vão desde diretrizes de tratamento até cálculos clínicos. Para cada pergunta, foram geradas as respostas usando *Almanac*, Bing, ChatGPT e Bard. Cada especialista avaliou as respostas geradas pelos quatro modelos dentro de três dimensões apresentadas na Tabela 6.2. Para cada pergunta dentro de cada dimensão o especialista numerou os modelos de 1 (melhor) a 4 (pior). Ao final, foi efetuada a média de cada modelo e *Almanac* superou significativamente os demais em todas as dimensões. Também foi avaliada a veracidade das citações geradas nas respostas, pois todos os modelos foram solicitados a fornecer as referências, neste aspecto *Almanac* também superou os demais.

### 6.2.3. Triagem para ensaios clínicos

A triagem de pacientes para ensaios clínicos é tradicionalmente realizada de forma manual, sendo propensa a erros e altamente demorada. Buscando minimizar estes problemas, [Unlu et al. 2024] propõem RECTIFIER (*RAG-Enabled Clinical Trial Infrastructure for Inclusion Exclusion Review*), um sistema RAG baseado no GPT-4 visando melhorar a precisão, a eficiência e a relação custo-benefício da determinação da elegibilidade para ensaios clínicos usando dados não estruturados de prontuários eletrônicos. O caso de

**Tabela 6.2. Métricas usadas na avaliação do *Almanac* e demais modelos**

| | |
|---|---|
| Factualidade | A resposta está de acordo com as práticas padrão e o consenso estabelecido pelos órgãos de autoridade em sua área de atuação?<br>Se aplicável, a resposta contém etapas de raciocínio corretas? |
| Completude | A resposta aborda todos os aspectos da questão?<br>A resposta omite algum conteúdo importante?<br>A resposta contém algum conteúdo irrelevante? |
| Preferência | Qual resposta você preferiu no geral? |

uso trata da triagem de pacientes para o estudo COPILOT-HF (focado em pacientes com insuficiência cardíaca), onde muitos critérios de elegibilidade dependem de prontuários clínicos em texto livre.

Para avaliar a elegibilidade dos pacientes, os pesquisadores desenvolveram um conjunto de perguntas a serem respondidas automaticamente com base em anotações clínicas existentes dos últimos dois anos. As anotações foram divididas em fragmentos com menos de 1000 tokens usando a estratégia de segmentação recursiva do LangChain [5] para preservar o contexto e evitar o truncamento de frases ou de palavras no meio. Em seguida foram geradas as representações vetoriais numéricas (embeddings) para cada fragmentos das anotações clínicas usando o modelo text-embedding-ada-002 da OpenAI. Para otimizar a recuperação durante a etapa de perguntas e respostas, utilizou-se a biblioteca AI Similarity Search (FAISS) do Facebook[6].

Foram criados *prompts* relacionados aos critérios alvo (perguntas de elegibilidade) para cada paciente. As perguntas foram transformadas em embeddings, que serviram como consultas de pesquisa no banco de dados de vetores, usando LangChain. As buscas recuperaram os três fragmentos das anotações clínicas mais relevantes com base na similaridade semântica de cada pergunta, juntamente com os links para as anotações da fonte original. O modelo não considerou a atualidade das anotações e simplesmente recuperou as mais relevantes. Esses fragmentos recuperados foram então combinados com uma instrução, conforme o exemplo da Tabela 6.3, que trata dos critérios exclusivos. A instrução e os fragmentos das anotações clínicas foram enviados para o modelo GPT-4 com temperatura de 0, o que gerou respostas binárias de "Sim"ou "Não".

Os pacientes foram divididos em 100 (desenvolvimento) e 282 (validação) para investigação dos melhores *prompts* a utilizar, e 1894 (conjunto de teste), dos quais 1509 possuíam anotações clínicas suficientes e que foram de fato utilizados na avaliação. Para comparar o desempenho do RECTIFIER com o desempenho da equipe do estudo para triagem, um clínico especialista efetuou uma revisão cega dos pacientes do conjunto de testes e respondeu às perguntas dos critérios-alvo (Sim/Não) sem acessar as respostas do RECTIFIER nem da equipe de estudo, estabelecendo assim respostas de referência (*gold standard*).

As respostas da equipe do estudo e do RECTIFIER foram bastante alinhadas com o *gold standard* com precisão variando entre 97,9% e 100% para o RECTIFIER e entre

[5]http://langchain.com/

[6]https://github.com/facebookresearch/faiss

**Tabela 6.3. Exemplo do *prompt* usado para as perguntas dos critérios de exclusão dos pacientes**

```
Can you please answer each of the following 12 questions
based on the information in the clinical notes with
only "Yes"or "No"?  Please return the answers in a comma
separated list.
1) Does the patient have unrepaired severe aortic stenosis,
or have unrepaired severe aortic valve insufficiency?
Repair includes aortic valve surgery and TAVR.
2) Does the patient have a history of known amyloid heart
disease?
3) Does the patient currently have WHO Group 1 pulmonary
arterial hypertension on disease-specific therapies like
Ambrisentan, Bosentan, Epoprostenol, Treprostinil, Iloprost
(do not include if the patient is on ONLY Sildenafil or
Tadalafil as disease-specific therapies)?
4) Is the patient currently getting chemotherapy or
hormonal therapy due to an active malignancy?
5) Is the patient currently receiving or will receive
hospice care?  Answer only for the patient, not for the
family members.
6) Does the patient have a history (Hx) of solid organ
transplant, being evaluated for transplant, or currently
on wait list above at the UNOS status level above 4?
7) Does the patient currently use a Ventricular Assist
Device?
8) Does the patient have a history of established
hypertrophic cardiomyopathy?
9) Does the patient have a history of Type 1 Diabetes?
10) Is the patient currently undergoing dialysis?
11) Is the patient currently pregnant or breastfeeding?
12) Does the patient have a history of Congenital Heart
Disease?
```

91,7% e 100% para a equipe do estudo. O RECTIFIER apresentou desempenho superior ao da equipe do estudo na determinação de insuficiência cardíaca sintomática, com precisão de 97,9% versus 91,7%. No geral, a sensibilidade e a especificidade para determinar a elegibilidade do paciente com o RECTIFIER foram de 92,3% e 93,9%, respectivamente, e 90,1% e 83,6% com a equipe do estudo.

### 6.2.4. Procedimentos Operacionais Padrão

Os profissionais de saúde frequentemente devem seguir procedimentos operacionais padrão (POPs) que estabelecem, por exemplo, protocolos internos na comunicação e disseminação de informações [Kuriki P. and R. 2024]. Os métodos de busca tradicionais têm dificuldades com consultas formuladas em linguagem natural, especialmente em meio a

grandes volumes de dados, o que destaca a necessidade de um sistema de recuperação mais eficaz. Neste contexto, a abordagem RAG aprimora o gerenciamento de informações, permitindo uma recuperação precisa e baseada em linguagem natural.

[Kuriki P. and R. 2024] apresentaram, durante a SIIM24[7] (Reunião Anual da Sociedade de Informática de Imagem em Medicina), um chatbot clínico baseado na abordagem RAG. O sistema demonstrou melhorias na precisão da recuperação de Procedimentos Operacionais Padrão (POPs), ilustrando seu potencial de aplicação na indústria farmacêutica para garantir conformidade, reduzir erros e agilizar atualizações de protocolos. Para sua implementação, foi utilizado o modelo de linguagem Mistral-7B, e o fluxo geral do sistema pode ser visualizado na Figura 6.5.

Os textos das POPs foram convertidos em embeddings que foram armazenados em um banco de dados vetorial chamado SOP2vec, permitindo a busca por similaridade. Uma interface web de chatbot permite que os usuários interajam com os documentos usando linguagem natural. As perguntas são convertidas em embeddings que são comparadas com o banco de dados vetorial, recuperando fragmentos de contexto que são adicionados ao *prompt*. O *prompt* é enviado a um modelo Mistral-7B implantado localmente, gerando uma resposta para o usuário. Os autores escolheram a abordagem local por questões de desempenho e segurança no manuseio de informações sensíveis.

Os autores relatam a implementação de um processo de avaliação automático usando GPT-4. Foram geradas de 10 a 20 perguntas por documento pertencentes a um conjunto de POPs. Essas perguntas foram aplicadas ao *pipeline* RAG+LLM, e cada resposta foi avaliada pelo GPT-4 quanto à acurácia da recuperação e à relevância da resposta. Segundo os autores, esse processo permitiu a identificação de falhas, levando a melhorias significativas, por exemplo, a otimização dos embeddings, refatoração de perguntas, aplicação de reranking, melhorias em metadados e correções de erros nas POPs.

## 6.3. Componentes do RAG

Nesta seção, serão apresentados os componentes e processos da abordagem RAG. O RAG tem duas macroetapas principais: a recuperação (*Retriever*) e a geração (*Generator*, também conhecida como *Reader* na literatura, mas referida neste capítulo como *Generator*) , conforme ilustrado na Figura 6.6.

O *Retriever* engloba todas as tarefas relacionadas à obtenção de contexto adicional de bases externas que possam auxiliar a responder à pergunta. O *Generator* recebe a pergunta e um conjunto de documentos ou partes de documentos (o contexto adicional) e gera a resposta.

O conjunto de documentos funciona como uma espécie de extensão (aumento) do conhecimento paramétrico do LLM, pois explora o aprendizado baseado no contexto (*In-Context Learning*, ICL). ICL refere-se à capacidade de modelos de linguagem, como o GPT, de aprender e realizar tarefas apenas a partir do contexto fornecido na entrada (*prompt*), sem a necessidade de ajustar os pesos do modelo por meio de um novo treinamento. Assim, em vez de treinar o modelo novamente, são fornecidos exemplos no próprio *prompt* (*few-shot learning*), e o modelo utiliza esses exemplos para gerar a res-

[7]https://siim.org/

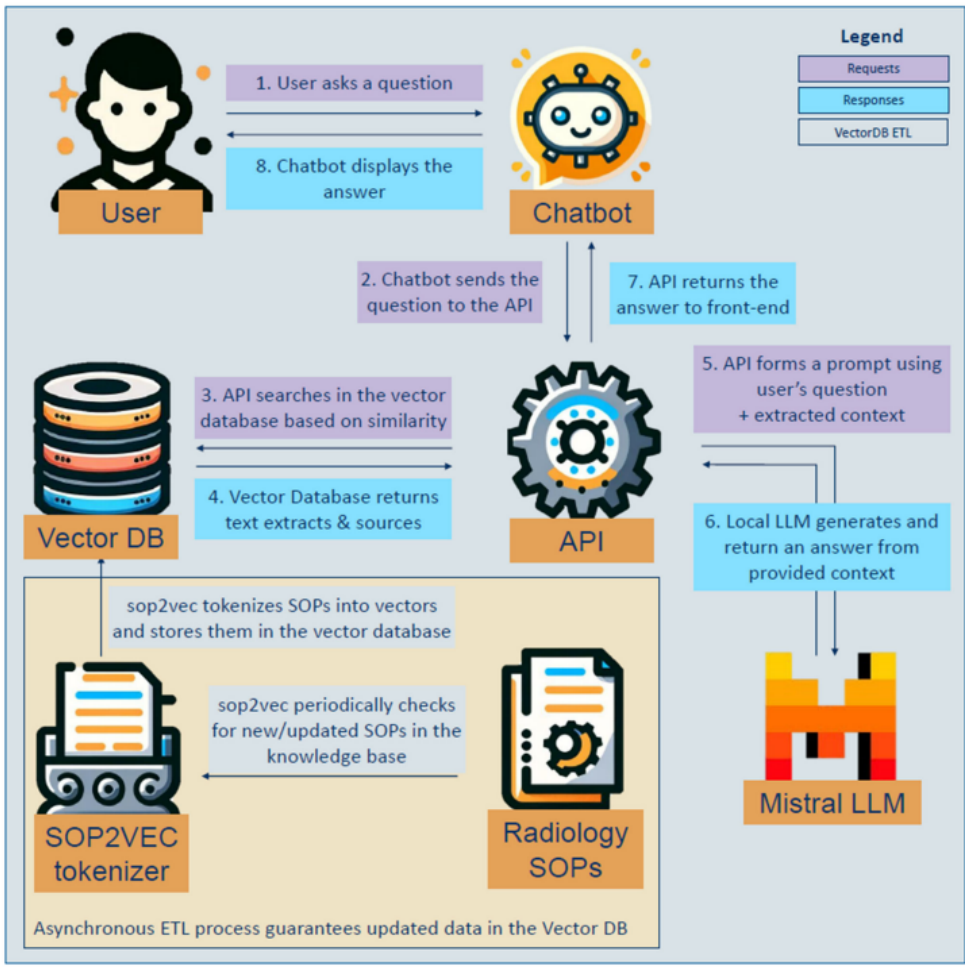



**Figura 6.5. RAG para Procedimentos Operacionais Padrão (POPs) na saúde. Extraído de [Kuriki P. and R. 2024]**

posta adequada [Brown et al. 2020].

Existem vários componentes dentro das macroetapas do *Retriever* e do *Generator*, utilizadas quando o sistema está "em produção". Mas também há uma série de etapas no estágio pré-produção, envolvendo especialmente o preparo dos documentos usados pelo RAG, como pode ser observado na Figura 6.7 que apresenta em mais detalhes um *pipeline* RAG. Cada etapa será explicada de forma geral nesta seção e mais detalhadamente nas seções subsequentes.

### 6.3.1. Segmentação do texto

Dividir grandes volumes de texto em blocos menores e de mais fácil processamento é uma prática conhecida como segmentação. Esse processo é amplamente conhecido pelo termo em inglês *chunking* e assim nos referiremos a ele neste capítulo. Embora seja uma etapa fundamental em diversas aplicações de processamento de linguagem natural e recuperação de informação, muitas vezes é deixada em segundo plano. Na Figura 6.7 o *chunking* faz parte da etapa de pré-produção, onde os documentos da base de dados serão processados e segmentados para posterior indexação. Um *chunk* se refere a um segmento

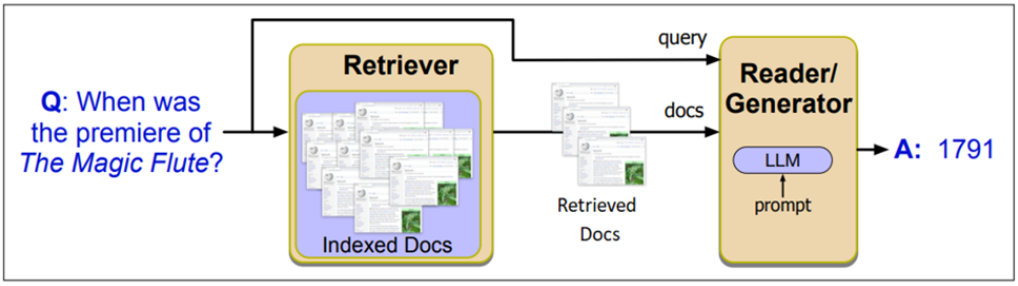



**Figura 6.6. As macroetapas do RAG. Extraído de [Jurafsky and Martin 2025]**

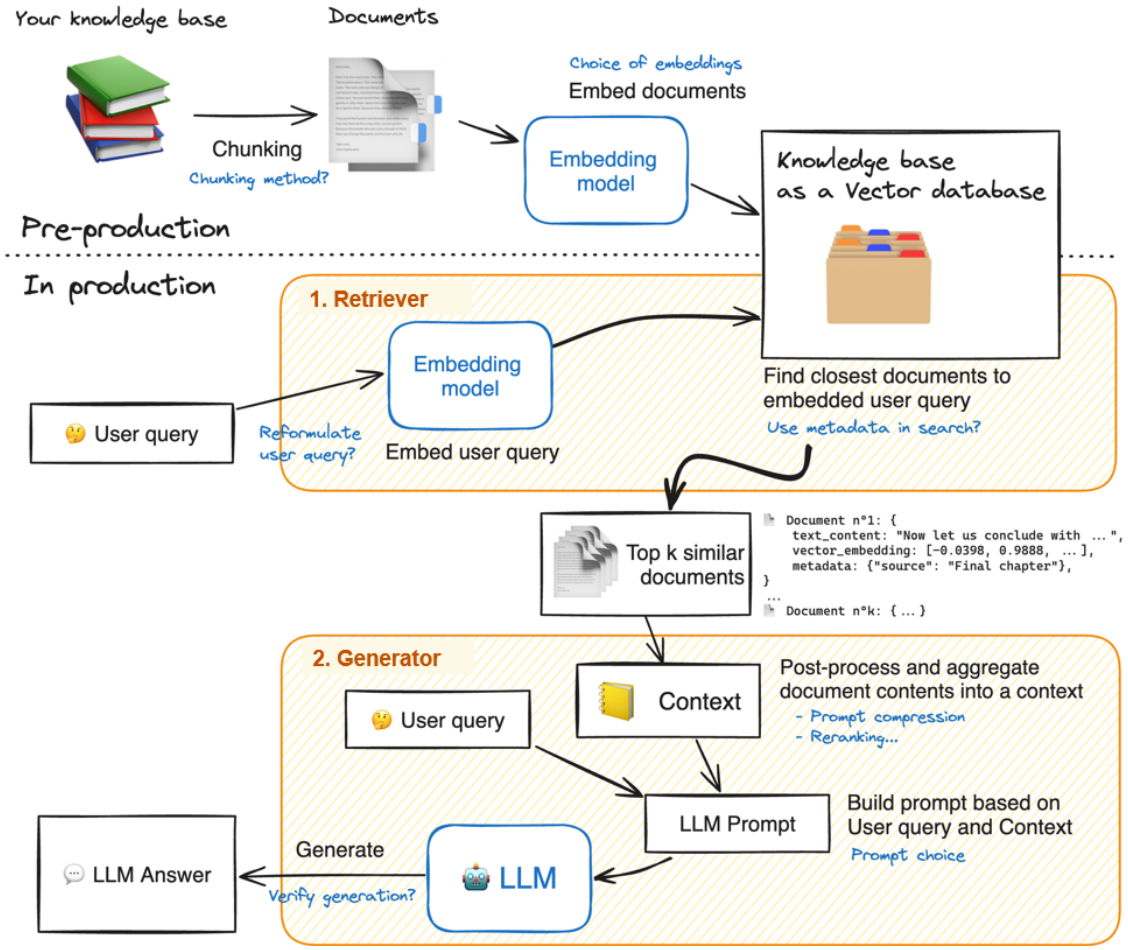



**Figura 6.7. As macroetapas do RAG. Adaptado de [Huggingface 2024]**

de texto.

A granularidade na qual um documento é segmentado desempenha um papel essencial, pois estratégias de *chunking* ineficazes podem levar a fragmentos com contexto incompleto ou excesso de informações irrelevantes, o que prejudica o desempenho dos modelos de recuperação [Duarte et al. 2024b]. Ao final, *chunks* adequados, coesos e semanticamente significativos melhoram a precisão da busca aumentando a probabilidade de respostas mais coerentes.

Neste contexto, uma das escolhas a fazer é o tamanho máximo do *chunk*, que impacta limites das janelas de contexto dos modelos usados tanto para representação do texto (modelos para geração de embeddings vistos na Seção 6.3.2) como dos LLMs na geração da resposta final (Seção 6.3.4.1). A janela de contexto é o tamanho máximo de entrada dos modelos, que é quantificada em número de tokens. Tokens podem ser palavras ou sub-palavras e representam as unidades mínimas de texto em que uma sequência textual é dividida para ser processada pelo modelo. Cada modelo utiliza um tokenizador (algoritmo responsável por dividir o texto em tokens) específico e portanto o número de tokens de um mesmo segmento de texto pode ser diferente para cada modelo. O E5[8] é um exemplo de modelo para extração de embeddings usado na recuperação densa (ver Seção 6.3.2) que possui 512 tokens de entrada, enquanto o modelo text-embedding-3-large[9] da Open AI tem 8.191 tokens.

As janelas de contexto dos LLMs usados na resposta também precisam ser consideradas. O GPT-4o possui uma janela de 128 mil tokens, o Mistral-7B-Instruct-v0.2 tem 32 mil tokens, já o LLamma 2 e GPT-3.5-turbo ambos tem 4.096 tokens. A ideia do RAG é recuperar *chunks* relevantes para a pergunta visando compor o contexto que será enviado ao LLM juntamente com a pergunta. Dessa forma, *chunks* muito longos podem trazer problemas na entrada dos modelos.

Uma abordagem bastante comum e simples é utilizar um comprimento fixo de *chunk* e o texto de um documento é dividido em blocos com no máximo esse tamanho. Geralmente essa técnica é usada considerando uma sobreposição. Na Figura 6.8 podem ser observadas diferentes estratégias de *chunking* para o texto "*A gripe é uma infecção aguda do sistema respiratório, provocado pelo vírus da influenza, com grande potencial de transmissão. Existem quatro tipos de vírus influenza/gripe: A, B, C e D. O vírus influenza A e B são responsáveis por epidemias sazonais, sendo o vírus influenza A responsável pelas pandemias.*". A estratégia "A"usa um tamanho de 15 tokens com nenhum recobrimento. A estratégia "B"usa o mesmo número de tokens e uma sobreposição de 20% (os 3 primeiros tokens do *chunk* são os 3 últimos do *chunk* imediatamente anterior). Já na estratégia "C"os *chunks* são separados por frases.

Uma estratégia de *chunking* comumente empregada em sistemas RAG é a fragmentação recursiva, cujo objetivo é preservar a integridade semântica dos trechos recuperáveis. O método utiliza marcadores sintáticos (por exemplo vírgulas, pontos finais e quebras de parágrafo) como delimitadores naturais para particionar o conteúdo. A recursividade está no fato de que se o trecho ainda for grande, tenta-se quebrar novamente em unidades menores, respeitando limites linguísticos naturais. O processo continua até que os *chunks* estejam dentro do limite de tamanho definido. Isso ajuda a manter coerência semântica dentro de cada segmento e evita que partes relacionadas fiquem separadas em *chunks* diferentes. Frameworks como LangChain disponibilizam *chunking* recursivo[10]. Por ser computacionalmente eficiente, mostra-se apropriado para *pipelines* em larga escala, sobretudo quando aplicado a textos bem estruturados. Ao preservar a coesão semântica entre segmentos adjacentes, essa abordagem contribui para a geração

[8]https://huggingface.co/intfloat/multilingual-e5-large

[9]https://platform.openai.com/docs/guides/embeddings/embedding-models#embedding-models

[10]https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/recursive_text_splitter/

**A) 15 tokens 0% de recobrimento:**

| | |
|---|---|
| 1 | A gripe é uma infecção aguda do sistema respiratório, provocado pelo vírus da influenza |
| 2 | , com grande potencial de transmissão. Existem quatro tipos de vírus influenza/gripe |
| 3 | : A, B, C e D. O vírus influenza A e B |
| 4 | são responsáveis por epidemias sazonais, sendo o vírus influenza A responsável pelas pandemias. |

**B) 15 tokens e 20% de recobrimento:**

| | |
|---|---|
| 1 | A gripe é uma infecção aguda do sistema respiratório, provocado pelo vírus da influenza |
| 2 | vírus da influenza, com grande potencial de transmissão. Existem quatro tipos de vírus |
| 3 | tipos de vírus influenza/gripe: A, B, C e D. |
| 4 | e D. O vírus influenza A e B são responsáveis por epidemias sazonais, |
| 5 | epidemias sazonais, sendo o vírus influenza A responsável pelas pandemias. |

**C) Separando por frases:**

| | |
|---|---|
| 1 | A gripe é uma infecção aguda do sistema respiratório, provocado pelo vírus da influenza, com grande potencial de transmissão. |
| 2 | Existem quatro tipos de vírus influenza/gripe: A, B, C e D. |
| 3 | O vírus influenza A e B são responsáveis por epidemias sazonais, sendo o vírus influenza A responsável pelas pandemias. |

**Figura 6.8. Exemplos de estratégias de *chunking***

de representações vetoriais mais informativas e, consequentemente, para a melhoria da qualidade da recuperação e da geração de respostas [Lee et al. 2021].

No trabalho de [Kamradt 2024], é adotada uma abordagem baseada em embeddings densos para identificar *chunks* semanticamente próximos. O processo inicia com a segmentação inicial do texto por meio de uma técnica de fragmentação recursiva, que delimita sentenças de forma estruturada. Em seguida, para cada *chunk*, são geradas representações vetoriais utilizando uma janela deslizante composta por múltiplas sentenças consecutivas. A partir dessa sequência vetorial organizada, o método percorre iterativamente o conjunto, comparando janelas adjacentes com o objetivo de detectar transições semânticas e, assim, estabelecer pontos adequados de segmentação.

A abordagem denominada *LumberChunker* [Duarte et al. 2024a] explora a capacidade de raciocínio dos modelos de linguagem para realizar a segmentação de textos de forma a maximizar a independência semântica entre os *chunks* gerados. Nesse método, o texto é inicialmente separado em parágrafos, os quais são submetidos iterativamente a um LLM. O modelo recebe a tarefa de identificar o ponto em que ocorre uma mudança substancial de conteúdo em relação ao que foi previamente apresentado. Os parágrafos anteriores a essa transição são agrupados e definidos como um *chunk*. O processo é repetido em sequência até que todo o conteúdo seja organizado.

A estratégia de mistura de *chunkers* (*Mixture-of-Chunkers*, MoC) [Zhao et al. 2025] busca otimizar o *chunking* em sistemas RAG usando uma rede de especialistas leves (meta-*chunkers*) gerenciados por um roteador sensível à granularidade. Ele se destaca por gerar regras de *chunking* (expressões regulares) em vez de texto completo, o que reduz o custo computacional, e utiliza um algoritmo de pós-processamento para garantir a precisão da extração dos blocos. Os experimentos mostraram que o MoC melhora o desempenho do RAG em comparação com métodos anteriores.

[Ke et al. 2024] implementaram um sistema RAG com GPT4-o para diretrizes pré-operatórias. As respostas do RAG foram comparadas a respostas geradas por humanos atingindo bons resultados. Esse trabalho utilizou embeddings do modelo text-embedding-ada-002 (modelo com entrada de 8191 tokens). Foi utilizado o *chunking* recursivo do Langchain configurando tamanho máximo de 1000 tokens e um overlap de 100 tokens.

### 6.3.2. Modelos de embeddings

De forma geral embeddings são vetores de números reais que visam representar dados discretos (palavras, itens, usuários ou imagens). Em aprendizado profundo, os embeddings podem ser definidos como representações vetoriais densas (a maioria ou praticamente todos os valores dentro do vetor são diferentes de zero) e de dimensão fixa utilizadas para codificar os dados em um espaço contínuo de alta dimensionalidade. O objetivo é capturar relações semânticas entre os dados que estes vetores representam, de forma que itens com significados ou comportamentos semelhantes fiquem próximos entre si nesse espaço vetorial.

Métodos como Word2Vec [Mikolov et al. 2013] e GloVe [Pennington et al. 2014] propuseram técnicas para o aprendizado de embeddings, atribuindo um vetor fixo a cada palavra no vocabulário do corpus utilizado durante o treinamento. Esses vetores são aprendidos automaticamente durante o treinamento destes modelos e capturam muito bem a sinonímia, entretanto não resolvem a polissemia (uma mesma palavra pode ter diferentes sentidos, como por exemplo a palavra "banco": banco da praça, banco de areia, banco como instituição financeira, banco de dados, banco do verbo bancar).

Nos modelos baseados na arquitetura Transformer [Vaswani et al. 2017], como o BERT [Devlin et al. 2019], os embeddings passaram a ser chamados de contextuais, pois o vetor atribuído a uma palavra pode variar conforme o contexto em que ela aparece, resolvendo assim o problema da polissemia. A capacidade destes modelos baseados em atenção de gerar embeddings contextuais reside em seu mecanismo de atenção que simula a forma como os humanos focam em diferentes partes de uma informação ao processá-la. Dessa forma, o modelo pondera a importância de cada palavra em uma sequência (por exemplo, uma frase) ao gerar a representação vetorial de outra palavra na mesma sequência. Em vez de tratar todas as palavras igualmente, a atenção permite que o modelo foque nas palavras mais relevantes do contexto da palavra que está sendo codificada.

Nos sistemas RAG, os modelos de embeddings são utilizados em dois momentos conforme apresentado na Figura 6.7: nos processos de pré-produção, onde são usados para gerar os vetores de toda a base de dados que será consultada posteriormente; e em produção, na codificação dos vetores da pergunta que serão utilizadas pelo *Retriever* para recuperar vetores semanticamente similares na base de dados.

Muitos modelos de embeddings modernos (como SBERT[11], E5[12], BGE[13], GTE[14]) são construídos sobre a mesma arquitetura Transformer que os LLMs. Eles são "gran-

[11]https://huggingface.co/sentence-transformers
[12]https://huggingface.co/intfloat/multilingual-e5-large
[13]https://huggingface.co/BAAI/bge-large-en-v1.5
[14]https://huggingface.co/Alibaba-NLP/gte-modernbert-base

des"no sentido de terem muitos parâmetros e treinados em vastos volumes de dados textuais. Os modelos de embeddings "modelam"a linguagem para criar suas representações. Assim, teoricamente, qualquer modelo que aprenda a partir de dados de linguagem e capture aspectos dela poder ser considerado um "modelo de linguagem". Neste capítulo, quando usamos o termo LLMs estamos nos referindo aos modelos utilizados na geração da resposta em sistemas RAG.

Existem diversos modelos de embeddings disponíveis. A capacidade semântica destes modelos são avaliadas por meio de tarefas como recuperação de informação, agrupamento, classificação, etc. O MTEB (*Massive Text Embedding Benchmark*) é um benchmark reconhecido onde é possível encontrar o ranking de modelos de embeddings. A maior parte dos modelos disponibilizados até este momento priorizam a língua inglesa, existindo alguns multilíngues. No painel do MTEB é possível verificar o desempenho destes modelos em datasets do domínio médico[15]. São encontrados datasets do domínio usados para avaliar estes modelos, como pode ser verificado na Tabela 6.4.

**Tabela 6.4. Datasets Médicos disponíveis no MTEB**

| **Dataset** | **Tipo de Tarefa** | **Linguagens** | **Domínios** |
|---|---|---|---|
| CMedQAv2-reranking | Reranking | mandarim | Médico |
| CUREv1 | Retrieval | inglês, árabe, chines, francês, coreano, russo, espanhol, vietnamita | Médico, Acadêmico |
| CmedqaRetrieval | Retrieval | mandarim | Médico |
| MedicalQARetrieval | Retrieval | inglês | Médico |
| Medrxiv-ClusteringP2P.v2 | Clustering | inglês | Acadêmico, Médico |
| Medrxiv-ClusteringS2S.v2 | Clustering | inglês | Acadêmico, Médico |
| NFCorpus | Retrieval | inglês | Médico, Acadêmico |
| PublicHealthQA | Retrieval | inglês, árabe, chines, francês, coreano, russo, espanhol, vietnamita | Médico, Governamental, Web |
| SciFact | Retrieval | inglês | Acadêmico, Médico |
| TRECCOVID | Retrieval | inglês | Médico, Acadêmico |

O dataset PublicHealthQA, por exemplo, é utilizado para medir a capacidade dos modelos em recuperar informações relevantes em contextos médicos e de saúde pública. O dataset contem perguntas em oito línguas (inglês, árabe, chines, francês, coreano, russo, espanhol, vietnamita). Na Tabela 6.5 alguns exemplos de perguntas e respostas em inglês e espanhol.

No RAG, os modelos de embeddings convertem os fragmentos de texto (chunks)

[15]https://mteb-leaderboard.hf.space/?benchmark_name=MTEB%28Medical%2C+v1%29

**Tabela 6.5. Exemplos do dataset PublicHealthQA**

| Pergunta | Resposta |
|---|---|
| What temperature kills the virus that causes COVID-19? | Generally coronaviruses survive for shorter periods at higher temperatures and higher humidity than in cooler or dryer environments. However, we don't have direct data for this virus, nor do we have direct data for a temperature-based cutoff for inactivation at this point. The necessary temperature would also be based on the materials of the surface, the environment, etc. Regardless of temperature please follow CDC's guidance for cleaning and disinfection. |
| ¿Cómo debo preparar a mis hijos por si se produce un brote de COVID-19 en nuestra comunidad? | Los brotes pueden ser estresantes tanto para los adultos como para los niños. Hable con sus hijos acerca del brote, intente mantener la calma y recuérdeles que se encuentran a salvo. Si lo considera apropiado, explíqueles que la mayoría de los casos COVID-19 parecen ser leves. Los niños responden de manera diferente a las situaciones estresantes que los adultos. Los CDC ofrecen recursos para conversar con los niños acerca del COVID-19. |

em vetores dentro de um espaço vetorial. Estes vetores são usados nas buscas semânticas baseadas numa métrica de similaridade, o que é discutido na Seção 6.3.3. Para isso ocorrer, os vetores precisam ser pré-processados e armazenados para estarem aptos a serem utilizados nas buscas. Existem uma série de desafios técnicos na gestão e uso dos vetores, entre eles estão [Pan et al. 2024]:

- Ambiguidade da similaridade semântica: é difícil definir de forma inequívoca o que é "semelhante", e diferentes funções de similaridade podem produzir rankings divergentes.
- Custo computacional elevado: comparações de similaridade são proporcionalmente caras.
- Ausência de propriedades estruturais: diferente de dados relacionais, vetores não são naturalmente ordenáveis.
- Consultas híbridas: dificuldade de combinar vetores e atributos tradicionais de forma eficiente.

Existem diversas opções de banco de dados vetoriais como Pinecone[16], Qdrant[17], CrhomaDB[18], Vespa[19], Milus[20], e ferramentas de busca que suportam o armazento veto-

[16]https://www.pinecone.io/
[17]https://qdrant.tech/
[18]https://www.trychroma.com/
[19]https://vespa.ai/
[20]https://milvus.io

rial como ElasticSearch[21], OpenSearch[22], etc. Segundo [Pan et al. 2024], para aplicações baseadas em RAG, a escolha de um sistema de banco de dados vetorial (*Vector Database Management Systems*, VDBMS) deve considerar uma combinação de requisitos que garantam desempenho, flexibilidade e integração com modelos de linguagem. Um VDBMS ideal para RAG precisa apresentar alta eficiência na execução de buscas por vizinhos mais próximos, tanto exatas (*k-Nearest Neighbors*, k-NN) quanto aproximadas (*Approximate Nearest Neighbors*, ANN), já que o processo de recuperação de contexto relevante depende diretamente da qualidade e velocidade dessas buscas. Além disso, é essencial que o sistema ofereça suporte eficiente a consultas híbridas, ou seja, consultas que combinem filtros estruturados com similaridade vetorial. A latência também é um fator crítico, pois o tempo de resposta da recuperação impacta diretamente a experiência do usuário em sistemas interativos, como chatbots e assistentes inteligentes.

Outro aspecto importante é a capacidade do VDBMS de lidar com atualizações dinâmicas. Idealmente, ele deve permitir inserções e remoções de vetores de forma eficiente, ou ao menos fornecer mecanismos claros de reindexação periódica que evitem degradação do desempenho ao longo do tempo. A existência de indexadores de alto desempenho como HNSW[23] (*Hierarchical Navigable Small World*), que combina eficiência de busca com boa qualidade de resultados é fundamental.

### 6.3.3. Recuperação da informação

O objetivo central da Recuperação da Informação (RI) é realizar buscas, isto é, encontrar documentos relevantes com base em uma consulta fornecida pelo usuário. Um dos principais desafios dessa tarefa é que os termos utilizados na consulta nem sempre correspondem aos termos presentes nos documentos relevantes. Segundo [Moreira 2024], esse problema é conhecido como incompatibilidade de vocabulário (*vocabulary mismatch*), e decorre de dois fenômenos linguísticos frequentes: a sinonímia, quando diferentes palavras expressam o mesmo significado, e a polissemia, quando uma mesma palavra possui múltiplos significados.

Essa etapa do RAG pode utilizar diferentes métodos, desde sistemas clássicos de recuperação de informações como BM25 (Best Match 25) até o uso de modelos de linguagem especializados em representações vetoriais (modelos de embeddings) viabilizando a recuperação densa [Jurafsky and Martin 2025]. A principal diferença entre a recuperação usando o modelo BM25 e a recuperação densa é a forma como a informação é representada e em como a similaridade entre a pergunta e os fragmentos de texto é calculada.

#### 6.3.3.1. BM25

O BM25 utiliza um algoritmo de recuperação de informação baseado na frequência de termos [Robertson et al. 2009]. Ele avalia a relevância de um documento para uma consulta com base na frequência com que os termos da consulta aparecem no documento. A

[21] https://www.elastic.co/

[22] https://opensearch.org/

[23] HNSW é um algoritmo baseado em grafos usado para busca aproximada de vizinhos mais próximos (ANN) em espaços de alta dimensão.

representação do texto é esparsa, assim cada termo é uma dimensão no vetor que representa um documento. O objetivo principal do BM25 é a correspondência exata de termos entre a consulta e os documentos e por isso não captura a semântica da linguagem.

Existem formas de minimizar esse problema expandindo os termos da consulta, mas isso depende da disponibilidade de grafos de conhecimento e dicionários especializados. Na Figura 6.9 estão representadas algumas dessas estratégias de expansão, como a utilizar de léxicos e até modelos de linguagem para gerar reformulações da consulta de forma a facilitar o sistema de recuperação em questão.

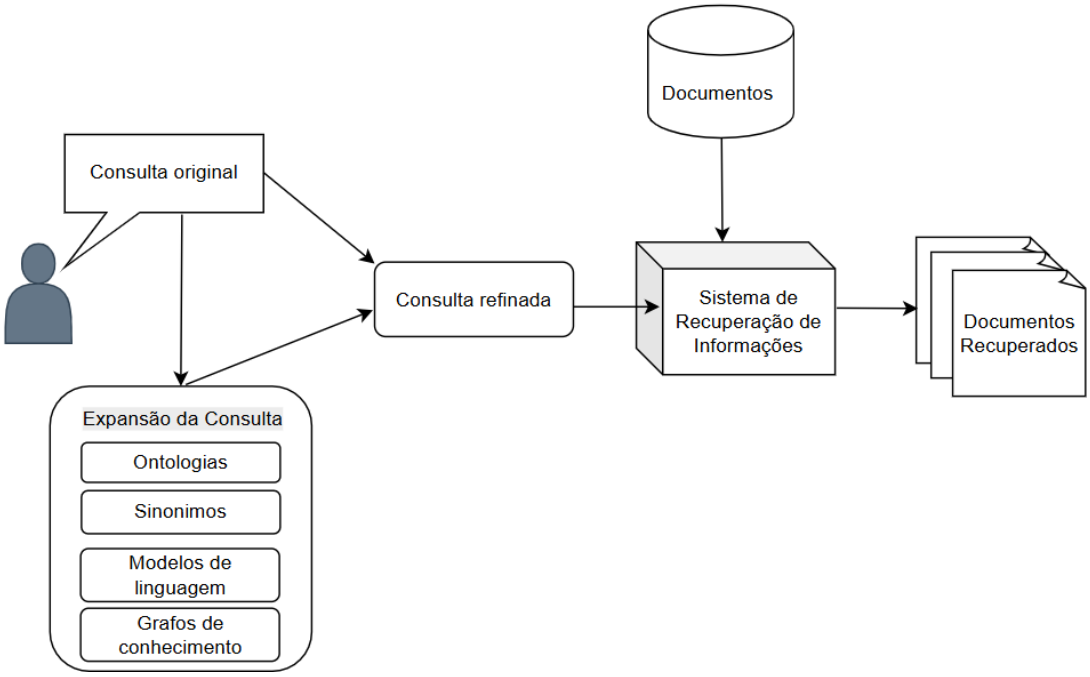



**Figura 6.9. Fluxo para expansão de consultas**

Outra maneira de expandir a consulta é denominada *Pseudo-Relevance Feedback*, onde o usuário faz uma consulta original, por exemplo "*asma crônica*", e o sistema executa BM25 e retorna os top-k documentos. A ideia baseia-se na suposição de que esses top-k documentos são relevantes — daí o termo "pseudo", já que não há uma verificação real com o usuário. A partir desses documentos, extraem-se os termos mais frequentes, como por exemplo: "*dispneia*", "*bronquite*", "*pulmão*" e "*tratamento*". Esses termos são combinados com a consulta original para formar uma nova consulta expandida: "*asma crônica bronquite pulmão tratamento*". Neste ponto, executa-se novamente o BM25, agora com a consulta expandida, e espera-se uma melhora na cobertura de sinônimos e contexto.

Como o BM25 utiliza diretamente os termos presentes nos documentos e nas consultas, o desempenho final do modelo depende da qualidade e relevância dessas palavras. Textos em geral apresentam muitos termos genéricos, flexões irrelevantes ou palavras não informativas (stop-words). Usar técnicas de pré-processamento do texto pode melhorar os resultados de forma significativa. As técnicas mais utilizadas são :

- Stemming: algoritmos [Orengo and Huyck 2001, Porter 1980] que visam unificar diferentes formas de uma mesma palavra ao reduzir seus sufixos, gerando uma representação comum. O principal benefício do stemming é aumentar a cobertura da busca, possibilitando a recuperação de um maior número de documentos potencialmente relevantes [Moreira 2024].

- Lematização: transforma uma palavra em sua forma canônica por exemplo: "correram"tem o lema "correr". Dessa forma, reduz-se o tamanho do vocabulário permitindo que variações de uma mesma palavra sejam reconhecidas como iguais. A qualidade do lematizador na língua em questão é fundamental para que se tenham bons resultados.

- Remoção de stop words: termos como “de”, “a”, “o”, “e”, etc. devem ser removidas por serem frequentes e pouco informativos. Mas em domínios específicos, há outras palavras altamente frequentes e genéricas (por exemplo: no domínio médico “paciente” e “sintoma”; em computação “usuário” e “sistema”). Assim, criar uma lista de stop words personalizada para cada corpus melhora a capacidade do BM25 de distinguir documentos relevantes.

- Normalização: se refere a remoção de acentos (diacríticos), padronização de caixa alta ou baixa. Esse processo visa garantir que termos semanticamente equivalentes não tratados de maneiras diferentes por variações ortográficas superficiais como versões em minúsculas e maiúsculas, ou com e sem acento.

#### 6.3.3.2. Recuperação Densa

A busca lexical, como observado com BM25, se baseia na correspondência exata de termos entre a consulta e os documentos, ignorando sinônimos e diferentes grafias. Por outro lado, a busca semântica (ou recuperação densa) transforma a consulta em um vetor e localiza os documentos cujos vetores são similares nesse espaço. Na Figura 6.10 há uma apresentação simplificada dessa ideia usando duas dimensões: cada ponto verde é um vetor de um documento no corpus. No momento da busca, a consulta é incorporada no mesmo espaço vetorial (ponto vermelho) e os embeddings mais próximos são encontrados (documento relevante).

A recuperação densa utiliza modelos de embeddings para representar consultas e documentos como vetores densos em um espaço de embeddings. Esses vetores capturam a semântica da linguagem, permitindo que o modelo encontre documentos relevantes mesmo que não haja correspondência exata de termos. A representação dos dados é dita densa porque cada dimensão do vetor de números reais que representa o documento não corresponde a um termo individual, mas sim a características latentes aprendidas pelo modelo [Jurafsky and Martin 2025]. Assim, essa representação é mais rica semanticamente, mas pode ser mais custosa em termos de armazenamento e computação.

Na recuperação densa, compara-se a consulta e os fragmentos de texto por meio da similaridade semântica entre seus vetores. Para isso, utilizam-se diferentes métricas de similaridade, como a similaridade do cosseno formulada na Equação 2 (com a normaliza-

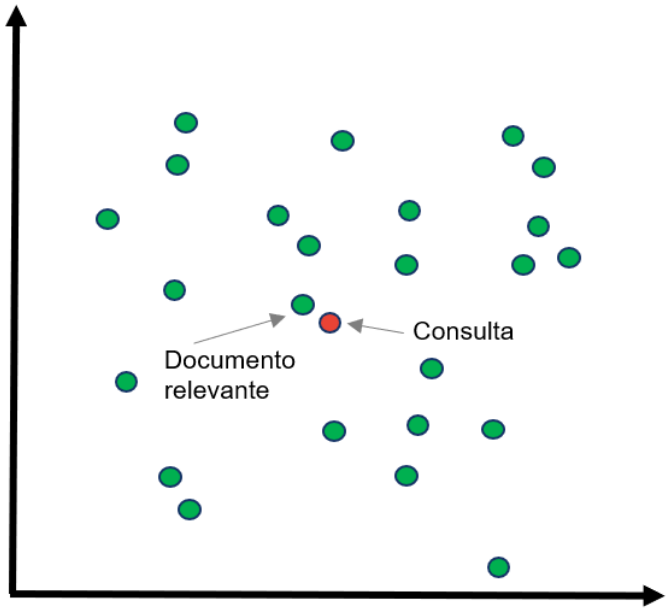



**Figura 6.10. Recuperação densa**

ção dos vetores pode ser transformada na métrica da distância do cosseno apresentada na Equação 3), a distância euclidiana na Equação 1, dentre outras [Eminagaoglu 2022].

$$d_{\text{euclidiana}}(a,b) = \|a-b\|_2 = \sqrt{\sum_{i=1}^{n}(a_i-b_i)^2} \tag{1}$$

$$\cos(\theta) = \frac{a \cdot b}{\|a\|\|b\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}} \tag{2}$$

$$d_{\text{cosseno}}(a,b) = 1 - \cos(\theta) \tag{3}$$

#### 6.3.3.3. Abordagens híbridas

A etapa de recuperação pode determinar o sucesso ou fracasso de um *pipeline* RAG. Se o *Retriever* não for capaz de localizar *chunks* relevantes, os resultados finais do RAG podem ser comprometidos e a chance do LLM gerador da resposta produzir alucinações aumenta, visto que o mesmo não receberá contexto adicional (aumentado) apropriado. Determinadas consultas podem ser melhor respondidas por técnicas de recuperação baseadas em palavras-chave, como o BM25. Enquanto outras apresentam melhor resultado com métodos de recuperação densos. Visando contornar as limitações individuais de cada abordagem, estratégias híbridas têm sido empregadas com bons resultados. Existem duas abordagens principais para construir um sistema de busca híbrido: fusão e reranking.

A **abordagem por fusão** combina os resultados obtidos por diferentes métodos de busca com base apenas nas pontuações fornecidas por cada um deles. Como essas

pontuações podem estar em escalas distintas, é comum aplicar algum tipo de normalização. Em seguida, uma fórmula agrega essas medidas de relevância para calcular uma pontuação final, que serve para reordenar os documentos [Qdrant 2025].

O *Reciprocal Rank Fusion* (RRF) [Cormack et al. 2009] é uma técnica de combinação de resultados que agrega listas de resultados ranqueadas retornadas por múltiplos sistemas de busca ou diferentes estratégias de recuperação. O objetivo é gerar uma lista final de resultados que seja superior a qualquer uma das listas individuais utilizadas. A ideia central é que documentos que aparecem no topo de várias listas têm alta probabilidade de serem relevantes de fato. Em vez de simplesmente somar ou ponderar as pontuações de relevância brutas dos diferentes sistemas (que podem estar em escalas diferentes e não ser diretamente comparáveis), a RRF considera a posição (rank) do documento em cada lista.

Na Tabela 6.6, observa-se que a abordagem híbrida com RRF apresenta desempenho superior em diversos conjuntos de dados, superando os métodos BM25, o denso e o híbrido convencional. A métrica utilizada na comparação é o nDCG@10, do inglês *Normalized Discounted Cumulative Gain* [Järvelin and Kekäläinen 2002], uma das principais métricas de RI para avaliar a qualidade de rankings gerados por sistemas de busca. Ela considera tanto a relevância dos documentos quanto sua posição no ranking, atribuindo maior peso aos documentos mais relevantes posicionados nas primeiras colocações. O valor do nDCG é normalizado entre 0 e 1, sendo 1.0 o valor correspondente ao ranking ideal.

**Tabela 6.6. Compara as pontuações do NDCG@10 com diferentes métodos de busca. Adaptado de [Bogan et al. 2025]**

| Dataset | BM25 | Denso | Híbrido | Híbrido com RRF | Diferença percentual |
|---|---|---|---|---|---|
| NFCorpus | 0,3065 | 0,2174 | 0,3076 | 0,2977 | 3,22% |
| ArguAna | 0,4258 | 0,4239 | 0,4507 | 0,4476 | 0,69% |
| FIQA | 0,2389 | 0,2004 | 0,2693 | 0,2474 | 8,13% |
| Trec-Covid | 0,6087 | 0,2718 | 0,5905 | 0,5877 | 0,47% |
| SciDocs | 0,155 | 0,1075 | 0,1602 | 0,1525 | 4,81% |
| Quora | 0,7424 | 0,8256 | 0,8452 | 0,796 | 5,82% |

A **abordagem por *reranking*** consiste na aplicação de modelos computacionalmente mais sofisticados e custosos — como modelos densos do tipo *bi-encoder* (por exemplo, Contriever [Izacard et al. 2021] ou E5 [Wang et al. 2022]) ou ainda *cross-encoders* como *ms-marco-MiniLM-L4-v2*[24] — sobre um subconjunto reduzido de documentos previamente recuperados. Inicialmente, utiliza-se um método de recuperação eficiente, como o BM25, para selecionar os top-N documentos mais relevantes com base em critérios lexicais. Em seguida, esse conjunto é reordenado por um modelo de *reranking* que realiza uma avaliação mais precisa da relevância, permitindo selecionar os top-k resultados finais ($k << N$). Essa estratégia em duas etapas equilibra desempenho computacional e precisão na recuperação da informação.

[24]https://huggingface.co/cross-encoder/ms-marco-MiniLM-L4-v2

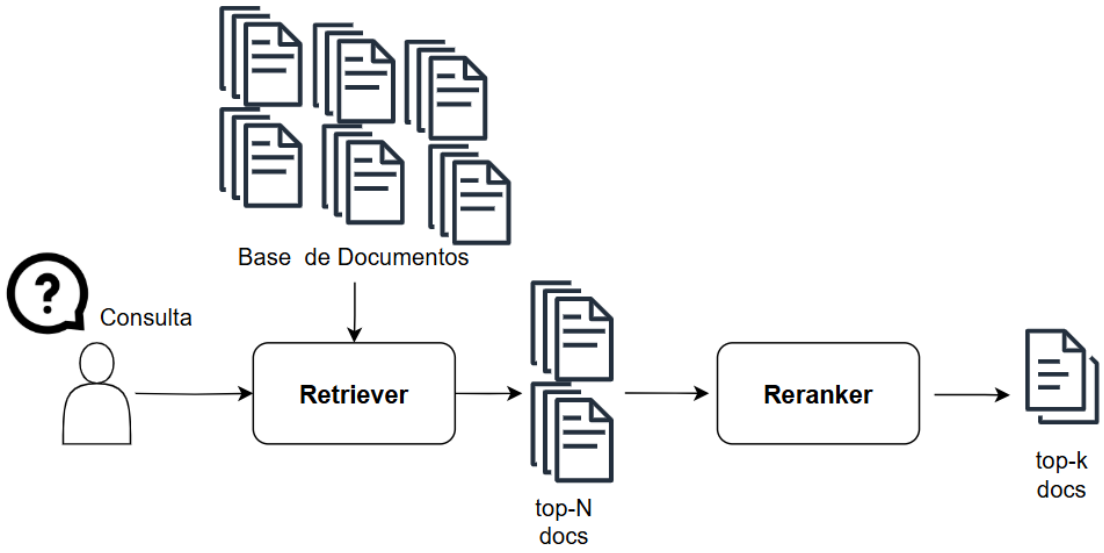



**Figura 6.11. Sistema de Recuperação usando *Reranking***

### 6.3.4. Geração da resposta

Nesta seção, serão abordados os processos e componentes do *Generator* apresentado na Figura 6.7. Essa etapa é fundamental no RAG pois é onde o que foi recuperado pelo *Retriever* é utilizado para gerar a resposta. Este processo envolve várias técnicas que devem ser cuidadosamente avaliadas e ajustadas para o caso de uso em questão: qual LLM será utilizado, qual a melhor estratégia de decodificação na geração do texto, qual o *prompt* ideal e como usar o contexto recuperado da melhor forma a evitar confundir o LLM e minimizar alucinações.

#### 6.3.4.1. Modelos de linguagem

Os LLMs revolucionaram o processamento de linguagem natural graças a avanços em arquiteturas baseadas em *deep learning*. A arquitetura transformer [Vaswani et al. 2017] trouxe como grande benefício a capacidade de processar sequências inteiras de dados (como textos) simultaneamente, por meio do mecanismo de autoatenção. Isso permitiu o processamento paralelo que acelera significativamente o treinamento destes modelos e beneficia a escalabilidade, pois facilita o treinamento de modelos maiores com mais dados. Também melhora muito a captura de dependências, ou seja, o modelo consegue identificar relações entre elementos distantes dentro da sequência com mais eficiência. Abaixo um exemplo destacado em [Paes et al. 2024]:

*"A garota de blusa amarela com uma frase em que os verbos estavam em letras pretas, que andava tão rápido e nunca em linha reta, a ponto de passar pelas nossas vistas como se fosse quase um furacão, tinha na parte de trás da sua blusa uma frase atribuída a Gandhi: "Acreditar em algo e não vivê-lo, é desonesto".*

Conforme destacam os autores do trecho, se quiséssemos descobrir a cor das letras

**Tabela 6.7. Principais LLMs na área médica. Adaptado de [Carchiolo and Malgeri 2025]**

| Modelo | Objetivo | Características |
|---|---|---|
| BioGPT [Lee et al. 2020] | PLN biomédico | Treinado com literatura biomédica para geração de texto e resposta a perguntas clínicas |
| GatorTron [Yang et al. 2022] | Análise de dados clínicos | Treinado com milhões de prontuários eletrônicos para melhorar a compreensão da linguagem clínica |
| MedPaLM [Singhal et al. 2025] | Diagnóstico e perguntas clínicas | Combina o PaLM com dados médicos para responder a perguntas clínicas |
| PubMedBERT [Gu et al. 2021] | PLN biomédico | Versão do BERT otimizada para artigos do PubMed, útil para classificação de texto e extração de informações |
| Galactica [Taylor et al. 2022] | Pesquisa científica | Treinado em textos científicos, incluindo publicações médicas |
| ClinicalBERT [Alsentzer et al. 2019] | Compreensão da linguagem clínica | Adaptação do BERT para registros eletrônicos de saúde (EHR) |
| BioBERT [Lee et al. 2020] | Biologia e aplicações médicas | Treinado com dados do PubMed e PMC para aprimorar a precisão do PLN na saúde |

com que a palavra "*Acreditar*" foi escrita, precisaríamos associá-la à palavra "*verbo*" e, então, retornar ao início da frase para notar que os verbos estão em preto. Além disso, há uma grande quantidade de palavras entre a informação sobre a cor e o primeiro verbo da frase de Gandhi. Uma rede recorrente (RNN), usada para esse tipo de problema antes do advento dos transformers, tem mais dificuldade em aprender essas conexões.

O Transformer é composto por duas grandes partes: o *encoder*, responsável por codificar o texto e gerar representações linguísticas, e o *decoder*, encarregado de decodificar essas representações e gerar texto. Modelos como GPT e LLaMA, são compostos por *decoders* e são otimizados para geração sequencial de texto — ou seja, predizem a próxima palavra com base nas anteriores. Por outro lado, arquiteturas bidirecionais, como o BERT e os modelos de embeddings especializados em gerar representações vetoriais do texto, tratados na Seção 6.3.2, utilizam *encoders* e são voltadas para compreensão contextual. Modelos BERT predizem palavras mascaradas considerando o contexto à esquerda e à direita e são amplamente utilizados em tarefas como classificação. Já os modelos *encoder-decoder*, como T5 e BART, combinam as capacidades de codificação e geração, sendo usados para tarefas como tradução automática e sumarização.

Para RAG os modelos precisam ter autorregressivos pois são usados na geração da resposta final. Mais detalhes sobre a geração do texto na Seção 6.3.4.2. Dentro da área médica alguns dos principais LLMs, seus objetivos e características estão listados na Tabela 6.7.

#### 6.3.4.2. Geração de texto

A geração de texto refere-se à produção de sequências textuais condicionadas a uma entrada textual. Essa capacidade é empregada em diversas tarefas de *PLN*, como sumarização, tradução, geração de texto em diálogo aberto, transcrição de fala para texto, conversão de imagens em descrições textuais, entre outras. Segundo [Jurafsky and Martin 2025], a utilização de modelos de linguagem para gerar texto representa uma das áreas de maior impacto dos modelos neurais no campo do *PLN*. A geração de texto, juntamente com a criação de imagens e a geração de código, compõe uma nova vertente da Inteligência Artificial (IA), frequentemente denominada **IA generativa**.

Os modelos de linguagem autorregressivos usados na geração automática de texto, seguem o princípio da **Modelagem de Linguagem Causal** (em inglês, *Casual Language Modeling*) , no qual cada palavra gerada depende do histórico das palavras anteriores. A geração ocorre de forma sequencial: a cada etapa, o token produzido é acrescentado ao contexto e utilizado como base para prever o próximo elemento da sequência.

Os modelos são treinados com grandes volumes de texto (livros, artigos, websites, etc.) para prever a próxima palavra/token em uma sequência. O modelo gera tokens um por um, escolhendo as palavras seguintes com base nas probabilidades aprendidas. Entrando a sequência "O céu está muito ___." o modelo provavelmente responderá com palavras como "azul" ou "nublado", pois são respostas prováveis. Assim, quando um modelo como o GPT recebe uma sequência de entrada, ele calcula a distribuição de probabilidade para o próximo token. Chamamos de **Estratégia de Decodificação** o mecanismo que transforma essa distribuição em uma escolha concreta de palavras, determinando como a sequência será construída.

Diferentes estratégias de decodificação podem levar a resultados significativamente diferentes em relação à qualidade, coerência e diversidade do texto. Algumas das estratégias de decodificação mais comumente usadas são as seguintes:

- Busca Gulosa: seleciona a palavra com a maior probabilidade como a próxima palavra. É o método mais simples, mas computacionalmente muito eficiente. Oferece baixa diversidade, levando a repetições [Jurafsky and Martin 2025].

- Busca por Feixes [Freitag and Al-Onaizan 2017]: explora múltiplas alternativas por etapa e seleciona a de maior probabilidade total, mas consome muitos recursos e pode gerar saídas repetitivas.

- Amostragem Top-k [Fan et al. 2018]: o modelo restringe a escolha da próxima palavra às $k$ mais prováveis e realiza uma seleção aleatória entre elas, conforme a distribuição de probabilidade. Essa limitação introduz variabilidade e criatividade no texto gerado, mas pode afetar a consistência semântica quando $k$ é excessivamente alto.

- Amostragem Top-p [Holtzman et al. 2019]: escolhe o menor conjunto possível de palavras cuja probabilidade cumulativa excede a probabilidade $p$. A massa de probabilidade é então redistribuída entre esse conjunto de palavras. O valor de $p$ con-

trola a diversidade do texto gerado. Valores mais altos de $p$ levam a um texto mais diverso, enquanto valores mais baixos geram um texto mais previsível.

- Temperatura: reduzir a temperatura do softmax significa tornar a distribuição de probabilidade mais nítida [Kamath et al. 2019]. A temperatura é um hiperparâmetro que controla a aleatoriedade do processo de decodificação no LLM. Valores mais baixos resultam em texto mais previsível e repetitivo, e o modelo se torna determinístico, enquanto uma temperatura mais alta gera texto mais diverso e criativo.

A escolha de uma estratégia de decodificação deve considerar um equilíbrio entre diversidade, coerência e eficiência computacional. Em muitos casos, uma combinação dessas estratégias ou modificações personalizadas podem ser usadas.

#### 6.3.4.3. Organização dos *prompts*

A aprendizagem em contexto (*In-Context Learning*, ICL), mencionada na Seção 6.3, refere-se à capacidade dos LLMs de realizar novas tarefas com base apenas nos exemplos e instruções fornecidos na entrada, sem necessidade de atualizar seus parâmetros durante a inferência. Nesse paradigma, o modelo não passa por nenhum processo de treinamento adicional, ou seja, não há atualização de gradientes, mas sim utiliza o contexto do *prompt* para simular o comportamento esperado. Assim, o LLM é capaz de se adaptar a tarefas ou conceitos que não foram explicitamente vistos durante o treinamento, generalizando a partir de poucos exemplos apresentados no momento da consulta (inferência).

A ICL pode ocorrer em um cenário onde não se fornece nenhum exemplo (*zero-shot*), ou seja, apenas a solicitação (a tarefa a ser resolvida) é enviada ao modelo com a instância de teste para a qual o modelo deve fazer gerar a resposta. Também é possível utilizar a abordagem com poucos exemplos (*few-shot*) onde utiliza-se um conjunto de demonstrações na tarefa-alvo, consistindo na entrada e na saída desejada. Portanto, além de enviar a descrição da tarefa, enviamos as demonstrações, seguidas de um único exemplo não rotulado para o qual a previsão deverá ser feita pelo LLM. A ideia é que o modelo utilize esses exemplos para fornecer a resposta à tarefa para a qual não foi especificamente treinado. Toda essa entrada é chamada de "prompt".

Assim, no contexto dos LLMs, um *prompt* é um enunciado em linguagem natural que orienta o modelo sobre qual tarefa executar. Ele pode assumir diferentes formas, como uma pergunta direta, uma afirmação contextual, uma instrução descritiva ou até uma solicitação de tarefa específica, como resumir ou classificar um conteúdo [Paes et al. 2024].

A engenharia de *prompts* é a prática de projetar e refinar *prompts* de entrada para aprimorar as respostas que obtemos do modelo de linguagem. É uma atividade empírica. As saídas do modelo variam de acordo com o *prompt* utilizado, os exemplos selecionados e a ordem dos exemplos. Embora os *prompts* possam ser redigidos de maneira flexível, os LLMs são altamente sensíveis à sua formulação. A Figura 6.12 mostra um processo de criação manual de *prompts*. A segunda etapa na Figura (criação do *prompt*) pode ser automatizada e/ou usar fontes externas para complementar o *prompt* e fornecer um contexto mais relevante para melhorar a resposta do LLM.

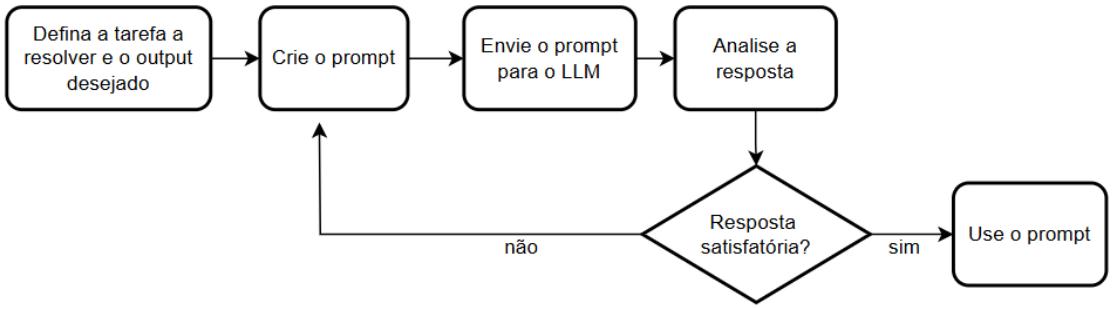



**Figura 6.12. Ciclo de Engenharia de *Prompts***

Diversas estratégias de criação de *prompts* foram desenvolvidas. Uma das mais utilizadas é a Cadeia de Pensamento (*Chain-of-Thought*, CoT) [Wei et al. 2022]. Ela aprimora a capacidade de realizar raciocínios complexos, incorporando etapas intermediárias. É uma estratégia que faz o modelo explicitar seu raciocínio antes de chegar à resposta final. Em vez de pedir uma resposta direta, o *prompt* instrui o modelo a "pensar passo a passo".

[Wu et al. 2024] desenvolveram um método baseado CoT para detecção e correção de erros médicos em notas clínicas, desenvolvido para o desafio MEDIQA-CORR 2024. Os autores identificam manualmente três tipos comuns de erro médico: diagnóstico, intervenção e manejo. Na Figura 6.13 é possível visualizar os três *prompts* distintos para cada categoria de erro. O modelo GPT-4 é guiado por etapas: primeiro detecta se há erro; caso afirmativo, identifica a sentença; por fim, gera a correção.

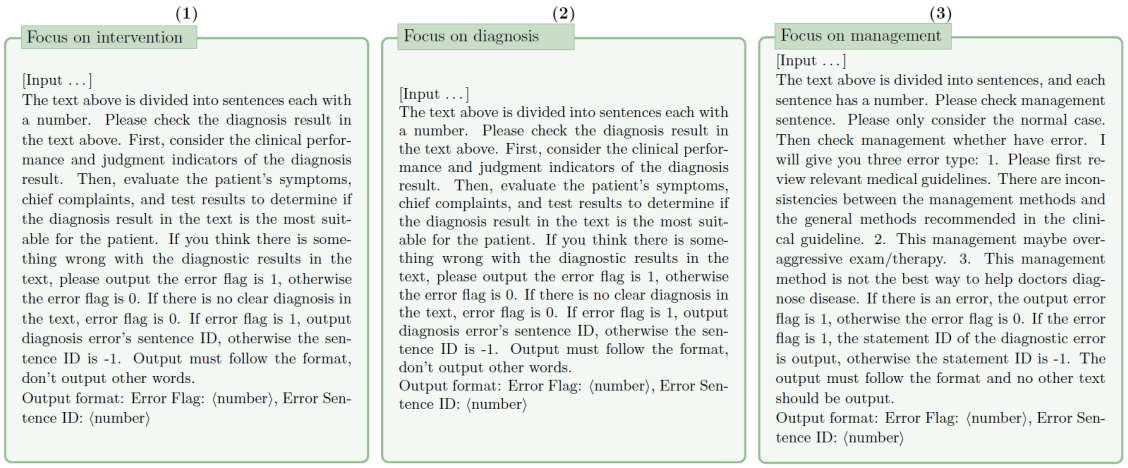



**Figura 6.13. Três tipos de *prompts* de Cadeia de Pensamento (CoT) utilizados para identificar erros de intervenção, diagnóstico e gestão, respectivamente. Extraído de [Wu et al. 2024]**

Uma estratégia eficaz na engenharia de *prompts* consiste em decompor tarefas complexas em subtarefas menores. Cada uma dessas subtarefas é apresentada ao LLM por meio de um*prompt* específico, cuja saída pode ser reutilizada como entrada em uma etapa posterior. Esse processo é conhecido como encadeamento de *prompts* (*prompt chaining*), no qual a resolução de uma tarefa ocorre de forma progressiva, por meio de uma

sequência estruturada de instruções. Essa abordagem se destaca em casos onde o modelo teria dificuldade em lidar com uma instrução muito extensa ou complexa de uma só vez. Ao aplicar o encadeamento, cada resposta intermediária pode ser refinada, complementada ou transformada por novos *prompts*, até que se alcance o resultado final esperado.

*Reflexion* [Shinn et al. 2023] é uma técnica de *prompting* composta, iterativa e reflexiva. Ela se enquadra em um tipo mais avançado de engenharia de *prompt*, onde o LLM é induzido a aprender com seus próprios erros por meio de autoavaliação e reutilização estratégica de contexto textual. *Reflexion* opera inteiramente via linguagem natural. Ele não treina o modelo, mas sim estrutura a interação com o modelo de forma estratégica, por meio de *prompts* que executam a tarefa, que avaliam o próprio desempenho, que integram a memória das reflexões passadas como contexto para a próxima tentativa.

No contexto do RAG, a escolha das estratégias de *prompting* dependem do domínio, mas também do LLM, alguns modelos tem uma capacidade maior de seguir instruções, por exemplo. Os *chunks* recuperados pelo *Retriever* são ordenados segundo a métrica de similaridade usada na recuperação e seleciona-se os top-k com melhores valores nesta métrica. A escolha do $k$ a utilizar é um ponto importante e para isso deve-se considerar a janela de contexto do modelo e o tamanho máximo do *chunks* (ver discussão na Seção 6.3.2) para evitar o estouro do limite de entrada do LLM.

O número de top-k documentos a considerar na etapa de recuperação de um sistema RAG é uma decisão crítica. Não existe um único valor ótimo de top-k, mas sim recomendações baseadas no tipo de aplicação, modelo usado e tolerância a erro. Valores baixos de $k$ (de 3 a 5) são recomendados para perguntas factuais ou quando se sabe que a base de dados tem alta qualidade. Usar $k > 10$ aumenta a diversidade e cobertura de possíveis evidências, mas eleva o risco de conflito (informações contraditórias) e exige mecanismos de fusão ou seleção. Para *chunks* longos um $k$ menor é recomendado.

Os *chunks* são recebidos pelo *Generator* e serão organizados e adicionados ao *prompt* que será enviado ao LLM. Neste ponto, as técnicas mencionadas e outras mais de engenharia de *prompt* podem ser aplicadas e incluindo variações dependendo do caso de uso do RAG, como por exemplo sumarização dos *chunks* para casos em que sejam muitos a considerar ou quando o LLM utilizando possui uma janela de contexto pequena. Na Figura 6.14 um exemplo de um *prompt* simples e genérico para RAG com $k = 3$.

#### 6.3.4.4. Fine-tuning de modelos

Em uma arquitetura RAG, a princípio, não é necessário fazer fine-tuning do LLM para gerar as respostas, mas em contextos mais especializados ou críticos como saúde, o fine-tuning pode melhorar significativamente a performance. O fine-tuning de modelos de linguagem (LLMs) abrange diversas técnicas para adaptar um modelo pré-treinado a tarefas ou domínios específicos.

O termo fine-tuning pode se referir à extensão do ajuste de parâmetros (quanto do modelo será treinado) ou ao objetivo do treinamento (o que o modelo aprende). Quanto à extensão do ajuste ele pode ser total ou parcial. O ajuste total atualiza todos os parâmetros do modelo, é computacionalmente custoso e apresenta o risco do esquecimento

```
Baseando-se exclusivamente nos documentos fornecidos abaixo,
responda à pergunta a seguir.

Se a informação solicitada não estiver presente, diga "As
informações fornecidas não são suficientes para responder com
segurança."

Inclua na resposta trechos que justifiquem suas afirmações.


Pergunta: [Pergunta]


Documentos:

[Chunk 1] ...

[Chunk 2] ...

[Chunk 3] ...
```

**Figura 6.14. Exemplo de *prompt* para RAG**

catastrófico (*catastrophic forgetting*) [Li et al. 2025], que se refere ao esquecimento do conhecimento prévio. Ajustes parciais são mais leves, exigindo menos memória e treino, por exemplo congelar camadas iniciais (que capturam características gerais) e ajustar apenas as camadas superiores. PEFT (*Parameter-Efficient Fine-Tuning*) [Ding et al. 2023] é um conjunto de técnicas projetadas para ajustar modelos de linguagem grandes de forma eficiente, modificando apenas uma pequena fração dos parâmetros do modelo original. Seu objetivo é reduzir custos computacionais, minimizando o risco de *catastrophic forgetting*. Técnicas como LoRA (*Low-Rank Adapter*) [Hu et al. 2022] e QLoRA (*quantized Low-Rank Adapter*) [Dettmers et al. 2023] congelam os pesos originais do modelo base e adicionam pequenas matrizes de baixo posto (*low-rank*) treináveis, que capturam os ajustes necessários para a tarefa.

Quanto ao objetivo, o *fine-tuning* pode adotar diferentes estratégias, dependendo da finalidade da adaptação do modelo. O ajuste supervisionado (*Supervised Fine-Tuning*, SFT) consiste no treinamento do modelo em um conjunto de dados anotado para tarefas específicas, como classificação. No ajuste por instrução (*Instruction Tuning*) o modelo é treinado para seguir instruções explícitas em linguagem natural, com o objetivo de torná-lo mais útil e alinhado a comandos humanos [Ouyang et al. 2022]. Neste caso, utiliza-se pares instrução–resposta para melhorar a capacidade do modelo de seguir comandos diversos. O método de aprendizado por reforço a partir do feedback humano (*Reinforcement Learning from Human Feedback, RLHF)* [Ouyang et al. 2022], ajusta o comportamento

do modelo com base em preferências humanas: as saídas do modelo são ranqueadas por humanos. O ChatGPT, por exemplo, foi ajustado com RLHF. Nessa linha, a Otimização Direta por Preferência (*Direct Preference Optimization, DPO*) [Rafailov et al. 2023] é uma alternativa ao RLHF, pois treina diretamente com base em pares de preferência (resposta preferida, não preferida), otimizando o modelo diretamente para gerar respostas mais alinhadas ao julgamento humano.

O RAFT (*Retrieval-Augmented Fine-Tuning*) [Zhang et al. 2024] é uma técnica de ajuste supervisionado que especializa um LLM para funcionar melhor dentro de um sistema RAG. Os autores comparam o desempenho de LLMs com o comportamento de estudantes em provas com consulta: o RAG tradicional é como um aluno que só olha o livro na hora da prova, sem ter estudado previamente; o RAFT é um aluno que estudou os livros antes da prova, ou seja, o modelo treina com os documentos antes de ser testado com recuperação.

Em vez de usar um LLM genérico e simplesmente adicionar os documentos recuperados no *prompt*, como no RAG tradicional, o RAFT treina o modelo com exemplos explícitos de como raciocinar com documentos relevantes e ignorar os irrelevantes. Os autores usam o modelo LLaMA 2 de 7 bilhões de parâmetros e criam um conjunto de dados sintético onde cada exemplo tem os seguintes elementos:

- Pergunta
- Conjunto de documentos recuperados que contém documentos relevantes (contêm a resposta) mas também documentos irrelevantes (distratores, que devem ser ignorados).
- Resposta correta gerada a partir dos documentos (A resposta usa conteúdo apenas dos documentos relevantes).
- Explicação com raciocínio (*Chain-of-Thought*): Inclui citações textuais dos documentos relevantes para justificar a resposta.

Depois de treinado o modelo é usado dentro do *pipeline* do RAG normalmente. Dessa forma, o objetivo do RAFT é ensinar o LLM usado no RAG a raciocinar com base em múltiplos documentos, ignorar informações irrelevantes e usar informações relevantes de forma correta.

Na mesma linha, [Jeong et al. 2024] apresentam o framework Self-BioRAG de ajuste do LLM usado no RAG visando melhorar a capacidade em responder perguntas médicas. Para isso, os autores realizam fine-tuning em dois modelos LLaMA2. Primeiro, fazem ajuste do modelo crítico *C* que aprende a prever quatro tokens reflexivos: (i) identificar se uma questão requer recuperação (RET); (ii) determinar se a evidência recuperada fornece informações úteis para resolver uma questão (REL); (iii) avaliar se todas as declarações de respostas podem ser apoiadas por evidências (SUP); (iv) avaliar se todas as declarações de respostas são uma resposta útil à questão (USE). O modelo *C* é usado para anotar um dataset maior com esses tokens reflexivos. Esse dataset é então usado para ajustar o LLM gerador.

O LLM ajustado é o modelo usado em produção, e ele aprendeu, com base nos exemplos anotados pelo modelo crítico *C*, a decidir se precisa recuperar ou não, a avaliar relevância e fundamentar a resposta. Na Figura 6.15 podemos ver a comparação de três abordagens para responder perguntas clínicas. O esquema C demonstra o LLM gerador atuando primeiramente na verificação da necessidade de recuperar contexto. Caso não seja necessário o LLM responde diretamente. Sendo necessário ele segue o fluxo de uma aplicação RAF e o contexto é recuperado da base PubMed Central[25] (PMC). Depois de recuperado o contexto é avaliado pelo LLM que seleciona a melhor evidencia a ser usada na geração da resposta.

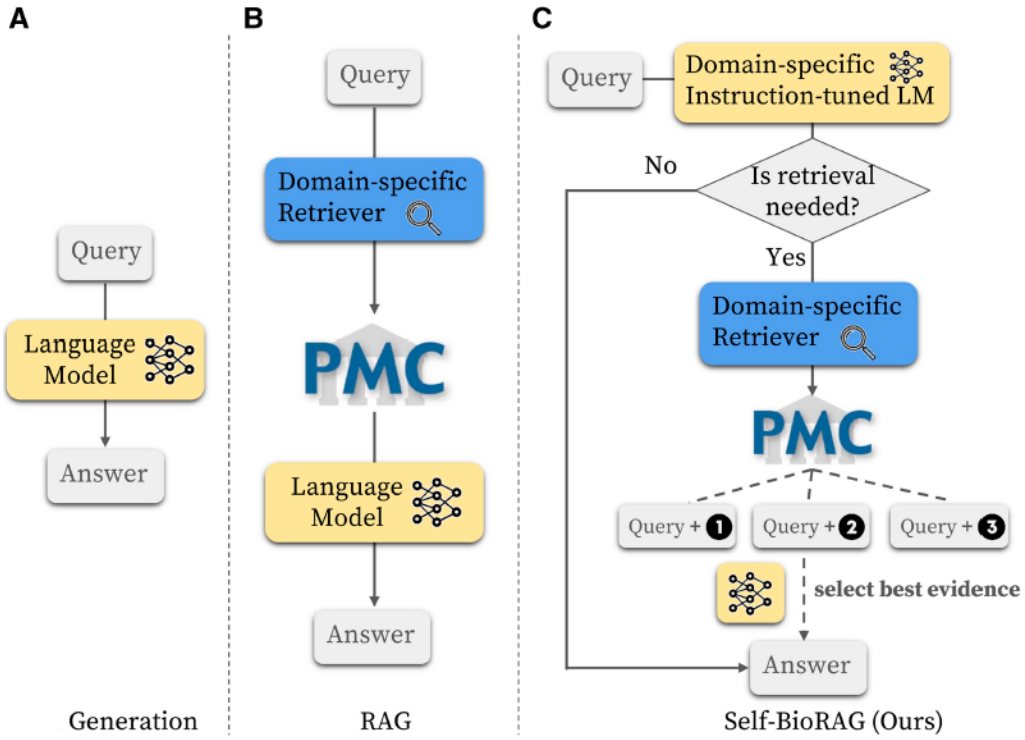



**Figura 6.15. Comparação entre três estruturas para gerar respostas a perguntas: A) usando somente LLM, B) geração aumentada de recuperação (RAG) e C) Usando modelo ajustado Self-BioRAG. Extraído de [Jeong et al. 2024]**

## 6.4. Avaliação dos resultados do RAG

Avaliar sistemas RAG apresenta desafios devido à sua estrutura híbrida que envolve a participação de múltiplos componentes. De maneira geral, o desempenho do *pipeline* RAG pode ser avaliado examinando os dois componentes principais: *Retriever* e *Generator*.

Esforços foram feitos para avaliar RAG no contexto clínico com desenvolvimento de benchmarks de desempenho que utilizam respostas de múltipla escolha ou verdades categóricas para respostas como MedRAG [Xiong et al. 2024]. Entretanto esse tipo de avaliação falha em capturar as complexidades e riscos associados com gerações de respostas abertas [Chowdhury et al. 2025]. Para avaliar RAG é necessário dispor de métricas

[25]https://pmc.ncbi.nlm.nih.gov/tools/textmining/

capazes de aferir tanto o desempenho global do sistema quanto o funcionamento individual de cada módulo de forma a diagnosticar as origens dos erros e compreender como eles se manifestam ao longo do processo [Ru et al. 2024].

### 6.4.1. Métricas tradicionais

No caso *Retriever*, métricas traicionais utilizadas na tarefa de Recuperação de Informação (RI) clássica como recall@k e MMR [Moreira 2024], entre outras, podem ser utilizadas. O problema que essas métricas dependem de esquemas fixos de segmentação (chunking) e anotações específicas para RI considerando estes esquemas, isso significa que variar o esquema de *chunking* compromete a anotação inicial que não é necessariamente mapeável para um novo esquema.

Já para o *Generator*, métricas usadas na geração de texto poderiam ser usadas desde que exista um texto de referência. Essas métricas comparam o texto gerado à referência. Por exemplo, em perguntas e respostas cada pergunta tem uma resposta esperada que será comparada à resposta gerada.

Métricas como BLEU e ROUGE são amplamente usadas na geração de texto. Elas são métricas de sobreposição de N-gramas[26], mas apresentam diversas limitações na avaliação de textos gerados por LLM, que é o caso de sistemas RAG. Estas métricas se baseiam na correspondência exatas de palavras ou frases, ignorando sinônimos, paráfrases e expressões semanticamente equivalentes, mas lexicalmente diferentes. Por exemplo, o texto gerado "*O felino descansou no carpete*"ao ser avaliado contra o texto de referência "*O gato sentou no tapete*"apresentará pontuação baixa mesmo que o significado esteja correto. Enquanto que o texto gerado "*O cachorro preguiçoso salta sobre a rápida raposa marrom*"avaliado contra a referência "*A rápida raposa marrom salta sobre o cachorro preguiçoso*"poderia obter uma pontuação enganosamente alta com base na sobreposição de palavras individuais (unigramas), mascarando a mudança fundamental de significado devido à ordem das palavras.

BERTScore [Zhang et al. 2020] é uma métrica que também pode ser usada na geração de texto. Ela é baseada em embeddings de modelos *encoder-only* como o BERT e tem mais sucesso em lidar com casos semanticamente equivalentes mas lexicalmente diferentes. Entretanto, em exemplos de negação ou de contradição onde pode atribuir alta similaridade, porque não é sensível a negação de forma explícita. No caso do texto gerado "*Tomar aspirina é excelente para pressão alta*"comparado ao texto de referência "*Tomar aspirina é um risco para pressão alta*"pode resultar em alta similaridade, pois as duas frases compartilham quase o mesmo vocabulário apesar de terem sentidos opostos. O BERTScore não entende negação nem oposição lógica explícita. Isso significa que, nesse caso, uma resposta perigosa (como recomendar aspirina indevidamente) pode ser avaliada pelo BERTScore como "boa". Além disso, há a questão do conhecimento do modelo usado no BERTScore sobre o domínio: utilizar modelos que não reconhecem o domínio da saúde pode não identificar a similaridade de exemplos como este: "*A pressão arterial elevada pode aumentar o risco de AVC*"e "*Hipertensão eleva a probabilidade de um derrame cerebral*".

[26]Um n-grama é uma sequência contígua de n itens de uma determinada amostra de texto ou fala

Em textos longos, o BERTScore perde precisão e interpretabilidade. Além disso, modelos BERT-base têm limite de 512 tokens em média. Textos maiores são truncados, o que pode fazer com que partes relevantes da resposta sejam ignoradas. Mesmo para textos longos menores que 512 tokens, BERTScore atribui peso igual para todos os tokens no calculo da similaridade. Não distingue entre um token irrelevante e um que expressa uma afirmação crítica (por exemplo: "não", "risco", "fatal"). Consequentemente, em textos longos, afirmações importantes podem se diluir no meio de conteúdo superficial. Por isso, em tarefas como RAG, é preferível complementar BERTScore com com métricas baseadas em inferência textual (*entailment*), que indicam contradição, implicação ou neutralidade entre duas sentenças. Para isso é necessário modelos de inferência textual também especializados no domínio desejado.

### 6.4.2. LLMs como avaliadores

Um um conjunto com perguntas e respostas de referência no domínio de interesse, muitas vezes não está disponível para averiguar se as configurações do sistema RAG projetado alcançam bons resultados. Além disso, a construção destes conjuntos é bastante custosa exigindo tempo considerável dos anotadores (humanos que avaliam os dados), que algumas vezes precisam ser especialistas no domínio. Adicionalmente, a avaliação do RAG precisa conectar os resultados da geração e da recuperação e isso não é simples de anotar. Por exemplo, até que ponto a resposta gerada se baseia efetivamente nos trechos recuperados? há informações incluídas que não estão presentes nos top-$k$ documentos retornados?

Existe um movimento de avaliar sistemas RAG usando LLMs que é aderente ao que ocorre para uma série de tarefas de IA que usam esses modelos como julgadores ou avaliadores. Essas iniciativas são classificadas sob a descrição "LLM as a judge" [Zheng et al. 2023]. São técnicas baseadas no paradigma do aprendizado em contexto (*In-Context Learning*) popularizado a partir do modelo GPT-3 [Brown et al. 2020].

Frameworks de avaliação RAG apresentam formas de avaliar automaticamente os resultados usando métricas que não exigem uma base anotada, como por exemplo RAGAS [Es et al. 2024] e RAGTriad [Trulens 2024] . Eles utilizam LLMs para avaliar a resposta geral, mas também buscam avaliar os resultados dos diversos componentes da arquitetura RAG. Estão integrados a outros frameworks para desenvolvimento de soluções RAG e agentes como LangChain[27] e Llama Index[28].

Recentemente, o framework de avaliação RAGChecker [Ru et al. 2024] apresentou uma ampla consolidação das métricas e seu principal diferencial em relação ao RAGAS é o de considerar para o cálculo das métricas verificações detalhadas no nível de afirmações dentro da resposta e dentro dos fragmentos. Os autores também destacam que as métricas são projetadas para fornecer *insights* claros sobre as fontes de erros dentro dos diversos componentes RAG. Muitas destas métricas não exigem uma base anotada e usam LLM como julgadores, que tem sido uma tendência em várias tarefas, em especial em recuperação de informações [Bencke et al. 2024].

A RAG Triad, ilustrada na Figura 6.16, é uma proposta de avaliação automática de

[27]https://www.langchain.com/
[28]https://www.llamaindex.ai/

sistemas RAG implementada na ferramenta TrueLens[29]. Essa plataforma de software foi desenvolvida com o objetivo de mensurar a qualidade e a eficácia de aplicações baseadas em LLMs por meio de funções de feedback. Essas funções avaliam a qualidade de entradas, saídas e resultados intermediários em diferentes contextos de uso, como sistemas de perguntas e respostas, sumarização, RAG e agentes inteligentes. As três dimensões da RAG Triad são avaliadas por *prompts* que solicitam notas em uma escala de 0 a 10, adotando a estratégia *Chain-of-Thought* (CoT) [Wei et al. 2022] para fundamentar as respostas.

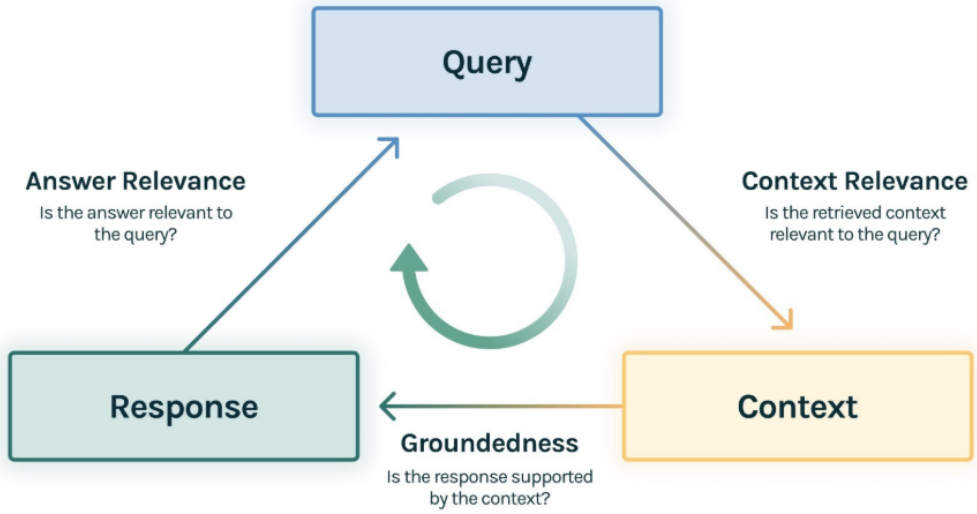



**Figura 6.16. As três dimensões da avaliação RAG Triad. Extraído de [Trulens 2024]**

O framework RAGAS proposto por [Es et al. 2024], implementa um conjunto de métricas para avaliar diferentes dimensões da resposta sem precisar de anotações humanas. Três métricas principais estão descritas a seguir.

**Fidelidade** (*Faithfulness*): a resposta deve estar fundamentada no contexto informado (*chunks* recuperados pelo *Retriever*). Isso é importante para evitar alucinações e garantir que o contexto recuperado possa servir de justificativa para a resposta gerada. Para esta métrica RAGAS usa LLM em dois passos: (1) envia a resposta gerada solicitando a extração de um conjunto de declarações (*claims*), 2) envia as declarações obtidas em (1) e pede que o LLM determine quais delas estão fundamentada pelas informações presentes no contexto inicial (conjuntos de *chunks* recuperados) e explique o porque de sua conclusão. O resultado final da Equação 4 corresponde ao percentual de declarações que estão fundamentadas pelo contexto de *chunks*. Quanto maior a fidelidade maior a aderência da resposta final ao contexto informado (*chunks*).

$$\text{Faithfulness} = \frac{\text{Nro de declarações na resposta gerada que podem ser inferidas do contexto}}{\text{Número total de declarações na resposta gerada}} \tag{4}$$

[29] https://www.trulens.org/

**Relevância da resposta** (*Answer Relevance*): refere-se à ideia de que a resposta gerada deve abordar a pergunta real fornecida. RAGAS solicita que o LLM gere $n$ perguntas potenciais com base na resposta final. Para cada pergunta $q_i$, calcula a similaridade $sim(q, q_i)$ com a questão original $q$, calculando o cosseno entre os embeddings correspondentes. A relevância da resposta, para a questão $q$ é então calculada de acordo com a Equação 5.

$$\text{Answer Relevance} = \frac{1}{n}\sum_{i=1}^{n} sim(q, q_i) \tag{5}$$

**Utilização do Contexto** (*Context Utilization*): A resposta gerada é dividida em declarações (*claims*) usando um modelo de extração semântica (por exemplo, via LLM ou heurística). Cada declaração é então verificada contra o contexto recuperado usando um modelo de inferência textual (NLI). A métrica é calculada como:

$$\text{Context Utilization@K} = \frac{\sum_{k=1}^{K}(\text{Precision@k} \times v_k)}{\text{Número total de item relevantes nos top-}K\text{ resultados}} \tag{6}$$

$$\text{Precision@k} = \frac{\text{true positives@k}}{\text{true positives@k} + \text{false positives@k}} \tag{7}$$

Onde $K$ é o número total de *chunks* recuperados (contexto) e $v_k \in \{0,1\}$ é o indicador de relevância do rank $k$. e pode ser julgado por um LLM.

## 6.5. Conclusões e oportunidades de pesquisa

O RAG na área da saúde apresenta um campo importante para pesquisa, com potencial para transformar a prática clínica, educação médica e pesquisa biomédica [Ng et al. 2025], especialmente por sua capacidade de integrar conhecimento externo confiável ao processo de geração textual. É fundamental que sistemas RAG para aplicações médicas foquem no acesso rápido e preciso a informações atualizadas e relevantes, minimizando os riscos de "alucinações"ou informações equivocadas que podem ser perigosas em um contexto clínico. Existem desafios significativos, mas as oportunidades para impacto positivo são maiores, exigindo colaboração entre pesquisadores de IA, profissionais da saúde e especialistas em informática médica.

Muitos sistemas RAG utilizam mecanismos genéricos de recuperação, como o BM25 ou embeddings treinados em domínios amplos, o que pode levar à seleção de documentos irrelevantes. Para aumentar a precisão e relevância no contexto clínico, é essencial desenvolver estratégias de recuperação especializadas como o fine-tuning de modelos de embeddings usando benchmarks do domínio da saúde no idioma em questão. Explorar técnicas de recuperação que utilizem ontologias médicas para enriquecer a busca pode melhorar buscas mais específicas.

Podem existir casos onde diferentes fontes apresentam recomendações múltiplas e divergentes dependendo do contexto clínico. Um caminho a ser explorado é projetar mecanismos de fusão de contexto que consigam identificar e ponderar múltiplas evidências, integrando critérios como qualidade da fonte, data de publicação e força da evidência.

A utilização de modelos de linguagem via API levanta preocupações quanto à privacidade de dados sensíveis, especialmente em domínios como a saúde. Uma alternativa interessante é a especialização de LLMs não tão grandes, mas com bom poder de resposta, capazes de operar em uma infraestrutura viável para instituições de saúde.

Destacam-se também várias oportunidades de aplicações do RAG no domínio da saúde:

- Sistemas de apoio à decisão clínica como sistemas de diagnóstico assistido que recuperam e resumem informações de guias médicos, literatura recente e prontuários eletrônicos para auxiliar no diagnóstico.

- Sistemas que contemplem alertas clínicos inteligentes que monitoram prontuários eletrônicos e recuperam informações relevantes para alertar sobre interações medicamentosas ou condições emergentes.

- Tutores virtuais baseados em RAG em sistemas que respondam a perguntas de médicos e estudantes recuperando as informações mais atualizadas.

- Melhorias no atendimento ao paciente como chatbots avançados que fornecem informações precisas recuperadas de fontes confiáveis e/ou a geração de explicações personalizadas para pacientes baseadas em seu perfil e nas evidências disponíveis.

Uma das principais preocupações na aplicação de RAG em contextos sensíveis como a saúde: o uso indevido do *self-knowledge*, ou seja, quando o modelo de linguagem utiliza conhecimentos "internos"adquiridos durante o pré-treinamento em vez de se restringir às informações recuperadas. Em tarefas clínicas, esse comportamento pode gerar alucinações factuais perigosas, pois o modelo pode gerar respostas convincentes baseadas em dados genéricos, desatualizados ou errôneos, mesmo quando o conteúdo recuperado não dá suporte àquela afirmação. Controlar o uso do conhecimento paramétrico (*self-knowledge*) dos LLMs em RAG na saúde é importante e demanda uma combinação de engenharia de *prompts*, arquitetura, técnicas de *reranking* e de fusão, métricas de avaliação e treinamento de modelos aptos a não responder como trabalhos vistos na Seção 6.3.4.4. O ideal é que o sistema não só gere respostas baseadas em fatos, mas também demonstre de onde a informação foi tirada.

## Referências

[Alsentzer et al. 2019] Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical BERT embeddings. In Rumshisky, A., Roberts, K., Bethard, S., and Naumann, T., editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

[Bencke et al. 2024] Bencke, L., Paula, F. S., dos Santos, B. G., and Moreira, V. P. (2024). Can we trust llms as relevance judges? In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 600–612. SBC.

[Bogan et al. 2025] Bogan, R., Gaievski, M., Shah, M., and Kolchina, F. (2025). Introducing reciprocal rank fusion for hybrid search.

[Brown et al. 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[Carchiolo and Malgeri 2025] Carchiolo, V. and Malgeri, M. (2025). Trends, challenges, and applications of large language models in healthcare: A bibliometric and scoping review. *Future Internet*, 17(2):76.

[Chowdhury et al. 2025] Chowdhury, M., He, Y. V., Higham, A., and Lim, E. (2025). Astrid–an automated and scalable triad for the evaluation of rag-based clinical question answering systems. *arXiv preprint arXiv:2501.08208*.

[Cormack et al. 2009] Cormack, G. V., Clarke, C. L., and Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

[Cortes et al. 2024] Cortes, E. G., Vieira, R., and Barone, D. A. C. (2024). Perguntas e respostas. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 16. BPLN, 2 edition.

[Dettmers et al. 2023] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

[Devlin et al. 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

[Ding et al. 2023] Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., et al. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

[Duarte et al. 2024a] Duarte, A., Marques, J., Graça, M., Freire, M., Li, L., and Oliveira, A. (2024a). Lumberchunker: Long-form narrative document segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6473–6486.

[Duarte et al. 2024b] Duarte, A. V., Marques, J. D., Graça, M., Freire, M., Li, L., and Oliveira, A. L. (2024b). LumberChunker: Long-form narrative document segmentation. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6473–6486, Miami, Florida, USA. Association for Computational Linguistics.

[Eminagaoglu 2022] Eminagaoglu, M. (2022). A new similarity measure for vector space models in text classification and information retrieval. *Journal of Information Science*, 48(4):463–476.

[Es et al. 2024] Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In Aletras, N. and De Clercq, O., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

[Fan et al. 2018] Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical neural story generation. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

[Freitag and Al-Onaizan 2017] Freitag, M. and Al-Onaizan, Y. (2017). Beam search strategies for neural machine translation. In Luong, T., Birch, A., Neubig, G., and Finch, A., editors, *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.

[Gao et al. 2023] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1.

[Gu et al. 2021] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

[Holtzman et al. 2019] Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The curious case of neural text degeneration. In *International Conference on Learning Representations*.

[Hu et al. 2022] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

[Huggingface 2024] Huggingface (2024). Rag worfflow.

[Izacard et al. 2021] Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. (2021). Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

[Järvelin and Kekäläinen 2002] Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

[Jeong et al. 2024] Jeong, M., Sohn, J., Sung, M., and Kang, J. (2024). Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

[Jurafsky and Martin 2025] Jurafsky, D. and Martin, J. H. (2025). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall.

[Kamath et al. 2019] Kamath, U., Liu, J., and Whitaker, J. (2019). *Deep learning for NLP and speech recognition*, volume 84. Springer.

[Kamradt 2024] Kamradt, G. (2024). Semantic chunking. `https://github.com/FullStackRetrieval-com/RetrievalTutorials/tree/main/tutorials/LevelsOfTextSplitting`.

[Ke et al. 2024] Ke, Y., Jin, L., Elangovan, K., Abdullah, H. R., Liu, N., Sia, A. T. H., Soh, C. R., Tung, J. Y. M., Ong, J. C. L., and Ting, D. S. W. (2024). Development and testing of retrieval augmented generation in large language models–a case study report. *arXiv preprint arXiv:2402.01733*.

[Kuriki P. and R. 2024] Kuriki P., Kay F., B. C. and R., P. (2024). Rag worfflow. Disponível em: https://annualmeeting.siim.org/wp-content/uploads/2024/04/4021_Kuriki_RadPointGPT-A-Generative-Chatbot.pdf, último acesso em 21.04.2025.

[Lee et al. 2021] Lee, J., Wettig, A., and Chen, D. (2021). Phrase retrieval learns passage retrieval, too. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3661–3672.

[Lee et al. 2020] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

[Li et al. 2025] Li, X., Ren, W., Qin, W., Wang, L., Zhao, T., and Hong, R. (2025). Analyzing and reducing catastrophic forgetting in parameter efficient tuning. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

[Li et al. 2023] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., and Zhang, Y. (2023). Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

[Mikolov et al. 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Moreira 2024] Moreira, V. P. (2024). Recuperação de informação. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 21. BPLN, 3 edition.

[Ng et al. 2025] Ng, K. K. Y., Matsuba, I., and Zhang, P. C. (2025). Rag in health care: a novel framework for improving communication and decision-making by addressing llm limitations. *NEJM AI*, 2(1):AIra2400380.

[Orengo and Huyck 2001] Orengo, V. M. and Huyck, C. R. (2001). A stemming algorithmm for the portuguese language. In *spire*, volume 8, pages 186–193.

[Ouyang et al. 2022] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

[Paes et al. 2024] Paes, A., Vianna, D., and Rodrigues, J. (2024). Modelos de linguagem. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 17. BPLN, 3 edition.

[Pan et al. 2024] Pan, J. J., Wang, J., and Li, G. (2024). Survey of vector database management systems. *The VLDB Journal*, 33(5):1591–1615.

[Pennington et al. 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[Porter 1980] Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

[Qdrant 2025] Qdrant (2025). Fusion vs reranking.

[Rafailov et al. 2023] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

[Robertson et al. 2009] Robertson, S., Zaragoza, H., et al. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

[Ru et al. 2024] Ru, D., Qiu, L., Hu, X., Zhang, T., Shi, P., Chang, S., Jiayang, C., Wang, C., Sun, S., Li, H., et al. (2024). Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *arXiv preprint arXiv:2408.08067*.

[Shinn et al. 2023] Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., and Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning.

[Singhal et al. 2023] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

[Singhal et al. 2025] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., et al. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

[Taylor et al. 2022] Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. (2022). Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

[Trulens 2024] Trulens (2024). The rag triad.

[Unlu et al. 2024] Unlu, O., Shin, J., Mailly, C. J., Oates, M. F., Tucci, M. R., Varugheese, M., Wagholikar, K., Wang, F., Scirica, B. M., Blood, A. J., et al. (2024). Retrieval-augmented generation–enabled gpt-4 for clinical trial screening. *NEJM AI*, 1(7):AIoa2400181.

[Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Wang et al. 2022] Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

[Wei et al. 2022] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

[Wu et al. 2024] Wu, Z., Hasan, A., Wu, J., Kim, Y., Cheung, J., Zhang, T., and Wu, H. (2024). Knowlab_aimed at mediqa-corr 2024: Chain-of-though (cot) prompting strategies for medical error detection and correction. In *proceedings of the 6th clinical natural language processing workshop*, pages 353–359.

[Xiong et al. 2024] Xiong, G., Jin, Q., Lu, Z., and Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

[Yang et al. 2022] Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., Flores, M. G., et al. (2022). A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

[Zakka et al. 2024] Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., et al. (2024). Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068.

[Zhang et al. 2020] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert.

[Zhang et al. 2024] Zhang, T., Patil, S. G., Jain, N., Shen, S., Zaharia, M., Stoica, I., and Gonzalez, J. E. (2024). Raft: Adapting language model to domain specific rag. In *First Conference on Language Modeling*.

[Zhao et al. 2025] Zhao, J., Ji, Z., Fan, Z., Wang, H., Niu, S., Tang, B., Xiong, F., and Li, Z. (2025). Moc: Mixtures of text chunking learners for retrieval-augmented generation system. *arXiv preprint arXiv:2503.09600*.

[Zheng et al. 2023] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

# Capítulo 7

# Explorando a Fotopletismografia por Imagem: Uma Abordagem Prática para Aplicações Biomédicas

Vitor Kauã Oliveira de Souza (UENP), Alan Floriano (IFPR), Teodiano Freire Bastos-Filho (UFES)

***Abstract***

*This chapter addresses the extraction of cardiac signals through video cameras using imaging photoplethysmography (iPPG). The main objective is to provide a comprehensive analysis of the topic, which has gained significant relevance in recent years. The text explores the theoretical foundations of iPPG signals, the algorithmic principles underlying its main methodologies, the proposed models, and the practical implementation of a model for vital sign extraction. Additionally, case studies and real-world applications of iPPG methods are presented, followed by final considerations on technical challenges and future applications.*

***Resumo***

*Este capítulo aborda a extração de sinais cardíacos por meio de câmeras de vídeo utilizando a fotopletismografia por imagem, conhecida em inglês como Imaging Photoplethysmography (iPPG). O principal objetivo é oferecer uma análise abrangente do tema, que ganhou relevância significativa nos últimos anos. O texto explora os fundamentos teóricos dos sinais de iPPG, os princípios algorítmicos que sustentam suas principais metodologias, os modelos propostos e a implementação prática de um modelo para extração de sinais vitais. Além disso, são apresentados estudos de caso e aplicações reais dos métodos de iPPG, seguidos de considerações finais sobre os desafios técnicos e as aplicações futuras.*

## 7.1. Introdução ao PPG e ao iPPG

A fotopletismografia, conhecida em inglês como *Photoplethysmography* (PPG), mede as variações no volume sanguíneo em camadas vasculares da pele [Allen 2007]. Para isso, utiliza-se uma fonte de luz que incide sobre o tecido, sendo parcialmente espalhada e

absorvida à medida que atravessa diferentes camadas. A luz atenuada, ao ser transmitida ou refletida de volta à superfície do tecido, é captada por um sensor óptico. Esse sensor converte a intensidade da luz detectada em um sinal elétrico, o qual pode ser analisado para a extração de informações fisiológicas relevantes [Kyriacou and Allen 2021].

Esse sinal apresenta dois tipos de variações: uma de alta frequência (componente AC), que reflete as mudanças no volume arterial a cada batimento cardíaco, e uma de variação lenta, quase estática (DC). A componente DC contém informações sobre respiração, fluxo venoso, atividades do sistema nervoso simpático e termorregulação. Essa tecnologia é essencial para monitoramento cardíaco e diversas aplicações médicas, pois permite avaliar de forma simples e não invasiva a circulação sanguínea [Allen 2007].

Os primeiros relatos sobre fotopletismografia datam da década de 1930, mas foi apenas no início da década de 1970 que a técnica passou a ser aplicada em ambientes clínicos, com o surgimento dos primeiros oxímetros de pulso [Kyriacou and Allen 2021]. Desde então, a fotopletismografia tem sido amplamente explorada em dispositivos médicos, como oxímetros e monitores de frequência cardíaca, além de ser incorporada em tecnologias vestíveis voltadas para o monitoramento da saúde.

A fotopletismografia pode operar em duas configurações principais: modo de transmissão e modo de reflexão, determinadas pelo posicionamento da fonte de luz e do fotodetector [Sun and Thakor 2016]. No modo de transmissão, o tecido é colocado entre a fonte e o detector. Dessa forma, esse método é restrito a locais com pouca espessura, como dedos e lóbulos das orelhas, mas é sensível a variações ambientais e pode interferir em atividades cotidianas [Sun and Thakor 2016]. No método de PPG por transmissão, a luz emitida por um LED atravessa o tecido e é detectada por um fotodiodo posicionado do lado oposto. A quantidade de luz transmitida varia conforme o volume sanguíneo nos vasos, que oscila conforme a pulsação cardíaca.

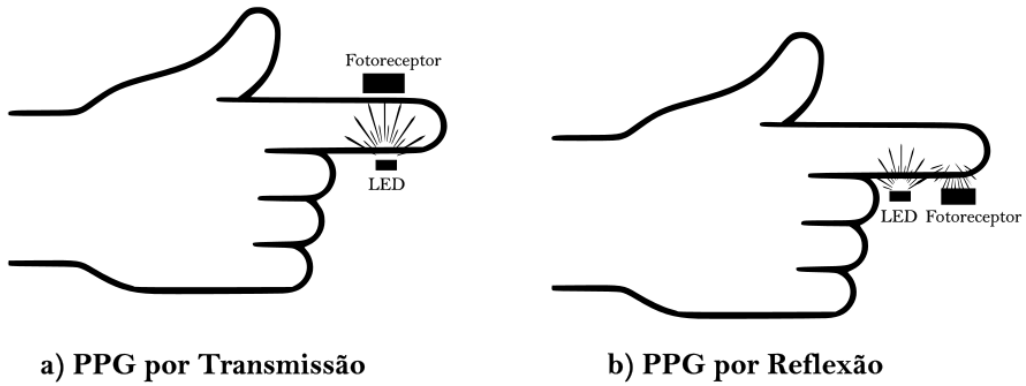



**Figura 7.1. (a) PPG no modo transmissão, (b) PPG no modo reflexão.**

Já no modo de reflexão, o fotodetector mede a luz que é refletida de volta, permitindo medições em praticamente qualquer área da pele. Esse método é mais utilizado em dispositivos vestíveis, como relógios inteligentes, que monitoram sinais vitais como a frequência cardíaca e a variabilidade do pulso [Sun and Thakor 2016]. Para evitar interferência direta da fonte de luz, é utilizado um escudo opaco entre os componentes

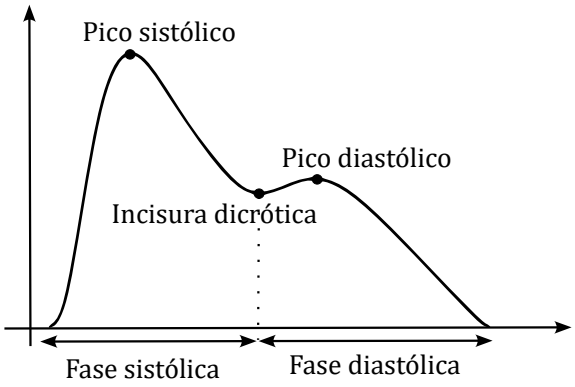



**Figura 7.2. Forma de onda do sinal de PPG do dedo.**

ópticos. Apesar da maior flexibilidade de aplicação, esse método exige uma boa fixação do sensor em superfícies cutâneas planas para garantir a precisão das medições [Sun and Thakor 2016].

A fotopletismografia possui ampla aplicação em dispositivos de monitoramento da saúde, sendo especialmente utilizada em oxímetros de pulso. Esses dispositivos ópticos empregam duas fontes de luz, tipicamente nas faixas do vermelho (660 nm) e do infravermelho (940 nm), para realizar a medição contínua da oxigenação arterial do sangue [Kyriacou and Allen 2021]. A partir dessa técnica, é possível estimar com precisão a saturação de oxigênio ($SpO_2$), a frequência cardíaca e outros parâmetros fisiológicos relevantes.

A Figura 7.2 ilustra uma forma de onda típica do sinal de PPG obtido na região do dedo. A partir dessa forma de onda, é possível extrair diversos parâmetros fisiológicos relevantes. Entre os principais, destacam-se a frequência cardíaca, obtida pela contagem dos picos sistólicos ao longo do tempo, e a variabilidade da frequência cardíaca, que reflete o equilíbrio do sistema nervoso autônomo [Allen 2007].

Além do dedo indicador, o sinal de PPG pode ser captado em diversas regiões do corpo, como o rosto, o lóbulo da orelha e o pulso, cada uma apresentando características distintas de perfusão sanguínea e resposta óptica. As variações na morfologia do sinal refletem diferenças anatômicas e vasculares específicas de cada local. A Figura 7.3 apresenta um gráfico com sinais de PPG registrados em diferentes regiões corporais, evidenciando essas variações fisiológicas.

Por se tratar de uma técnica não invasiva, a fotopletismografia oferece maior conforto aos usuários, além de ser uma tecnologia de baixo custo, o que amplia seu potencial de acessibilidade no monitoramento de sinais vitais. Nesse cenário, dispositivos vestíveis baseados em PPG, como relógios e pulseiras inteligentes, têm ganhado ampla popularidade, permitindo o acompanhamento contínuo da saúde em situações do dia a dia. Esses dispositivos são capazes de monitorar parâmetros fisiológicos em tempo real e detectar alterações significativas, como arritmias cardíacas, taquicardias e bradicardias, contribuindo para a identificação precoce de possíveis disfunções cardiovasculares.

No entanto, esse método ainda apresenta algumas limitações, como a sensibilidade ao movimento, que pode comprometer a precisão das medições, e a interferência da luz

ambiente, especialmente relevante nos métodos baseados em reflexão [Kyriacou and Allen 2021].

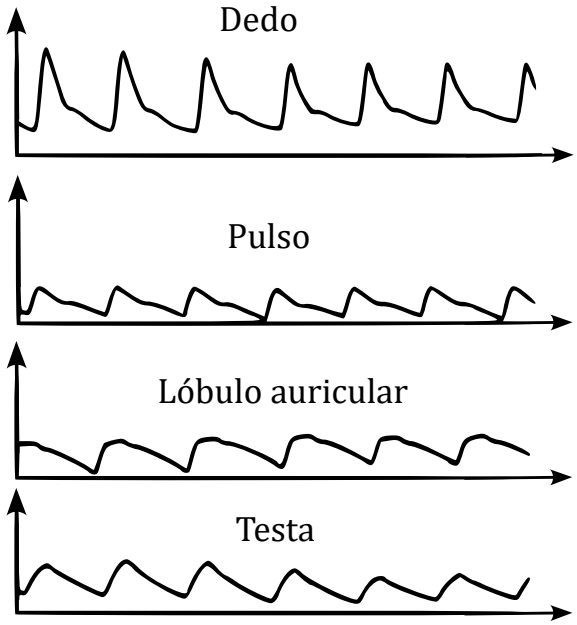



**Figura 7.3. Gráfico representando os sinais de PPG em diferentes regiões do corpo, mostrando as variações na perfusão e características vasculares do corpo humano.**

Nos últimos anos, a fotopletismografia por imagem, conhecida em inglês como *Imaging Photoplethysmography* (iPPG), emergiu como uma evolução dessa tecnologia [Wang et al. 2017a]. Diferente da técnica tradicional de PPG, que utiliza sensores de contato com a pele, a fotopletismografia por imagem usa câmeras de vídeo para capturar variações no volume sanguíneo [Xiao et al. 2024]. A técnica envolve a captura de sequências de imagens da pele e a análise das variações nas propriedades ópticas da pele, que são causadas pelas flutuações no volume de sangue nos vasos sanguíneos. O processamento das imagens permite a medição de parâmetros fisiológicos como a frequência cardíaca e a saturação de oxigênio, sem necessidade de contato direto com o paciente [Wang et al. 2024].

Entre os sensores ópticos utilizados, as câmeras digitais convencionais ganham destaque por sua ampla disponibilidade, baixo custo e boa sensibilidade às variações cromáticas provocadas pelo pulso sanguíneo. Essas características fazem com que sejam amplamente aplicadas em soluções práticas de iPPG, especialmente em contextos clínicos e dispositivos móveis. Por outro lado, câmeras infravermelhas (IR) vêm sendo exploradas como alternativa em ambientes com baixa iluminação ou onde há forte interferência da luz ambiente, apresentando maior robustez, embora exijam técnicas específicas de processamento.

Apesar das vantagens do monitoramento remoto por imagem, a técnica de iPPG ainda enfrenta desafios importantes. A qualidade das medições pode ser comprometida por variações na iluminação ambiente, exigindo condições controladas para garantir maior precisão. Além disso, o movimento do indivíduo ou da câmera pode introduzir ruídos nos sinais capturados, dificultando a extração confiável das informações fisiológicas

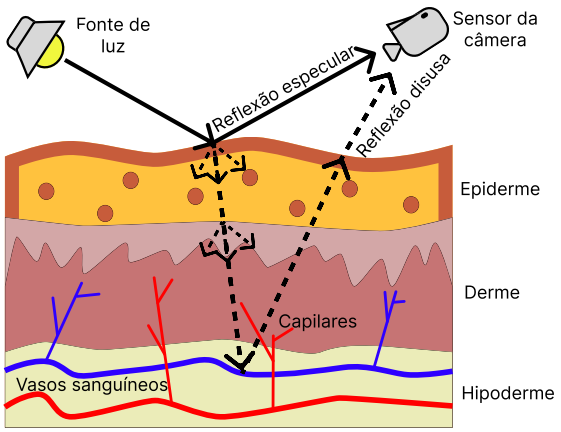



**Figura 7.4. Modelo de Reflexão da Pele [Wang et al. 2017a].**

[Sun and Thakor 2016].

Ainda assim, a fotopletismografia, especialmente em sua vertente baseada em imagem, representa um avanço significativo no monitoramento de saúde não invasivo. Com o contínuo progresso tecnológico, essa técnica oferece um caminho promissor para a coleta remota e em tempo real de sinais vitais, ampliando as possibilidades de cuidados médicos mais acessíveis, confortáveis e eficientes.

## 7.2. Fundamentação teórica do processamento de sinais de iPPG

A fundação teórica do processamento de sinais de iPPG se baseia no Modelo de Reflexão da Pele, ou do inglês, *Skin Reflection Model*. Esse modelo emprega representações matemáticas com matrizes e vetores para descrever os canais de cor, vetores unitários e outros elementos ópticos relevantes. Para compreender adequadamente a extração de sinais cardíacos por meio de métodos de iPPG, é essencial partir de uma formulação básica que considere as propriedades ópticas e fisiológicas envolvidas na reflexão da luz pela pele. Esse modelo fornece uma base sólida para a análise dos desafios encontrados e permite entender como diferentes abordagens de iPPG abordam e solucionam esses problemas [Wang et al. 2017a].

Para o modelo, ilustrado na Figura 7.4, considera-se uma fonte de luz que ilumina uma área de tecido humano que contém fluxo sanguíneo, enquanto uma câmera de vídeo registra essa região. Assume-se que a fonte de luz possui uma composição espectral constante, embora sua intensidade possa variar. A intensidade da luz captada pela câmera depende da distância entre a fonte de luz, o tecido da pele e o sensor da câmera. A pele observada pela câmera apresenta uma coloração que varia ao longo do tempo, devido a movimentos que causam mudanças na reflexão especular e na intensidade da luz, bem como às variações no fluxo sanguíneo, que alteram a cor da pele. Essas variações temporais são proporcionais à intensidade da luz refletida.

Assim, com base no modelo descrito, a reflexão de cada pixel correspondente à pele em uma sequência de imagens pode ser representada como uma função do tempo nos canais RGB, como:

$$C_k(t) = I(t) \cdot (v_s(t) + v_d(t) + v_n(t)), \tag{1}$$

onde $C_k(t)$ representa os canais RGB do $k$-ésimo pixel da pele, e $I(t)$ indica o nível de intensidade da iluminação, o qual incorpora variações de intensidade provocadas tanto pela fonte de luz quanto pelas mudanças na distância entre a fonte, o tecido da pele e a câmera. Essa intensidade $I(t)$ é modulada por dois componentes no modelo: a reflexão especular $v_s(t)$ e a reflexão difusa $v_d(t)$. A dependência temporal desses termos se deve tanto aos movimentos corporais quanto às variações induzidas pelo fluxo sanguíneo. Por fim, o componente $v_n(t)$ representa o ruído introduzido pela quantização do sensor da câmera.

A reflexão especular age como um espelho no tecido da pele, refletindo a luz que ilumina a superfície da pele. Essa luz que é refletida não possui nenhuma informação do pulso cardíaco do indivíduo. Dessa forma, a composição espectral é equivalente à fonte de luz. Ela é dependente do tempo no sentido que o movimento do corpo influencia a estrutura geométrica entre a fonte de luz, a superfície da pele e a câmera. A reflexão especular $v_s(t)$ é denotada por:

$$v_s(t) = u_s \cdot (s_0 + s(t)) \tag{2}$$

onde $u_s$ denota o vetor unitário correspondente à cor da luz no espectro. As variáveis $s_0$ e $s(t)$ representam as componentes da reflexão especular: $s_0$ é a parte estacionária, enquanto $s(t)$ é a parte variável, induzida pelo movimento.

Já a reflexão difusa está associada à absorção e dispersão da luz no tecido cutâneo. A presença de hemoglobina e melanina nesse tecido confere uma cromaticidade característica à componente difusa $v_d$. Essa componente varia ao longo do tempo devido às alterações no volume sanguíneo, sendo, portanto, uma função do tempo. A reflexão difusa $v_d(t)$ é representada por:

$$v_d(t) = u_d \cdot d_0 + u_p \cdot p(t), \tag{3}$$

onde $u_d$ denota o vetor unitário correspondente à cor do tecido da pele, $d_0$ representa a intensidade da componente difusa estacionária, $u_p$ indica a contribuição pulsátil relativa nos canais RGB e $p(t)$ representa o sinal de pulso ao longo do tempo. Substituindo esses valores na equação do modelo, obtém-se:

$$C_k = I(t) \cdot (u_s \cdot (s_0 + s(t)) + u_d \cdot d_0 + u_p \cdot p(t)) + v_n(t). \tag{4}$$

As componentes estacionárias das reflexões difusa e especular podem ser combinadas em um único termo, representando a reflexão estacionária da pele da seguinte forma:

$$u_c \cdot c_0 = u_s \cdot s_0 + u_d \cdot d_0, \tag{5}$$

onde $u_c$ denota o vetor unitário correspondente à cor da reflexão da pele e $c_0$ representa a intensidade dessa reflexão. Substituindo esses termos na fórmula do modelo, tem-se:

$$C_k(t) = I(t) \cdot (u_c \cdot c_0 + u_s \cdot s(t) + u_p.p(t)) + v_n(t). \quad (6)$$

A intensidade da luz capturada, $I(t)$, também pode ser expressa como a combinação de uma componente estacionária $I_0$ e uma componente variante no tempo, modelada como $I_0 \cdot i(t)$ — por exemplo, variações de intensidade induzidas por movimento, observadas pela câmera, sendo proporcionais ao nível de intensidade. Os sinais $i(t)$, $s(t)$ e $p(t)$ são considerados com média zero. Substituindo a modelagem na fórmula do modelo, tem-se:

$$C_k(t) = I_0 \cdot (1 + i(t)) \cdot (u_c \cdot c_0 + u_s \cdot s(t) + u_p.p(t)) + v_n(t). \quad (7)$$

Observa-se que a reflexão especular tende a ser o componente dominante na equação, frequentemente ofuscando as demais contribuições. Assim, assume-se a existência de mecanismos capazes de rejeitar as regiões em que a reflexão especular é predominante. Dessa forma, consideram-se apenas os pixels $k$ nos quais $u_d$ contribui de maneira significativa para a reflexão difusa, ou seja, sua influência não é desprezível.

De modo geral, o objetivo dos métodos iPPG é extrair o sinal $p(t)$ a partir de $C_k(t)$, filtrando as contribuições da reflexão especular para isolar a informação pulsátil da reflexão difusa.

## 7.3. Métodos para a extração do sinal de iPPG

Os métodos convencionais de extração de sinal de iPPG baseiam-se em modelos matemáticos cujo objetivo é eliminar artefatos de ruído e movimento presentes nas imagens capturadas pela câmera. Esses métodos geralmente consistem em calcular a média dos valores das componentes RGB dentro de regiões de interesse (ROIs) em cada quadro do vídeo, construindo, assim, sinais temporais RGB que são posteriormente processados para a extração do sinal de pulso.

A etapa de média dos pixels contribui para a redução do erro introduzido pela câmera. Com base na equação completa do modelo, assume-se que uma quantidade suficiente de pixels esteja focada em regiões de pele com propriedades ópticas comparáveis. Dessa forma, a média $C_k(t)$ sobre os pixels de pele pode ser aproximada como [Wang et al. 2017a]:

$$C_k(t) \approx I_0 \cdot (1 + i(t)) \cdot (u_c \cdot c_0 + u_s \cdot s(t) + u_p \cdot p(t)) \quad (8)$$

A representação fornece um sinal $C_k(t)$ em que a quantização de ruído $v_n(t)$ se torna insignificante, desde que haja um número suficientemente grande de pixels de pele na média. No entanto, observa-se que, quando essa média é calculada em uma área pequena, contendo poucos pixels de pele, o erro de quantização introduzido pela câmera permanece elevado e, portanto, não pode ser desprezado.

Além disso, observa-se que os vetores de cor envolvidos na equação não dependem diretamente da posição dos pixels de pele na imagem. O sinal $C(t)$ resultante é essencialmente uma trajetória no espaço RGB ao longo do tempo $t$, a qual pode ser expandida e simplificada da seguinte forma:

$$\begin{aligned} C(t) =& u_c \cdot I_0 \cdot c_0 + u_s \cdot I_0 \cdot s(t) + u_p \cdot I_0 \cdot p(t) + \\ & u_c \cdot I_0 \cdot c_0 \cdot i(t) + u_s \cdot I_0 \cdot s(t) \cdot i(t) + \\ & u_p \cdot I_0 \cdot p(t) \cdot i(t) \\ & \approx u_c \cdot I_0 \cdot c_0 + u_c \cdot I_0 \cdot c_0 \cdot i(t) + u_s \cdot I_0 \cdot s(t) + \\ & u_p \cdot I_0 \cdot p(t) \end{aligned} \tag{9}$$

A aproximação se mantém por conta de que todos os termos da modulação AC são muito menores em ordem de magnitude do que o termo DC e portanto os termos de modulação do produto da equação podem ser ignorados. A aproximação apresentada no formato de equação mostra que a observação $C(t)$ é uma mistura linear de três fontes de sinais $i(t)$, $s(t)$ e $p(t)$. Isso implica que usando de uma projeção linear é possível separar essas três fontes de sinais. Dito isso, o objetivo de extrair o sinal de pulso dos sinais RGB observados pode ser traduzido como a definição de um sistema de projeção para decompor $C(t)$.

Dessa forma, diversas abordagens têm sido propostas para realizar a decomposição do sinal observado $C(t)$ e extrair o sinal de pulso. Entre as abordagens mais simples, destacam-se os métodos que utilizam diretamente os canais de cor da imagem, como o método **G**, que considera apenas o canal verde [Verkruysse et al. 2008], e o G-R [Hülsbusch 2008], que calcula a diferença entre os canais verde e vermelho. Esses métodos baseiam-se na observação de que o canal verde apresenta maior sensibilidade às variações do volume sanguíneo, sendo, portanto, um bom indicador do sinal de pulso.

Além desses, existem métodos mais sofisticados, como os baseados em *separação cega de fontes*, que empregam técnicas como a Análise de Componentes Independentes [Poh et al. 2010] e a Análise de Componentes Principais [Lewandowska and Nowak 2012] para isolar os sinais de interesse sem conhecimento prévio sobre as fontes. Outra categoria importante são os *métodos baseados em modelos* (*Model-Based Methods*), que utilizam descrições analíticas da interação entre a luz e o tecido biológico para estimar o sinal de pulso. Dentro dessa categoria, destacam-se os métodos CHROM [De Haan and Jeanne 2013] e POS [Wang et al. 2017a], amplamente utilizados por sua robustez e bom desempenho em condições desafiadoras de iluminação e movimento [Xiao et al. 2024].

Todas as abordagens e seus respectivos métodos convencionais têm o mesmo objetivo: extrair o sinal de pulso de $C(t)$ por meio de sua decomposição. No entanto, cada abordagem se distingue das demais pela metodologia específica empregada para alcançar esse objetivo. A seguir, serão abordados os métodos e suas metodologias matemáticas, incluindo também técnicas baseadas em aprendizado profundo utilizando redes neurais.

### 7.3.1. Métodos convencionais de iPPG

Nesta parte do capítulo serão descritos os principais métodos convencionais de extração de sinais fisiológicos por meio da fotopletismografia por imagem. O objetivo é apresentar de forma clara os princípios algorítmicos fundamentais que nortearam o desenvolvimento dessas técnicas. Todos os métodos recebem como entrada uma sequência temporal de sinais RGB normalizados, denotada por $\mathbf{x}(t) = (x_r(t), x_g(t), x_b(t))^\top$, obtida a partir da média espacial da imagem e de filtros para remoção de tendências e ruídos. O objetivo final de cada abordagem é gerar uma sequência monovariada $y(t)$ que represente uma estimativa do sinal de volume de pulso.

#### 7.3.1.1. Método ICA

A Análise de Componentes Independentes, conhecida em inglês como *Independent Component Analysis* (ICA), é uma técnica estatística que visa decompor uma mistura linear de sinais sob a suposição de independência estatística e não-gaussianidade [Poh et al. 2010]. No contexto da iPPG, considera-se que os sinais temporais RGB $\mathbf{x}(t)$ representam medições multivariadas resultantes da mistura de três fontes latentes $\mathbf{z}(t) = (z_1(t), z_2(t), z_3(t))^\top$. Esse processo de mistura instantânea pode ser descrito por:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{z}(t), \tag{10}$$

onde $\mathbf{A} \in \mathbb{R}^{3\times 3}$ é uma matriz de mistura desconhecida. O objetivo da ICA é estimar uma matriz de separação $\mathbf{W}$ tal que:

$$\hat{\mathbf{z}}(t) = \mathbf{W}\mathbf{x}(t) \approx \mathbf{z}(t). \tag{11}$$

Esse problema é conhecido como separação cega de fontes, e diferentes abordagens podem ser utilizadas para estimar $\mathbf{W}$ com base na não-gaussianidade das componentes. No entanto, as soluções obtidas por ICA apresentam indeterminações típicas: as fontes podem ser recuperadas com escala, permutação e atrasos arbitrários. Apesar disso, a forma de onda original costuma ser preservada, o que é suficiente para análise no domínio tempo-frequência, como requerido na iPPG.

Uma limitação importante é que, após a separação, não é possível saber diretamente qual das três componentes contém a informação de pulso mais relevante. Para lidar com isso, adota-se uma abordagem simples: calcula-se a densidade espectral de potência (PSD) normalizada de cada componente separada, e escolhe-se aquela com o maior pico de frequência ou maior razão sinal-ruído (SNR) na faixa de 40 a 220 BPM.

**Entrada:** $\mathbf{x}(t) = [x_r(t), x_g(t), x_b(t)]^\top$ com $N$ amostras

Aplicar ICA: $\hat{\mathbf{z}}(t) \leftarrow \text{ICA}(\mathbf{x}(t))$

**Para** (cada componente $\hat{z}_i(t)$)

Calcular PSD e identificar pico de frequência

**Fim-Para**

**Saída:** $y(t) = \hat{z}_i(t)$ com maior pico na faixa de 40–220 BPM

#### 7.3.1.2. Método PCA

A Análise de Componentes Principais, conhecida em inglês como *Principal Component Analysis* (PCA), é uma técnica amplamente utilizada em estatística multivariada, que busca maximizar a variância, minimizar as covariâncias e reduzir a dimensionalidade dos dados. Ao aplicar PCA ao vetor $\mathbf{x}(t)$, obtêm-se componentes ortogonais que explicam a variabilidade do sinal. A saída $y(t)$ é escolhida como a componente cuja energia espectral está concentrada na faixa cardíaca [Lewandowska and Nowak 2012].

No contexto da iPPG, assim como no método ICA, é preciso realizar a escolha da componente que melhor representa o sinal de pulso. Para resolver isso, calcula-se a densidade espectral de potência normalizada de cada componente principal e seleciona-se a componente com maior pico de frequência ou maior razão sinal-ruído dentro da faixa de 40 a 220 BPM.

**Entrada:** Sinal RGB $\mathbf{x}(t) = [x_r(t), x_g(t), x_b(t)]^\top$ com $N$ amostras

Aplicar PCA: $\hat{\mathbf{z}}(t) \leftarrow \text{PCA}(\mathbf{x}(t))$

**Para** (cada componente $\hat{z}_i(t)$)

Calcular PSD($\hat{z}_i(t)$)

**Fim-Para**

**Saída:** $y(t) = \hat{z}_i(t)$ com maior pico na faixa de 40–220 BPM

#### 7.3.1.3. Método G

Este método simples considera que o canal verde da imagem apresenta a maior quantidade de informação relacionada ao variação do volume do sangue, devido à absorção da luz verde pela hemoglobina [Verkruysse et al. 2008]. Assim, o sinal extraído é diretamente:

$$y(t) = x_g(t). \tag{12}$$

**Entrada:** Sinal RGB temporal $\mathbf{x}(t) = [x_r(t), x_g(t), x_b(t)]^\top$ com $N$ amostras

**Saída:** Sinal de pulso $y(t)$

**Para** ($t = 1$ até $N$)

$y(t) = x_g(t)$

**Fim-Para**

**Retornar** $y(t)$

#### 7.3.1.4. Método G-R

O método G-R é uma técnica simples para extração do sinal de pulso a partir do vídeo, baseada na diferença entre os canais verde e vermelho do sinal RGB [Hülsbusch 2008]. A justificativa para essa abordagem está no fato de que o canal verde contém uma maior informação pulsátil, pois a hemoglobina absorve mais luz nesta faixa espectral, enquanto o canal vermelho apresenta menor conteúdo de pulsatilidade e pode ser usado para atenuar ruídos comuns ao sinal.

Assim, o sinal de pulso é obtido pela subtração do canal vermelho do canal verde:

$$y(t) = x_g(t) - x_r(t). \tag{13}$$

Devido a sua simplicidade, o método G-R pode apresentar limitações em ambientes com variações de iluminação intensa ou movimentos.

**Algorithm 1** Método G-R

**Entrada:** Sinal RGB temporal $\mathbf{x}(t) = [x_r(t), x_g(t), x_b(t)]^\top$ com $N$ amostras

**Saída:** Sinal de pulso $y(t)$

**Para** ($t = 1$ até $N$)

$y(t) = x_g(t) - x_r(t)$

**Fim-Para**

**Retornar** $y(t)$

Em aplicações práticas, recomenda-se complementar o método G-R com técnicas de pré-processamento, como suavização temporal e remoção de artefatos, para melhorar a qualidade do sinal extraído. Ainda assim, por sua simplicidade computacional, é uma alternativa eficiente para sistemas com recursos limitados.

#### 7.3.1.5. Método CHROM

O método CHROM elimina o componente de reflexão especular da pele usando sinais de crominância derivados de combinações lineares dos canais RGB [De Haan and Jeanne 2013].

De forma simplificada, a luz refletida da pele é composta por dois componentes, segundo o modelo dicromático de reflexão: um componente de reflexão difusa, cujas variações estão relacionadas ao ciclo cardíaco, e um componente de reflexão especular, que apresenta a cor da fonte luminosa, mas não contém sinal de pulso. A contribuição relativa desses dois componentes varia ao longo do tempo devido ao movimento da pessoa e à geometria entre câmera, pele e fonte de luz, causando dificuldades para os algoritmos de iPPG que não eliminam o componente especular aditivo. O método CHROM elimina o componente especular utilizando diferenças de cor, ou seja, sinais de crominância.

Dado o sinal RGB $\mathbf{x}(t)$, o método CHROM realiza inicialmente uma normalização por desvio padrão zero e projeta os valores normalizados em dois vetores de crominância ortogonais, definidos por:

$$X_{\text{CHROM}}(t) = 3x_r(t) - 2x_g(t), \tag{14}$$

$$Y_{\text{CHROM}}(t) = 1.5x_r(t) + x_g(t) - 1.5x_b(t). \tag{15}$$

O sinal iPPG final é então calculado por:

$$y(t) = X_{\text{CHROM}}(t) - \alpha\, Y_{\text{CHROM}}(t), \tag{16}$$

onde

$$\alpha = \frac{\sigma(X_{\text{CHROM}}(t))}{\sigma(Y_{\text{CHROM}}(t))}, \tag{17}$$

e $\sigma(\cdot)$ representa o desvio padrão.

Para lidar com variações temporais e manter a consistência do sinal ao longo do tempo, o sinal $y(t)$ é processado em janelas temporais com sobreposição (*overlap*). O resultado de cada janela é centralizado (remoção da média) e somado ao sinal final com sobreposição, técnica conhecida como *overlap-adding*. Para reduzir artefatos na junção entre janelas, pode se aplicar janelas suavizantes como a janela de Hamming no processo de soma com sobreposição. Isso permite atenuar descontinuidades entre janelas e preservar melhor o conteúdo espectral do sinal cardíaco.

**Entrada:** Vetor RGB temporal $x(t) = [x_r(t), x_g(t), x_b(t)]^\top$ com $N$ amostras

**Saída:** Sinal de pulso $H$

**Inicializar** $H = \text{zeros}(1, N)$, $l = 48$ (para câmera de 30 fps)

**Para** ($n = 1$ até $N$):

**Se** ($m = n - l + 1 > 0$):

$R = x_r(m:n),\ G = x_g(m:n),\ B = x_b(m:n)$

$R = \frac{R}{\mu(R)} - 1, \quad G = \frac{G}{\mu(G)} - 1, \quad B = \frac{B}{\mu(B)} - 1$ {Normalizar temporalmente}

$X = 3 \cdot R - 2 \cdot G$

$Y = 1.5 \cdot R + G - 1.5 \cdot B$

$\alpha = \frac{\sigma(X)}{\sigma(Y)}$

$h = X - \alpha \cdot Y$ {Sinal bruto}

$H(m:n) = H(m:n) + (h - \mu(h))$ {Soma com sobreposição}

**Fim-Se**

**Fim-Para**

**Retornar:** $H$

### 7.3.1.6. Método POS

O método *Plane-Orthogonal-to-Skin* (POS) tem como objetivo, assim como o método CHROM, eliminar os reflexos especulares na superfície da pele. Para isso, ele define um plano perpendicular ao tom de pele dominante dentro do espaço RGB normalizado temporalmente.

A Figura 7.5 busca ilustrar a distribuição da força pulsátil dentro do plano ortogonal à tonalidade da pele como uma função de um vetor de projeção **z**. São apresentados exemplos de vetores $\mathbf{z}_1$ e $\mathbf{z}_2$, que geram sinais pulsáteis de polaridades opostas (anti-fásicos), e um vetor $\mathbf{z}_3$ que resulta em um sinal com baixo conteúdo pulsátil, ou seja, dominado por ruído.

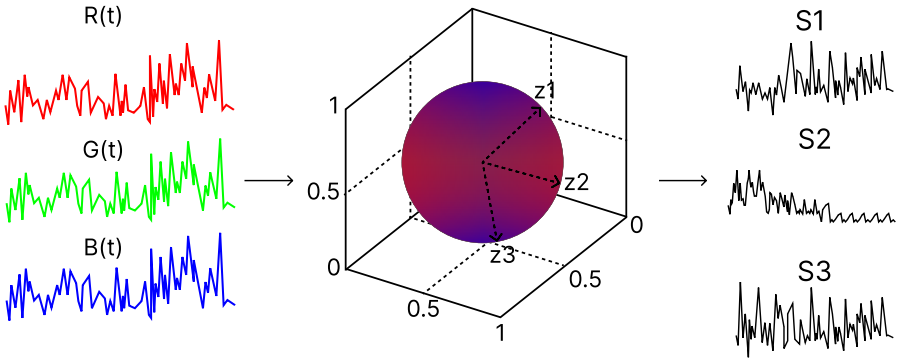



**Figura 7.5. Plano ortogonal da pele [Wang et al. 2017a].**

Especificamente, dado o vetor RGB temporal $\mathbf{x}(t)$, o método POS é composto por três etapas principais. A primeira etapa consiste em uma normalização temporal dos sinais de cor, que remove variações de iluminação constante. Em seguida, os sinais são projetados sobre um plano ortogonal à tonalidade da pele, utilizando combinações lineares dos canais RGB:

$$X_{\text{POS}}(t) = x_g(t) - x_b(t), \tag{18}$$

$$Y_{\text{POS}}(t) = x_g(t) + x_b(t) - 2x_r(t). \tag{19}$$

Por fim, realiza-se um ajuste da direção de projeção dentro da região delimitada pelas componentes anteriores, resultando no sinal iPPG:

$$y(t) = X_{\text{POS}}(t) + \alpha, Y_{\text{POS}}(t), \tag{20}$$

onde o fator de ponderação $\alpha$ é dado por:

$$\alpha = \frac{\sigma(X_{\text{POS}}(t))}{\sigma(Y_{\text{POS}}(t))}, \tag{21}$$

e $\sigma(\cdot)$ representa o desvio padrão, como no método CHROM.

O método POS apresenta uma diferença fundamental em relação ao CHROM: enquanto este utiliza sinais projetados em antifase, o POS define diretamente dois eixos de projeção que produzem sinais em fase, o que tende a melhorar a consistência da pulsação extraída.

Além disso, para aumentar a relação sinal-ruído (SNR), o sinal é extraído em janelas temporais menores da sequência de vídeo. Cada segmento é processado separadamente e, ao final, os sinais parciais são recombinados utilizando a técnica de soma com sobreposição (*overlap-adding*), o que preserva a continuidade do sinal final e melhora a qualidade espectral.

**Entrada:** Vetor RGB temporal $x(t) = [x_r(t), x_g(t), x_b(t)]^\top$ com $N$ amostras

**Saída:** Sinal de pulso $H$

**Inicializar** $H = \text{zeros}(1, N)$, $l = 48$ (para câmera de 30 fps)

**Para** ($n = 1$ até $N$):

**Se** ($m = n - l + 1 > 0$):

$R = x_r(m:n)$, $G = x_g(m:n)$, $B = x_b(m:n)$

$R = \frac{R}{\mu(R)} - 1, \quad G = \frac{G}{\mu(G)} - 1, \quad B = \frac{B}{\mu(B)} - 1$ {Normalizar temporalmente}

$X = G - B$

$Y = G + B - 2 \cdot R$

$\alpha = \frac{\sigma(X)}{\sigma(Y)}$

$h = X + \alpha \cdot Y$ {Sinal bruto}

$H(m:n) = H(m:n) + (h - \mu(h))$ {Soma com sobreposição}

**Fim-Se**

**Fim-Para**

**Retornar:** $H$

### 7.3.2. Métodos de extração de sinal de iPPG com *Deep Learning*

O Aprendizado Profundo, ou do inglês, *Deep Learning*, é uma subárea da Inteligência Artificial que utiliza arquiteturas de redes neurais profundas, compostas por múltiplas camadas não lineares, para modelar padrões complexos em grandes volumes de dados [Xiao et al. 2024]. Diferentemente das abordagens tradicionais de aprendizado de máquina, que dependem de engenharia manual de atributos, os métodos de *Deep Learning* aprendem representações hierárquicas diretamente a partir de dados brutos, como imagens e vídeos [Xiao et al. 2024]. Essa capacidade os torna particularmente adequados para tarefas como a extração de sinais de iPPG.

No contexto do *Deep Learning*, destacam-se as abordagens supervisionadas para estimativa da frequência cardíaca, nas quais os modelos são treinados com dados rotulados. Isso exige um volume considerável de exemplos anotados para garantir um treinamento eficaz [Xiao et al. 2024].

Considerando a ampla variedade de técnicas atualmente disponíveis para a extração de sinais de iPPG por meio de métodos de *Deep Learning* [Xiao et al. 2024], verifica-se que esta é uma área de pesquisa em contínua expansão, com elevado potencial de inovação e aplicação. Em virtude da extensa quantidade de abordagens propostas na literatura, este capítulo se concentra especificamente nos principais métodos supervisionados, com ênfase naqueles que empregam redes neurais convolucionais, conforme descrito nos estudos de referência [Chen and McDuff 2018, Lin et al. 2019, Yu et al. 2019, Zhan et al. 2020, Lampier et al. 2022].

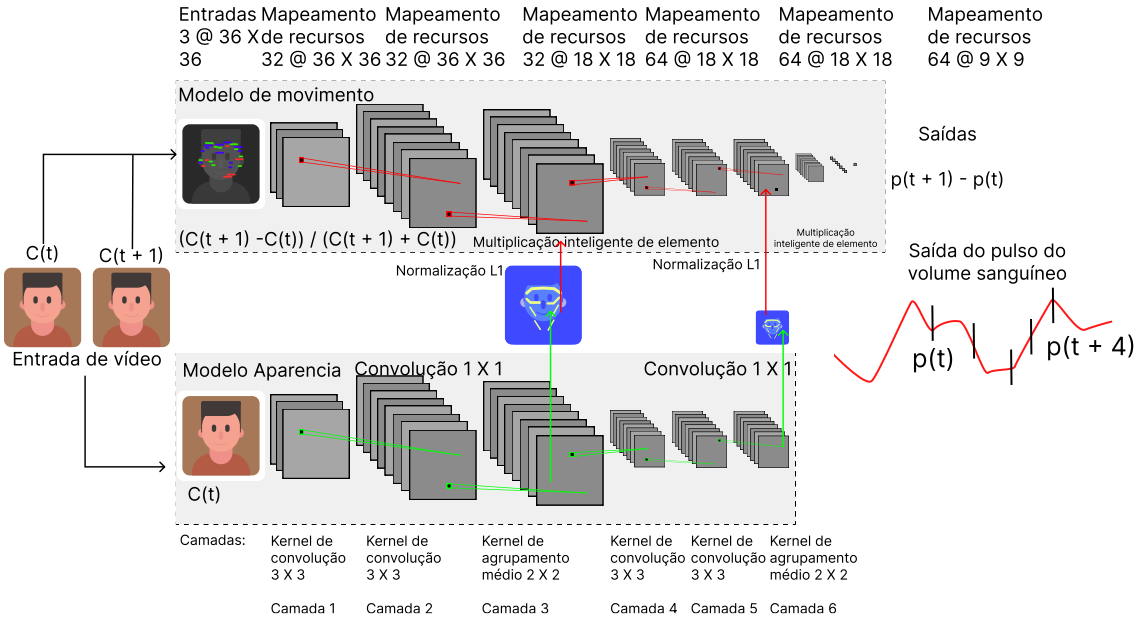



**Figura 7.6. Arquitetura do modelo de convolução DeepPhys [Chen and McDuff 2018].**

#### 7.3.2.1. DeepPhys

O método DeepPhys é um método de extração iPPG com aprendizado supervisionado que utiliza de redes neurais convolucionais bidimensionais (2D CNN) para a extração do sinal. A Figura 7.6 apresenta a arquitetura do modelo DeepPhys, ilustrando uma das inovações do modelo que foi a criação de dois módulos distintos: um modelo de aparência e um modelo de movimento [Chen and McDuff 2018]. O modelo de aparência orienta o modelo de movimento a aprender representações de movimento por meio de um mecanismo de atenção. O modelo de movimento, por sua vez, utiliza a diferença normalizada entre quadros consecutivos como entrada para representar movimentos e mudanças de cor, o que melhora a robustez do DeepPhys frente a movimentos [Chen and McDuff 2018].

Além disso, o DeepPhys incorpora um mecanismo de atenção que gera máscaras de atenção suaves a partir dos quadros brutos do vídeo, atribuindo pesos maiores às regiões da pele que apresentam sinais fisiológicos mais intensos. Esse mecanismo também permite visualizar a distribuição espaço-temporal desses sinais no rosto. Embora o DeepPhys apresente um bom desempenho, em situações com variações de iluminação e artefatos de movimento, ele apresenta uma limitação importante: não consegue capturar adequadamente informações temporais dos sinais de iPPG devido à natureza estática das redes 2D CNN [Chen and McDuff 2018].

### 7.3.3. PhysNet

O método PhysNet é um método proposto que utiliza de redes neurais convolucionais tridimensionais (3D CNN) [Yu et al. 2019]. O modelo PhysNet foi considerado um novo marco entre os modelos 3D CNN.

Isso por que apesar de ignorar o pré-processamento da imagem, trabalhando diretamente com quadros RGB brutos como entrado, a sua arquitetura é capaz de apren-

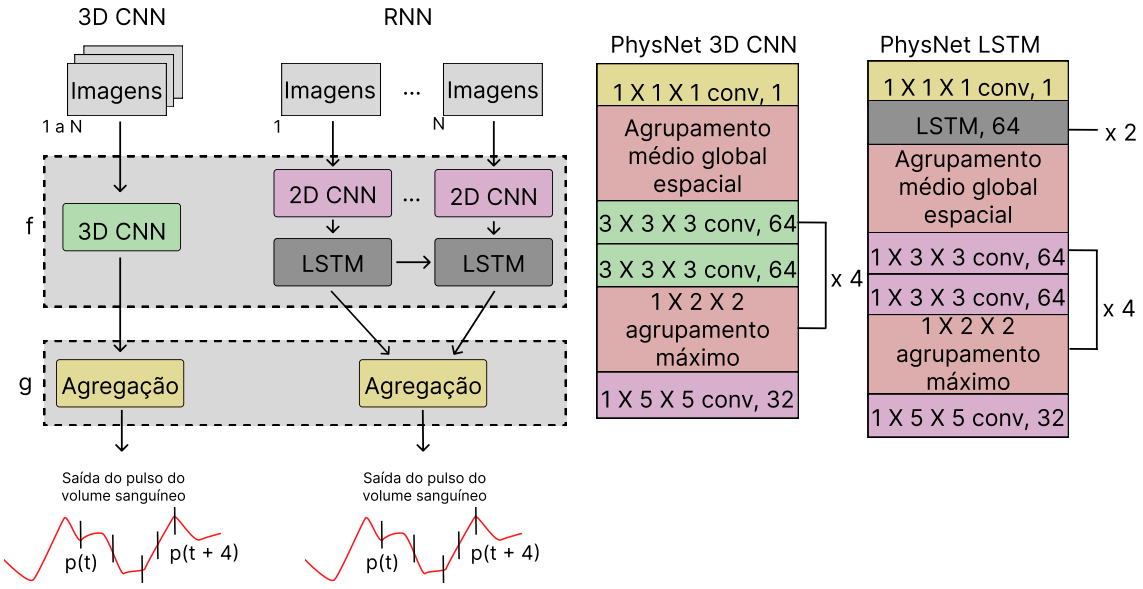



**Figura 7.7. Arquitetura do modelo de convolução PhysNet [Yu et al. 2019].**

der de forma eficiente a característica espaço-temporais presentes nas sequências faciais, produzindo diretamente o sinal de iPPG, sem a necessidade do pós-processamento [Yu et al. 2019]. Na Figura 7.7 é ilustrada os modelos de 3D CNN, RNN e as suas respectivas arquiteturas.

#### 7.3.3.1. TS-CAN

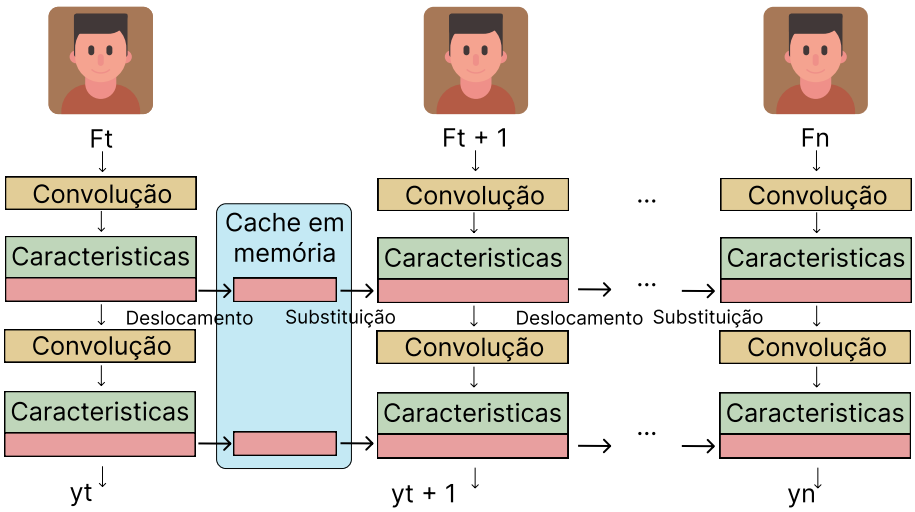



**Figura 7.8. Arquitetura do modelo TSM unidirecional [Lin et al. 2019].**

O método TS-CAN é uma rede neural convolucional bidimensional (2D CNN) proposta para mitigar a limitação do DeepPhys, que não consegue capturar informações temporais de forma adequada. Também conhecido como MTTS-CAN, esse método é construído com base no DeepPhys, incorporando o *Temporal Shift Module* (TSM) para

recuperar a informação temporal que era antes negligenciada. Na Figura 7.8 é ilustrada a arquitetura do modelo TSM unidirecional e as propriedades de recuperação de informação temporal em cache. O princípio do TSM permite a troca de informações entre quadros adjacentes sem o uso de operações convolucionais complexas, apenas movendo blocos nos tensores ao longo do eixo temporal. Isso permite ao modelo capturar, mesmo que parcialmente, a dinâmica temporal dos sinais fisiológicos.

Diferentemente do modelo original DeepPhys, a entrada da rede de aparência no MTTS-CAN não é composta por quadros brutos do vídeo capturado, mas por quadros gerados a partir da média de múltiplos quadros adjacentes. Essa estratégia favorece a extração de informações temporais e melhora a robustez do modelo em relação a variações dinâmicas [Lin et al. 2019].

### 7.3.4. Modelo de CNN baseado na diferença de quadros normalizada

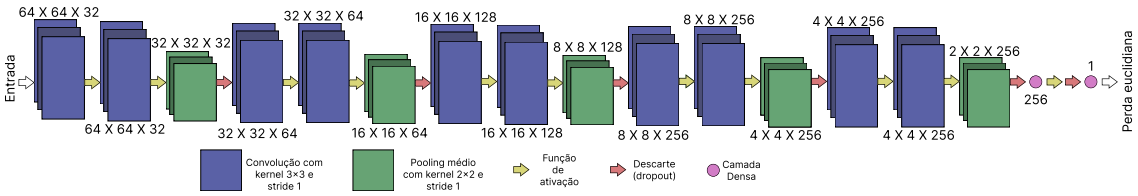



**Figura 7.9. Modelo de CNN baseado na diferença de quadros normalizada [Zhan et al. 2020].**

A arquitetura da rede neural convolucional (CNN) utilizada neste trabalho é ilustrada na Figura 7.9. Foi desenvolvido um modelo baseado na diferença de quadros normalizada, ou do inglês, *normalized frame difference model*, semelhante ao proposto em [Chen and McDuff 2018], com o objetivo de aprender a relação entre variações temporais na imagem obtidas por meio da normalização de diferenças entre quadros e os sinais de referência utilizados como rótulos durante o treinamento.

Diferentemente do modelo [Chen and McDuff 2018], esta arquitetura proposta opta por não empregar o módulo de atenção. Para garantir que a rede aprenda variações de cor associadas aos pixels da pele, a região facial dos sujeitos foi selecionada antes e utilizada como entrada da rede. A arquitetura CNN projetada contém dez camadas convolucionais, todas com kernels de tamanho $3 \times 3$. A cada duas camadas convolucionais, é inserida uma camada de *average pooling* com kernel $2 \times 2$, seguida por uma camada de *dropout*, com o intuito de mitigar o risco de sobreajuste. A função de ativação adotada após cada convolução é a tangente hiperbólica (*tanh*).

Em consonância com tarefas de regressão utilizando CNNs, a função de perda adotada é a distância euclidiana, que mede a discrepância entre a saída da camada totalmente conectada e o rótulo de treinamento, que corresponde ao sinal PPG diferenciado.

#### 7.3.4.1. U-Net RGB-to-PPG

A arquitetura, denominada RGB-to-PPG [Lampier et al. 2022], foi inspirada na U-Net, uma rede convolucional originalmente desenvolvida para segmentação de imagens biológicas [Ronneberger et al. 2015]. No entanto, nesta abordagem, uma camada LSTM é

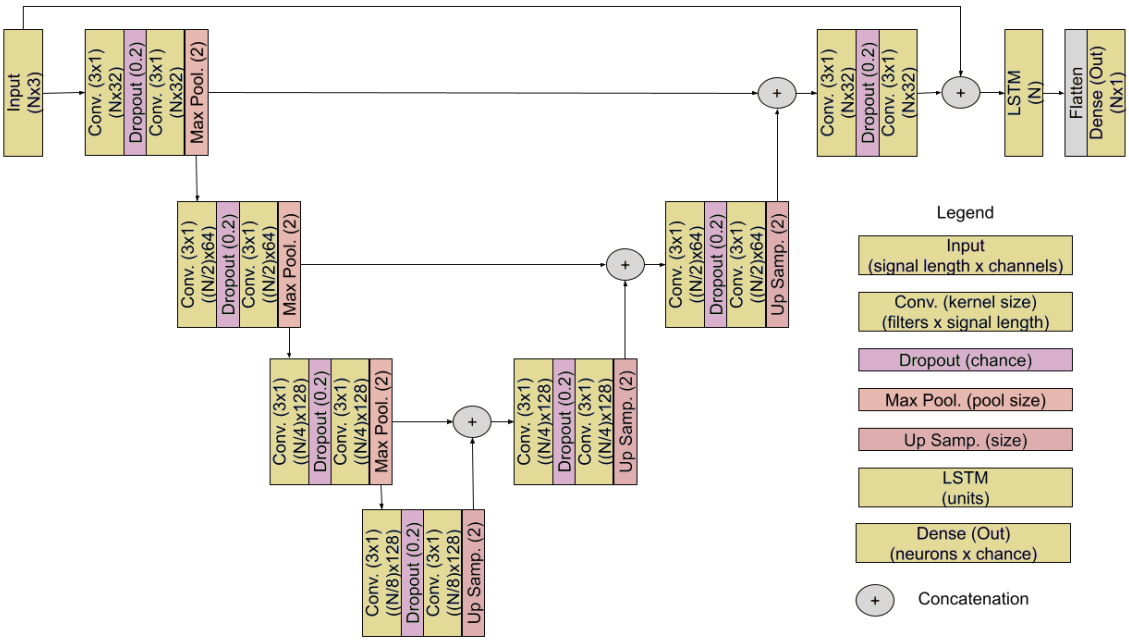



**Figura 7.10. Arquitetura U-Net RGB-to-PPG [Lampier et al. 2022].**

adicionada ao final da U-Net, com o objetivo de aprimorar o desempenho na tarefa de regressão. O modelo proposto recebe como entrada sinais RGB e os transforma em um sinal de pulso, a partir do qual é estimada a frequência de pulso.

A Figura 7.10 ilustra a arquitetura completa da rede proposta. As camadas convolucionais utilizam a função de ativação ReLU, enquanto as camadas LSTM empregam a tangente hiperbólica. A camada densa de saída, por sua vez, utiliza ativação linear. Os círculos representados na figura indicam operações de concatenação, responsáveis por empilhar duas entradas.

A entrada da rede consiste em uma série temporal RGB de dimensão $N \times 3$, em que $N$ representa o número de amostras do sinal de entrada. Esses valores são previamente normalizados para o intervalo de 0 a 1, por meio da divisão por 255, correspondente à intensidade máxima em imagens com resolução de cor de 8 bits.

Para a estimativa da frequência de pulso, identificam-se os picos no sinal de iPPG gerado pelo modelo. A partir desses picos, calcula-se o inverso do intervalo entre ocorrências consecutivas. A mediana dos valores obtidos é então multiplicada por 60, convertendo a frequência de Hertz para batimentos por minuto.

## 7.4. Conjunto de Dados

Os conjuntos de dados, ou *datasets*, desempenham um papel essencial no desenvolvimento de métodos. Eles são utilizados para treinar, validar e testar os modelos, além de permitir a comparação entre diferentes abordagens. Neste capítulo, serão apresentados os *datasets* mais populares da área [Pirzada et al. 2024, Xiao et al. 2024].

Esses *datasets* são amplamente utilizados por fornecerem vídeos faciais acompa-

nhados de sinais fisiológicos sincronizados, como os obtidos por oxímetros, registrados em diferentes condições de iluminação, movimento e qualidade de imagem. A variedade de cenários e de participantes contida nesses conjuntos é essencial para avaliar a robustez e a capacidade de generalização dos modelos de aprendizado profundo. Dessa forma, esses *datasets* se consolidaram como referência na pesquisa e no desenvolvimento de técnicas de extração de sinais de iPPG.

**Tabela 7.1. Conjuntos de dados para estudos de iPPG.**

| Dataset | Indivíduos | Vídeos | Resolução | Padrão-ouro | Público |
|---|---|---|---|---|---|
| DEAP | 32 | 874 | 720 × 576 @ 56 fps | ECG | Sim |
| MAHNOB-HC | 27 | 527 | 1040 × 1392 @ 24 fps | ECG | Sim |
| UBFC-rPPG | 50 | 50 | 640 × 480 @ 30 fps | PPG, HR | Sim |
| PURE | 10 | 60 | 640 × 480 @ 30 fps | PPG, SpO2 | Sim |
| SCAMPS | 2800 | 2800 | 320 × 240 @ 30 fps | PPG, PR, RR | Sim |
| MMPD | 22 | 55 | 1280 × 720 @ 30 fps | PPG, HR | Sim |
| BP4D+ | 140 | 1400 | 1040 × 1392 @ 25 fps | PPG, HR, BP, RR | Sim |
| UBFC-Phys | 56 | 168 | 1024 × 1024 @ 30 fps | PPG, HR | Sim |
| COHFACE | 40 | 160 | 640 × 480 @ 20 fps | PPG | Sim |
| ECG-Fitness | 17 | 204 | 1920 × 1080 @ 30 fps | PPG, ECG | Sim |
| VIPL-HR | 107 | 3130 | 960 × 720, 1920 × 1080, 640 × 480 @ 60, 30 fps | PPG, HR, SpO2 | Sim |
| MR-NIRP | 19 | 190 | 640 × 640 @ 60 fps | PPG | Sim |
| VicarPPG-2 | 50 | 50 | 1280 × 720 @ 30 fps | PPG, HR | Sim |
| V4V | 179 | 1358 | 1720 × 720 @ 25 fps | PPG, HR, BP | Sim |

### 7.4.1. DEAP

O conjunto de dados *DEAP* [Koelstra et al. 2011] foi inicialmente desenvolvido para a análise de emoções, mas também pode ser utilizado na avaliação de métodos de *rPPG*, uma vez que inclui sinais *PPG* autênticos. O conjunto contém dados de 32 participantes, totalizando 874 vídeos gravados com resolução de 720 × 576 e taxa de 50 quadros por segundo. Cada participante assistiu a um videoclipe musical de 1 minuto, com o objetivo de induzir diferentes estados emocionais, o que resulta em variações na frequência cardíaca. O *DEAP* coletou sinais *PPG* reais, a partir dos quais é possível calcular valores verdadeiros de frequência cardíaca.

### 7.4.2. MAHNOB-HCI

O *MAHNOB-HCI* [Soleymani et al. 2011] é um banco de dados multimodal composto por 27 participantes, sendo que cada um gravou 20 vídeos, totalizando 527 gravações. Os vídeos foram capturados com resolução de 780 $\times$ 580 e taxa de 61 quadros por segundo. Embora seu propósito original tenha sido o reconhecimento de emoções e pesquisa em marcação implícita, o *MAHNOB-HCI* também é adequado para avaliação de métodos de medição remota da frequência cardíaca baseados em *rPPG*, devido à inclusão de sinais fisiológicos reais, como o eletrocardiograma (ECG). Todos os participantes realizaram experimentos de indução emocional e marcação implícita, durante os quais a frequência cardíaca variou em resposta às emoções. Além disso, foram utilizadas seis câmeras para capturar diferentes ângulos dos participantes (visão frontal, perfil, grande angular e close-up), tornando este conjunto adequado para testar o desempenho de métodos frente a variações de pose e ângulo.

### 7.4.3. UBFC-rPPG

O conjunto de dados UBFC-rPPG foi desenvolvido especificamente para a avaliação de métodos de extração de sinais de iPPG (também conhecido como *remote PPG* ou rPPG). Ele contém 50 vídeos, cada um com um indivíduo diferente, com resolução de 640 x 480 e taxa de quadros de 30 quadros por segundo. As gravações consideram variações de iluminação, tanto solar quanto interna. O UBFC-rPPG é composto por dois subconjuntos: o primeiro é uma versão reduzida, com 8 vídeos, em que os indivíduos foram instruídos a permanecer imóveis durante a captura; o segundo é um subconjunto maior, com 42 vídeos, onde os indivíduos participaram de um jogo matemático com limite de tempo, com o objetivo de aumentar sua frequência cardíaca.

O UBFC-rPPG é amplamente utilizado por pesquisadores da área devido à sua alta qualidade de vídeo e à presença de dados reais, como sinais de frequência cardíaca e PPG. Embora o conjunto de dados contenha ambos os subconjuntos, a maioria dos pesquisadores opta por utilizar o subconjunto maior devido à sua maior quantidade de dados e à qualidade superior dos vídeos gravados [Bobbia et al. 2019].

### 7.4.4. PURE

O conjunto de dados PURE é composto por gravações de 10 indivíduos, sendo 8 homens e 2 mulheres, com cada participante gravando 6 vídeos, totalizando 60 vídeos. As gravações foram realizadas com resolução de 640 × 480 pixels, taxa de 30 quadros por segundo e duração de um minuto por vídeo. Durante os experimentos, os participantes realizaram seis diferentes atividades, com o objetivo de gerar variações nos movimentos da cabeça. As tarefas incluíram: permanecer sentado e imóvel, conversar, movimentar a cabeça lentamente, movimentar a cabeça rapidamente, girar a cabeça em um ângulo de 20 graus e girar a cabeça em um ângulo de 35 graus. Esses movimentos foram projetados para avaliar a robustez dos métodos frente a diferentes níveis de movimentação.

Além disso, o conjunto levou em consideração variações nas condições de iluminação, utilizando luz natural proveniente de uma grande janela, sujeita à interferência de nuvens, para introduzir mudanças realistas na iluminação durante a gravação dos vídeos. Os sinais fisiológicos de referência foram obtidos com um oxímetro de dedo, com taxa de

amostragem de 60 Hz, garantindo medições precisas da frequência cardíaca real. Cabe destacar que todas as imagens do conjunto PURE estão armazenadas no formato PNG sem perdas, o que assegura a fidelidade visual necessária para uma estimativa precisa dos sinais de iPPG [Stricker et al. 2014].

### 7.4.5. SCAMPS

O conjunto de dados SCAMPS é uma coleção de dados fisiológicos sintéticos de larga escala, composta por 2800 vídeos com resolução de 320 × 240 pixels e taxa de 30 quadros por segundo [McDuff et al. 2022]. Ele fornece rótulos de referência no nível de quadro, incluindo sinais de PPG, intervalos de pulso, formas de onda respiratórias, intervalos respiratórios e 10 ações faciais distintas. Além disso, o SCAMPS também oferece rótulos no nível de vídeo, abrangendo diversos indicadores fisiológicos.

Os parâmetros fornecidos são utilizados para gerar sinais de PPG com duração de 20 segundos e frequência de 300 Hz, juntamente com as intensidades das unidades de ação facial. Cada vídeo é sintetizado com base nesses sinais, combinados com intensidades de ação facial e atributos de aparência escolhidos aleatoriamente, como textura da pele, cor do cabelo, vestimentas, condições de iluminação e cenários de fundo.

O grande volume e a diversidade dos dados sintéticos presentes no SCAMPS são extremamente úteis para a pesquisa na área, especialmente em tarefas onde a obtenção de dados reais com tal riqueza de informações seria complexa e onerosa. No entanto, vale destacar que o SCAMPS é principalmente utilizado para fins de treinamento de modelos, e não tanto para validação ou testes finais [McDuff et al. 2022].

### 7.4.6. MMPD

O MMPD é o primeiro conjunto de dados desenvolvido inteiramente a partir de gravações feitas com câmeras de celulares ou *smartphones* [Tang et al. 2023]. A base é composta por 33 indivíduos e um total de 660 vídeos, com duração de um minuto cada, gravados originalmente em resolução 1280 × 720 e 30 quadros por segundo. Para facilitar o compartilhamento dos dados coletados, os vídeos foram comprimidos para uma resolução de 320 × 240.

O conjunto foi cuidadosamente projetado para abranger uma diversidade de tons de pele, com quatro categorias diferentes, e diversas condições de iluminação, incluindo LED com alta e baixa intensidade, luz incandescente e luz natural. Além disso, ele inclui diferentes atividades, como repouso, rotação de cabeça, conversação e caminhada, criando um cenário variado que permite aos pesquisadores avaliar a robustez de seus métodos em diferentes contextos ambientais.

Adicionalmente, o MMPD contém quatro experimentos focados no impacto de movimentos mais intensos e bruscos. Nesses testes, os participantes realizaram atividades físicas vigorosas, como elevações de joelhos, para aumentar sua frequência cardíaca antes da gravação dos vídeos. Após cada sessão de exercício, os indivíduos recebiam um período adequado de descanso para que a frequência cardíaca se estabilizasse antes de iniciar o próximo experimento.

O MMPD também disponibiliza rótulos reais, incluindo valores de frequência car-

díaca e sinais de PPG de referência, que são recursos valiosos para estudos de estimação de sinais fisiológicos [Tang et al. 2023].

### 7.4.7. BP4D+

O conjunto de dados BP4D+ é uma base multimodal projetada para análise de emoções espontâneas, com foco na estimativa remota de sinais fisiológicos, como frequência cardíaca e pressão arterial. Ele contém gravações de vídeos de 140 indivíduos, cada um participando de 10 tarefas emocionais, resultando em 1400 vídeos no total. As gravações são feitas a uma taxa de 25 quadros por segundo e incluem tanto vídeos RGB quanto térmicos, ambos capturados na mesma frequência de quadros [Zhang et al. 2016].

Além dos vídeos, o BP4D+ fornece dados fisiológicos, como medições de pressão arterial (sistólica, diastólica e média), frequência cardíaca (batimentos por minuto), taxa de respiração e atividade galvânica da pele (EDA), que são coletados simultaneamente aos vídeos. Essas medições são acompanhadas por modelos 3D dinâmicos da face dos participantes, o que permite uma análise detalhada das expressões faciais durante a execução das tarefas [Zhang et al. 2016].

As tarefas emocionais foram projetadas para induzir respostas emocionais específicas nos participantes, proporcionando um cenário controlado para a análise de emoções e suas correlações com os sinais fisiológicos. O BP4D+ tem sido amplamente utilizado em pesquisas que envolvem a estimativa de sinais fisiológicos a partir de vídeos faciais, como a fotopletismografia por imagem, análise de expressões faciais para reconhecimento de emoções e diagnóstico de condições psicológicas, além do desenvolvimento de interfaces afetivas que respondem às emoções dos usuários. Sua riqueza multimodal e a diversidade de dados tornam o BP4D+ um recurso valioso para o avanço da pesquisa nessa área [Zhang et al. 2016].

### 7.4.8. UBFC-Phys

O conjunto de dados UBFC-Phys foi desenvolvido inicialmente para o reconhecimento de emoções e é composto por 56 indivíduos, sendo 46 do sexo feminino e 10 do sexo masculino. Cada participante foi instruído a realizar três tarefas distintas: descansar, conversar e resolver problemas matemáticos, resultando em um total de 168 gravações de vídeo. As gravações foram realizadas com resolução de $1024 \times 1024$ pixels e taxa de 35 quadros por segundo [Meziatisabour et al. 2021].

Além dos vídeos, o conjunto UBFC-Phys utiliza um dispositivo de pulseira inteligente para coletar sinais PPG, considerados como referência padrão-ouro para medições de frequência cardíaca. Para complementar as gravações, os participantes preencheram questionários antes e após os experimentos, com o objetivo de registrar dados relacionados ao nível de ansiedade, fornecendo informações adicionais sobre o estado emocional durante as tarefas [Meziatisabour et al. 2021].

### 7.4.9. COHFACE

O *dataset* COHFACE [Heusch et al. 2017] é um conjunto de dados público criado pelo com o intuito de permitir que pesquisadores avaliem seus métodos de RPPG/ IPPG de maneira padronizada e justa. A base contém dados de 40 participantes, sendo 28 homens

e 12 mulheres, e cada indivíduo contribuiu com quatro gravações de vídeo, totalizando 160 vídeos. As gravações foram feitas com resolução de 640 × 480 pixels e uma taxa de 20 quadros por segundo. Para obter os sinais fisiológicos reais, todos os participantes utilizaram sensores de PPG de contato durante as filmagens.

Durante a coleta dos dados, foram consideradas duas condições distintas de iluminação. Em uma delas, os vídeos foram gravados com iluminação artificial de estúdio, onde as janelas foram mantidas fechadas para bloquear a luz natural e garantir uma iluminação estável com luzes artificiais. Na outra condição, os vídeos foram gravados com iluminação natural, mantendo as janelas abertas e desligando todas as luzes artificiais.

Apesar de sua importância e ampla utilização, a principal limitação do COHFACE é o fato de que os vídeos foram fortemente comprimidos, o que introduz uma quantidade significativa de ruído. Esse ruído pode afetar negativamente a extração precisa dos sinais rPPG, comprometendo a acurácia dos métodos avaliados com essa base de dados [Heusch et al. 2017].

### 7.4.10. ECG-Fitness

ECG-Fitness [Spetlik et al. 2018] é um *dataset* público composto por registros de 17 participantes, sendo 14 do sexo masculino e 3 do sexo feminino, que realizaram quatro tipos distintos de atividades: fala, remo, exercícios em bicicleta ergométrica e em aparelho elíptico. As gravações foram feitas com o uso de duas webcams Logitech C920 e uma câmera térmica FLIR, sob três condições de iluminação diferentes: luz natural proveniente de janelas próximas, iluminação com lâmpadas halógenas de 400 W e luzes LED de 30 W. Para cada participante, foram gerados 12 vídeos, cobrindo todas as combinações entre os três tipos de iluminação e os quatro estados de atividade, totalizando assim 204 vídeos no conjunto. As gravações foram feitas em resolução Full HD(1920 × 1080 pixels) a uma taxa de 30 quadros por segundo, com duração de um minuto cada. Um dos aspectos mais notáveis do ECG-Fitness é que ele é o único *dataset* conhecido que inclui a atividade de remo entre os registros [Spetlik et al. 2018].

### 7.4.11. VIPL-HR

O VIPL-HR [Niu et al. 2018] é um *dataset* público multimodal de grande escala e alta complexidade, composto por gravações de vídeos de 107 indivíduos. Ele inclui três tipos distintos de vídeos: vídeos RGB, vídeos no espectro infravermelho próximo (NIR) e vídeos gravados por câmeras de celulares smartphones. Essas gravações foram realizadas utilizando câmeras RGB, câmeras RGB-D e câmeras de smartphones. No total, o *dataset* reúne 3.130 vídeos faciais em luz visível. Os vídeos RGB foram obtidos tanto por câmeras RGB quanto por câmeras RGB-D, com resoluções de 960 × 720 pixels a 25 fps e 1920 × 1080 pixels a 30 fps, respectivamente. Já os vídeos NIR foram capturados com câmeras RGB-D, com resolução de 640 × 480 pixels a 30 fps. Os vídeos de smartphone, por sua vez, foram registrados em 1920 × 1080 pixels a 30 fps [Niu et al. 2018].

O uso desses diferentes tipos de dispositivos busca permitir a avaliação da robustez dos métodos propostos dada as diferentes modalidades de vídeo. Além disso, o VIPL-HR introduz dois fatores que influenciam a coleta dos dados: movimento da cabeça (estável, movimento intenso, falando) e condições de iluminação (ambiente de laboratório, escuro

e claro), fornecendo um cenário mais realista para os testes. O conjunto também disponibiliza rótulos fisiológicos reais, como frequência cardíaca (HR), oxigenação do sangue (SPO2) e volume de pulso sanguíneo (BVP) [Niu et al. 2018].

### 7.4.12. MR-NIRP

O MR-NIRP [Koelstra et al. 2011] é o primeiro *dataset* de vídeos de dados fisiológicos que inclui cenários de direção, oferecendo uma abordagem mais realista em comparação com os tradicionais ambientes controlados. Ele é composto por 190 vídeos de 19 indivíduos, capturados tanto durante a condução de um veículo quanto dentro de um carro estacionado. Durante as gravações, os sujeitos também realizaram atividades como falar e mover a cabeça aleatoriamente, simulando situações comuns ao dirigir. Os vídeos foram capturados em uma resolução de 640 × 640 pixels com uma taxa de 60 quadros por segundo.

O objetivo principal do MR-NIRP é permitir a avaliação do desempenho de diferentes métodos de RPPG em contextos de direção, ampliando os testes para além das condições controladas da realização da coleta de dados. Os vídeos são sincronizados com sinais reais de PPG, obtidos por meio de um oxímetro de pulso colocado no dedo dos participantes. As gravações incluem simultaneamente dados em RGB e infravermelho próximo (NIR), embora muitos estudos optem por utilizar os dados NIR para treinar e testar os modelos de deep learning. No entanto, o *dataset* apresenta algumas limitações, como a presença de valores nulos (zeros) nos sinais de PPG, o que pode dificultar a avaliação precisa dos métodos de rPPG aplicados nesses dados [Koelstra et al. 2011].

### 7.4.13. VicarPPG-2

O VicarPPG-2 [Gudi et al. 2020] é um conjunto de dados público composto por gravações de vídeos de 10 voluntários, com idade média de 29 anos. Foram registrados 40 vídeos, cada um com duração de 5 minutos, gravados em resolução de 1280 × 720 pixels e a uma taxa de 60 quadros por segundo. Cada participante realizou a gravação de vídeos distintos. No primeiro, os indivíduos permaneciam em estado estático. No segundo, realizavam cinco movimentos planejados de corpo e cabeça: inclinação lateral da cabeça , movimento vertical da cabeça , combinação dos dois , movimentação dos olhos com a cabeça imóvel, e movimentos naturais da cabeça enquanto ouviam música. No terceiro vídeo, os participantes eram expostos a um jogo projetado para gerar estresse, enquanto no quarto, apareciam em estado de relaxamento, logo após terem passado por exercícios físicos que induzem à fadiga. Para obter os sinais fisiológicos reais, o *dataset* utilizou oxímetros de pulso CMS50E, conectados aos dedos dos participantes, registrando sinais PPG autênticos [Gudi et al. 2020].

### 7.4.14. V4V

O conjunto de dados V4V [Revanur et al. 2021] compreende uma coleção de gravações de vídeos e dados fisiológicos e conta com vídeos de 179 indivíduos, abrangendo diferentes etnias, como afro-americanos, caucasianos e asiáticos. Cada voluntário participou de até 10 tarefas experimentais, planejadas para provocar emoções específicas, totalizando 1.358 vídeos. As gravações têm duração variável, entre 5 e 206 segundos, e foram realizadas com resolução de 1280 × 720 pixels e 25 quadros por segundo. Para a coleta

de dados fisiológicos reais, o V4V utiliza o sistema de aquisição BIOPAC MP150, que registra sinais como PPG, frequência cardíaca, pressão arterial e outras medidas vitais. Apesar de seu grande volume de dados e da variedade de desafios emocionais propostos, o conjunto mantém condições de iluminação consistentes em todos os vídeos, garantindo uniformidade na qualidade das imagens [Revanur et al. 2021].

## 7.5. Métricas de Avaliação

As métricas de avaliação em fotopletismografia por imagem são fundamentais para quantificar a precisão e a qualidade dos modelos aplicados na estimativa de parâmetros fisiológicos, como a frequência cardíaca. Como as técnicas de iPPG estimam sinais contínuos ao longo do tempo, é necessário utilizar métricas de regressão para avaliar o desempenho dos métodos. A seguir, são apresentadas as principais métricas utilizadas na avaliação de modelos baseados em iPPG.

### 7.5.1. Relação Sinal-Ruído

A Relação Sinal-Ruído, conhecida do inglês como *Signal-to-Noise Ratio* (SNR) é uma métrica amplamente empregada para avaliar a qualidade dos sinais de iPPG. Sua função principal é quantificar o quanto da informação contida no sinal está efetivamente associada à atividade cardíaca, em comparação com os componentes considerados ruído [De Haan and Jeanne 2013].

Para o cálculo da SNR, equação 22, adota-se a razão entre a energia espectral concentrada na vizinhança da frequência fundamental do pulso e a energia remanescente no intervalo de 40 a 240 batimentos por minuto (bpm). A frequência fundamental é determinada com precisão por meio de um sinal de ECG registrado simultaneamente, servindo como referência confiável.

$$\mathrm{SNR} = 10 \cdot \log_{10} \left( \frac{\sum_{f=40}^{240} \left( U_t(f) \cdot \hat{S}(f) \right)^2}{\sum_{f=40}^{240} \left( (1 - U_t(f)) \cdot \hat{S}(f) \right)^2} \right) \tag{22}$$

Onde:

- $\hat{S}(f)$ é o espectro do sinal de pulso (obtido por iPPG);
- $f$ representa a frequência em batimentos por minuto (bpm), no intervalo de 30 a 240 bpm;
- $U_t(f)$ é uma janela binária (template), que assume valor 1 nas regiões em torno da frequência fundamental do pulso e sua primeira harmônica, e 0 fora dessas regiões.

Dado que a frequência cardíaca varia ao longo do tempo, especialmente durante atividades físicas, a análise é feita em janelas temporais curtas, utilizando uma abordagem de janela deslizante. A SNR final é obtida pela média dos valores calculados em cada janela, permitindo uma avaliação dinâmica e realista da qualidade dos sinais de iPPG ao longo do tempo.

### 7.5.2. Erro Absoluto Médio

O Erro Absoluto Médio, do inglês, *Mean Absolute Error* (MAE) é uma das métricas tradicionais utilizadas na avaliação de modelos de regressão. Ele calcula a média das diferenças absolutas entre os valores preditos e os valores reais. O MAE fornece uma visão clara do erro médio cometido pelo modelo em suas previsões, sem considerar a direção do erro (se positivo ou negativo).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{23}$$

Onde:

- $y_i$ são os valores reais,
- $\hat{y}_i$ são os valores preditos,
- $n$ é o número total de observações.

### 7.5.3. Erro Quadrático Médio

O Erro Quadrático Médio, do inglês *Mean Squared Error* (MSE) é uma métrica que calcula a média dos quadrados das diferenças entre os valores reais e preditos. O MSE é útil quando se deseja enfatizar grandes desvios entre as previsões e os valores reais.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{24}$$

### 7.5.4. Raiz do Erro Quadrático Médio

A Raiz do Erro Quadrático Médio, do inglês *Root Mean Squared Error* (RMSE) é a raiz quadrada do MSE. Essa métrica expressa o erro na mesma unidade dos dados originais, o que facilita a interpretação dos resultados e permite uma análise direta da magnitude dos desvios entre os valores estimados e os reais. A métrica RMSE é especialmente útil para comparar diferentes modelos de previsão, já que permite uma interpretação direta em termos da magnitude do erro.

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{25}$$

### 7.5.5. Coeficiente de Determinação

O Coeficiente de Determinação, também conhecido como $R^2$, mede a proporção da variabilidade dos dados explicada pelo modelo. Ele varia entre 0 e 1, onde um valor de 1 indica que o modelo explica toda a variabilidade dos dados, e um valor de 0 indica que o modelo não explica nada da variabilidade. Um $R^2$ mais alto indica um modelo melhor ajustado aos dados.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{26}$$

Onde $\bar{y}$ é a média dos valores reais.

## 7.6. Gráficos de Avaliação

Além das métricas numéricas, a análise por meio de gráficos é uma etapa complementar indispensável para verificar a consistência das estimativas em relação aos sinais de referência. A seguir, são apresentadas as principais representações gráficas utilizadas na avaliação de modelos baseados em iPPG.

### 7.6.1. Espectrograma

O espectrograma é uma representação tempo-frequência que permite observar como o conteúdo espectral de um sinal varia ao longo do tempo. Na análise de sinais de iPPG, ele desempenha um papel importante ao revelar a frequência cardíaca predominante e seus harmônicos, bem como sua estabilidade e variações dinâmicas. Um exemplo ilustrativo é apresentado na Figura 7.11.

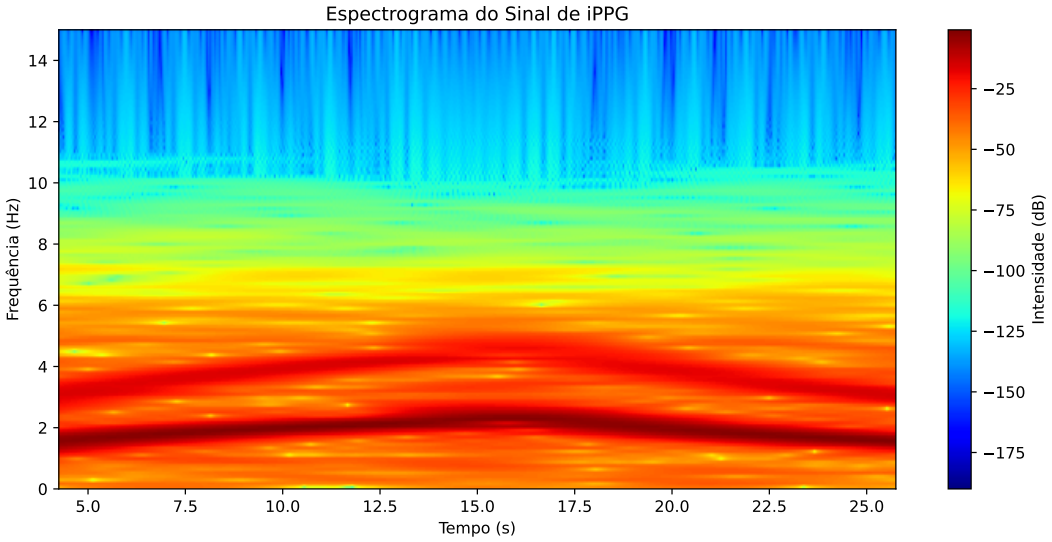



**Figura 7.11. Exemplo de um gráfico de espectrograma.**

Essa representação é especialmente relevante na avaliação comparativa entre diferentes métodos de extração de iPPG. Ao aplicar espectrogramas aos sinais obtidos por distintas abordagens, é possível visualizar não apenas a presença da frequência cardíaca estimada, mas também o nível de ruído, a ocorrência de variações abruptas ou artefatos e a fidelidade espectral em relação ao comportamento fisiológico esperado.

Por exemplo, um método robusto tende a produzir espectrogramas com uma faixa espectral bem definida e contínua ao longo do tempo, enquanto métodos mais suscetíveis a interferências podem exibir múltiplas bandas incoerentes, interrupções ou espalhamento espectral. Além disso, a presença clara de harmônicos pode indicar que a morfologia do pulso foi preservada, reforçando a qualidade da estimativa.

### 7.6.2. Gráfico de Dispersão

O gráfico de dispersão, também conhecido como Scatter Plot, é uma ferramenta fundamental para comparar visualmente os valores reais com os valores preditos por um modelo. Esse gráfico é especialmente útil na avaliação de modelos de regressão, pois permite observar se há uma correlação entre os dois conjuntos de dados, facilitando a identificação de padrões e a avaliação da precisão do modelo.

No gráfico de dispersão, os valores reais ($y_i$) são representados no eixo vertical (eixo $y$), enquanto os valores preditos ($\hat{y}_i$) são plotados no eixo horizontal (eixo $x$). Para um modelo de regressão ideal, espera-se que os pontos do gráfico estejam distribuídos ao longo da linha $y = x$, o que indica que as previsões do modelo estão próximas dos valores reais. Quanto mais próximo o conjunto de pontos estiver da linha de identidade (linha reta onde $y = x$), melhor será o desempenho do modelo.

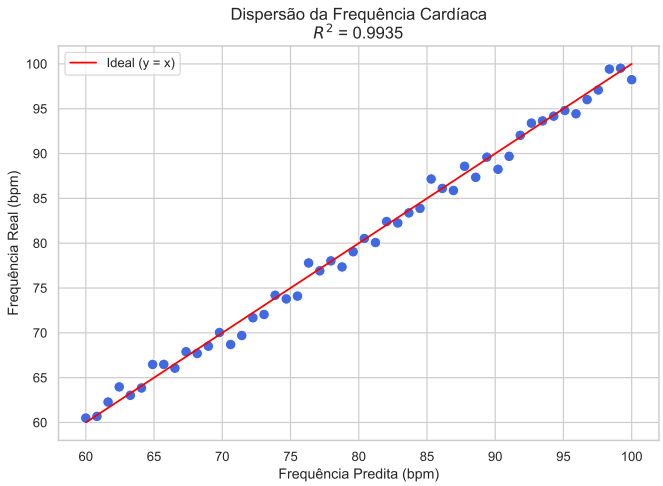



**Figura 7.12. Exemplo de um gráfico de dispersão.**

A análise do gráfico de dispersão oferece diversas informações sobre a qualidade do modelo, como:

- Correlações: Se os pontos formam uma linha reta ou uma curva, isso indica a presença de uma correlação entre os valores reais e preditos. No caso de uma correlação linear, os pontos estarão próximos da linha $y = x$. Em casos de correlação não linear, os pontos podem formar uma curva.
- Erros sistemáticos: A presença de desvios sistemáticos, como agrupamentos de pontos em certas áreas ou uma distribuição não uniforme ao longo da linha $y = x$, pode indicar que o modelo está cometendo erros em certas faixas de valores. Por exemplo, se os pontos tendem a se concentrar mais em uma parte do gráfico, pode ser um sinal de que o modelo tem dificuldades em prever valores em outra parte do intervalo.
- Homocedasticidade: A dispersão dos pontos ao longo do gráfico pode revelar informações sobre a variância dos erros. Se os pontos estão igualmente distribuídos

ao longo de toda a linha $y = x$, isso sugere que o modelo tem uma variância constante nos erros, ou seja, homocedasticidade. No entanto, se a dispersão dos pontos for maior em algumas áreas e menor em outras, pode indicar heterocedasticidade (variância não constante).

### 7.6.3. Histograma dos erros

O histograma é uma ferramenta visual fundamental para avaliar a qualidade de modelos preditivos, especialmente na estimativa de sinais fisiológicos por iPPG. Quando aplicado aos erros do modelo, ele exibe a frequência de ocorrência de diferentes valores de erro, permitindo observar a forma da distribuição. No eixo horizontal são representados os valores dos erros (diferença entre os valores preditos e os reais), enquanto o eixo vertical indica a frequência com que esses erros ocorrem.

Uma distribuição simétrica e centrada em torno de zero sugere que o modelo não possui viés sistemático e que os erros são majoritariamente aleatórios, o que é desejável. Por outro lado, uma concentração de erros positivos ou negativos pode indicar viés, enquanto a presença de caudas longas ou picos incomuns pode apontar para *outliers* ou falhas na suposição de normalidade dos resíduos. Essas observações são essenciais para verificar a robustez do modelo e orientar ajustes ou melhorias. A Figura 7.13 apresenta um exemplo desse tipo de análise.

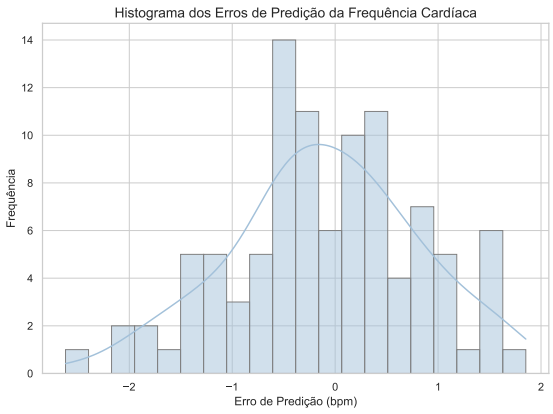



**Figura 7.13. Exemplo de histograma dos erros.**

### 7.6.4. Boxplot dos erros

O boxplot é uma representação gráfica compacta que resume a distribuição dos erros de um modelo, permitindo visualizar rapidamente sua dispersão, simetria e presença de *outliers*. A Figura 7.14 ilustra um exemplo desse tipo de gráfico.

No gráfico, a caixa representa o intervalo interquartil (IQR), entre o primeiro (Q1) e o terceiro quartil (Q3), com a linha interna indicando a mediana. As extremidades dos "bigodes" mostram os limites inferiores e superiores (até 1,5 vezes o IQR), enquanto pontos fora desses limites são considerados *outliers*.

Na análise dos erros, o boxplot permite verificar:

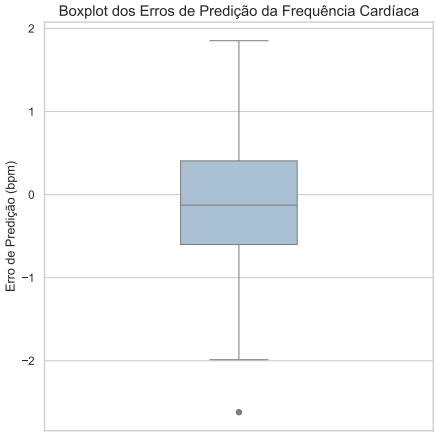



**Figura 7.14. Exemplo de boxplot dos erros.**

- **Centralidade**: A mediana próxima de zero sugere ausência de viés sistemático.
- **Dispersão**: Uma caixa estreita indica baixa variabilidade nos erros; uma caixa larga, maior incerteza nas previsões.
- **Outliers**: Erros extremos podem sinalizar limitações do modelo em certos casos ou a presença de dados atípicos.

Na avaliação de modelos de extração de sinais de iPPG, onde a precisão nas estimativas de parâmetros fisiológicos é fundamental, o boxplot dos erros é uma ferramenta eficaz para avaliar se o modelo é consistente e se há pontos que exigem ajustes ou maior atenção.

### 7.6.5. Gráfico de Bland-Altman

O gráfico de Bland-Altman é uma ferramenta visual amplamente utilizada para avaliar a concordância entre valores reais ($y_i$) e preditos ($\hat{y}_i$) [Giavarina 2015], especialmente em sinais de iPPG. Ele representa a diferença entre os valores preditos e reais em função da média entre eles, permitindo identificar possíveis desvios sistemáticos ou erros concentrados em determinadas faixas de valor.

$$\text{Erro} = \hat{y}_i - y_i \tag{27}$$

Esse tipo de análise é útil para verificar a presença de viés (quando os erros não se distribuem em torno de zero) e para detectar variações na precisão do modelo ao longo do intervalo de estimativas. Uma distribuição uniforme e simétrica dos erros ao redor de zero indica um bom desempenho, enquanto padrões sistemáticos sugerem limitações do modelo em certas condições.

A Figura 7.15 mostra um exemplo típico desse gráfico, utilizado para avaliar o desempenho de estimativas de frequência cardíaca a partir de sinais de iPPG. Este gráfico complementa outras métricas ao oferecer uma perspectiva visual direta sobre a consistência das previsões em diferentes regiões do sinal analisado.

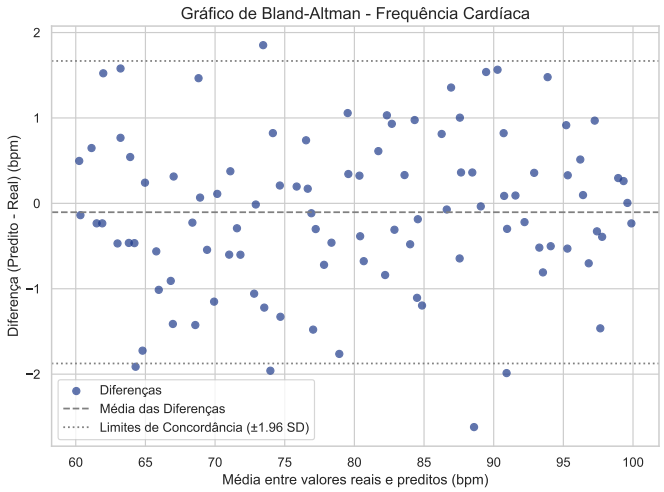



**Figura 7.15. Exemplo de gráfico de Bland-Altman.**

## 7.7. Implementação prática de um modelo de iPPG

A implementação prática de um modelo de extração de sinais de iPPG envolve uma série de etapas, que vão desde a configuração da câmera e iluminação até a obtenção do sinal do pulso extraído. Os métodos de iPPG são variados, com diferenças significativas entre eles. O modelo que será descrito a seguir utiliza o método tradicional de extração baseado no método POS, amplamente utilizado para a extração de sinais cardíacos a partir de variações de cor na pele, especialmente na região facial.

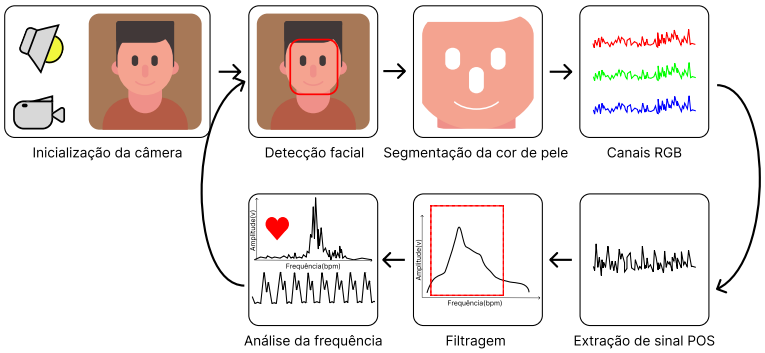



**Figura 7.16. Visão geral da captura e processamento do sinal de iPPG.**

### 7.7.1. Aquisição dos dados

A primeira etapa no processo de implementação de um modelo de iPPG é a captura de imagens da região facial do participante. A qualidade da aquisição do sinal depende fortemente das condições do captura de vídeo. É recomendado:

- **Câmera:** Utilize uma câmera RGB com uma taxa de quadros adequada (por exemplo de 30 fps) e resolução suficiente para capturar detalhes faciais (por exemplo,

640x480 pixels). Webcams, câmeras DSLR, câmeras de smartphones ou sensores como o Kinect podem ser utilizados. Além disso, para garantir uma captura estável e precisa, recomenda-se desativar o foco automático e o equilíbrio de branco automático da câmera. O foco automático pode causar variações no sinal devido ao ajuste constante da lente, enquanto o equilíbrio de branco automático pode alterar as cores do sinal, comprometendo a precisão da extração. Ajuste manualmente o foco e o equilíbrio de branco para obter uma imagem estável e consistente ao longo da captura.

- **Iluminação:** É essencial garantir uma fonte de luz estável e difusa, com variação temporal mínima. Flutuações na iluminação ambiente podem introduzir ruídos consideráveis, prejudicando a qualidade do sinal extraído. A iluminação homogênea melhora a precisão da detecção dos sinais fisiológicos.

- **Posicionamento do Participante:** O indivíduo deve estar posicionado de frente para a câmera, com o rosto claramente visível e evitando movimentos excessivos durante a captura. Movimentos significativos da cabeça ou do corpo podem comprometer a estabilidade e a qualidade do sinal. Para resultados ótimos, recomenda-se uma distância entre 0,9 m e 1,2 m entre o participante e a câmera [Han et al. 2015].

### 7.7.2. Detecção da face e seleção da região de interesse

A detecção da face é realizada por meio de algoritmos especializados, disponíveis em bibliotecas de processamento de imagens como OpenCV [Bradski and Kaehler 2008] e Mediapipe [Lugaresi et al. 2019], que apresentam alto desempenho, especialmente em tempo real. Após a identificação da face, é determinada a Região de Interesse (ROI) para a extração do sinal. As áreas mais comumente selecionadas para essa finalidade na literatura incluem a testa, as bochechas ou toda a face, dependendo do contexto e da precisão desejada. A ROI pode ser ajustada automaticamente a cada novo quadro utilizando algoritmos de rastreamento, como o Kanade-Lucas-Tomasi [Wu et al. 2013], que seguem pontos característicos do rosto ao longo do tempo, reduzindo os efeitos de pequenos movimentos e garantindo a precisão da extração do sinal.

### 7.7.3. Segmentação da pele

A segmentação da pele é uma etapa essencial para isolar a área relevante e minimizar a influência de outras que não são relacionadas à pele. A segmentação pode ser feita utilizando algoritmos baseados em limiar de cor ou em técnicas mais elaboradas de aprendizado de máquina.

Um método simples envolve a utilização do espaço de cores YCbCr para identificar os pixels da pele [Saxen and Al-Hamadi 2014]. Para isso, são aplicados limiares nas componentes de cor, como mostrado abaixo:

$$\text{MáscaraPele}(x,y) = \begin{cases} 1, & \text{se } Cb_{\min} \leq Cb(x,y) \leq Cb_{\max} \text{ e } Cr_{\min} \leq Cr(x,y) \leq Cr_{\max} \\ 0, & \text{caso contrário} \end{cases} \tag{28}$$

Os limiares $Cb_{\min}, Cb_{\max}, Cr_{\min}, Cr_{\max}$ são definidos com base em estudos empíricos sobre a distribuição dos valores da cor da pele humana [Saxen and Al-Hamadi 2014]. Alternativamente, a segmentação pode ser feita utilizando algoritmos de agrupamento tradicionais ou, mais recentemente com redes neurais convolucionais (CNN).

### 7.7.4. Construção dos sinais de cores

Uma vez capturadas as regiões de interesse das imagens, realiza-se a extração dos sinais brutos de cor (canais R, G e B). Para cada quadro, os valores médios dos pixels da ROI são calculados para formar três séries temporais, representadas como:

$$s_R(t) = \frac{1}{N}\sum_{i=1}^{N} I_R(x_i, y_i, t) \tag{29}$$

$$s_G(t) = \frac{1}{N}\sum_{i=1}^{N} I_G(x_i, y_i, t) \tag{30}$$

$$s_B(t) = \frac{1}{N}\sum_{i=1}^{N} I_B(x_i, y_i, t) \tag{31}$$

onde $s_R(t), s_G(t)$ e $s_B(t)$ são as séries temporais para os canais vermelho, verde e azul, respectivamente, $I_R(x_i, y_i, t)$, $I_G(x_i, y_i, t)$ e $I_B(x_i, y_i, t)$ são os valores de intensidade dos pixels da ROI em cada quadro $t$, e $N$ é o número de pixels na ROI. Esse processo é importante para atenuar o ruído presente em pixels individuais.

### 7.7.5. Extração do sinal com o método POS

Os sinais extraídos de cada canal RGB são normalizados com base na média de uma janela temporal de análise escolhida. A normalização é realizada dividindo-se o valor de cada canal pela média da série dentro da janela. Os sinais normalizados são então organizados em um vetor tridimensional, representando os canais de cor normalizados. Depois, o método transforma os sinais RGB normalizados em um novo espaço de sinais pulsáteis utilizando uma projeção ortogonal. Após a projeção, o sinal obtido é então ajustado com base em um parâmetro de ponderação ($\alpha$), e o sinal de pulso final é gerado por meio da soma com sobreposição das estimativas anteriores, como visto anteriormente na explicação do método.

### 7.7.6. Filtragem dos sinais

O sinal extraído por meio do método POS é, posteriormente, submetido a um processo de filtragem utilizando um filtro passa-faixa, cuja principal finalidade consiste em isolar as componentes espectrais associadas à frequência cardíaca, atenuando simultaneamente ruídos de baixa e alta frequência que não apresentam relevância fisiológica. A faixa espectral de interesse para a frequência cardíaca humana situa-se, em geral, entre 0,75 Hz e 4,0 Hz, correspondendo a um intervalo aproximado de 45 a 240 batimentos por minuto (bpm). Entre os filtros passa-faixa comumente adotados nesse contexto, destacam-se os modelos de Butterworth, Chebyshev e Elíptico, sendo a escolha do tipo de filtro deter-

minada por critérios como a taxa de atenuação fora da banda passante, a linearidade da resposta em fase e a complexidade computacional, conforme as exigências da aplicação.

### 7.7.7. Extração da frequência cardíaca

Concluída a etapa de filtragem, procede-se à conversão do sinal para o domínio da frequência, com o emprego da Transformada Rápida de Fourier (FFT). Essa técnica permite decompor o sinal temporal em suas componentes harmônicas, fornecendo uma representação espectral detalhada. A partir dessa análise espectral, identifica-se o pico de maior amplitude dentro da faixa de interesse, o qual corresponde à frequência cardíaca dominante, expressa em Hertz (Hz). Para a conversão dessa medida para batimentos por minuto, utiliza-se a relação direta $\text{bpm} = \text{Hz} \times 60$, valor padronizado na quantificação da frequência cardíaca.

## 7.8. Aplicações

Os avanços nas pesquisas relacionadas à área de iPPG têm possibilitado uma expansão significativa no campo de aplicações dessas abordagens. A seguir, são discutidas algumas das possíveis aplicações emergentes dessas abordagens, incluindo aquelas que já estão sendo investigadas por estudos recentes.

### 7.8.1. Medição de múltiplos sinais vitais

A técnica de iPPG surgiu inicialmente com o propósito de monitorar remotamente a frequência cardíaca. Com o amadurecimento das pesquisas na área, essa tecnologia expandiu suas possibilidades de aplicação, permitindo a estimativa de diversos sinais fisiológicos de forma não intrusiva.

Atualmente, a técnica de iPPG tem sido adaptada para a medição de parâmetros como pressão arterial [Zeng et al. 2025], frequência respiratória, variabilidade da frequência cardíaca e saturação de oxigênio [Lampier et al. 2023]. No caso da pressão arterial, seu monitoramento remoto representa uma alternativa promissora aos métodos convencionais, sobretudo para a detecção de quadros de hipertensão sem a necessidade de dispositivos de contato direto com a pele.

A saturação de oxigênio no sangue, por sua vez, é um indicador crítico da capacidade do organismo de transportar oxigênio adequadamente. Níveis reduzidos podem sugerir hipóxia e demandam atenção clínica imediata. Embora a técnica de iPPG já venha sendo utilizada para estimar esse parâmetro, os resultados ainda são moderadamente precisos, sendo necessário o aprimoramento de técnicas para maior confiabilidade diagnóstica [Lewandowska and Nowak 2012].

### 7.8.2. Monitoramento em hospitais

O monitoramento de sinais vitais sem contato apresenta grande potencial em ambientes hospitalares, principalmente por eliminar a necessidade de sensores tradicionais, como os utilizados em eletrocardiogramas e na fotopletismografia convencional, que exigem fixação direta à pele do paciente. Essa característica é especialmente vantajosa para indivíduos com pele sensível ou comprometida, como pacientes internados em Unidades de Terapia Intensiva (UTI), vítimas de queimaduras ou recém-nascidos em Unidades de

Terapia Intensiva Neonatal (UTIN) [Wang et al. 2023, Zeng et al. 2024].

Além do uso em ambientes clínicos tradicionais, a técnica de iPPG pode ser aplicada em triagens hospitalares, permitindo o monitoramento simultâneo de múltiplos pacientes, e também em contextos cirúrgicos, onde pode ser empregada para avaliar a perfusão sanguínea em tempo real.

### 7.8.3. Monitoramento de treino físico

O monitoramento da frequência cardíaca é fundamental em exercícios físicos, especialmente para indivíduos que buscam controlar sua frequência cardíaca dentro da zona alvo para treinamento cardiovascular ou queima de gordura. Até o momento, os métodos tradicionais de monitoramento de frequência cardíaca baseados em contato, como sensores de ECG em faixas torácicas e sensores de PPG em pulseiras de pulso, têm sido amplamente utilizados em aplicações de treino físico para consumidores. No entanto, essas medições baseadas em contato são geralmente desconfortáveis, inconvenientes, e difíceis de visualizar durante o treino, o que limita sua eficácia informativa.

Como alternativa, um sistema de monitoramento baseado em câmeras, incorporando a técnica de iPPG e um monitor, pode ser integrado diretamente em equipamentos de academia, como esteiras ou bicicletas ergométricas [Wang et al. 2017b]. Esse sistema permite a medição, visualização e análise da frequência cardíaca do usuário de forma contínua e automática, sem a necessidade de qualquer dispositivo corporal. Esse método sem contato oferece uma solução mais eficaz para o acompanhamento da frequência cardíaca ao longo do exercício, otimizando o treino de forma mais prática e conveniente. Além disso, o sistema pode avaliar a recuperação da frequência cardíaca imediatamente após o exercício, fornecendo informações sobre a saúde cardiovascular do indivíduo, como a presença de arritmias, ou até mesmo prever sua mortalidade.

### 7.8.4. Monitoramento de saúde em casa

O método de iPPG pode ser aplicado no ambiente doméstico para acompanhar e avaliar a saúde de uma pessoa (mesmo que não seja paciente) no cotidiano, com o intuito de promover melhorias em seu estilo de vida, como no caso de atividades físicas para o coração, hábitos alimentares, controle de estresse mental e qualidade do sono. Um exemplo disso é o monitoramento da frequência cardíaca: é possível integrar uma câmera RGB comum a um espelho, criando um sistema inteligente de monitoramento de saúde. Ao se posicionar em frente ao espelho, o indivíduo tem sua frequência cardíaca medida automaticamente, com o valor sendo exibido no espelho.

Além disso, esse sistema pode ser instalado acima da cama para monitorar continuamente a frequência cardíaca (e sua variabilidade) durante o sono [Vogels et al. 2018]. Com base nessas medições, é possível analisar os diferentes estágios do sono (como sono leve, sono profundo e movimento rápido dos olhos (REM)), ajudando a melhorar a qualidade do sono, por exemplo, ao configurar um alarme inteligente para o momento ideal de despertar. A tecnologia também pode ser integrada a sistemas de monitoramento em vídeo em residências, usados para monitorar idosos ou bebês, com funcionalidades como avaliação de marcha, detecção de quedas, monitoramento da saída da cama ou identificação de outras emergências.

### 7.8.5. Computação afetiva

Os métodos de iPPG têm se destacado na computação afetiva, sobretudo pela capacidade de extrair sinais fisiológicos de forma não invasiva a partir de vídeos. Estudos iniciais demonstraram que a variação da frequência cardíaca extraída medida pela técnica de iPPG pode ser usada para estimar níveis de estresse com acurácia de até 85% [Meziati Sabour et al. 2021].

Desde então, diversas pesquisas vêm expandindo sua aplicação, incluindo a criação de bases de dados como a UBFC-Phys, voltada à análise de estresse e emoções [Meziati Sabour et al. 2021]. O reconhecimento de emoções, por sua vez, tem se beneficiado do uso de iPPG, inicialmente em propostas para detectar microexpressões faciais. Estudos mais recentes também exploram o uso de grafos e redes neurais para combinar informações fisiológicas e visuais na detecção de estados emocionais [Liu et al. 2024]. Além disso, há investigações em andamento sobre o uso de iPPG para reconhecimento de dor. Espera-se que, em um futuro próximo, essas abordagens avancem em aplicações como interfaces humano-máquina e avaliações fisiológicas automatizadas [Li et al. 2014].

### 7.8.6. Vigilância por vídeo

Um grande número de câmeras foi instalado em cidades, como em estações centrais, ruas e bares, com o objetivo de monitorar pedestres para fins de segurança. Essas câmeras são geralmente conectadas a sistemas inteligentes que utilizam algoritmos de visão computacional para analisar o fluxo de pessoas, entender comportamentos dos pedestres, detectar ações agressivas ou violentas, e até prever outras emergências. O princípio é utilizar padrões físicos (como características espaciais ou de movimento temporal) dos pedestres para reconhecer comportamentos ou eventos perigosos.

De forma semelhante, a técnica de iPPG pode ser integrada a esses sistemas de câmeras para medir e analisar os padrões fisiológicos dos pedestres à distância, facilitando a identificação e a classificação de comportamentos. Como a vigilância por vídeo geralmente ocorre em ambientes externos, vários desafios adicionais precisam ser considerados, como as condições de iluminação, a resolução facial e os movimentos corporais, em comparação com as aplicações internas abordadas anteriormente neste capítulo.

### 7.8.7. Entretenimento

A técnica de iPPG oferece um grande potencial para inovar sistemas de entretenimento, como televisores inteligentes, realidade virtual, realidade aumentada e dispositivos de jogos cinéticos. Ao monitorar variáveis fisiológicas do usuário, como a frequência cardíaca e a variabilidade da frequência cardíaca, a técnica de iPPG pode adaptar a experiência de interação de maneira mais personalizada e responsiva.

Em contextos como jogos ou filmes, por exemplo, a tecnologia de iPPG pode analisar as respostas emocionais do usuário e ajustar a intensidade, o ritmo ou o conteúdo da experiência com base nas reações detectadas. Isso resulta em uma interação mais imersiva, ajustando automaticamente os estímulos para maximizar o prazer e o engajamento. Além disso, o monitoramento das emoções pode ser integrado a plataformas de mídia para compreender melhor as preferências e reações dos usuários, ajudando a prever gostos e aprimorar a personalização do conteúdo. Com isso, cria-se uma experiência

mais dinâmica e alinhada ao estado emocional do usuário, transformando a forma como interagimos com a tecnologia no universo do entretenimento.

### 7.8.8. Monitoramento de motoristas

A prevenção de acidentes de trânsito causados por fatores comportamentais e cognitivos do motorista tem motivado o desenvolvimento de tecnologias baseadas em visão computacional. Sistemas modernos embarcados em veículos utilizam câmeras para monitorar a postura da cabeça, piscadas de olhos, direção do olhar, movimentação da boca e expressões faciais, emitindo alertas diante de sinais de fadiga, sonolência, distração, estresse ou ansiedade [Smart Eye 2023].

Mais recentemente, técnicas de iPPG têm ganhado destaque como solução não invasiva para o monitoramento contínuo de sinais vitais, como frequência cardíaca e variabilidade da frequência cardíaca, indicadores fortemente relacionados ao estado emocional e cognitivo do condutor [Ahmed et al. 2025]. Algoritmos de aprendizado profundo têm mostrado resultados animadores na extração de sinais de iPPG em cenários reais com iluminação variável, movimento do veículo e mudanças na expressão facial do condutor. Esses sistemas também têm sido propostos para a detecção de direção sob influência de álcool, associando dados de expressões faciais e iPPG a modelos de redes neurais com bons resultados [Keshtkaran 2025]. Além do uso automotivo, o potencial da técnica de iPPG se estende ao monitoramento de pilotos em aeronaves ou operadores de máquinas pesadas, favorecendo intervenções preventivas em contextos críticos.

Apesar dos avanços, desafios persistem quanto à generalização dos modelos para diferentes perfis demográficos e à robustez em ambientes altamente dinâmicos. Por isso, o desenvolvimento de sistemas de monitoramento de motoristas baseados em iPPG exige não apenas inovações em algoritmos, mas também a construção de bases de dados diversificadas e a integração com sistemas de assistência à condução em tempo real.

### 7.8.9. Detecção de *Deepfake*

*Deepfake*, uma combinação das palavras *deep learning* (aprendizado profundo) e *fake* (falso), refere-se a algoritmos de aprendizado profundo utilizados para simular e criar conteúdos de áudio e vídeo [Rana et al. 2022]. Atualmente, *Deepfake* tornou-se um campo altamente popular, sendo amplamente aplicado em técnicas de inteligência artificial para troca de rostos, síntese de voz, geração de rostos e vídeos [Rana et al. 2022].

O surgimento dessa tecnologia tornou possível manipular e gerar conteúdos de vídeo e áudio incrivelmente realistas, com uma difícil detecção, o que torna desafiador para os consumidores desses conteúdos sintetizados discernirem o que é verdadeiro e o que é falso. Em função disso, pesquisadores têm se dedicado ao desenvolvimento de métodos para distinguir esses conteúdos. Estudos demonstraram que a medição da frequência cardíaca a partir de vídeos faciais pode ser uma ferramenta eficaz para determinar e diferenciar se um vídeo é real ou falso. Como resultado, os métodos de iPPG começaram a ser empregados na detecção de *Deepfakes*, com o surgimento de abordagens mais específicas para essa aplicação, alcançando resultados promissores e demonstrando o grande potencial dos métodos iPPG nesta área. Destaca-se que a pesquisa em reconhecimento de *Deepfake* por meio de métodos iPPG continua a ser uma das áreas mais promissoras da

pesquisa científica [Xiao et al. 2024].

### 7.8.10. *Anti-spoofing* Facial

Com o avanço contínuo da tecnologia ao longo dos anos, os métodos de segurança e criptografia têm se desenvolvido de maneira significativa. Um dos métodos de segurança mais comuns atualmente são as chaves biométricas, como impressões digitais e características faciais. Celulares, aplicativos bancários e de mensagens já adotam a biometria como uma chave de acesso. No entanto, como mencionado anteriormente, tecnologias como o *deepfake* de vídeo utilizam inteligência artificial para realizar a troca de rostos, o que torna a biometria facial vulnerável a ataques *spoofing*. Além do *deepfake*, agentes maliciosos podem obter fotos ou vídeos contendo o rosto de um indivíduo alvo e usá-los para contornar esse tipo de segurança, acessando assim os dados da vítima.

Em razão disso, o interesse por parte dos pesquisadores em desenvolver técnicas para combater esse tipo de ataque tem crescido consideravelmente. Com o rápido avanço dos métodos de iPPG, estudiosos têm identificado o potencial dessas técnicas para aprimorar os sistemas de reconhecimento facial, oferecendo uma solução mais robusta contra ataques *spoofing* [Xiao et al. 2024, Wang et al. 2024].

## 7.9. Limitações, desafios e estudos futuros

As aplicações dos métodos e abordagens de fotopletismografia por imagem (iPPG) são amplas, mas, conforme discutido ao longo do trabalho, ainda existem desafios significativos em sua implementação. Esses desafios estão presentes tanto nas abordagens tradicionais quanto nas que utilizam aprendizado profundo, indicando a necessidade contínua de estudos e melhorias.

Um dos principais obstáculos enfrentados está relacionado à sensibilidade ao movimento durante a captura do vídeo. Essa limitação afeta a estabilidade do sinal extraído, pois movimentos do indivíduo podem comprometer o processo de detecção facial e, consequentemente, a área de interesse utilizada na extração do sinal. Quando a detecção facial é feita apenas no primeiro quadro, por exemplo, qualquer movimentação fora da área inicial pode resultar em quadros inutilizáveis. Mesmo quando a detecção ocorre a cada quadro, o movimento contínuo pode demandar maior processamento e causar diferenças entre o tempo de gravação e o número de quadros válidos coletados.

Outro fator importante é a iluminação durante a etapa de gravação. A presença de sombras causadas por movimentações, especialmente quando há segmentação da cor da pele, pode interferir na extração precisa do sinal. Além disso, a qualidade da luz ambiente influencia diretamente no desempenho do método, sendo observada a necessidade de uma iluminação estável e de intensidade adequada. Em experimentos, foi verificado que luzes muito intensas podem prejudicar a obtenção do sinal de iPPG, o que reforça a importância de um controle cuidadoso das condições de iluminação. Além disso, as características ópticas da pele influenciam na eficácia da extração, assim, diferenças na refletância cutânea podem dificultar a detecção das variações de cor associadas ao pulso [Wang and Shan 2020].

Em resumo, apesar do grande potencial da técnica de iPPG como ferramenta não

invasiva de monitoramento fisiológico, seu uso efetivo ainda depende da superação de limitações relacionadas ao movimento, à iluminação e à adaptação às características visuais dos indivíduos. Esses pontos representam áreas importantes para pesquisas futuras, com foco na melhoria da robustez, da precisão e da aplicabilidade em diferentes cenários.

## Referências

[Ahmed et al. 2025] Ahmed, S. G., Verbert, K., Zaki, N., Khalil, A., Aljassmi, H., and Alnajjar, F. (2025). Ai innovations in rppg systems for driver monitoring: Comprehensive systematic review and future prospects. *IEEE Access*, 13:22893–22913.

[Allen 2007] Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, 28(3):R1.

[Bobbia et al. 2019] Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., and Dubois, J. (2019). Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90. Award Winning Papers from the 23rd International Conference on Pattern Recognition (ICPR).

[Bradski and Kaehler 2008] Bradski, G. and Kaehler, A. (2008). *Learning OpenCV: Computer vision with the OpenCV library*. "O'Reilly Media, Inc.".

[Chen and McDuff 2018] Chen, W. and McDuff, D. (2018). Deepphys: Video-based physiological measurement using convolutional attention networks. pages 356–373.

[De Haan and Jeanne 2013] De Haan, G. and Jeanne, V. (2013). Robust pulse rate from chrominance-based rppg. *IEEE transactions on biomedical engineering*, 60(10):2878–2886.

[Giavarina 2015] Giavarina, D. (2015). Understanding bland altman analysis. *Biochemia medica*, 25(2):141–151.

[Gudi et al. 2020] Gudi, A., Bittner, M., and van Gemert, J. (2020). Real-time webcam heart-rate and variability estimation with clean ground truth for evaluation. *CoRR*, abs/2012.15846.

[Han et al. 2015] Han, B., Ivanov, K., Wang, L., and Yan, Y. (2015). Exploration of the optimal skin-camera distance for facial photoplethysmographic imaging measurement using cameras of different types. In *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*, pages 186–189.

[Heusch et al. 2017] Heusch, G., Anjos, A., and Marcel, S. (2017). A reproducible study on remote heart rate measurement. *CoRR*, abs/1709.00962.

[Hülsbusch 2008] Hülsbusch, M. (2008). *An Image-Based Functional Method for Opto-Electronic Detection of Skin-Perfusion*. Phd thesis, RWTH Aachen University.

[Keshtkaran 2025] Keshtkaran, E. (2025). *Automated Methods for Estimating Blood Alcohol Concentration Level from Facial Cues*. Ph.d. thesis, Edith Cowan University.

[Koelstra et al. 2011] Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. (2011). Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31.

[Kyriacou and Allen 2021] Kyriacou, P. A. and Allen, J. (2021). *Photoplethysmography: technology, signal analysis and applications*. Academic Press.

[Lampier et al. 2023] Lampier, L. C., Floriano, A., Valadão, C. T., Silva, L. A., Caldeira, E. M. D. O., and Bastos-Filho, T. F. (2023). A deep learning approach for estimating spo 2 using a smartphone camera. *IEEE Transactions on Instrumentation and Measurement*, 72:1–8.

[Lampier et al. 2022] Lampier, L. C., Valadão, C. T., Silva, L. A., Delisle-Rodríguez, D., Caldeira, E. M. d. O., and Bastos-Filho, T. F. (2022). A deep learning approach to estimate pulse rate by remote photoplethysmography. *Physiological Measurement*, 43(7):075012.

[Lewandowska and Nowak 2012] Lewandowska, M. and Nowak, J. (2012). Measuring pulse rate with a webcam. *Journal of Medical Imaging and Health Informatics*, 2(1):87–92.

[Li et al. 2014] Li, X., Chen, J., Zhao, G., and Pietikainen, M. (2014). Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4264–4271. IEEE.

[Lin et al. 2019] Lin, J., Gan, C., and Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093.

[Liu et al. 2024] Liu, I., Liu, F., Zhong, Q., Ma, F., and Ni, S. (2024). Your blush gives you away: detecting hidden mental states with remote photoplethysmography and thermal imaging. *PeerJ Computer Science*, 10:e1912.

[Lugaresi et al. 2019] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M., Lee, J., et al. (2019). Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, volume 2019.

[McDuff et al. 2022] McDuff, D., Wander, M., Liu, X., Hill, B. L., Hernandez, J., Lester, J., and Baltrusaitis, T. (2022). Scamps: Synthetics for camera measurement of physiological signals.

[Meziati Sabour et al. 2021] Meziati Sabour, R., Benezeth, Y., De Oliveira, P., Chappé, J., and Yang, F. (2021). Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*.

[Meziatisabour et al. 2021] Meziatisabour, R., Benezeth, Y., Oliveira, P., Chappé, J., and Yang, F. (2021). Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, PP:1–1.

[Niu et al. 2018] Niu, X., Han, H., Shan, S., and Chen, X. (2018). VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. *CoRR*, abs/1810.04927.

[Pirzada et al. 2024] Pirzada, P., Wilde, A., and Harris-Birtill, D. (2024). Remote photoplethysmography for heart rate and blood oxygenation measurement: A review. *IEEE Sensors Journal*, 24(15):23436–23453.

[Poh et al. 2010] Poh, M.-Z., McDuff, D. J., and Picard, R. W. (2010). Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express*, 18(10):10762–10774.

[Rana et al. 2022] Rana, M. S., Nobi, M. N., Murali, B., and Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE access*, 10:25494–25513.

[Revanur et al. 2021] Revanur, A., Li, Z., Ciftci, U. A., Yin, L., and Jeni, L. A. (2021). The first vision for vitals (V4V) challenge for non-contact video-based physiological estimation. *CoRR*, abs/2109.10471.

[Ronneberger et al. 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

[Saxen and Al-Hamadi 2014] Saxen, F. and Al-Hamadi, A. (2014). Color-based skin segmentation: an evaluation of the state of the art. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4467–4471. IEEE.

[Smart Eye 2023] Smart Eye (2023). Driver monitoring system (dms). https://www.smarteye.se/solutions/automotive/driver-monitoring-system/.

[Soleymani et al. 2011] Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M. (2011). A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1):42–55.

[Spetlik et al. 2018] Spetlik, R., Cech, J., Franc, V., and Matas, J. (2018). Visual heart rate estimation with convolutional neural network.

[Stricker et al. 2014] Stricker, R., Müller, S., and Gross, H.-M. (2014). Non-contact video-based pulse rate measurement on a mobile service robot. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 2014:1056–1062.

[Sun and Thakor 2016] Sun, Y. and Thakor, N. (2016). Photoplethysmography revisited: From contact to noncontact, from point to imaging. *IEEE Transactions on Biomedical Engineering*, 63(3):463–477.

[Tang et al. 2023] Tang, J., Chen, K., Wang, Y., Shi, Y., Patel, S., McDuff, D., and Liu, X. (2023). Mmpd: Multi-domain mobile video physiology dataset.

[Verkruysse et al. 2008] Verkruysse, W., Svaasand, L. O., and Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. In *Optics Express*, volume 16, pages 21434–21445. Optical Society of America.

[Vogels et al. 2018] Vogels, T., Van Gastel, M., Wang, W., and De Haan, G. (2018). Fully-automatic camera-based pulse-oximetry during sleep. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1349–1357.

[Wang et al. 2023] Wang, H., Huang, J., Wang, G., Lu, H., and Wang, W. (2023). Contactless patient care using hospital iot: Cctv-camera-based physiological monitoring in icu. *IEEE Internet of Things Journal*, 11(4):5781–5797.

[Wang et al. 2024] Wang, J., Shan, C., Liu, L., and Hou, Z. (2024). Camera-based physiological measurement: Recent advances and future prospects. *Neurocomputing*, 575.

[Wang et al. 2017a] Wang, W., den Brinker, A. C., Stuijk, S., and de Haan, G. (2017a). Algorithmic principles of remote-ppg. In *IEEE Transactions on Biomedical Engineering*, volume 64, pages 1479–1491. Institute of Electrical and Electronics Engineers (IEEE).

[Wang et al. 2017b] Wang, W., den Brinker, A. C., Stuijk, S., and de Haan, G. (2017b). Robust heart rate from fitness videos. *Physiological measurement*, 38(6):1023.

[Wang and Shan 2020] Wang, W. and Shan, C. (2020). Impact of makeup on remote-ppg monitoring. *Biomedical Physics & Engineering Express*, 6(3):035004.

[Wu et al. 2013] Wu, Y., Lim, J., and Yang, M.-H. (2013). Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418.

[Xiao et al. 2024] Xiao, H., Liu, T., Sun, Y., Li, Y., Zhao, S., and Avolio, A. (2024). Remote photoplethysmography for heart rate measurement: A review. *Biomedical Signal Processing and Control*, 88.

[Yu et al. 2019] Yu, Z., Li, X., and Zhao, G. (2019). Recovering remote photoplethysmograph signal from facial videos using spatio-temporal convolutional networks. *CoRR*, abs/1905.02419.

[Zeng et al. 2024] Zeng, Y., Yu, D., Song, X., Wang, Q., Pan, L., Lu, H., and Wang, W. (2024). Camera-based cardiorespiratory monitoring of preterm infants in nicu. *IEEE Transactions on Instrumentation and Measurement*.

[Zeng et al. 2025] Zeng, Y., Zhu, Y., Song, X., Wang, Q., Yang, J., and Wang, W. (2025). Camera-based neonatal blood pressure estimation from multisite and multiwavelength pulse transit time-a proof of concept in nicu. *IEEE Internet of Things Journal*.

[Zhan et al. 2020] Zhan, Q., Wang, W., and De Haan, G. (2020). Analysis of cnn-based remote-ppg to understand limitations and sensitivities. *Biomedical optics express*, 11(3):1268–1283.

[Zhang et al. 2016] Zhang, Z., Girard, J., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., Cohn, J., Ji, Q., and Yin, L. (2016). Multimodal spontaneous emotion corpus for human behavior analysis. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December.

# Capítulo 8

# Big Data Linkage no Brasil: Aspectos metodológicos e práticos

Robespierre Pita (IC-UFBA/Cidacs-Fiocruz), Roberto P. Carreiro (Cidacs-Fiocruz), Carlos J. C. Santos (IC-UFBA/Cidacs-Fiocruz), Laianne dos S. Protasio (IC-UFBA), Marcos E. Barreto (London School of Economics and Political Science/Cidacs-Fiocruz), Victor B. Orrico (IC-UFBA), José A. D. Gomes (Cidacs-Fiocruz), Fernanda S. Eustáquio (Cidacs-Fiocruz), Samila Sena (Cidacs-Fiocruz), Mauricio L. Barreto (Cidacs-Fiocruz), Pablo I. P. Ramos (Cidacs-Fiocruz), Denis Rangel (Cidacs-Fiocruz), Bethânia de A. Almeida (Cidacs-Fiocruz)

***Abstract***

*Big Data Linkage (BDL) is a key step in the integration of large-scale datasets, addressing the challenges and solutions related to the management and linkage of diverse, high-dimensional records. This short course aims to present the fundamental concepts of BDL, with a special focus on the Brazilian context. Key methodological challenges—such as accuracy, scalability, and privacy preservation—are discussed, highlighting tools developed in Brazil, such as AtyImo and CIDACS-RL. The course also explores current research gaps, opportunities for further investigation, and emerging initiatives that are driving progress in the BDL field.*

***Resumo***

*O Big Data Linkage (BDL) é uma etapa essencial na integração de grandes volumes de dados, lidando com os desafios e soluções relacionados à gestão e ao cruzamento de registros diversos e de alta dimensionalidade. Este minicurso tem como objetivo apresentar os conceitos fundamentais do BDL, com ênfase no contexto brasileiro. São discutidos os principais desafios metodológicos — como acurácia, escalabilidade e preservação da privacidade —, com destaque para ferramentas desenvolvidas no Brasil, como o AtyImo e o CIDACS-RL. O minicurso também aborda lacunas atuais, oportunidades de pesquisa e as iniciativas emergentes que vêm impulsionando o avanço da área.*

## 8.1. Introdução

A Integração de Dados (ID) abriga uma classe de aplicações científicas com inequívoco potencial para unificar informações provenientes de múltiplas fontes, permitindo análises mais robustas [Dong and Srivastava 2013]. O Record Linkage (RL) consiste numa classe de soluções computacionais capazes de vincular registros de uma mesma entidade que se encontram em bases de dados diferentes [Christen 2019]. Estas bases de dados, que frequentemente dão suporte a sistemas de informação, formulados para atender propósitos distintos, não compartilham uma chave única capaz de identificar estas entidades.

Considere duas bases de dados fictícias do domínio da saúde. A Base de Dados A ($BD_A$), ilustrada na Tabela 1, tem propósito e variáveis similares ao Sistema de Informação que suporta a notificação compulsória dos casos de tuberculose em qualquer estabelecimento de saúde (SINAN-TB). Esta fonte de dados administrativos tem o potencial de habilitar diversas análises para auxiliar o planejamento da saúde, monitorar surtos e emergências em saúde, políticas interventórias, além da avaliação do impacto destas intervenções. A Tabela 1 reúne cinco registros fictícios, atribuídos a pessoas distintas, descrevendo eventos de notificação. As variáveis nome, sexo, mãe, dn e cidade podem ser consideradas quase identificadoras (QI). As variáveis categóricas hemoptise e dispneia suportam a caracterização dos casos suspeitos, podendo assumir os valores 1= 'Sim', 2= 'Não', e 9= 'Ignorado'. Os valores no atributo cid correspondem aos códigos da Classificação Internacional de Doenças, utilizados para classificar e codificar doenças investigadas para cada caso. Por fim, a variável criterio indica qual foi o método utilizado para confirmar ou descartar o caso suspeito. Os valores possíveis para esta variável são 1=Laboratorial, 2=Clínico-epidemiológico, ou 3=Por imagem.

**Tabela 8.1. Base de dados sintética A, simulando o Sistema de Informação de notificação e agravo da tuberculose (SINAN-TB)**

| nome | sexo | mae | dn | hemoptise | dispneia | cid10 | criterio |
|---|---|---|---|---|---|---|---|
| KEIDSON RIBEIRO | 1 | MARIA SANTOS | 2008-02-18 | 2 | 1 | A15.0 | 3 |
| CAMILA SARAIVA | 2 | KAREN DA SILVA | 2007-01-30 | 1 | 1 | None | 1 |
| VINICIUS DA COSTA | 1 | MAGNA LOPEZ | 2007-05-17 | 2 | 1 | A16.0 | None |
| VIVIAN GOMES | 2 | ARLENE LABORDA | 2006-12-18 | 2 | 2 | A15.0 | 2 |
| MYCHEL OLIVEIRA | 1 | ANA OLIVEIRA | 2007-05-10 | 2 | 2 | None | 4 |

A Base de Dados B ($BD_B$), por sua vez, simula os registros do Sistema de Informação de Mortalidade (SIM). Além dos QI, a Tabela 2 apresenta outros dois atributos para BDB, o circobito e causabas. Os valores possíveis para circobito denotam as circunstâncias do óbito, podendo ser 1=Natural, 2=Acidente, 3=Suicídio, 4=Homicídio, e 5=Outros. Por sua vez, o CID10 que representa a causa básica da morte é armazenado em causabas. Esta fonte de dados tem o potencial de suportar a construção de indicadores de mortalidade geral, mortalidade infantil, tendências temporais ou impacto de epidemias.

Uma integração bem sucedida das bases $BD_A$ e $BD_B$ pode habilitar tarefas analíticas ainda mais robustas, tais como o perfil populacional associado à morbimortalidade da tuberculose, o risco de morte pela doença dos diferentes estratos demográficos, o impacto de políticas públicas e o comportamento espacial ou temporal dos desfechos. Considerando que não existe uma chave que identifique os indivíduos existentes nestas duas bases

**Tabela 8.2. Base de dados sintética B, simulando o Sistema de Informação de Mortalidade (SIM)**

| nome | sexo | mae | dn | circobito | causabas |
|---|---|---|---|---|---|
| JULIA S LIMA | M | MARIA SANTOS | 9-out.-2007 | 2 | 3 |
| VIVIAN GOMES | F | ARLENE LABORDA | 8-dez.-2006 | 2 | 2 |
| RN DE MAGNA | F | MAGNA LOPEZ | 15-jun.-2008 | 1 | 2 |
| KEIDSON R | M | MARIA R SANTOS | 18-fev.-2008 | 1 | 3 |
| KAMILA S SARAIVA | F | KAREN SILVA | 30-jan.-2007 | 5 | 1 |

de dados, o *pipeline* de RL deve cumprir seis etapas fundamentais. A primeira etapa consiste na seleção das variáveis QI capazes de identificar de forma unívoca os indivíduos e que estão presentes nas duas bases de dados envolvidas. Em seguida, diversas estratégias de pré-processamento são utilizadas para aprimorar a qualidade dos dados envolvidos.

A próxima etapa é a indexação/blocagem das bases com o intuito de reduzir a quantidade de comparações entre os registros dos bancos. A quarta consiste na comparação par-a-par através de medidas de similaridade das variáveis selecionadas para representar o indivíduo. Na próxima etapa, cada par comparado é classificado como i) "provável ligação", que possivelmente correspondem a mesma entidade, correspondentes, não correspondentes ou potenciais correspondentes, de acordo com o valor de similaridade, ii) "ligação em potencial", que alcançaram grau de semelhança ou probabilidade suficiente para que sejam submetidos a uma nova rodada de comparação (frequentemente manual), e iii) "provável não-ligação", que possivelmente não correspondem à mesma entidade. Por fim, a sexta etapa agrupa os esforços para avaliação da qualidade do RL.

Segundo [Harron et al. 2016], o desenvolvimento e pesquisa em RL são ameaçados por quatro desafios. O principal desafio metodológico associado com a minimização dos erros de vinculação ocasionados pelo preenchimento imperfeito ou incorreto dos atributos QI. O segundo desafio consiste na dificuldade em endereçar os requisitos trazidos pelo contexto do Big Data. Em terceiro lugar, a autora aponta para as adversidades de manter um plano de governança de dados e engajamento público alinhados com os mais novos desdobramentos e técnicas disponíveis na literatura. Por fim, o quarto desafio se refere a escassez de acesso a dados identificados que impacta diretamente na formação de novos pesquisadores e analistas capazes de executar estas vinculações. As seções a seguir explora os dois primeiros desafios mencionados, focando em iniciativas brasileiras que implementam soluções de RL capazes de produzir grandes volumes de dados integrados com alta qualidade.

Este capítulo tem o objetivo de discutir as particularidades metodológicas e tecnológicas das tarefas de integração de coleções massivas de registros que demandam *Big Data Linkage* (BDL), com foco especial nas experiências e soluções nacionais. Será feita uma discussão sobre a evolução das ferramentas brasileiras para integrar bases de dados administrativas, suportando a pesquisa, tomada de decisão ou a formulação de políticas públicas. Entre estas iniciativas nacionais, destacam-se as duas soluções implementadas no escopo do Centro de Integração de Dados para a Saúde (CIDACS/Fiocruz). Desde sua criação em 2016, o CIDACS desenvolveu uma infraestrutura robusta, além de frameworks de governança e compartilhamento de dados e protocolos operacionais para a aquisição, gestão e vinculação de registros eletrônicos de saúde e sociais de abrangência nacional,

coletados administrativamente no Brasil. Como fruto desta evolução, o CIDACS, em parceria com a UFBA, lançou o AtyImo [Pita et al. 2018], uma ferramenta baseada em Apache Spark capaz de prover preservação de privacidade, integrando as primeiras bases de dados da Coorte de 100 Milhões de Brasileiros [Barreto et al. 2022]. A segunda ferramenta, o CIDACS-RL [Barbosa et al. 2020], implementa uma sofisticada estratégia de indexação baseada em Elasticsearch para diminuir drasticamente o custo computacional da comparação par-a-par.

## 8.2. Big Data Linkage

A arquitetura da Integração de Dados (ID) [Dong and Srivastava 2013] tradicional inclui três principais componentes, chamados *Schema Alignment* (SA), *Record Linkage* (RL) e *Data Fusion* (DF). As estratégias compreendidas no SA têm o objetivo de mediar a modelagem nas bases de dados de mesmo domínio que estão envolvidas no ID. As soluções computacionais no componente de RL aplicam técnicas de classificação para identificar registros que correspondem a uma mesma entidade, indivíduo ou evento em mais de uma base de dados [Christen 2019]. Por fim, os métodos do DF produzem datamarts contendo o subconjunto dos valores de interesse de cada entidade integrada. Com foco em suportar processos analíticos cada vez mais sofisticados, pesquisadores das áreas de saúde pública, estatística e computação se dedicam a elaborar métodos de RL capazes de endereçar diversos desafios relacionados à gestão de dados, desempenho e preservação de privacidade [Dunn and Chief 1946, Harron et al. 2016].

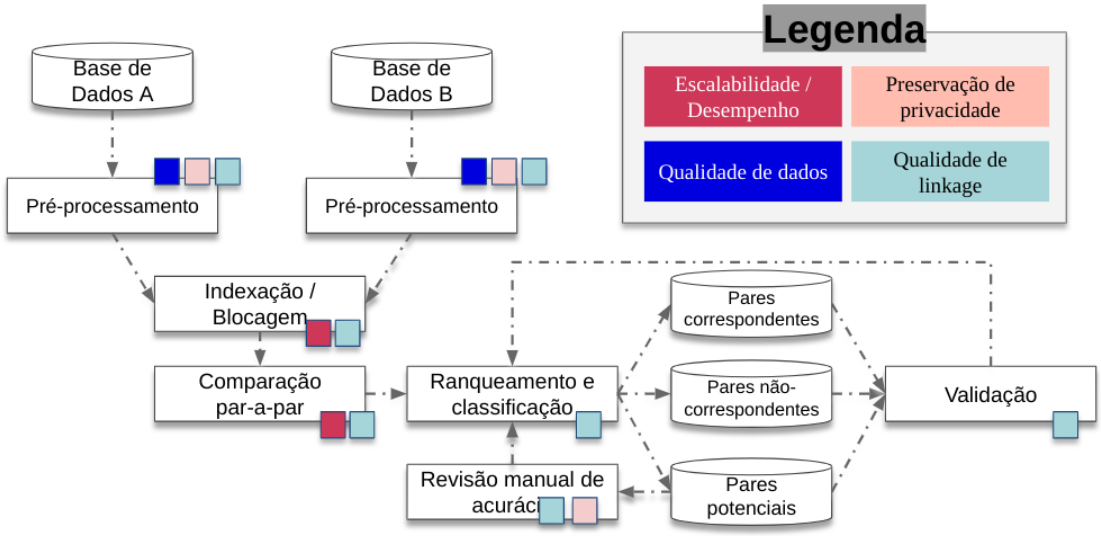



**Figura 8.1. Fluxo genérico de Integração de Dados e seus desafios associados**

Segundo [Harron et al. 2016], o desenvolvimento e pesquisa em RL são ameaçados por quatro desafios. O principal desafio metodológico associado com a minimização dos erros de vinculação ocasionados pelo preenchimento imperfeito ou incorreto dos atributos QI. O segundo desafio consiste na dificuldade em endereçar os requisitos trazidos pelo contexto do Big Data. Em terceiro lugar, a autora aponta para as adversidades de

manter um plano de governança de dados e engajamento público alinhados com os mais novos desdobramentos e técnicas disponíveis na literatura. Por fim, o quarto desafio se refere a escassez de acesso a dados identificados que impacta diretamente na formação de novos pesquisadores e analistas capazes de executar estas vinculações. Endereçar estes desafios ao longo do fluxo de ID, como ilustrado na Figura 8.2, requer a incorporação de algorítmos otimizados e adequados para a distribuição de dados envolvidas no processo.

O *Big Data Linkage* (BDL) refere-se ao processo de conectar registros de múltiplas fontes de dados em larga escala, garantindo que informações relacionadas a uma mesma entidade sejam corretamente identificadas [Dong and Srivastava 2013]. Diferente da vinculação tradicional, que lida com um número limitado de registros armazenados utilizando o modelo estruturado, o linkage em Big Data deve ser capaz de processar milhões de fontes dinâmicas, muitas vezes altamente heterogêneas, com diferentes níveis de qualidade, esquemas, escalas de mensuração ou modalidades.

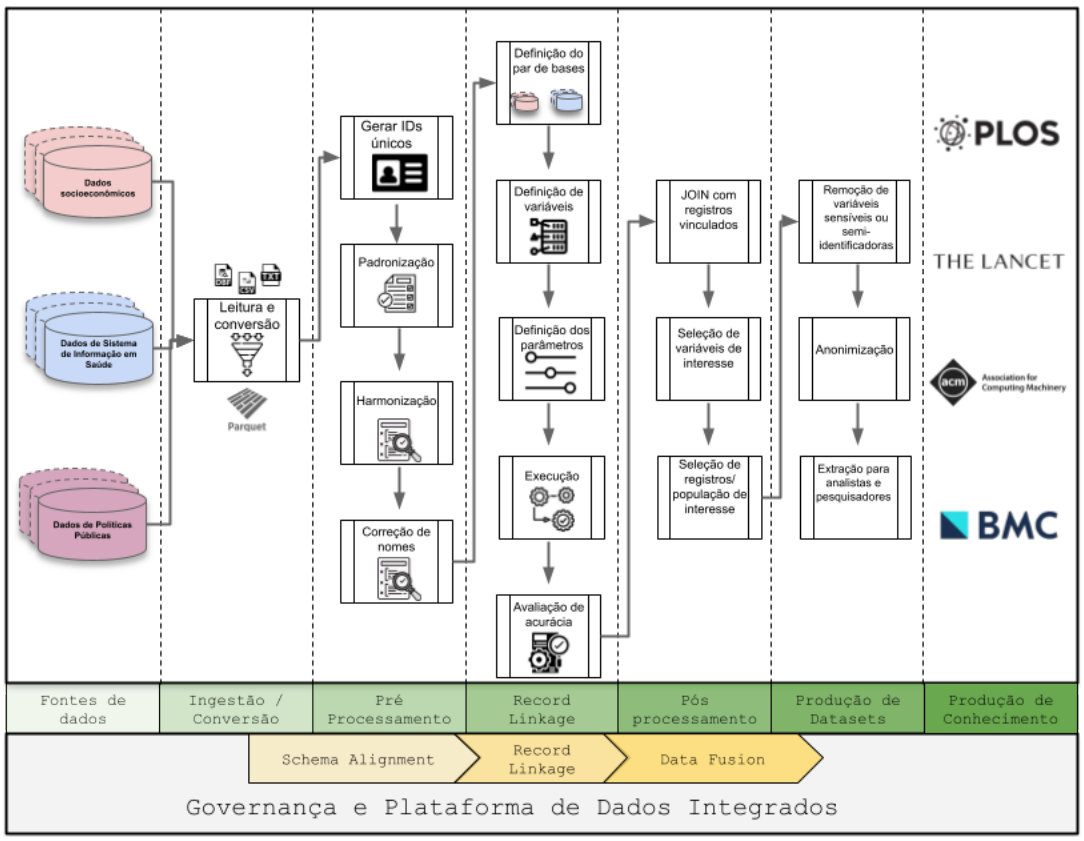



**Figura 8.2. *Pipeline* Genérico de Integração de dados**

No Brasil, diversos sistemas de informação suportam os processos de gestão do Sistema Único de Saúde (SUS). Estes sistemas diariamente coletam, armazenam e organizam dados capazes de subsidiar a medicina de precisão, o ajuste e a formulação de políticas públicas. Isoladamente, uma base de dados administrativa que registra eventos de hospitalizações pode oferecer *insights* capazes de subsidiar o processo de repasse de verbas federais, o provisionamento de insumos, o planejamento de estratégias para atenção primária, etc. Contudo, o enriquecimento destas informações com dados socio-

econômicos, notificação compulsória de doenças, ou morbimortalidade, permite a criação de um 'livro da vida', como proposto por [Dunn and Chief 1946], consolidando fatos relacionados a um mesmo indivíduo em várias fontes. Neste contexto, o uso de rotinas de RL tem chamado a atenção de pesquisadores, técnicos e gestores das áreas de gestão, governança, engenharia, análise, ciência e sistemas de dados [Harron et al. 2016].

A produção de conhecimentos em saúde através da ID dependem, portanto, da implementação de um *pipeline* maduro para gestão, manipulação, integração e uso destas bases massivas. A Figura 8.1 ilustra uma organização genérica deste *pipeline*, diluindo as três etapas propostas por [Dong and Srivastava 2013] em cinco fases de um modelo de Gestão e Governança numa Plataforma de Dados Integrados. Estas cinco fases são antecedidas de uma aquisição de dados e suportam a análise e descoberta de conhecimentos. É importante ressaltar que todo este processo de suporte a análises complexas, avaliação de políticas públicas e formulação de modelos para vigilância em saúde consiste num processo retroalimentado. Desta forma, as questões de pesquisa, hipóteses ou projetos elaborados impulsionam o pré-processamento, integração e extração destes dados da plataforma. Nas próximas subseções detalharemos essas fases, apresentando aspectos práticos de sua implementação através de pseudoalgoritmos comentados.

### 8.2.1. Técnicas de pré-processamento

Processos de vinculação de dados usam vários atributos que descrevem informações pessoais para classificar se um determinado par de registros se refere à mesma pessoa. As principais causas para baixos níveis de qualidade de dados estão relacionados com o fato de que bases de dados provenientes de diferentes sistemas de informação e mantenedores. Estes sitemas frequentemente definem diferentes padrões de formato, sistemas de dados e armazenamento físico. Do ponto de vista de rotina, os níveis de qualidade de dados são diretamente impactados pela supressão de valores, registros ou atributos para garantir segurança e privacidade [Christen et al. 2020].

As métricas de qualidade de dados são categorizadas entre aquelas capazes de medir a acurácia, validação, e completude [Christen 2012]. A acurácia é comumente avaliada de acordo com um documento auxiliar que mapeia a corretude dos dados coletados, tais como um dicionário de dados. A validade é aferida pela existência de valores plausíveis. Por exemplo, não é razoável haver indivíduos com idade maior que 200 anos. Para medir a completude, deve-se considerar a proporção de valores nulos ou ausentes.

Outras medidas de qualidade de dados incluem temporalidade, disponibilidade e credibilidade [Harron et al. 2016, Christen et al. 2020]. A temporalidade denota a distância entre os valores mensurados no tempo, suas versões mais atuais e a versão vinculada via ID. A disponibilidade indica a presença de atributos quase-identificadores com mesmo significado em ambas as bases. Por fim, para além do formato e da plausibilidade das métricas anteriores, a avaliação de credibilidade aferem se os registros armazenados numa coleção de dados são de origem confiável ou descrevem um fato.

Um dos principais desafios do BDI se refere a capacidade de conduzir um processo adequado de SA. produção de conhecimentos em saúde através da ID dependem, portanto, da implementação de um *pipeline* maduro para gestão, manipulação, integração e uso destas bases massivas. A Figura 8.1 ilustra uma organização genérica deste *pipe-*

**Algorithm 1** Conversão Genérica de Formatos de Dados

**Parâmetros:**

- $dir\_entrada$: Diretório fonte (suporta múltiplos formatos - ver comentários)
- $dir\_saida$: Diretório destino
- $formato\_saida \in \{\text{CSV}, \text{Parquet}, \text{ORC}\}$

**procedure** CONVERTERDADOS($dir\_entrada$, $dir\_saida$, $formato\_saida$)
    $spark \leftarrow iniciar\_sessao()$ ▷ Suporta .dbf, .dbc, .csv, .txt, etc.
    $arquivo$ in $listar\_arquivos(dir\_entrada)$
    $df \leftarrow carregar\_qualquer\_formato(spark, arquivo)$
    $df \leftarrow tratar\_dados(df)$
    $escrever\_saida(df, dir\_saida, formato\_saida)$

    $spark.stop()$
**end procedure**

▷ Detalhes de implementação:
▷ - Suporta formatos: DBF, DBC, CSV, JSON, TXT e similares
▷ - Processamento otimizado para big data via Spark
▷ - Conversão preserva metadados e estrutura original

*line*, diluindo as três etapas propostas por [Dong and Srivastava 2013] em cinco fases de um modelo de Gestão e Governança de uma Plataforma de Dados Integrados. A seguir, faremos um detalhamentos das principais rotinas representadas na Figura 8.1, expondo os principais aspectos práticos que afetam a escalabilidade e acurácia de uma tarefa de integração de dados. Os conceitos e estratégias apresentadas a seguir estão de acordo com a experiência acumulada pelo Núcleo de Produção de Dados do Cidacs, que historicamente é liderado por pesquisadoras da estatística e formada tecnicamente por pessoas engenheiras de dados de diferentes níveis de senioridade. Esta equipe é, portanto, co-responsável, na Plataforma de Dados, pelas peças de software que implementam as rotinas nucleares da Figura 8.1.

As principais tarefas *upstream* que abordaremos aqui são as de maior impacto no contexto de BDI, principalmente para assegurar um bom resultado na etapa de BDL. Detalharemos as tarefas de i) Conversão, Geração de IDs únicos, ii) Padronização, e iii) Correção de nomes.

### 8.2.2. Conversão

Como ilustrado no Algorítmo 1, um primeiro esforço para garantir um bom resultado de integração consiste em estabelecer um formato inicial padronizado para todas as bases de origem (*raw*). Desta forma, arquivos com extensões suportadas por ferramentas analíticas tradicionais (csv, dat, dbf, json, sas, sql, txt, xls, xlsx ou xml) devem ser convertidos para uma única extensão/formato. Esta conversão objetiva, portanto, compatibilizar a leitura dos arquivos *raw*, permitindo a estruturação do *schema* em um formato fortemente tipado que suporte o processamento paralelo e distribuido, como o *parquet*.

Idealmente, este processo também deve incluir a substituição de caracteres que

possam interferir na quantidade de colunas e limitações de campos em um arquivo csv (contrabarra, vírgula e aspas duplas), remoção de acentuações e de caracteres binários não visíveis. Adicionalmente, converte-se a codificação dos caracteres do arquivo de UTF-8 (*8-bit Unicode Transformation Format* – padrão de codificação mais completo e complexo existente hoje, que inclui acentuações) para ASCII (*American Standard Code for Information Interchange*).

**Algorithm 2** Geração de IDs Únicos para Bases de Dados

**Parâmetros:**

- $nome\_projeto$: Nome do projeto (ex: 'base_sim_2019_2020')
- $caminho\_entrada$: Caminho para os dados de entrada
- $caminho\_saida$: Caminho para os dados processados

**procedure** GERARIDSUNICOS
  Configurar ambiente: $spark.conf.set(\text{"}spark.sql.shuffle.partitions\text{"}, 288)$
  Listar arquivos de dados: $arquivos \leftarrow sorted(listar\_arquivos(caminho\_entrada))$
  Carregar dados:
  **for all** $arquivo$ in $arquivos$ **do**
    $dataframes \leftarrow spark.read.csv(arquivo, header = True, multiLine = True)$
  **end for**
  **Função** $adicionar\_id\_sequencial(dataframes, nome\_id, inicio)$:
  $dataframes\_com\_id \leftarrow []$
  $acumulador \leftarrow 0$
  $s \leftarrow inicio$
  **for all** $df$ in $dataframes$ **do**
    $inicio \leftarrow acumulador + s$
    $acumulador \leftarrow acumulador + df.count()$
    $rdd\_com\_indice \leftarrow df.rdd.zipWithIndex()$
    Adicionar ID único:
    $dataframes\_com\_id.append(spark.createDataFrame(\ldots))$
  **end for**
  **return** $dataframes\_com\_id$
  Aplicar função: $dfs\_com\_id \leftarrow adicionar\_id\_sequencial(dataframes, \text{"}id\_unico\text{"})$
  **Validação**:
  **for all** $df$ in $dfs\_com\_id$ **do**
    $df.select(\text{"}id\_unico\text{"}).describe().show()$
  **end for**
  **Organização dos resultados**:
  **for all** $i$ in $range(len(dfs\_com\_id))$ **do**
    $dfs\_com\_id[i].repartition(144).write.csv(\ldots)$
  **end for**
**end procedure**

Para garantia da integridade, a conversão deve ser sucedida de uma inspeção onde se compara a estrutura, o número de colunas, o nome das colunas e a quantidade de registros entre a base convertida e a *raw*.

### 8.2.3. Criação de IDs únicos

Em seguida, é importante atribuir um código identificador sequencial único controlado pela equipe de engenharia de dados envolvida na integração. Esta estratégia objetiva cumprir com requisitos de pseudonimização, onde o compartilhamento de identificadores internos não podem ser relacionados a registros idenficados fora do ambiente seguro. A inspeção do resultado do Algoritmo 2 consiste na verificação se o número de IDs únicos $X$ coincide com a quantidade de registros $N$ e se os valores vão entre um número inicial e $X + N$.

### 8.2.4. Padronização baseada em dicionário

A padronização explicitada no Algoritmo 3 tem papel relevante no SA. Além das bases de dados, deve-se considerar o dicionário de dados para compatibilizar os tipos de variáveis, sintaxe dos valores e nomes de atributos. Para fins de simplificação, consideraremos variáveis numéricas, categóricas, textuais (nomes) e de data. Considerando os tipos de dados suportados pelo Apache Spark, pode-se utilizar *LongType* para variáveis numéricas com mais de 13 dígitos e IntegerType para valores inteiros com menos de 13 dígitos que não são categóricas. A existência de caracteres não numéricos poderá indicar a nulidade daquele valor em determinados registros.

Para variáveis categóricas, os valores com correspondentes no dicionário deverão ser mantidos e os demais transformados. Registros nulos são mapeados para 0 (zero) e os inconsistentes (os que não estão contidos no dicionário) são mapeados para 99.

As variáveis de datas devem cumprir com algum formato suportado pelas ferramentas de processamento e análise e, para tanto, pode-se considerar a conversão destes valores para um padrão AAAA-MM-DD. Antes de padronizar cada variável de data, é utilizado uma Expressão Regular (regex) para substituir eventuais nomes de meses escritos ou abreviados em Inglês ou em Português, como mostrado na Tabela 8.2, para o valor numérico correspondente do mês.

Caso o campo de data contenha apenas 7 dígitos, a partir do formato da data informado no dicionário, faz-se 3 tentativas de adicionar um 0 à esquerda de cada um dos elementos da data (dia, mês, ano) até que uma data válida de 8 dígitos seja gerada correspondente com o formato no dicionário. Caso um valor não se enquadre como uma data ou contenha menos de 7 dígitos, ele é considerado inválido e seu campo será entendido como nulo (vazio). Para os casos de campo de data com 6 dígitos, deve-se fazer uma análise prévia e verificar se o ano não está abreviado em somente dois dígitos. Caso esteja, a depender do contexto da variável, o analista pode preencher este ano com '20' ou com '19', transformando o campo para 8 dígitos e, desta forma, a Padronização não irá anular estas datas.

### 8.2.5. Limpeza de nomes

Segundo [Christen et al. 2020], diferente de palavras comuns, nomes podem ser escritos e pronunciados de várias formas. Adicionalmente, os erros de grafia são os mais comuns neste tipo de variáveis, sendo que 80% acontecem apenas em um caracter [Damerau 1964]. Para que a base de dados seja submetida ao processo de Vinculação, é essencial que as variáveis utilizadas (nome, nome da mãe, data de nascimento, código do município de

**Algorithm 3** Padronização de Dados com Base em Dicionário

**Entrada:**

- *dicionario*: Arquivo CSV com especificação dos campos (nome, tipo, transformação)
- *bases_dados*: Lista de DataFrames com os dados brutos
- *caminho_saida*: Diretório para armazenar dados padronizados

**procedure** PADRONIZARDADOS
  Definir funções de padronização:
  **function** PADRONIZARDATA(*valor*)
    **if** *valor* é nulo **then return** 0
    **end if**
    Tentar converter para formato AAAA-MM-DD
    **if** conversão falhar **then return** 0
    **elsereturn** data convertida
    **end if**
  **end function**
  **function** PADRONIZARNUMERO(*valor*)
    **if** *valor* é nulo **then return** 0
    **else if** *valor* é numérico **then return** inteiro(*valor*)
    **elsereturn** 0
    **end if**
  **end function**
  **function** PADRONIZARCATEGORIA(*valor*, *mapeamento*)
    **if** *valor* é nulo **then return** 0
    **else if** *valor* existe em *mapeamento* **then return** código correspondente
    **elsereturn** 99 ▷ Valor desconhecido
    **end if**
  **end function**
  Aplicar padronização:
  **for all** $df$ in *bases_dados* **do**
    **for all** *coluna* in $df.colunas$ **do**
      Obter função de padronização correspondente
      $df[coluna] \leftarrow aplicar\_funcao(df[coluna])$
    **end for**
    Adicionar coluna *ano_base*
  **end for**
  Escrever resultados:
  **for all** $df$ in *bases_dados* **do**
    $df.escrever\_parquet(caminho\_saida)$
  **end for**
**end procedure**

residência e sexo) estejam devidamente tratadas. Considerando que a etapa de Padronização já garante a remoção de inconsistências das variáveis de data de nascimento, sexo e município, faz-se necessário checar e remover inconsistências das variáveis de nome e nome da mãe.

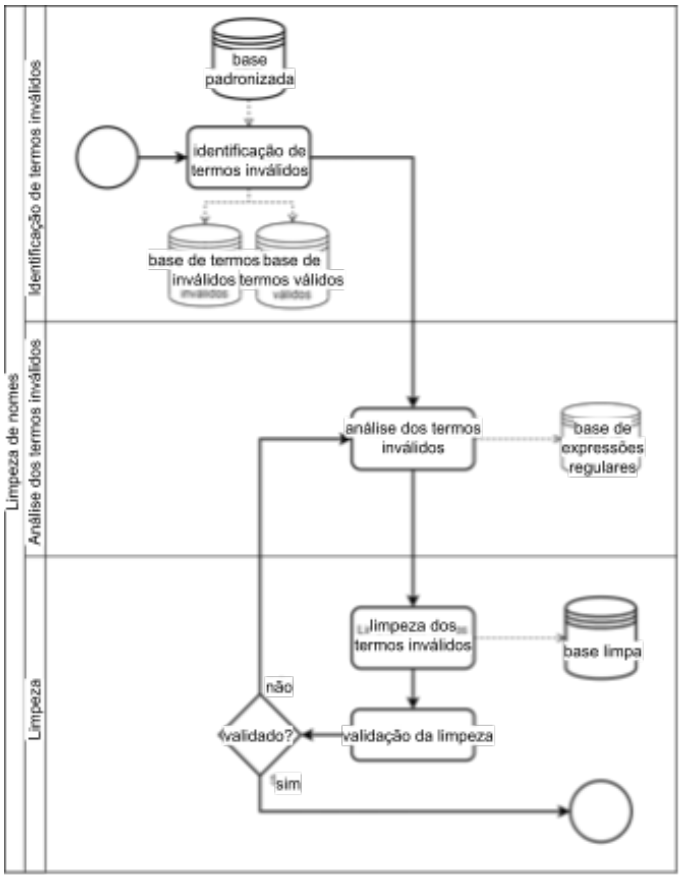



**Figura 8.3. Fluxo de limpeza de nomes**

Para as duas variáveis restantes o tratamento é mais complexo, pois se trata de variáveis do tipo String que possuem inconsistências diversas, incluindo o mau preenchimento por parte das pessoas que alimentam os sistemas de informação da base. A Figura 8.3 apresenta o fluxo da Limpeza de Nomes.

O processo de limpeza de nomes é iniciado com uma identificação dos tipos de inconsistência que ocorrem na base, das quais destacam-se três tipos de erros: i) gerados por ortografia ou uso de caracteres inválidos, ii) gerados por falta de informação, e iii) gerados por problemas na decriptação/dificação. No caso de erros ocasionados por problemas na digitação, um caractere pode ser trocado por outro, ou haver uma sequência de caracteres que não apresentam significado dentro do contexto da variável.

Muitas vezes, quando o indivíduo não possui documento de identificação no ato do cadastro no sistema, o campo nome ou nome da mãe são preenchidos com descrições da situação. Como, por exemplo, “IGNORADO”, “FILHO DE MARIA DA SILVA”; alguma caracterização do indivíduo como “DESCONHECIDA TRAJANDO SAIA ROSA” ou até algum apelido como “JOÃO DA BARBEARIA”.

Em alguns casos, a base, antes de ser cedida ao CIDACS, precisa passar por processos de criptografia e, quando é decriptada, revela alguns valores “corrompidos". Esses erros são bem específicos e com poucos padrões de repetição, o que torna difícil recuperar a informação original. Alguns exemplos de erros desse tipo são: "$EB", "&EBTJM", "TJM", "RHKUX", etc.

Uma vez mapeados os erros, é preciso definir o procedimento de tratamento para esse tipo de campo. Alguns procedimentos que devem ser avaliados e tratados são:

- Remoção de caracteres numéricos e especiais;
- Remoção de nome com somente um ou dois caracteres ou nomes com somente uma letra repetida *n* vezes, como, por exemplo: XXXX XXXXX;
- Identificação e remoção de inconsistências através de palavras-chave, por exemplo: IGNORADA, DESCONHECIDO, RECÉM-NASCIDO, RN DE, GEMELAR, LACTANTE, INDIGENTE;
- Identificação e análise de nomes que identifiquem a mãe do indivíduo como “RN DE <NOME_DA_MÃE>” ou “FILHO DE <NOME_DA_MÃE>”;
- Remoção de valores corrompidos por problemas na decriptação/codificação.

Em todos os casos, retira-se os termos que representam inconsistências e que eventualmente estejam no lugar do nome ou em meio aos nomes. Existem casos em que o nome do indivíduo está preenchido com algo que faz referência ao nome da mãe e o campo do nome da mãe está nulo. Neste caso, é possível recuperar o nome da mãe através da informação que está no nome do indivíduo,

### 8.2.6. Códigos fonéticos

Algoritmos fonéticos, incluindo Soundex e Metaphone, são frequentemente empregados no relacionamento de registros para identificar correspondências mesmo diante de divergências ortográficas ou erros de digitação. Essas técnicas exercem influência em quatro aspectos principais: proteção de dados, acurácia na vinculação, desempenho computacional e potenciais distorções. Embora contribuam para fortalecer a privacidade e aumentar a resiliência do processo de matching, podem acarretar certas perdas de exatidão e introduzir tendências sistemáticas [Karakasidis and Koloniari 2017, Herzog et al. 2007].

A transformação de nomes em códigos fonéticos irreversíveis reduz significativamente o risco de exposição de identificadores pessoais durante processos de record linkage, fortalecendo a proteção de dados sensíveis. Essa abordagem é particularmente valiosa em contextos que demandam privacidade por design, como na vinculação de registros de saúde ou dados governamentais. Contudo, para maior robustez contra tentativas de re-identificação, recomenda-se a combinação com técnicas complementares, tais como Filtros de Bloom ou injeção de ruídos [Schnell 2015].

**Algorithm 4** Correção de Nomes com Regex e Validação via Base Auxiliar

**Entrada:**

- *bases_dados*: Lista de DataFrames Spark com dados de saúde
- *padroes_invalidos*: Lista de expressões regulares (`regex`) para nomes claramente inválidos
- *padroes_dubios*: Lista de expressões regulares (`regex`) para nomes de validade duvidosa
- *nomes_validos*: DataFrame Spark com nomes validados manualmente em validações históricas
- *caminho_saida*: Diretório para armazenar dados corrigidos

**procedure** CORRIGIRNOMES
    Carregar bases usando Spark
    Unir todas as bases em um único DataFrame $df\_unificado$
    **Etapa 1: Pré-limpeza dos Nomes**
    $df\_unificado \leftarrow$ aplicar `regexp_replace` para:

- Remover acentuação e caracteres especiais
- Eliminar múltiplos espaços
- Converter para letras maiúsculas

    **Etapa 2: Marcação de Nomes Inválidos**
    **for all** *regex* in *padroes_invalidos* **do**
        Marcar registros cujo *nome* corresponde a *regex* usando `rlike`
    **end for**
    **Etapa 3: Identificação de Nomes Duvidosos**
    **for all** *regex* in *padroes_dubios* **do**
        Marcar registros cujo *nome* corresponde a *regex* usando `rlike`
    **end for**
    **Etapa 4: Validação dos Nomes Duvidosos**
    $df\_dubios \leftarrow$ registros marcados como duvidosos
    $df\_validados \leftarrow$ resultado da junção de $df\_dubios$ com *nomes_validos* usando Spark `join`
    Adicionar novos nomes validados manualmente ao DataFrame *nomes_validos*
    **Etapa 5: Aplicação das Correções**
    **for all** *registro* in $df\_unificado$ **do**
        **if** *registro.nome* é inválido (marcado na etapa 2) **then**
            Corrigir para `NULL` ou valor padrão
        **else if** *registro.nome* está em $df\_validados$ **then**
            Manter nome original
        **else if** *registro.nome* é duvidoso e não validado **then**
            Marcar como "nome pendente para revisão"
        **else**
            Manter nome pré-limpo
        **end if**
    **end for**
    Escrever resultados no formato Parquet em *caminho_saida*
**end procedure**

**Tabela 8.3. Codificação fonética de nomes com diferentes algoritmos**

| Nome | Soundex [Herzog et al. 2007] | Metaphone ptBR [Jordão and Rosa 2012] | NYSIIS [Grannis et al. 2002] |
|---|---|---|---|
| MARIA | M600 | MARIA | MARY |
| MARIANA | M650 | MARIANA | MARYAN |
| MARIANNA | M655 | MARIANA | MARYAN |
| MARIA ANA | M655 | MARIANA | MARYAN |
| MARINA | M650 | MARINA | MARYN |
| MIRIAN | M650 | MIRIAN | MARYAN |
| MILLIAM | M450 | MILIAM | MALLYAN |
| MARIAH | M600 | MARIA | MARY |

## 8.3. Iniciativas de Linkage no Brasil

O Brasil tem avançado significativamente no uso de dados administrativos para pesquisa em saúde, impulsionado pelo desenvolvimento de infraestruturas especializadas e metodologias de linkage [Sanni Ali et al. 2019]. Iniciativas como o Centro de Integração de Dados e Conhecimentos para Saúde (CIDACS), a Universidade Federal de Minas Gerais (UFMG) e a Universidade do Estado do Rio de Janeiro (UERJ) têm promovido a integração de bases como o Sistema de Informações sobre Mortalidade (SIM) e o Cadastro Único para Programas Sociais (CadÚnico). Esses esforços possibilitam a realização de estudos epidemiológicos e avaliações de políticas públicas baseadas em evidências.

O CIDACS [Barreto et al. 2019], criado em 2016 e vinculado à Fiocruz, abriga a Coorte dos 100 Milhões de Brasileiros [Barreto et al. 2022] e a Coorte de Nascimentos [Paixao et al. 2021], dois dos maiores bancos de dados do país voltados à pesquisa populacional. Além de consolidar informações socioeconômicas e de saúde, o centro desenvolveu os algoritmos AtyImo e CIDACS-RL, que aprimoram a vinculação de registros administrativos, garantindo alta precisão na identificação de indivíduos em diferentes bases. O centro opera com um parque computacional robusto, que inclui procedimentos rigorosos de governança, anonimização de dados e ferramentas de linkage, que permite a análise e processamento de grandes volumes de dados.

Outros centros de pesquisa também contribuíram significativamente para a inovação no linkage de dados no Brasil. A UFMG tem avançado no campo de linkage de dados por meio de algoritmos como o FERAPARDA e o PAREIA [Sanni Ali et al. 2019]. Essas ferramentas utilizam técnicas de linkage probabilístico e computação de alto desempenho para lidar com grandes volumes de dados, permitindo a integração de bases como SIH, SIA, SIM, SINASC e SINAN. Estes esforços resultaram num banco de dados nacional que integra registros individuais do SUS ao longo de 15 anos, permitindo análises aprofundadas em saúde pública. A UERJ, por sua vez, desenvolveu o RecLink, uma das ferramentas de linkage probabilístico mais utilizadas no Brasil. Baseado no modelo de Fellegi-Sunter, o RecLink [Camargo Jr and Coeli 2000] tem sido amplamente empregado em estudos epidemiológicos e na avaliação de subnotificações de doenças, como a tuberculose. Além disso, a UERJ tem trabalhado na criação de um data warehouse para integrar sistemas de informação relacionados ao câncer, como SIH-SUS, APAC-ONCO e

SIM, facilitando a análise de dados clínicos e administrativos. Essas iniciativas demonstram o avanço da ciência de dados aplicada à pesquisa em saúde e à gestão pública.

### 8.3.1. Métodos probabilísticos

O *Record Linkage* é uma ferramenta essencial para integrar dados administrativos ou clínicos e utiliza métodos, que podem ser divididos em dois tipos principais: determinísticos e probabilísticos. A escolha do método depende da qualidade dos dados e dos recursos disponíveis [Dusetzina 2014]. O método determinístico utiliza regras exatas para comparar registros. Os pares são analisados de acordo com identificadores comuns, e apenas os que coincidem completamente são considerados correspondências. Esse processo é razoavelmente simples, mas não lida bem com erros ou variações nos dados [Dusetzina 2014].

Os métodos probabilísticos, por sua vez, são indicados quando os dados possuem erros ou variações frequentes. Eles estimam a probabilidade de dois registros representarem a mesma entidade. Esses métodos permitem uma abordagem mais flexível e robusta para lidar com incertezas nos dados [Dusetzina 2014]. A integração de dados probabilística é utilizada quando não há a disponibilidade de atributos identifidores capazes de discriminar indivíduos de forma unívoca em duas bases de dados [Blake et al. 2022]. Eles funcionam calculando as probabilidades de correspondência baseadas nos dados observados.

De acordo com a formulação teórica proposta por [Fellegi and Sunter 1969], o método de RL probabilístico tem o objetivo de calcular a probabilidade de dois registros representarem a mesma entidade (*match*) ou não (*non-match*) através de uma comparação de seus atributos. Desta forma, considere uma tarefa de integração das bases expressas nas Tabelas 8.1 e 8.2, utilizando os atributos em comum (nome, mãe, dn e sexo). Inicialmente, deve-se estabelecer pesos $\boldsymbol{m}$ e $\boldsymbol{u}$ para cada atributo. Os pesos $\boldsymbol{m}$ correspondem à probabilidade de os valores concordarem, dado que se referem à mesma pessoa. De forma análoga, os pesos $\boldsymbol{u}$ expressam a probabilidade da concordância entre um par de valores ser mera coincidência, não se referindo à mesma entidade. Os passos subsequentes à definição dos parâmetros consistem em calcular os *scores* para cada par de registros e realizar a classificação dos pares utilizando limiares (*thresholds*). Desta forma, a Tabela considere este exemplo de pesos para as bases das Tabelas 8.1 e 8.2.

**Tabela 8.4. Exemplo de Parâmetros *m* e *u* para linkage probabilístico sobre os dados nas Tabelas 8.1 e 8.2**

| Atributo | $m$ | $u$ |
|---|---|---|
| nome | 0,95 | 0,10 |
| mae | 0,90 | 0,05 |
| dn | 0,85 | 0,01 |
| sexo | 0,80 | 0,15 |

A etapa dedicada ao cálculo de *score* em tarefas de RL probabilístico pode utilizar a equação de **log-odds ratio**

$$Peso = \log_2 \left( \frac{m}{u} \right) \textbf{ (se concordam)} \quad (1)$$

$$Peso = \log_2 \left( \frac{1-m}{1-u} \right) \textbf{ (se discordam)} \quad (2)$$

. Então, a Tabela 8.5 apresenta um exemplo de aplicação das Equações 1 e 2 entre os registros 2 e 4 das Tabelas 8.1 e 8.2, condiderando que as etapas de *Schem Alignment* e pré-processamento já foram realizadas. O resultado deste caso, é expresso por $3.25 + 4.17 - 2.72 + 2.42 = 7.12$. Por fim, o método determina que todos estes pares de registros e seus scores sejam submetidos a uma etapa de classificação, onde subsidiarão a decisão final do RL.

**Tabela 8.5. Exemplo de cálculo de pesos para linkage probabilístico**

| Campo | Base 1 | Base 2 | Concorda? | Peso (cálculo) |
|---|---|---|---|---|
| nome | VIVIAN GOMES | VIVIAN GOMES | Sim | $\log_2 \left( \frac{0,95}{0,10} \right) = 3,25$ |
| mae | ARLENE LABORDA | ARLENE LABORDA | Sim | $\log_2 \left( \frac{0,90}{0,05} \right) = 4,17$ |
| dn | 2006-12-18 | 2006-12-08 | Não | $\log_2 \left( \frac{0,15}{0,99} \right) = -2,72$ |
| sexo | 2 | 2 | Sim | $\log_2 \left( \frac{0,80}{0,15} \right) = 2,42$ |

A fase final do RL probabilístico envolve a definição dos limiares superior $(T+)$ e inferior $(T-)$. Desta forma, aqueles pares que produziram *scores* acima de $(T+)$ são imediatamente classificados com *matches*. Abaixo de $(T-)$ estão os pares de registros classificados como não correspondentes (*non-matctches*). O intervalo entre $(T+)$ e $(T-)$, portanto, configuram uma "zona cinzenta"que demandará um processo de revisão. A parametrizaçao adequada deste procedimento está associado a limiares que reduzam a zona cinzenta e consiga otimizar uma métrica de qualidade de linkage (especificidade, sensibilidade, precisão, etc).

Escolher variáveis adequadas para o linkage é fundamental. Elas devem ser discriminantes e possuir boa qualidade. Isso ajuda a derivar probabilidades mais precisas e confiáveis [Blake et al. 2022]. Os métodos probabilísticos se diferenciam dos determinísticos, pois consideram que alguns identificadores têm maior poder discriminativo. Por exemplo, uma coincidência em sobrenome e data de nascimento pode ser mais significativa do que uma coincidência no sexo [Dusetzina 2014].

As principais limitações do RL probabilístico circulam em três principais categorias: desempenho computacional, uso de recursos e a necessidade de informações de referência para definição de pesos e limiares. Considerando a complexidade do pareamento, expressa por $|A| \times |B|$, onde $|A|$ e $|B|$ se referem a quantidade de registros nas bases envolvidas, executar o RL em grandes volumes de dados é proibitivo. Para contornar isso, utiliza-se a indexação/blocagem. Estes métodos implementam estruturas de dados robustas capazes de suportar a organização e recuperação eficiente de subconjuntos de registros. Este filtro tem o potencial de diminuir drasticamente as comparações desnecessárias e diminuir o tempo total de execução do método de RL. Contudo, uma definição adequada do algorítmo de indexação e seus parâmetros vai garantir que pares legítimos sejam retirados do escopo de busca [Dusetzina 2014, Christen 2011].

Outra limitação latente associada a este tipo de RL é a ausência de informações que suportem uma parametrização adequada para cada bar de bases de dados. Por exemplo, sem um conjunto de documentos que suporte as definições dos pesos $\boldsymbol{m}$ e $\boldsymbol{u}$ na integração dos dados que estão nas Tabelas 8.1 e 8.2, o procedimento tende a falhar no *link* de muitos registros. Uma abordagem eficiente é usar as frequências empíricas para estimar as probabilidades de correspondência. Essa técnica utiliza a proporção real de coincidências dentro dos próprios dados. Ela simplifica o processo e torna o método mais adaptado à realidade dos conjuntos de dados [Blake et al. 2022]. A eficácia dos métodos probabilísticos depende da suposição de independência entre os identificadores. Caso essa suposição seja violada, as estimativas podem perder validade. No entanto, esses métodos são flexíveis o suficiente para lidar com configurações de dados complexas [Blake et al. 2022]. A seguir, abordaremos de forma resumida algumas iniciativas brasileiras que se dedicam a endereçar alguns destes desafios.

### 8.3.2. Reclink

RecLink [Camargo Jr and Coeli 2000] é um sistema de relacionamento de bases de dados que implementa o método probabilístico de record linkage, desenvolvido em C++ utilizando o ambiente Borland C++ Builder 3.0. Desde sua criação, o RecLink tem passado por atualizações relevantes [Camargo Júnior and Coeli 2002, Camargo Junior and Coeli 2006] e hoje possui uma versão opensource [Camargo Jr and Coeli 2015, Camargo and Coeli 2011] que potencializou seu uso por pesquisadores da saúde pública, estatística e computação. Dentre suas evoluções, destaca-se também a adoção de algorítmos para a estimação de pesos e limiares [Junger 2006]. Inúmeros trabalhos relataram o uso de bases de dados integradas com esta solução [Coeli et al. , Oliveira et al. 2016, Guillen et al. 2017, Vidal et al. 2006, Vieira et al. 2017], aprimorando a completude e qualidade das informações para pesquisa e vigilância epidemiológica.

O funcionamento do RecLink é similar ao descrito nas seções anteriores deste capítulo, seguindo as três etapas fundamentais do record linkage probabilístico: padronização, blocagem e pareamento. A padronização trata campos como nomes e datas de forma uniforme, removendo variações que poderiam prejudicar a comparação (por exemplo, convertendo letras minúsculas em maiúsculas e eliminando pontuação). A blocagem utiliza o algoritmo soundex aplicado aos nomes para agrupar registros com pronúncias semelhantes, limitando as comparações aos pares dentro de um mesmo bloco e otimizando o desempenho. O pareamento se baseia na construção de um escore ponderado para cada par de registros, calculado a partir das comparações dos campos selecionados (como nome, data de nascimento e sexo), utilizando pesos baseados na razão de verossimilhança entre as probabilidades de concordância em pares verdadeiros e falsos (segundo a metodologia de Fellegi  Sunter). O programa permite que os pares sejam classificados como "verdadeiros", "falsos"ou "duvidosos", conforme seus escores em relação a limiares definidos.

Em suas versões mais recentes, o RecLink aprimorou sua usabilidade e oferece uma interface interativa onde o usuário pode configurar os campos a serem usados, os algoritmos de comparação, e os limiares de decisão. Desde sua primeira versão, o programa permite diversas rodadas de blocagem com diferentes chaves, aumentando a chance de encontrar pares verdadeiros mesmo diante de erros ou variações nos dados. Com essas

características, o RecLink se estabelece como uma ferramenta acessível e eficiente para deduplicação e integração de registros em contextos reais.

### 8.3.3. PAREIA/ FERAPARDA

Em seu trabalho [Santos et al. 2007, dos Santos Filho 2008] descreve o "Filtro de Entidades para Remoção Automática Paralela de Registros Duplicados e Agressivamente Distribuída"(FERAPARDA). Esta ferramenta de deduplicação paralela, posteriormente referida como PAREIA [Junior et al. 2018], foi desenvolvida com foco em alto desempenho e escalabilidade. Ela foi projetada para identificar e remover registros duplicados em grandes bases de dados de forma eficiente, explorando arquiteturas distribuídas. O PAREIA foi originalmente construído sobre a plataforma Anthill [Ferreira et al. 2005], que permite a execução de aplicações baseadas em filtros encadeados (modelo de *pipeline*), onde cada filtro executa uma etapa do processo e pode ser replicado para mitigar a complexidade computacional da deduplicação. Trabalhos mais recentes relatam o uso do PAREIA para a criação de grandes ativos de dados integrados, produzidos a partir da deduplicação de mais de 1 bilhão de regisros [Junior et al. 2018].

Desde sua primeira publicação, o PAREIA propõe uma arquitetura em pipeline composta por filtros especializados. O filtro Reader divide a base de dados entre os processos e gera as chaves de blocagem. Em seguida, o Blocking agrupa os registros semelhantes, enquanto o Merge elimina pares redundantes. O Scheduler organiza os pares para reduzir o tráfego de dados entre processos, e o Comparator — a etapa mais custosa — aplica as funções de similaridade, podendo ser replicado para ganho de desempenho. Por fim, o Classifier avalia os escores de similaridade e decide se os pares são réplicas, com base em limiares pré-definidos. A comunicação entre os filtros é gerenciada pela biblioteca PVM (Parallel Virtual Machine). Projetado para suportar escalabilidade horizontal, o sistema permite aumentar dinamicamente o número de instâncias dos filtros mais exigentes. Essa arquitetura torna o FERAPARDA uma solução robusta e escalável para deduplicação de dados em ambientes distribuídos e de grande volume.

Seu funcionamento é dividido em estágios, implementados como filtros: o Reader divide a base entre os processos e gera as chaves de blocagem; o Blocking agrupa registros semelhantes; o Merge remove pares redundantes; o Scheduler organiza os pares para minimizar a comunicação; o Comparator aplica funções de similaridade e representa o estágio mais custoso (por isso pode ser replicado); e o Classifier determina se os pares são réplicas com base em limiares. A comunicação entre os filtros é gerenciada pela biblioteca PVM (Parallel Virtual Machine), e o sistema foi concebido para permitir a escalabilidade horizontal, aumentando o número de instâncias dos filtros críticos conforme a demanda. Essa arquitetura torna o FERAPARDA uma proposta inovadora para deduplicação de dados em ambientes distribuídos e com alto volume de informação.

### 8.3.4. AtyImo: Privacy-Preserving Record Linkage (PPRL)

AtyImo [Pita et al. 2018] é uma metodologia de record linkage com preservação de privacidade (PPRL - na sigla em inglês) escalável usada para criar a primeira versão de uma coorte de 100 milhões de indivíduos, integrando o CadÚnico e vários dados relacionados à saúde. O nome da ferramenta tem origem na concatenação de duas palavras,

o "*aty*"significa união em Tupy e "*imo*"se refere a conhecimento na língua Iorubá. Um 'átimo' também pode se referir a um evento que acontece num tempo bastante rápido.

Os Filtros de Bloom (BF, do inglês *Bloom Filters*) [Bloom 1970], originalmente desenvolvidos para verificar se a existência de um elemento pertence a um grupo, consistem em um vetor com $M$ Bytes, todos iniciados com 0. Utilizando uma ou mais funções hash, define-se quais das $M$ posições serão alteradas para 1. Na literatura, diversos estudos comunicam os avanços da capacidade dos Filtros de Bloom para assegurar a privacidade de atributos quasi-identificadores (QIDs) em tarefas de integração de dados [Franke et al. 2021, Ranbaduge and Schnell 2020, Vaiwsri et al. 2018]. A adoção do BF no AtyImo tem o objetivo de endereçar os requisitos originais de privacidade no acesso e integração de dados administrativos em ambientes computacionais com implementação limitada de segurança física e lógica. Outra propriedade relevante dos BF para o PPRL é o suporte ao cálculo de similaridade entre o par de registros codificados utilizando medidas como o Sorensen Dice [Cha 2008] ou o Jaccard [Fletcher et al. 2018].

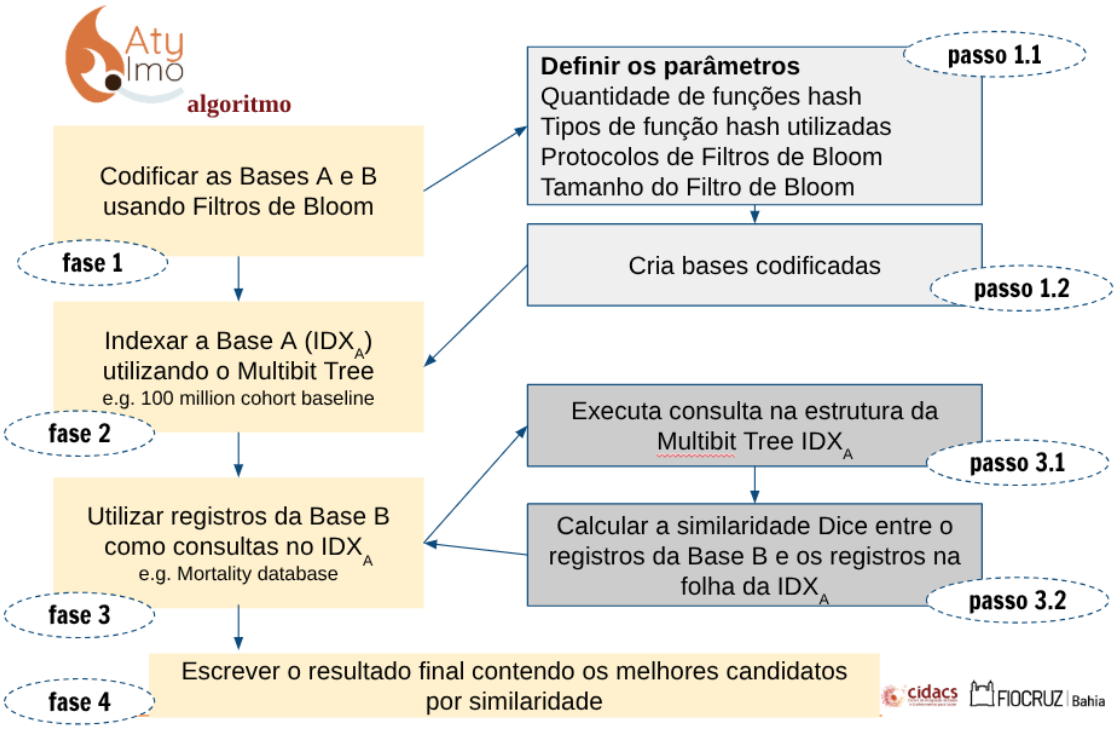



**Figura 8.4. Fluxo de execução do AtyImo**

### 8.3.4.1. Hashing e protocolos de Filtros de Bloom

Com o intuito de fortalecer a preservação de privacidade, é usual que o objeto usado na construção do BF passe por várias funções hash, alterando posições diferentes do vetor do BF para 1. Ao invés de utilizar várias funções hash diferentes, foi proposto [Dillinger and Manolios 2004] [Kirsch and Mitzenmacher 2006] o uso de duas funções hash (Double Hashing) dentro de uma fórmula, na qual uma das funções é multiplicada

por um fator que muda a cada iteração, simulando várias funções hash. Esse mesmo efeito pode ser alcançado através da Triple Hashing e Enhanced Double Hash.

```
import hashlib
# ============================
# Hashing Duplo (Double Hashing)
# ============================
def hash_duplo(token, repeticao, tamanho_filtro):
    """
    Calcula a posi  o no filtro com base em duas fun  es hash.
    Usa fator linear para dispers o das posi  es.
    """
    hash1 = hashlib.md5(token.encode()).hexdigest()
    hash2 = hashlib.sha256(token.encode()).hexdigest()
    posicao = (int(hash1, 16) + repeticao * int(hash2, 16)) %
   tamanho_filtro
    return posicao
# ============================
# Hashing Triplo (Triple Hashing)
# ============================
def hash_triplo(token, repeticao, tamanho_filtro):
    """
    Usa tres funcoes hash e termo quadratico para maior dispers o.
    """
    hash1 = hashlib.md5(token.encode()).hexdigest()
    hash2 = hashlib.sha256(token.encode()).hexdigest()
    hash3 = hashlib.sha512(token.encode()).hexdigest()
    termo_quadratico = int((repeticao * (repeticao - 1)) / 2)
    posicao = (
        int(hash1, 16)
        + repeticao * int(hash2, 16)
        + termo_quadratico * int(hash3, 16)
    ) % tamanho_filtro
    return posicao
# ====================================
# Hashing Duplo Aprimorado (Enhanced Double Hashing)
# ====================================
def hash_duplo_aprimorado(token, repeticao, tamanho_filtro):
    """
    Usa duas funcoes hash e termo cubico para aumentar a entropia.
    """
    hash1 = hashlib.md5(token.encode()).hexdigest()
    hash2 = hashlib.sha256(token.encode()).hexdigest()
    termo_cubico = int((repeticao**3 - repeticao) / 6)
    posicao = (
        int(hash1, 16)
        + repeticao * int(hash2, 16)
        + termo_cubico
    ) % tamanho_filtro
    return posicao
```

Para construir o BF em PPRL, geralmente os quase identificadores são divididos em q-gramas, que são strings de tamanho q. Por exemplo, usando q=3, é possível separar o nome João em dois q-gramas: "Joã"e "ão". Usando q-gramas, algumas técnicas foram desenvolvidas para preenchimento do BF. Em seu trabalho, Schnell et al. (2009)

sugeriu a codificação Attribute level Bloom filter (ABF), na qual cada atributo é utilizado para preencher um BF diferente. Como forma de tornar o BF mais resistentes a ataques, [Schnell et al. 2011] desenvolveu o Cryptographic Long-term Key (CLK), no qual todos os quase identificadores são inseridos no mesmo BF. Outra técnica é o record level Bloom filter (RBF), proposta por [Durham et al. 2013], na qual alguns bits do BF original são selecionados aleatoriamente para formar um novo BF.

Considere um registro $R = [$"JOAO PITA, "19870505, "1, "JANETE SANTOS$]$ com os seguintes valores para nome, data de nascimento, sexo e nome da mãe, respectivamente. Utilizando a estratégia de *Double Hashing*, deve cumprir com três etapas. Na primeira etapa, deve-se decompor o registro em n-gramas. Por exemplo, o valor "JOÃO PITA"deve ser inicialmente transformado em ['_J', 'JO', 'OA', 'AO', 'O ',' P', 'PI', 'IT', 'TA', 'A_']. Em seguida, aplica-se

$$\text{posição}_i = (\text{int}(h_1(\text{bigrama})) + i \cdot \text{int}(h_2(\text{bigrama}))) \mod m, \tag{3}$$

onde $i$ é o índice da função hash simulada, variando de 1 até $k$ (número total de funções hash), $h_1$ e $h_2$ são funções de hash distintas (por exemplo, MD5 e SHA256) e $m$ é o tamanho do filtro de Bloom (número total de bits). Desta forma, se o valor produzido pela função $h_1(\texttt{"JO}) = 146379$ e $(h_2(\texttt{"JO})) = 94809$, com $m = 20$, temos para $i = 1$:

$$\begin{aligned}\text{posição}_1 &= (146379 + 1 \cdot 94809) \mod 20 \\ &= 241188 \mod 20 \\ &= 8,\end{aligned}$$

para $i = 2$:

$$\begin{aligned}\text{posição}_2 &= (146379 + 2 \cdot 94809) \mod 20 \\ &= 335997 \mod 20 \\ &= 17\end{aligned}$$

e para $i = 3$:

$$\begin{aligned}\text{posição}_3 &= (146379 + 3 \cdot 94809) \mod 20 \\ &= 430806 \mod 20 \\ &= 6.\end{aligned}$$

O vetor correspondente ao registro $R$ iniciado com 20 zeros vai ter, portanto, as posições 8, 18, 10 alterados para 1. No final, codificar todo o registro $R$ exigirá que estes passos sejam sucessivamente repetidos até que se esgotem os bi-gramas em $R$.

O *Triple Hashing* consiste na estratégia de construir BF de $R$ utilizando três funções de hash distintas ($h_1$, $h_2$, $h_3$), e a posição $i$-ésima no vetor é obtida pela fórmula

$$\text{posição}_i = \left(\text{int}(h_1(\text{bigrama})) + i \cdot \text{int}(h_2(\text{bigrama})) + \frac{i \cdot (i-1)}{2} \cdot \text{int}(h_3(\text{bigrama}))\right) \mod m. \tag{4}$$

Portanto, considerando que $\text{int}(h_1(\texttt{"JO})) = 146379$, $\text{int}(h_2(\texttt{"JO})) = 94809$ e $\text{int}h_3(\texttt{"JO}))$ = 30721, então, para $i = 1$:

$$\begin{aligned}\text{posição}_1 &= \left(146379 + 1 \cdot 94809 + \frac{1 \cdot (1-1)}{2} \cdot 30721\right) \mod 20 \\ &= (146379 + 94809 + 0) \mod 20 \\ &= 241188 \mod 20 \\ &= 8\end{aligned}$$

para $i = 2$:

$$\begin{aligned}\text{posição}_2 &= \left(146379 + 2 \cdot 94809 + \frac{2 \cdot (2-1)}{2} \cdot 30721\right) \mod 20 \\ &= (146379 + 189618 + 30721) \mod 20 \\ &= 366718 \mod 20 \\ &= 18\end{aligned}$$

e para $i = 3$:

$$\begin{aligned}\text{posição}_3 &= \left(146379 + 3 \cdot 94809 + \frac{3 \cdot (3-1)}{2} \cdot 30721\right) \mod 20 \\ &= (146379 + 284427 + 2 \cdot 30721) \mod 20 \\ &= (146379 + 284427 + 61442) \mod 20 \\ &= 492248 \mod 20 \\ &= 8.\end{aligned}$$

Neste caso, enquanto o *Double Hashing* produzirá Bytes 1 nas posições 8, 18 e 10 do BF de $R$, o *Triple Hashing* vai incidir apenas nas posições 8 e 18.

A formulação do *Enhanced Double Hashing*, incluindo um termo cúbico dependente do índice $i$, envolve o seguinte calculo de posição:

$$\text{posição}_i = \left(\text{int}(h_1(\text{bigrama})) + i \cdot \text{int}(h_2(\text{bigrama})) + \frac{i^3 - i}{6}\right) \mod m. \qquad (5)$$

O termo cúbico expresso em $\frac{i^3-i}{6}$ tem o objetivo de i) reduzir colisões como as produzidas na demonstração do *Double Hashing*, ii) introduzir uma variação adicional que simula o uso de várias funções hash consecutivas, e iii) amplia o alcance das Equações 3 e 4 para atribuir valores 1 em posições mais distribuídas ao longo do BF. Então, considerando que $\text{int}(h_1(\texttt{"JO})) = 146379$ e $\text{int}(h_2(\texttt{"JO})) = 94809$, com $m = 20$, temos para $i = 1$:

$$\begin{aligned}\text{posição}_1 &= (146379 + 1 \cdot 94809 + \frac{1^3 - 1}{6}) \mod 20 \\ &= (146379 + 94809 + 0) \mod 20 \\ &= 241188 \mod 20 \\ &= 8,\end{aligned}$$

para $i = 2$:

$$\begin{aligned}\text{posição}_2 &= (146379 + 2 \cdot 94809 + \frac{8-2}{6}) \mod 20 \\ &= (146379 + 189618 + 1) \mod 20 = 335998 \mod 20 \\ &= 18,\end{aligned}$$

e para $i = 3$:

$$\begin{aligned}\text{posição}_3 &= (146379 + 3 \cdot 94809 + \frac{27-3}{6}) \mod 20 \\ &= (146379 + 284427 + 4) \mod 20 \\ &= 430810 \mod 20 \\ &= 10.\end{aligned}$$

A decisão sobre qual a melhor estratégia de composição de funções Hash para a decisão das posições que receberão o valor 1 em BFs deve ser tomada com base em avaliação experimental. Deve-se levar em consideração as bases de dados envolvidas, a quantidade média de q-gramas e a qualidade associada a estes registros.

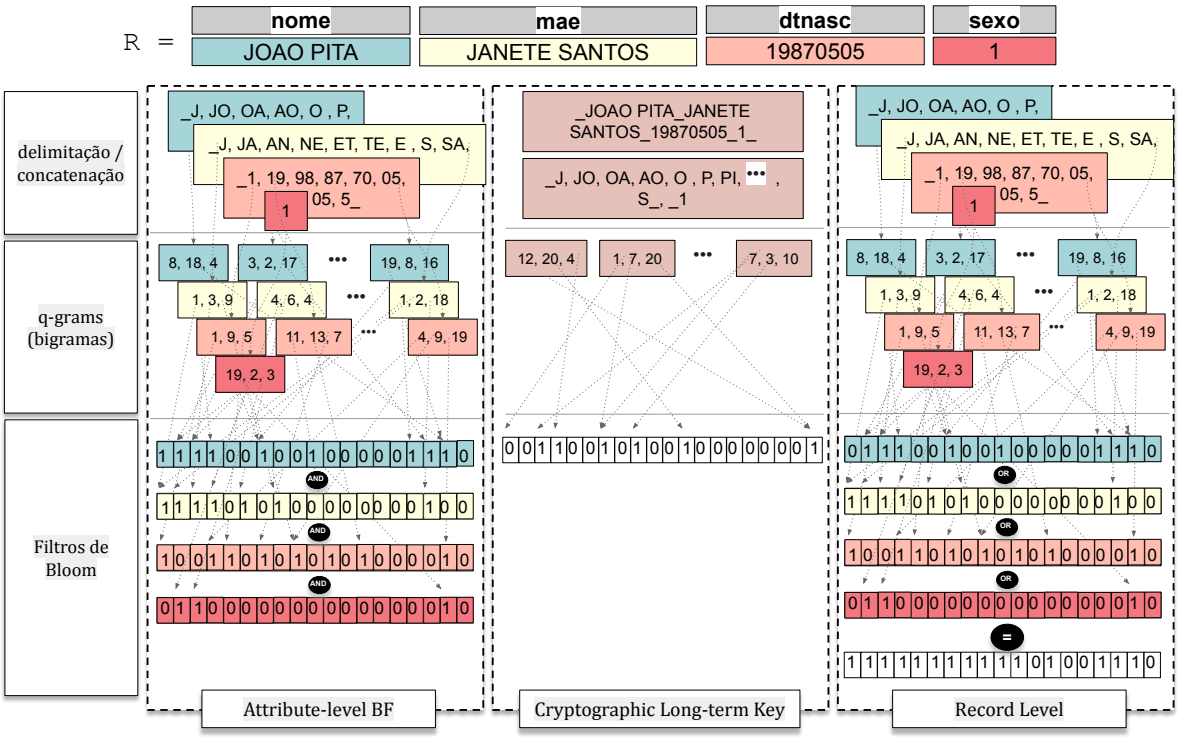



**Figura 8.5. Exemplos de Protocolos de Bloom para o Registro $R$ com Double Hashing com 20 posições**

Outra decisão relevante no projeto de um PPRL consiste na escolha do Protocolo BF que será utilizado. As alternativas mais citadas na literatura são o *Attribute level Bloom filter* (ABF) [Schnell et al. 2009], *Cryptographic Long-term Key* (CLK) [Schnell et al. 2011] e o *record level Bloom filter* (RLB) [Durham et al. 2013]. Considere um cenário hipotético em que pretende codificar o registro *R* usando o *Triple Hashing* e em um vetor de 20

posições, como ilustrado na Figura 8.5. O protocolo ABF implementa três passos. No primeiro passo, cada atributo é decomposto em q-grams - em bigramas neste caso. Cada conjunto de bigramas de um respectivo atributo será submetido à Equação 4 para definir qual posição de um vetor independente receberá o valor 1. Portanto, o resultado do ABF é um vetor com em todos os bigramas produzidos será de $a \times m = 80$, sendo $a$ e $m$ a quantidade de atributos e o tamanho do BF, respectivamente.

```
# ABF UDF: Geracao do filtro de Bloom em nivel de atributo
def abf(registro, m, k, t_hash):
    """
    Cria um Attribute Level Bloom Filter (ABF) para cada atributo do registro.
    Para cada valor, gera bigramas e aplica a funcao de hash definida.

    Parametros:
    - registro: lista de valores de atributos
    - m: tamanho do filtro de Bloom (numero de bits)
    - k: numero de iteracoes (funcoes hash simuladas)
    - t_hash: tipo de funcao de hash ("DoubleHash", "TripleHash" ou outro)

    Retorna:
    - Filtro de Bloom concatenado para todos os atributos.
    """
    filtro_bloom_final = []
    for item in registro:
        filtro = [0] * m
        if item is not None:
            item = '_' + str(item) + '_'
            for j in range(len(item) - 1):
                cont = 1
                while cont <= k:
                    substring = item[j:j+2]
                    if t_hash == "DoubleHash":
                        pos = dbhash(substring, cont, m)
                    elif t_hash == "TripleHash":
                        pos = tphash(substring, cont, m)
                    else:
                        pos = endbhash(substring, cont, m)
                    filtro[pos] = 1
                    cont += 1
        # Concatena o filtro do atributo ao vetor final
        filtro_bloom_final.extend(filtro)
    return filtro_bloom_final
```

O CLK, por sua vez, concatena todos os atributos como uma única *string* e produz os bigramas que serão submetidos à Equação 4, resultando em um BF de 20 posições.

```
# CLK UDF: Geracao do filtro de Bloom baseado no registro completo (Cryptographic Long-term Key)
def clk(registro, m, k, t_hash):
    """
    Cria um filtro de Bloom unico para um registro completo, aplicando a tecnica
    Cryptographic Long-term Key (CLK). Todos os atributos sao
```

```
    considerados juntos.

    Para cada valor do registro, sao gerados bigramas com delimitadores
     e, para cada bigrama,
    aplicam-se funcoes de hash simuladas para definir posicoes no
    filtro de Bloom.

    Parametros:
    - registro: lista de valores de atributos do registro
    - m: tamanho do filtro de Bloom (numero total de bits)
    - k: numero de funcoes hash simuladas
    - t_hash: tipo da funcao de hash a ser utilizada ("DoubleHash", "
    TripleHash" ou outro)

    Retorna:
    - Um vetor binario representando o filtro de Bloom do registro
    inteiro.
    """
    # Inicializa o vetor do filtro de Bloom com zeros
    filtro_bloom = [0] * m

    # Itera sobre cada atributo do registro
    for item in registro:
        if item is not None:
            # Converte o valor para string e adiciona delimitadores
            item = '_' + str(item) + '_'

            # Gera bigramas e aplica as funcoes de hash simuladas
            for j in range(len(item) - 1):
                bigrama = item[j:j+2]
                for cont in range(1, k + 1):
                    # Escolhe a funcao de hash conforme o tipo indicado
                    if t_hash == "DoubleHash":
                        pos = dbhash(bigrama, cont, m)
                    elif t_hash == "TripleHash":
                        pos = tphash(bigrama, cont, m)
                    else:
                        pos = endbhash(bigrama, cont, m)

                    # Define a posicao correspondente como 1
                    filtro_bloom[pos] = 1

    return filtro_bloom
```

Similarmente ao ABF, o RLB gera um BF independente para cada produto. Contudo, o RLB computa funções "OR"entre os bytes de mesma posição para definir quais posições do BF resultante receberão o valor 1. A decisão sobre a combinação mais adequada entre as estratégias de hashing e o protocolo de BF depende de várias características do projeto de integração de dados, a exemplo da capacidade de armazenamento e processamento, impactados pelo tamanho dos vetores e dos requisitos de privacidade.

```
# RLB UDF: Geracao do filtro de Bloom em nivel de registro (Record-
    level Bloom Filter)
def rlb(registro, a, m, k, t_hash, positions, vetor):
```

```
    """
    Cria um Record-Level Bloom Filter (RLB) a partir de um ABF e um
    conjunto
    de posicoes sorteadas, aplicando embaralhamento controlado.

    Parametros:
    - registro: lista de valores de atributos
    - a: tamanho final do filtro de Bloom (numero de bits no RLB)
    - m: tamanho do filtro de Bloom por atributo (usado no ABF)
    - k: numero de iteracoes (funcoes hash simuladas)
    - t_hash: tipo de funcao de hash ("DoubleHash", "TripleHash" ou
    outro)
    - positions: vetor com posicoes sorteadas para cada bloco do ABF
    - vetor: vetor de indices usados para embaralhar o resultado final

    Retorna:
    - Um vetor binario com 'a' bits representando o RLB.
    """
    filtro_intermediario = []
    filtro_final = [0] * a
    inicio = 0
    deslocamento = 0

    # Gera o ABF completo concatenado para todos os atributos
    abf_completo = abf(registro, m, k, t_hash)

    # Para cada bloco de tamanho m, seleciona posicoes sorteadas
    while inicio < a:
        for i in range(inicio, m + inicio):
            posicao = positions[i] + deslocamento
            filtro_intermediario.append(abf_completo[posicao])
        inicio += m
        deslocamento += m

    # Embaralha o vetor intermediario conforme o vetor de indices
    for i in range(len(filtro_intermediario)):
        filtro_final[i] = filtro_intermediario[vetor[i]]

    return filtro_final
```

Por fim, pode-se utilizar UDFs em rotinas Apache Spark para garantir a adaptação destas tarefas em uma infraestrutura distribuída e escalável.

```
# ABF - Aplicacao da funcao ABF como UDF no PySpark
# Referencia: https://sparkbyexamples.com/pyspark/pyspark-udf-user-
    defined-function/
abfUDF = udf(lambda z: abf(z, m, k, t_hash), ArrayType(IntegerType()))
df_a_ABF = df_a.withColumn("bloom", abfUDF(df_a.vet))
df_b_ABF = df_b.withColumn("bloom", abfUDF(df_b.vet))

# CLK - Aplicacao da funcao CLK como UDF no PySpark
clkUDF = udf(lambda z: clk(z, a, k, t_hash), ArrayType(IntegerType()))
df_a_CLK = df_a.withColumn("bloom", clkUDF(df_a.vet))
df_b_CLK = df_b.withColumn("bloom", clkUDF(df_b.vet))

```

```
# RLB - Aplicacao da funcao RLB como UDF no PySpark
rlbUDF = udf(lambda z: rlb(z, a, m, k, t_hash, positions, vetor),
    ArrayType(IntegerType()))
df_a_RLB = df_a.withColumn("bloom", rlbUDF(df_a.vet))
df_b_RLB = df_b.withColumn("bloom", rlbUDF(df_b.vet))
```

## 8.4. CIDACS-RL

O CIDACS-RL é uma ferramenta de BDL que implementa uma abordagem baseada em pontuação de similaridade para pareamento de bases de dados em larga escala. Desenvolvido para operar em ambientes distribuídos com Apache Spark e Elasticsearch, o sistema foi projetado para lidar com desafios típicos de ID em saúde e registros administrativos, como variabilidade nos identificadores, erros de digitação e ausência de chaves únicas universais. O núcleo metodológico do CIDACS-RL baseia-se em um processo de duas fases: uma etapa inicial de busca exata (exact match) seguida por uma fase de correspondência aproximada (fuzzy match), otimizada por técnicas de indexação e bloqueio para reduzir o espaço de busca.

Na fase de busca exata, o sistema constrói consultas estruturadas no Elasticsearch utilizando campos pré-definidos como obrigatórios (must_match), como nome completo e data de nascimento. Essas consultas são geradas dinamicamente a partir de um arquivo de configuração que especifica os campos relevantes, pesos e limiares de similaridade. Quando um registro atinge um score de similaridade acima do limiar configurado (cutoff_exact_match), ele é considerado uma correspondência válida e removido do pool de registros a serem processados na fase seguinte. Essa abordagem garante eficiência computacional ao resolver casos inequívocos precocemente.

Para registros não pareados na primeira fase, como ilustrado na Figura 8.6, o CIDACS-RL aplica algoritmos de correspondência aproximada com tolerância a erros. A ferramenta utiliza medidas de similaridade como Jaro-Winkler para nomes, Hamming para datas e overlap para campos categóricos, com pesos ajustáveis para cada atributo. Um aspecto inovador é a integração nativa com recursos de fuzzy matching do Elasticsearch, como consultas com parâmetros de fuzziness e boosting, que permitem capturar variações ortográficas sem sacrificar desempenho. A combinação linear ponderada dessas medidas produz um escore final de similaridade, utilizado para selecionar a melhor correspondência entre os candidatos pré-filtrados.

O CIDACS-RL é uma ferramenta avançada que implementa estratégias de indexação baseadas em Elasticsearch, reduzindo drasticamente o custo computacional do pareamento de registros. Desenvolvido pelo CIDACS/Fiocruz, o CIDACS-RL tem sido utilizado em projetos nacionais, demonstrando alta eficiência e escalabilidade [Barbosa et al. 2020].

A versão atual, publicamente disponível para uso inclui: i) o provisionamento local de infraestrutura usando conteinerização através do Docker e Podman, ii) a conexão com o serviço de computação em nuvem de preferência do usuário, ou iii) a conexão com a infraestrutura on-premises do pesquisador.

| valores_de_variaveis | consulta_elasticsearch | melhor candidato exato | similaridade | vetor_candid | linked_from |
|---|---|---|---|---|---|
| ... | ... | ... | ... | ... | ... |
| ['JOAO PITA', '19870505', '1', 'JANETE SANTOS'] | { "size": "50", "query": { "bool": { "should": [ {'match': {'nome_a': {'query': 'JOAO PITA', 'fuzziness':'AUTO', 'operator':'or', 'boost':'3.0'}}}, {match": {"birthdate":"19870505"}} ] } } }, "term": {"sexo_a":"1"}} ] } } } | 13 | 0.94 | {7: 0.98, 3: 0.8... | non_exact_match |
| ['MARIA SENA, 19891112', '2', NAIRA DILMA'] | { "size": "50", "query": { "bool": { "should": [ {'match': {'nome_a': {'query': 'JOAO PITA', 'fuzziness':'AUTO', 'operator':'or', 'boost':'3.0'}}}, {match": {"birthdate":"19870505"}} ] } } }, "term": {"sexo_a":"1"}} ] } } } | null | null | null | exact_match |
| ... | ... | ... | ... | ... | ... |

**Figura 8.6. Exemplo de Dataframe Spark operado pelo CIDACS-RL**

### 8.4.1. Algorítmo do CIDACS-RL

O CIDACS-RL, descrito em detalhes na Figura 8.7, é baseado em quatro fases. A primeira fase consiste numa indexação da base de busca (geralmente a base com maior quantidade de registros) usando a biblioteca Elasticsearch-Hadoop [1]. Nesta fase, um Dataframe Spark pré-processado é transformado num Índice Elasticsearch que poderá ser alvo de consultas posteriormente. O Elasticsearch [Shukla 2015] é uma plataforma de indexação e busca distribuída que oferece uma API, linguagem de consulta própria e ferramentas nativas que suportam análises em bancos de dados de alta dimensionalidade. Apesar de implementar o modelo de dados orientado a documentos, é possível usar, em bases estruturadas advindas de bancos relacionais, representar abstração de tuplas em documentos através de dicionários Python ou arquivos JSON. Desta forma, é possível navegar ou consultar coleções indexadas pelo Elasticsearch usando estuturas de dados chave-valor.

```
# Funcao python para indexacao
def index_dataframe(dataframe, es_index_name):
    # creating new index
    dataframe.write.format("org.elasticsearch.spark.sql") \
                .option("es.resource", es_index_name).mode('overwrite'
   ).save()
```

Para garantir a escalabilidade, permitindo o paralelismo ou distribuição plena das suas instruções, a implementação do CIDACS-RL foca em funções nativas ou definidas por usuário (UDF, do inglês *User Defined Function*) que operem diretamente em Spark Dataframes. Isso implica em evitar funções e procedimentos classificados como "Ações", permitindo um acúmulo de "Transformações"para garantir uma otimização maior do plano de execução da rotina.

A segunda fase do algorítmo do CIDACS-RL tem o objetivo de endereçar a complexidade computacional imposta pelo alto volume de dados. Sua implementação consiste em construir dois tipos de consultas a partir dos valores nas colunas que serão utilizadas na comparação par-a-par. Por exemplo, para uma base com atributos "nome", "dtnasc",

[1] https://mvnrepository.com/artifact/org.elasticsearch/elasticsearch-hadoop

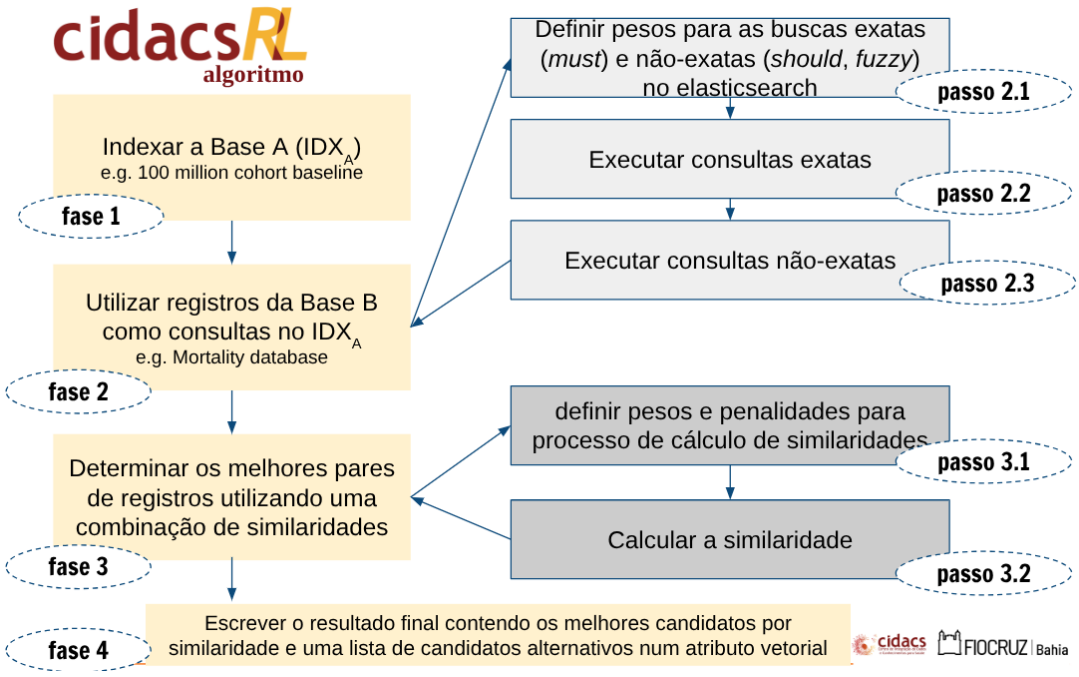



**Figura 8.7. Fluxo de execução do CIDACS-RL**

"sexo"e "nome_mae", como descrito na Figura 8.6. Os valores de uma tupla qualquer, neste Dataframe, pode ser, portanto, transformado em [‘JOAO PITA’, ‘19870505’, ‘1’, ‘JANETE SANTOS’] usando a Linha 7 do código a seguir.

```
1  # Trecho com implementacao da criacao de busca exata
2  def cidacs_rl_exact_phase(tolink_dataset):
3  # [...]
4      # selecting columns
5      tolink_dataset = tolink_dataset.select(tolink_columns)
6      # building array of variable values
7      tolink_dataset = tolink_dataset.withColumn('vars', F.array(
       tolink_columns))
8      # building exact queries
9      tolink_dataset = tolink_dataset.withColumn('exact_queries',
       udf_build_exact_queries(F.col('vars')))
10     # finding the best candidate for each tolink record
11     tolink_dataset = tolink_dataset.withColumn('result_exact_search', F
       .explode(F.array(udf_find_elasticsearch_exact_best_candidate(F.col(
       'vars'), F.col('exact_queries')))))
12 # [...]
13     # exploding array columns from the last function into 4 atomic cols
14     tolink_dataset = tolink_dataset.withColumn('best_candidate_exact',
       tolink_dataset.result_exact_search['best_candidate_exact'])
15     tolink_dataset = tolink_dataset.withColumn('
       sim_best_candidate_exact', tolink_dataset.result_exact_search['
       sim_best_candidate_exact'])
16     tolink_dataset = tolink_dataset.withColumn('
       similarity_exact_candidates', tolink_dataset.result_exact_search['
       similarity_exact_candidates'])
17
```

```
18      tolink_dataset = tolink_dataset.withColumn('
        sim_best_candidate_exact', F.col('sim_best_candidate_exact').cast('
        float'))
19
20      # dropping array columns
21      cols_to_drop = ['result_exact_search']
22      tolink_dataset = tolink_dataset.drop(*cols_to_drop)
23  # [...]
```

O resultado da UDF udf_build_exact_queries, expresso na Linha 9 do código acima seria algo parecido com o seguinte JSON, utilizado para efetivar a busca exata, exposta no Passo 2.2 da Figura 8.7.

```
1  # Exemplo de busca exata
2  { "size": "50",
3   "query": {
4    "bool": {
5     "must": [
6      {"match": {"nome":"JOAO PITA"}},
7      {"match": {"dtnasc":"19870505"}}] } } }
```

Perceba que, neste exemplo hipotético de busca exata apenas foram considerados o nome e a data de nascimento. Isso se dá porquê na busca exata, sugere-se utilizar um conjunto menor de variáveis que, conjuntamente, tem o potencial de identificar univocamente uma pessoa nas duas bases de dados envolvidas. Esta definição deve ser feita por um arquivo de configuração, onde deve-se definir também os pesos para valores correspondentes no Elasticsearch. Em seguida, produzidos os dicionários de busca exata para cada linha do Spark Dataframe, executa-se a UDF udf_find_elasticsearch_exact_best_candidate, exposta na Linha 11 do código acima. Esta UDF é responsável por enviar, de forma distribuída, as consultas a um cluster Elasticsearch. A latência associada a estas buscas dependem das configurações de *tunning* destes serviços e servidores. Uma vez finalizados, os *hits* de cada consulta são retornados na coluna result_exact_search(). Em seguida, estes candidatos são submetidos para um cálculo *ad hoc* de similaridade entre os candidatos e o registro que originou a consulta. São considerados pares "exatos"aqueles trazidos pelo Elasticserch que tenham obtido, pelo menos, 0$\dot{9}$5 de similaridade. Todos os registros que não obtiverem um candidato "exato"devem, portanto, ser submetidos a um outro passo da Etapa 2, que implementa uma busca mais flexível.

Similarmente ao passo 2.2, o passo 2.3 produz uma consulta utilizando um conjunto mais amplo de atributos utilizados na comparação par-a-par. Neste passo, deve-se considerar os parâmetros de *boosting* do Elasticseaerch, especialmente para atributos textuais nominais com diretiva de busca "*fuzzy*".

```
1  # Exemplo de busca exata
2  { "size": "50",  "query": {  bool ": {  should ": [
3
4  {'match': {'nome_a':
5      {'query': 'JOAO PITA', 'fuzziness':'AUTO', 'operator':'or', 'boost'
       :'3.0'}}},
6
7    {match": {"birthdate":"19870505"}} ] } } },
8      term ": {"sexo_a":"1"}} ] } } }
```

A consulta Elasticseach acima produzirá no máximo 50 candidatos para o registro da Base B utilizado no passo 2.2, flexibilizando erros na grafia do nome ou nos valores das demais variáveis envolvidas. Desta forma, as operações subsequentes são determinadas por uma UDF específica, considerando apenas os registros sem resultados no passo 2.2 (busca exata).

```
# Implementa  o do Passo 2.3 (busca nao - exata)
# [...]
    # building linked_from column. Non-null values on
    sim_best_candidate_exact must be filled
    # as 'exact_match', otherwise as 'non_exact_match'.
    filter_isnull = F.col('sim_best_candidate_exact').isNull()
    tolink_dataset = tolink_dataset.withColumn('linked_from', F.when(
    filter_isnull, 'non_exact_match').otherwise('exact_match'))

    # preparing filters for debug and non-debug executions
    filter_exact = F.col('linked_from') == 'exact_match'
    filter_non_exact = F.col('linked_from') == 'non_exact_match'

# [...]

    # creating, for remainder dataframe, the cols created in this
    function to ensure union
    tolink_dataset = tolink_dataset.withColumn('
    best_candidate_non_exact', F.lit(None))
    tolink_dataset = tolink_dataset.withColumn('
    sim_best_candidate_non_exact', F.lit(None))
    tolink_dataset = tolink_dataset.withColumn('
    similarity_non_exact_candidates', F.lit(None))
    tolink_dataset = tolink_dataset.withColumn('non_exact_queries', F.
    lit(None))
# [...]
```

Por fim, na função principal do CIDACS-RL, é estabelecida a ordem de execução do algorítimo, tal como ilustrado na Figura 8.7.

```
# Implementacao do Passo 2.3 (busca nao - exata)
# [...]
    tolink_dataset = cidacs_rl_exact_phase(tolink_dataset)

    tolink_dataset = cidacs_rl_non_exact_phase(tolink_dataset)

    tolink_dataset = tolink_dataset.withColumn('final_cidacs_rl_score',
     F.when(F.col('linked_from') == 'exact_match', F.col('
    sim_best_candidate_exact')).otherwise(F.col('
    sim_best_candidate_non_exact')))

    tolink_dataset = tolink_dataset.withColumn('final_cidacs_rl_id', F.
    when(F.col('linked_from') == 'exact_match', F.col('
    best_candidate_exact')).otherwise(F.col('best_candidate_non_exact')
    ))
# [...]
```

## 8.5. Iniciativas de Linkage emergentes

Nesta seção, classificaremos as iniciativas emergentes em quatro categorias: i) preservação de privacidade, ii) integração federada, iii) habilitada por modelos de linguagem ou inteligência artificial iv) alto desempenho e escalabilidade, e v) justiça algorítmica.

### 8.5.1. Métodos de PPRL emergentes

Avaliações experimentais recentes [Vidanage et al. 2023, Christen et al. 2025, Christen 2014] relatam que diversas estratégias de codificação, inclusive os Bloom Filters [Vidanage et al. 2019], nos diferentes modelos adversariais, podem ser explorados por ameaças amplamente discutidas na literatura. Uma iniciativa que visa superar a maior parte destes desafios contemporâneos foi apresentada por [Christen et al. 2024] que propõe um framework criptográfico que permite a comparação de similaridade entre pares de registros em ambientes não-confiáveis com o compartilhamento de apenas um parâmetro.

### 8.5.2. RL federado

A integração de dados federada é outra tendência emergente, impulsionada por projetos colaborativos que buscam combinar registros de saúde, educação e assistência social de diferentes países ou regiões. Plataformas como o OpenSAFELY, no Reino Unido, e o Population Data BC, no Canadá, demonstram como é possível realizar linkage em larga escala de forma segura e eficiente, mesmo em ambientes regulatórios distintos [Williamson 2020, Population Data BC 2024].

### 8.5.3. LLM e IA para RL

Nesse contexto de evolução, destaca-se também o uso de modelos de linguagem avançados (LLMs) aplicados à vinculação de registros, como exemplificado pelo LinkTransformer (LT), proposto por [Arora and Dell 2023].

O LT é uma ferramenta que oferece uma forma intuitiva de integrar Large Language Models (LLMs) nas análises de dados, promovendo eficiência e precisão na vinculação de registros. O LT preenche a lacuna entre os métodos tradicionais de correspondência de strings e os recursos avançados dos LLMs. A proposta do LT inclui uma API intuitiva, que permite aos usuários realizar tarefas de limpeza de dados de maneira simplificada, suportando múltiplos idiomas e oferecendo integração com modelos pré-treinados do Hugging Face ou OpenAI [Arora and Dell 2023]. A ferramenta visa democratizar o acesso a métodos sofisticados de vinculação de registros, tornando-os acessíveis mesmo para pesquisadores sem experiência em aprendizado profundo.

Ao superar as limitações das técnicas convencionais de correspondência de strings, o LT demonstra uma melhoria significativa na precisão, especialmente em tarefas como a vinculação de dados históricos [Arora and Dell 2023]. Com isso, o LT se destaca ao permitir que os pesquisadores realizem análises de dados complexas de maneira mais eficiente, sem necessidade de um profundo conhecimento em codificação ou estruturas avançadas de aprendizado de máquina.

### 8.5.4. RL de alto desempenho

Nos últimos anos, as pesquisas em Record Linkage têm se concentrado na área de ciência da computação, com ênfase na melhoria da eficiência computacional, seja por meio da paralelização de processos e de algoritmos avançados de busca ou na aplicação de modelos sofisticados de machine learning e estratégias para estimar erros de vinculação mesmo na ausência de dados rotulados. Apesar do avanço técnico, diversas análises apontam que nenhum método superou de forma consistente o modelo clássico de Fellegi-Sunter, especialmente em aplicações práticas de larga escala que envolvem dezenas de milhões de registros [Christen 2012]. Contudo, a demanda por modelos de alto desempenho tem motivados os pesquisadores a modelar soluções de RL baseados em sistemas computacionais especialistas ou placas gráficas [Rasch et al. 2019].

### 8.5.5. Quantificação e mitigação de vieses em RL

Os erros de RL podem provocar vieses importantes na análise *downstream* quando não são aleatórios e potencialmente favorecem um grupo demográfico em detrimento de outro [Doidge and Harron 2019]. Estes erros podem ocasionar grandes alterações nos achados, especialmente nas ferramentas mais modernas de linkage, que implementam modelos inteligentes ou mesmo baseados em heurísticas no seu *pipeline*. Faz-se, então, necessária uma reflexão que produza soluções capazes de vincular registros em bases de dados distintas que objetive a otimização da justiça algorítmica em conjunto com a acurácia ou precisão [Vatsalan et al. 2020].

## 8.6. Tendências e desafios de pesquisa

Dentre as as principais tendências apontadas por [Harron et al. 2016] e [Dong and Srivastava 2013], a quantificação e estudo dos impactos causados pelos vieses de vinculação são temas prioritários que devem ser endereçados por pesquisadores da área. Em seu trabalho, [Christen 2019] aponta, como tendências atuais no BDL, o uso de aprendizado de máquina para melhorar a acurácia do pareamento e o desenvolvimento de técnicas de PPRL para garantir a privacidade. No entanto, desafios como a integração de dados em tempo real e a heterogeneidade das fontes de dados ainda persistem, exigindo avanços metodológicos e tecnológicos. Além destes tópicos, nesta seção, discutiremos a avaliação quantitativa de vieses e as evoluções no estado-da-arte nas disciplinas de gestão de dados e inteligência artificial que podem ser incorporadas ao BDL.

Tradicionalmente, os modelos de Record Linkage (RL) eram treinados de forma específica para cada conjunto de dados, exigindo um novo processo de treinamento para cada domínio [Tang 2022, Li 2020, Brunner and Stockinger 2020]. No entanto, estudos como o de [Tang 2022] propõem o uso de modelos genéricos que, treinados sobre múltiplos domínios, conseguem generalizar para novos datasets com desempenho competitivo. Essa abordagem baseia-se na utilização de modelos de linguagem pré-treinados (PLMs) como BERT e RoBERTa, que encapsulam conhecimento semântico transferível e reduzem a necessidade de engenharia de atributos.

Com a chegada dos LLMs, como GPT-3, tornou-se viável realizar tarefas de RL com pouca ou nenhuma anotação supervisionada. [Peeters and Bizer 2023] demonstram que, por meio de prompt engineering e in-context learning, é possível obter resultados

comparáveis a modelos supervisionados, mesmo em cenários desafiadores. Esse modelo é promissor, especialmente para aplicações em que a anotação manual de dados é inviável.

O trabalho de [Arora and Dell 2023] apresenta o LinkTransformer, um pacote unificado que permite realizar RL com diversos tipos de codificadores (bi-encoders, cross-encoders e prompting com LLMs). Essa padronização do pipeline de RL representa um passo importante em direção à democratização da tecnologia, facilitando a reprodutibilidade e a comparabilidade entre métodos.

Pesquisas aplicadas como as de [Barbosa et al. 2020] e [Coeli 2021] reforçam a importância de desenvolver sistemas robustos e escaláveis. Em especial, eles destacam os desafios enfrentados ao aplicar RL em bases de dados públicas de saúde no Brasil, com dados ruidosos, ausentes ou inconsistentes. Essas iniciativas evidenciam a necessidade de soluções eficientes, voltadas à realidade dos sistemas de informação governamentais e científicos.

Apesar dos avanços notáveis nas técnicas de RL, inúmeros desafios ainda persistem e demandam atenção da comunidade científica. Um dos obstáculos mais recorrentes refere-se à escalabilidade computacional. Modelos baseados em cross-encoders, que processam pares de registros de forma conjunta para capturar interações contextuais, são computacionalmente custosos [Li 2020, Arora and Dell 2023]. Especialmente em bases de dados com milhões de combinações possíveis. Alternativas mais eficientes, como os bi-encoders, permitem maior paralelização e pré-cálculo de embeddings, mas frequentemente resultam em perda de precisão [Li 2020, Arora and Dell 2023]. Equilibrar desempenho preditivo e viabilidade computacional é um desafio da pesquisa contemporânea.

Além disso, embora *benchmarks* amplamente utilizados sejam úteis para validação inicial, eles frequentemente não refletem a complexidade dos dados reais [Coeli 2021, Tang 2022]. A maioria desses conjuntos é bem curada, com atributos bem definidos e baixo nível de ruído. Entretanto, aplicações reais, como aquelas documentadas por [Coeli 2021] no contexto do sistema de saúde brasileiro, mostram que a ausência de identificadores únicos, a presença de erros tipográficos e a heterogeneidade de esquemas são fatores comuns. A carência de benchmarks realistas compromete a avaliação generalizável dos modelos e evidencia a necessidade de conjuntos de dados mais desafiadores e representativos [Abramitzky 2021, Barbosa et al. 2020].

Outro desafio significativo é a falta de explicabilidade dos modelos modernos. À medida que técnicas baseadas em deep learning e modelos de linguagem se tornam dominantes, os critérios utilizados para classificar registros como correspondentes tornam-se menos transparentes. Sendo especialmente preocupante em domínios sensíveis, como justiça, saúde pública e análise demográfica, onde decisões automatizadas precisam ser auditáveis e justificáveis [Bhattacharya and Getoor 2007, Wilson 2011].

Além disso, a questão do viés algorítmico vem ganhando relevância. Modelos de linguagem pré-treinados são treinados em grandes corpora da web, sujeitos a estereótipos e desigualdades históricas [Gehman 2020, Peeters et al. 2023]. Quando utilizados em tarefas de RL, esses vieses podem se propagar e até se intensificar, impactando negativamente populações marginalizadas ou regiões com baixa representação nos dados de treinamento. Estratégias para mitigação de viés e avaliação de equidade nos out-

puts dos sistemas ainda são escassas e representam uma lacuna importante na literatura [Peeters and Bizer 2023, Peeters et al. 2023, Tang 2022].

Em síntese, a pesquisa em RL atravessa um momento de transformação. A incorporação de modelos de linguagem pré-treinados trouxe ganhos expressivos em termos de acurácia e adaptabilidade, ao mesmo tempo em que exigiu novas abordagens para lidar com escalabilidade, explicabilidade e robustez. O futuro do campo dependerá da capacidade da comunidade em enfrentar esses desafios de forma ética e técnica, desenvolvendo soluções que sejam não apenas eficazes, mas também confiáveis, transparentes e socialmente responsáveis.

## Referências

[Abramitzky 2021] Abramitzky, R. e. a. (2021). Automated linking of historical data. *Journal of Economic Literature*, 59(3):865–918.

[Arora and Dell 2023] Arora, R. and Dell, R. (2023). Linktransformer: A unified package for record linkage with transformer language models.

[Barbosa et al. 2020] Barbosa, G. C. G., Ali, M. S., Araujo, B., Reis, S., Sena, S., Ichihara, M. Y. T., Pescarini, J., Fiaccone, R. L., Amorim, L. D., Pita, R., Barreto, M. E., Smeeth, L., and Barreto, M. L. (2020). Cidacs-rl: a novel indexing search and scoring-based record linkage system for huge datasets with high accuracy and scalability. *BMC Medical Informatics and Decision Making*, 20(1).

[Barreto et al. 2019] Barreto, M. L., Ichihara, M. Y., Almeida, B. A., Barreto, M. E., Cabral, L., Fiaccone, R. L., Carreiro, R. P., Teles, C. A. S., Pitta, R., Penna, G. O., Barral-Netto, M., Ali, M. S., Barbosa, G., Denaxas, S., Rodrigues, L. C., and Smeeth, L. (2019). The centre for data and knowledge integration for health (cidacs): Linking health and social data in brazil. *International Journal of Population Data Science*, 4(2).

[Barreto et al. 2022] Barreto, M. L., Ichihara, M. Y., Pescarini, J. M., Ali, M. S., Borges, G. L., Fiaccone, R. L., Ribeiro-Silva, R. D. C., Teles, C. A., Almeida, D., Sena, S., Carreiro, R. P., Cabral, L., Almeida, B. A., Barbosa, G. C. G., Pita, R., Barreto, M. E., Mendes, A. A. F., Ramos, D. O., Brickley, E. B., and Smeeth, L. (2022). Cohort profile: The 100 million brazilian cohort. *International Journal of Epidemiology*, 51(2):E27–E38.

[Bhattacharya and Getoor 2007] Bhattacharya, I. and Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):5–es.

[Blake et al. 2022] Blake, H., Sharples, L., Harron, K., Van der Meulen, J., and Walker, K. (2022). Linkage of national clinical datasets without patient identifiers using probabilistic methods. *International Journal of Population Data Science*, 7(3).

[Bloom 1970] Bloom, B. H. (1970). Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426.

[Brunner and Stockinger 2020] Brunner, U. and Stockinger, K. (2020). Entity matching with transformer architectures-a step forward in data integration. In *23rd International Conference on Extending Database Technology*, pages 463–473. OpenProceedings.

[Camargo and Coeli 2011] Camargo, K. and Coeli, C. (2011). Openreclink a free and open source solution for probabilistic record linkage. In *AMERICAN JOURNAL OF EPIDEMIOLOGY*, volume 173, pages S108–S108. OXFORD UNIV PRESS INC JOURNALS DEPT, 2001 EVANS RD, CARY, NC 27513 USA.

[Camargo Jr and Coeli 2000] Camargo Jr, K. R. d. and Coeli, C. M. (2000). Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage. *Cadernos de Saúde Pública*, 16:439–447.

[Camargo Jr and Coeli 2015] Camargo Jr, K. R. d. and Coeli, C. M. (2015). Going open source: some lessons learned from the development of openreclink. *Cadernos De Saude Publica*, 31:257–263.

[Camargo Júnior and Coeli 2002] Camargo Júnior, K. and Coeli, C. (2002). Reclink ii: Guia do usuário.

[Camargo Junior and Coeli 2006] Camargo Junior, K. R. d. and Coeli, C. M. (2006). Reclink 3: nova versão do programa que implementa a técnica de associação probabilística de registros (probabilistic record linkage). *Cad. saúde colet.,(Rio J.)*, pages 399–404.

[Cha 2008] Cha, S.-H. (2008). Taxonomy of nominal type histogram distance measures. *City*, 1(2):1.

[Christen 2011] Christen, P. (2011). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE transactions on knowledge and data engineering*, 24(9):1537–1555.

[Christen 2012] Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media.

[Christen 2014] Christen, P. (2014). Privacy aspects in big data integration: Challenges and opportunities. In *Proceedings of the First International Workshop on Privacy and Secuirty of Big Data*, PSBD '14, page 1, New York, NY, USA. Association for Computing Machinery.

[Christen 2019] Christen, P. (2019). Data linkage: The big picture. *Harvard Data Science Review*, 1(2).

[Christen et al. 2020] Christen, P., Ranbaduge, T., and Schnell, R. (2020). *Linking Sensitive Data: Methods and Techniques for Practical Privacy-Preserving Information Sharing*. Springer Cham.

[Christen et al. 2025] Christen, P., Schnell, R., and Vidanage, A. (2025). Information leakage in data linkage. *arXiv preprint arXiv:2505.08596*.

[Christen et al. 2024] Christen, P., Ziyad, S., Vidanage, A., Nanayakkara, C., and Schnell, R. (2024). A single parameter method for secure privacy preserving record linkage. *International Journal of Population Data Science*, 9(5).

[Coeli et al. ] Coeli, C., Pinheiro, R., Camargo Jr, K., et al. Achievements and challenges for employing record linkage techniques in health research and evaluation in brazil. epidemiol e serviços saúde. 2015.

[Coeli 2021] Coeli, C. M. e. a. (2021). Record linkage under suboptimal conditions for data-intensive evaluation of primary care in rio de janeiro, brazil. *BMC Medical Informatics and Decision Making*, 21(1):190.

[Damerau 1964] Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

[Dillinger and Manolios 2004] Dillinger, P. C. and Manolios, P. (2004). Fast and accurate bitstate verification for spin. In Graf, S. and Mounier, L., editors, *Model Checking Software*, pages 57–75, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Doidge and Harron 2019] Doidge, J. C. and Harron, K. L. (2019). Reflections on modern methods: linkage error bias. *International journal of epidemiology*, 48(6):2050–2060.

[Dong and Srivastava 2013] Dong, X. L. and Srivastava, D. (2013). Big data integration. In *Proceedings of the International Conference on Data Engineering*, pages 1245–1248.

[dos Santos Filho 2008] dos Santos Filho, W. (2008). Algoritmo paralelo e eficiente para o problema de pareamento de dados.

[Dunn and Chief 1946] Dunn, H. L. and Chief, F. (1946). Record linkage.

[Durham et al. 2013] Durham, E. A., Kantarcioglu, M., Xue, Y., Toth, C., Kuzu, M., and Malin, B. (2013). Composite bloom filters for secure record linkage. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2956–2968.

[Dusetzina 2014] Dusetzina, S. B. e. a. (2014). *Linking Data for Health Services Research: A Framework and Instructional Guide*. Agency for Healthcare Research and Quality (US), Rockville (MD). Report No.: 14-EHC033-EF. Available at: https://pubmed.ncbi.nlm.nih.gov/25392892/. Accessed: 6 May 2025.

[Fellegi and Sunter 1969] Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

[Ferreira et al. 2005] Ferreira, R. A., Meira, W., Guedes, D., Drummond, L. M., Coutinho, B., Teodoro, G., Tavares, T., Araujo, R., and Ferreira, G. T. (2005). Anthill: A scalable run-time environment for data mining applications. In *17th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD'05)*, pages 159–166. IEEE.

[Fletcher et al. 2018] Fletcher, S., Islam, M. Z., et al. (2018). Comparing sets of patterns with the jaccard index. *Australasian Journal of Information Systems*, 22.

[Franke et al. 2021] Franke, M., Sehili, Z., Rohde, F., and Rahm, E. (2021). Evaluation of hardening techniques for privacy-preserving record linkage. In *EDBT*, pages 289–300.

[Gehman 2020] Gehman, S. e. a. (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

[Grannis et al. 2002] Grannis, S. J., Overhage, J. M., and McDonald, C. J. (2002). Analysis of identifier performance using a deterministic linkage algorithm. In *Proceedings of the AMIA Symposium*, page 305.

[Guillen et al. 2017] Guillen, L. C., Domenico, J., Camargo, K., Pinheiro, R., and Coeli, C. (2017). Match quality of a linkage strategy based on the combined use of a statistical linkage key and the levenshtein distance to link birth to death records in brazil.: Ijpds (2017) issue 1, vol 1: 036, proceedings of the ipdln conference (august 2016). *International Journal of Population Data Science*, 1(1).

[Harron et al. 2016] Harron, K., Goldstein, H., and Dibben, C. (2016). *Methodological developments in data linkage*. Wiley.

[Herzog et al. 2007] Herzog, T. N., Scheuren, F. J., Winkler, W. E., Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007). Phonetic coding systems for names. *Data Quality and Record Linkage Techniques*, pages 115–121.

[Jordão and Rosa 2012] Jordão, C. C. and Rosa, J. L. G. (2012). Metaphone-pt_br: the phonetic importance on search and correction of textual information. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 297–305. Springer.

[Junger 2006] Junger, W. L. (2006). Estimação de parâmetros em relacionamento probabilístico de bancos de dados: uma aplicação do algoritmo em para o reclink. *Cad. saúde colet.,(Rio J.)*, pages 225–232.

[Junior et al. 2018] Junior, A. A. G., Pereira, R. G., Gurgel, E. I., Cherchiglia, M., Dias, L. V., Ávila, J. D., Santos, N., Reis, A., Acurcio, F. A., and Junior, W. M. (2018). Building the national database of health centred on the individual: administrative and epidemiological record linkage-brazil, 2000-2015. *International Journal of Population Data Science*, 3(1):446.

[Karakasidis and Koloniari 2017] Karakasidis, A. and Koloniari, G. (2017). Phonetics-based parallel privacy preserving record linkage. pages 179–190.

[Kirsch and Mitzenmacher 2006] Kirsch, A. and Mitzenmacher, M. (2006). Less hashing, same performance: Building a better bloom filter. In *Proceedings of the European Symposium on Algorithms*, volume 4168, pages 456–467.

[Li 2020] Li, Y. e. a. (2020). Deep entity matching with pre-trained language models. *arXiv preprint arXiv:2004.00584*.

[Oliveira et al. 2016] Oliveira, G. P. d., Bierrenbach, A. L. d. S., Camargo Júnior, K. R. d., Coeli, C. M., and Pinheiro, R. S. (2016). Accuracy of probabilistic and deterministic record linkage: the case of tuberculosis. *Revista de saude publica*, 50:49.

[Paixao et al. 2021] Paixao, E. S., Cardim, L. L., Falcao, I. R., Ortelan, N., Silva, N. D. J., Rocha, A. D. S., Sena, S., Almeida, D., Ramos, D. O., Alves, F. J. O., Bispo, N., Ali, S., Fiaccone, R., Rodrigues, M., Smeeth, L., Brickley, E. B., Cabral, L., Teles, C., Costa, M. C. N., and Teixeira, M. G. (2021). Cohort profile: Centro de integração de dados e conhecimentos para saúde (cidacs) birth cohort. *International Journal of Epidemiology*, 50(1):37–38H.

[Peeters and Bizer 2023] Peeters, R. and Bizer, C. (2023). Using chatgpt for entity matching. In *European Conference on Advances in Databases and Information Systems*, pages 221–230. Springer Nature Switzerland.

[Peeters et al. 2023] Peeters, R., Steiner, A., and Bizer, C. (2023). Entity matching using large language models. *arXiv preprint arXiv:2310.11244*.

[Pita et al. 2018] Pita, R., Pinto, C., Sena, S., Fiaccone, R., Amorim, L., Reis, S., Barreto, M. L., Denaxas, S., and Barreto, M. E. (2018). On the accuracy and scalability of probabilistic data linkage over the brazilian 114 million cohort. *IEEE Journal of Biomedical and Health Informatics*, 22(2):346–353.

[Population Data BC 2024] Population Data BC (2024). About us.

[Ranbaduge and Schnell 2020] Ranbaduge, T. and Schnell, R. (2020). Securing bloom filters for privacy-preserving record linkage. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2185–2188.

[Rasch et al. 2019] Rasch, A., Schulze, R., Gorus, W., Hiller, J., Bartholomäus, S., and Gorlatch, S. (2019). High-performance probabilistic record linkage via multi-dimensional homomorphisms. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 526–533, New York, NY, USA. Association for Computing Machinery.

[Sanni Ali et al. 2019] Sanni Ali, M., Ichihara, M. Y., Lopes, L. C., Barbosa, G. C. G., Pita, R., Carreiro, R. P., dos Santos, D. B., Ramos, D., Bispo, N., Raynal, F., Canuto, V., de Araujo Almeida, B., Fiaccone, R. L., Barreto, M. E., Smeeth, L., and Barreto, M. L. (2019). Administrative data linkage in brazil: Potentials for health technology assessment. *Frontiers in Pharmacology*, 10(SEP).

[Santos et al. 2007] Santos, W., Teixeira, T., Machado, C., Meira Jr, W., Ferreira, R., Guedes, D., and Da Silva, A. S. (2007). A scalable parallel deduplication algorithm. In *19th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD'07)*, pages 79–86. IEEE.

[Schnell 2015] Schnell, R. (2015). Privacy-preserving record linkage. *Methodological developments in data linkage*, pages 201–225.

[Schnell et al. 2009] Schnell, R., Bachteler, T., and Reiher, J. (2009). Privacy-preserving record linkage using bloom filters. *BMC Medical Informatics and Decision Making*, 9:1–11.

[Schnell et al. 2011] Schnell, R., Bachteler, T., and Reiher, J. (2011). A novel error-tolerant anonymous linking code. Available at SSRN 3549247.

[Shukla 2015] Shukla, V. (2015). *Elasticsearch for Hadoop*. Packt Publishing Ltd.

[Tang 2022] Tang, J. e. a. (2022). Generic entity resolution models. In *NeurIPS 2022 First Table Representation Workshop*.

[Vaiwsri et al. 2018] Vaiwsri, S., Ranbaduge, T., and Christen, P. (2018). Reference values based hardening for bloom filters based privacy-preserving record linkage. In *Australasian Conference on Data Mining*, pages 189–202. Springer.

[Vatsalan et al. 2020] Vatsalan, D., Yu, J., Henecka, W., and Thorne, B. (2020). Fairness-aware privacy-preserving record linkage. In *International Workshop on Data Privacy Management*, pages 3–18. Springer.

[Vidal et al. 2006] Vidal, E., Coeli, C., Pinheiro, R., and Camargo, K. (2006). Mortality within 1 year after hip fracture surgical repair in the elderly according to postoperative period: a probabilistic record linkage study in brazil. *Osteoporosis international*, 17:1569–1576.

[Vidanage et al. 2023] Vidanage, A., Christen, P., Ranbaduge, T., and Schnell, R. (2023). A vulnerability assessment framework for privacy-preserving record linkage. *ACM Transactions on Privacy and Security*, 26(3):1–31.

[Vidanage et al. 2019] Vidanage, A., Ranbaduge, T., Christen, P., and Schnell, R. (2019). Efficient pattern mining based cryptanalysis for privacy-preserving record linkage. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1698–1701. IEEE.

[Vieira et al. 2017] Vieira, C., Coeli, C., Aguiar, F., Camargo Jr, K., Pinheiro, R., and Flores, P. (2017). Effect of short interdelivery interval between the first and second pregnancies in adolescence on low birth weight. *International Journal of Population Data Science*, 1(1):37.

[Williamson 2020] Williamson, E. J. e. a. (2020). Opensafely: Factors associated with covid-19 death in 17 million patients. *Nature*, 584(7821):430–436.

[Wilson 2011] Wilson, D. R. (2011). Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage. In *The 2011 International Joint Conference on Neural Networks*, pages 9–14. IEEE.

## PATROCINADORES

## APOIO

# SBCAS25

XXV SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO APLICADA À SAÚDE

**CELEBRANDO 25 ANOS DE HISTÓRIA**