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

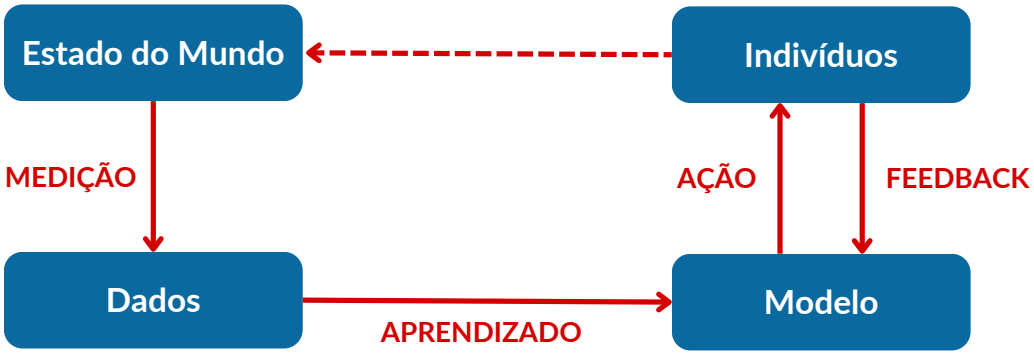

Figura 4.1. O ciclo fundamental no uso de aprendizado de máquina. Adaptado de [Barocas et al. 2023].

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

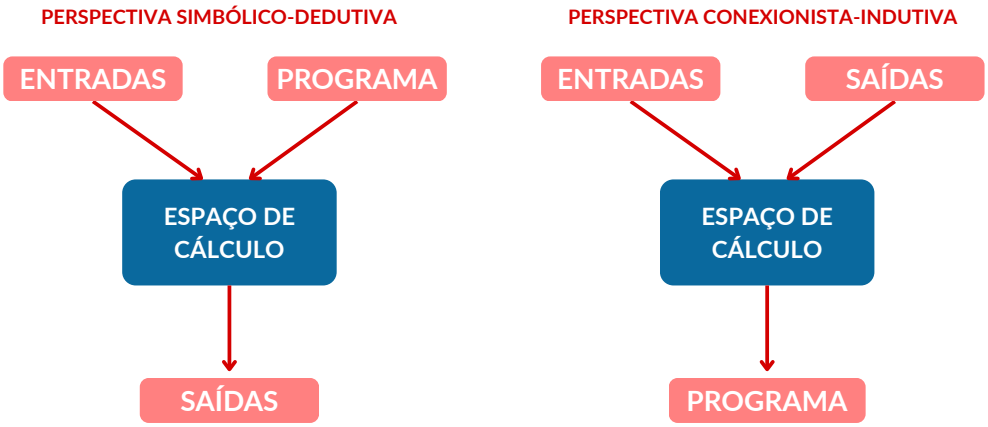

Figura 4.2. Paradigmas simbólico-dedutivo × conexionista-indutivo na inteligência artificial. Adaptado de [Silva 2022].

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

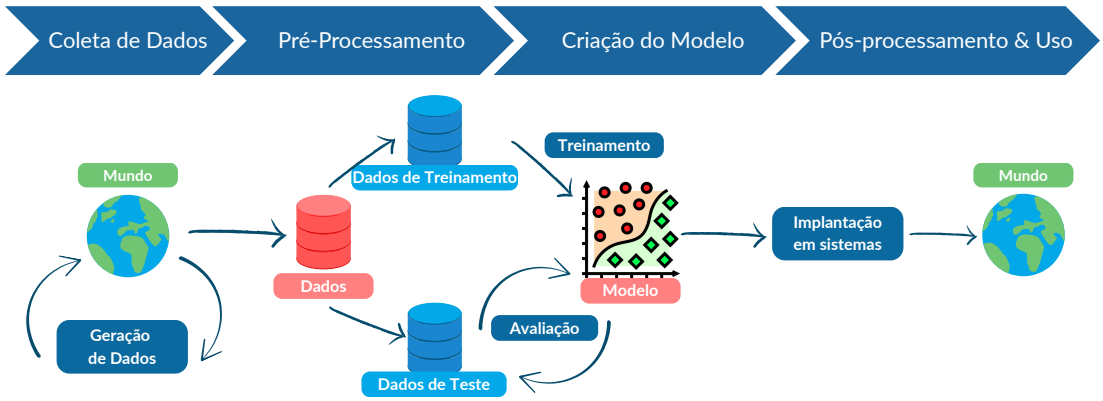

Figura 4.3. Etapas do aprendizado de máquina supervisionado. Adaptado de [Ruback et al. 2022].

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

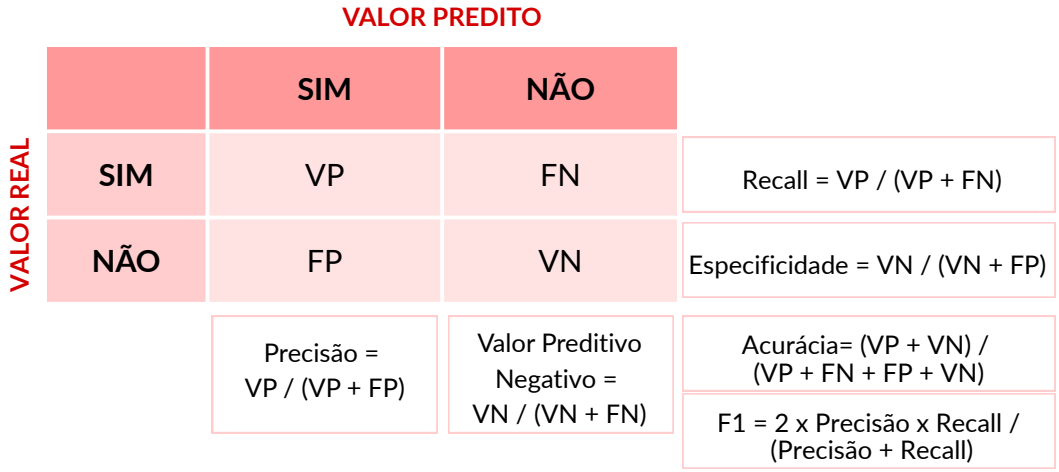

Figura 4.4. Matriz de confusão e principais métricas de avaliação para classificadores.

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 ∞, 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

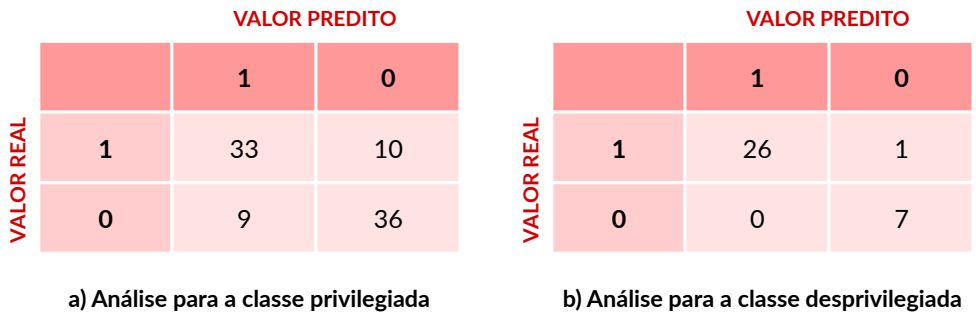

Figura 4.5. Matrizes de confusão para o problema de detecção de doença cardíaca, com análise para a classe (a) privilegiada e (b) desprivilegiada.

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, η 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 η 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 λ 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 λ 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 λ 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 λ. 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 θ são os parâmetros do modelo preditor, ϕ são os parâmetros do modelo adversário, e λ 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 θ e ϕ 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 (Δacc) 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 θ.

$$max(P(C^+|X), 1 - P(C^+|X)) < \theta \tag{24}$$

Onde $P(C^+|X$ é a probabilidade da instância X ser classificada como pertencente à classe positiva e θ 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 θ.

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.

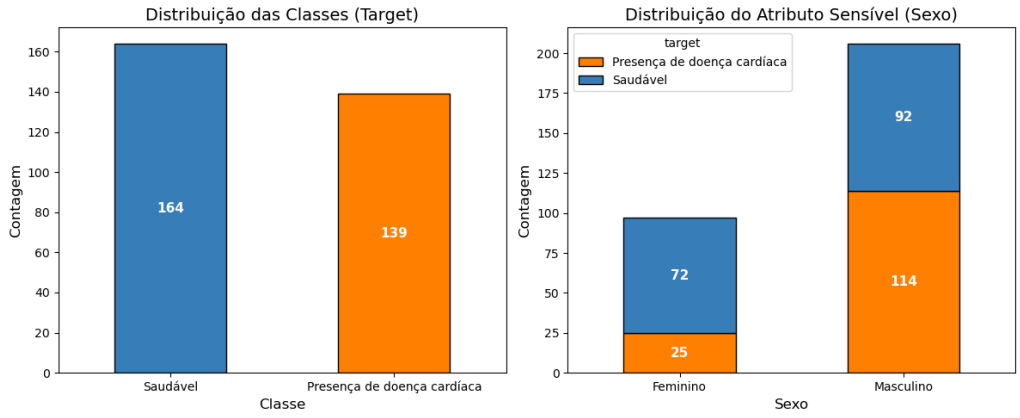

Figura 4.6. Balanceamento da variável *target* (binária) e do atributo sensível *sex* (Fonte: Os autores)

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

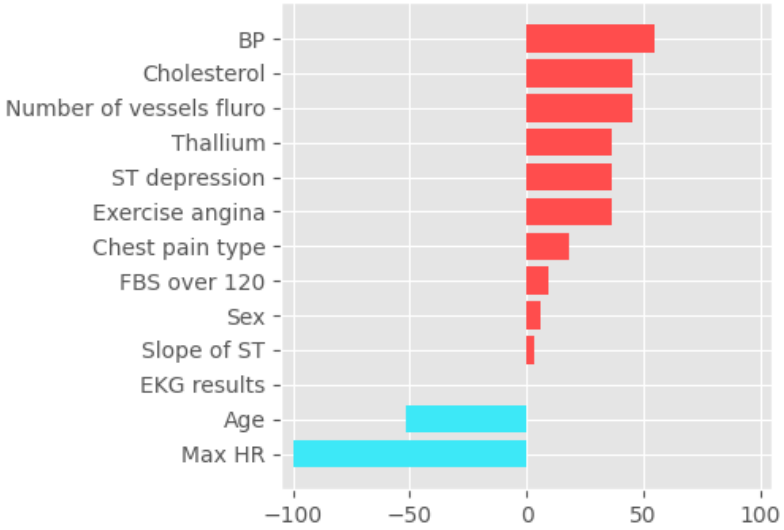

Figura 4.7. Importância dos atributos para conjunto de dados de predição de doença cardíaca utilizando uma regressão logística e a ferramenta FairML. (Fonte: Os autores)

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.

