

Capítulo

2

Inteligência Artificial para a educação

Laura O. Moraes, Carla A. D. M. Delgado

Abstract

The belief that new Artificial Intelligence (AI) technologies are important resources for development and improving quality of life is widely spread. Nowadays, much is being said about this topic, specifically regarding the threats and opportunities it brings to education. The impacts of accessible AI on learning practices and resource management in formal education are already noticeable. Furthermore, students' impressions, expectations, and experiences increasingly demand that AI be present in their learning practices. In this Chapter, we will explore some pathways for designing intelligent solutions for the educational field. Examples of AI techniques and their potential applications to well-known educational challenges will be examined, such as personalized feedback generation, the prevention of failure and disengagement, data protection and ethical use of AI in educational contexts.

Resumo

A crença de que as novas tecnologias de Inteligência Artificial (IA) são recursos importantes para o desenvolvimento e a melhoria da qualidade de vida está bastante difundida. Atualmente, muito se fala sobre isso, especificamente sobre as ameaças e oportunidades para a educação. Já são perceptíveis os impactos da IA acessível nas práticas de aprendizagem e de gestão de recursos na educação formal. Além disso, as impressões, expectativas e experiências dos estudantes demandam que a IA esteja presente em suas práticas formativas. Neste Capítulo vamos abordar alguns caminhos para o desenho de soluções inteligentes para a área da educação. Serão exploradas algumas técnicas de IA e suas potenciais aplicações em problemas reconhecidos do contexto educacional - construção de feedback personalizado, prevenção de situações de fracasso e desengajamento, proteção de dados e uso ético de IA no contexto educativo.

2.1. Introdução

A inteligência artificial (IA) tem se consolidado como uma das mais influentes e disruptivas inovações do século XXI, transformando diversos setores da sociedade, dentre eles

a educação. Porém, o uso da IA no contexto educacional não é um fenômeno recente. Desde as primeiras tentativas de automatizar o ensino por meio de sistemas tutoriais inteligentes nas décadas de 1960 e 1970, a aplicação da IA na educação teve momentos de mais ou menos interesse por parte dos atores desse contexto, mas nunca parou de evoluir. Inicialmente voltada para a instrução individual e o acompanhamento do desempenho dos estudantes, a IA tem ampliado seu escopo de aplicação na educação, principalmente em cenários onde há intermediação computacional da situação de aprendizagem ou nos quais grande parte dos dados da situação educativa já está digitalizada. Essa trajetória reflete uma combinação dos avanços no campo tecnológico e das mudanças nos paradigmas educacionais, que passaram a valorizar cada vez mais as pessoas e as situações de aprendizagem, abrindo caminho para uma educação mais equitativa, colaborativa, empoderadora e inclusiva.

Além das especificidades dos contextos tecnológicos e educacionais, outro aspecto a ser considerado quando se fala em educação é a implementação das trajetórias formativas, desde a infância até a plena atuação profissional dos indivíduos, e para além desta. No Brasil, como em vários outros países, a educação básica está amplamente a cargo do Estado e é por ele organizada. A pressão por um processo formativo eficiente em termos de tempo e de custos é também uma demanda válida, porém ortogonal ao entendimento de que a educação deve valorizar o indivíduo e sua integração livre e crítica na sociedade e no mundo. A qualidade requer investimento, e investimento demanda vontade política.

Com a ascensão de tecnologias como aprendizado de máquina, processamento de linguagem natural e, mais recentemente, as inteligências artificiais generativas, novos desafios e possibilidades emergem. A IA passa a ser vista como um dos instrumentos que pode contribuir para atender a essa ampla gama de demandas e expectativas consideradas relevantes atualmente, mas traz riscos que precisam ser entendidos e mitigados. Compreender a história da IA na educação é importante para que possamos refletir criticamente sobre seus usos atuais e futuros, suas implicações pedagógicas e seus impactos. É a partir desse entendimento que nos preparamos para esboçar os futuros projetos robustos de uso de IA na educação.

O restante dessa seção aborda a história da IA na educação. O capítulo segue abordando temáticas relevantes atualmente para viabilizar o uso de IA em contextos educativos. A Seção 2.2 apresenta a metodologia para a criação de modelos preditivos. A Seção 2.3 apresenta os fundamentos da IA Generativa e versa sobre a sua utilização no contexto educacional. Por fim, a Seção 2.4 conclui com os aspectos éticos relacionados ao uso de IA na educação.

2.1.1. Inteligência Artificial na Educação ao longo do tempo

Embora o tema “IA na educação” tenha se tornado muito popular nos últimos anos, devido à popularização do uso das inteligências artificiais generativas (IAs generativas) em meados de 2023, há iniciativas do uso de IA na educação desde os anos 50, quando o foco eram os tutores inteligentes.

A evolução das aplicações da IA na educação é o resultado do contexto de cada tempo, influenciado pela cultura vigente e suas demandas, com expectativas e propostas

calcadas nos recursos tecnológicos então disponíveis. A seguir, um panorama simplificado dos anos 50 até os dias atuais, buscando evidenciar como cada avanço da IA educacional dialoga com os desafios, expectativas e valores da política educacional vigente em cada época. Ao final do capítulo, discutimos as oportunidades e desafios contemporâneos e como esse entendimento nos ajuda a desenhar as futuras aplicações de IA na educação.

2.1.1.1. Anos 50 a 70

A pedagogia ainda seguia modelos tradicionais e autoritários, calcados na distribuição de conteúdo [Cunha 1980, Libâneo 1994]. A proposta tecnológica cunhada na época foi a “instrução programada” – uma espécie de programação da distribuição do conteúdo, baseada na teoria comportamental de Skinner [Skinner 1953], que enfatiza que o comportamento é aprendido e moldado pelas consequências de nossas ações, através, por exemplo, de reforço das ações consideradas “certas” e repressão das “erradas”. Apresentava-se ao aluno um pequeno módulo de conteúdo seguido de um teste. O próximo conteúdo era escolhido de acordo com o desempenho do aluno nesse teste, permitindo uma adequação à velocidade do aluno. A situação de aprendizagem era individual e solitária. O material era cuidadosamente estruturado para minimizar os erros, garantindo que o aluno mantivesse um alto índice de acertos — o que aumentava a motivação e a confiança.

Nos anos 60 surgem movimentos de contestação da ordem política e social que impactam a educação. Em 1970, Paulo Freire publica a obra “Pedagogia do Oprimido” [Freire 1970], influenciando toda a América Latina. Defende-se uma pedagogia dialógica, crítica e emancipadora, com foco na realidade dos estudantes e no papel da educação como transformação social. A visão tecnicista do processo educativo passa a ser criticada. No panorama tecnológico destacam-se os primeiros avanços da IA, incluindo a Conferência de Dartmouth (1956) [McCarthy et al. 1955], onde o termo “Inteligência Artificial” foi cunhado, o desenvolvimento do “Logic Theorist” [Newell et al. 1956], considerado o primeiro programa de IA, e o *chatbot* ELIZA [Weizenbaum 1966]. As expectativas tornaram-se muito altas e, por não serem correspondidas, seguiu-se o “inverno da IA” – período de desânimo e declínio no investimento e interesse na pesquisa em IA, causado por promessas não cumpridas e resultados limitados frente ao entusiasmo inicial.

2.1.1.2. Anos 80 e 90

As práticas pedagógicas são fortemente influenciadas pela Psicologia Histórico-Cultural de Vygotsky [Vygotsky 1984] e pela Epistemologia Genética de Piaget [Piaget 1973], em reação à corrente comportamental dos anos anteriores. Reforça-se o papel da escola como espaço de construção do conhecimento. Destacam-se os modelos cognitivos – representações computacionais da maneira como os seres humanos aprendem, pensam e resolvem problemas. Surgem corretores ortográficos baseados em Processamento de Linguagem Natural. No contexto tecnológico, surgem os sistemas especialistas marcando a “IA prática” (ou *practical AI*, em inglês) – termo que se refere ao uso aplicado da inteligência artificial para resolver problemas concretos do dia a dia, em ambientes como empresas, escolas, hospitais, etc. Nas escolas, aparecem os ambientes de aprendizagem – sistemas

que integram diferentes tecnologias, como gráficos, animação e simulação, criando espaços de aprendizagem interativos. As expectativas sobre a velocidade do aumento do poder computacional sobem, influenciadas pela repercussão do *Deep Blue* – supercomputador desenvolvido pela IBM, famoso por ser a primeira máquina a derrotar um campeão mundial de xadrez [Campbell et al. 2000]. Os jogos eletrônicos educativos ganham popularidade com a chegada do computador pessoal (PC) às residências. No final da década de 90 é criado o primeiro curso de Licenciatura em Computação no Brasil, na Universidade de Brasília.

2.1.1.3. Anos 2000 e 2010

O final dos anos 90 e a transição para os anos 2000 foram tempos muito influenciados pelo pensamento neoliberal. Houve o avanço de políticas educacionais de cunho gerencialista – estratégias e diretrizes adotadas na educação que se inspiram em modelos de gestão do setor privado. Um marco é a criação de sistemas de avaliação em larga escala [Luckesi 2011] como o SAEB [INEP 2023], a Prova Brasil e o teste PISA, o que justificou investimentos em mineração de dados educacionais. No contexto tecnológico, destacam-se as pesquisas na internet e o avanço dos algoritmos de busca. Pairava a crença de que, com a internet, o acesso à informação de qualidade seria gradativamente universalizado. A personalização da aprendizagem volta à cena com o conceito de “aprendizagem adaptativa” – ajuste inteligente, através de algoritmos, do conteúdo do curso de acordo com o desempenho do aluno. Aumenta o interesse em robótica educacional. No contexto tecnológico, é importante ressaltar os avanços de técnicas de reconhecimento de imagens, a popularização das redes sociais e os primeiros passos em carros autônomos e assistentes virtuais como a Siri.

2.1.1.4. Anos 2020 e expectativas para 2030

Questões de equidade e inclusão ganham força na pauta educativa. Ocorre a expansão da pedagogia inclusiva, da educação antirracista, decolonial e dos debates sobre gênero e diversidade. O apelo ao uso de tecnologias digitais nos contextos educativos cresce, já que essas tecnologias estão muito presentes em praticamente todos os contextos sociais urbanos. Destacam-se os ambientes e recursos de aprendizagem distribuídos e ubíquos, o ensino a distância e híbrido, cursos online abertos e gratuitos (MOOCs), gamificação, preditores de desempenho escolar. Emergem discussões sobre o uso ético de dados.

Na contemporaneidade, a pedagogia é marcada por uma pluralidade de correntes de pensamento que dialogam entre si e com os desafios sociais, culturais, políticos e tecnológicos da atualidade. Em particular, a corrente pedagógica das competências tem forte presença em políticas curriculares como a Base Nacional Comum Curricular (BNCC) [Ministério da Educação do Brasil 2017]. A autonomia, o autorrespeito, a ética e o protagonismo do estudante estão valorizados nas correntes pedagógicas contemporâneas, assim como as competências socioemocionais, fortalecendo a relação entre o indivíduo e o seu entorno, incluindo o contexto tecnológico, que passa a suscitar muitos questionamentos e incertezas quanto ao futuro do trabalho e da convivência.

Destacam-se a IA generativa e as discussões sobre sua interferência nas práticas didáticas e na construção de um comportamento acadêmico e profissional ético. As expectativas para o uso das novas tecnologias na educação envolvem tutores dialógicos, assistentes de avaliação de aprendizagem, gestão de padrões de aprendizagem e ecossistemas educacionais – sistemas compostos por diferentes atores, instituições, recursos e tecnologias que interagem de forma integrada para promover o aprendizado ao longo da vida. O conceito parte de uma visão mais ampla e sistêmica da educação, que vai além da escola tradicional.

2.1.2. Desafios atuais e desenho de futuras soluções

As aplicações de Inteligência Artificial (IA) na educação, ao longo do tempo, acompanham as pautas educacionais predominantes em cada período histórico. Recentemente, o advento das IAs generativas — como o ChatGPT — ampliou ainda mais as possibilidades de uso da IA na educação, permitindo a geração automática de textos, resolução de problemas complexos, correções automatizadas e a simulação de diálogos educacionais. Essas tecnologias têm potencial para apoiar o ensino personalizado em larga escala, tornando a aprendizagem mais acessível e interativa. A chegada dessa nova tecnologia na esfera escolar, porém, ameaça as práticas didáticas vigentes. Isso não é novidade no âmbito de tecnologias educacionais ou tecnologias que tiveram impactos na educação, como a calculadora ou a internet.

Compreender as tensões que envolvem a incorporação da inteligência artificial na educação é essencial para o desenvolvimento de aplicações educativas que sejam coerentes, contextualizadas e socialmente relevantes. Esse processo de concepção começa pela análise crítica das demandas educacionais contemporâneas, que se concentram, sobretudo, em três grandes eixos: a melhoria da qualidade da educação, com foco na aprendizagem significativa e no desenvolvimento de competências para o século XXI; o aumento da eficiência dos processos de ensino, visando otimizar o tempo e os recursos disponíveis; e o fortalecimento do engajamento dos estudantes, especialmente em um cenário de alta evasão, exclusão e desmotivação¹. Paralelamente, é fundamental reconhecer as oportunidades que as tecnologias baseadas em IA oferecem atualmente, como a personalização do aprendizado, permitindo trajetórias adaptadas ao ritmo e estilo de cada aluno; a promoção da dialogicidade, por meio de interações mais ricas e responsivas com os sistemas educativos; e o uso de modelos preditivos, capazes de identificar padrões de comportamento e desempenho que podem antecipar dificuldades, apoiando intervenções pedagógicas mais precisas e oportunas por parte dos professores e até dos próprios alunos.

Enquanto esses três pilares – tensões da comunidade escolar, necessidades da educação e potencialidades da IA – são o ponto de partida para o projeto de soluções educativas, é também indispensável a análise dos riscos envolvidos, como dependência excessiva de suporte tecnológico, a perpetuação de desigualdades, vieses algorítmicos e questões éticas relacionadas à privacidade e segurança dos dados dos alunos. Um plano de mitigação desses riscos deve fazer parte de qualquer projeto tecnológico educativo relevante. Além disso, a universalização do ensino e a busca por maior eficiência da educação formal permanecem como metas centrais, sobretudo em contextos de alta vulnerabilidade

¹ Vale notar que a universalização do ensino é uma demanda latente a esses 3 eixos.

social.

A trajetória da inteligência artificial na educação evidencia avanços significativos, mas também demonstra que a tecnologia, por si só, não é a protagonista das soluções para os desafios do campo educacional. Para que seu potencial se concretize de forma efetiva, é imprescindível que esteja alinhado a princípios éticos, amparado por políticas públicas inclusivas e sustentado por investimentos consistentes em infraestrutura e formação docente. Para que seus efeitos sejam significativos e duradouros, a IA deve integrar soluções educacionais amplas, concebidas de forma intersetorial, com foco no ser humano, na equidade e na promoção da justiça social.

2.2. Análise dados educacionais e sucesso escolar - Modelos estatísticos preditivos na educação

Esta seção apresenta a metodologia para a criação de modelos preditivos, trazendo uma abordagem que combina a inferência estatística, visando a interpretabilidade das variáveis envolvidas, e técnicas avançadas de aprendizado de máquina para alcançar o melhor desempenho dos modelos. Os principais recursos bibliográficos desta seção serão os livros [de Carvalho et al. 2024], [James et al. 2023] e [Provost and Fawcett 2016].

2.2.1. Contextualização

A utilização de recursos e sistemas virtuais para apoio ao aprendizado é uma realidade comum em aulas remotas e presenciais. Com isso, é possível criar grandes repositórios de dados educacionais, contendo informações das interações dos alunos com essas plataformas. Além disso, esses repositórios podem ser enriquecidos pela integração com dados provenientes de sistemas acadêmicos, como informações sociodemográficas, histórico de disciplinas cursadas, notas, frequência, e participação dos alunos em atividades acadêmicas, como bolsas e eventos, entre outros dados disponíveis. Esses sistemas produzem informações valiosas para o meio educacional, mas a quantidade de dados gerada torna inviável a análise manual. Um dos maiores desafios das instituições de ensino atualmente é transformar essa multitude de dados em conhecimento e ações que beneficiem alunos, professores e administradores.

A análise de dados educacionais é uma área de pesquisa que abrange desde a coleta dos dados, passando pela análise, até a formulação de conclusões e recomendações de ações para melhorar o aprendizado dos alunos, a distribuição de recursos institucionais, o desenvolvimento de atividades didáticas, a organização dos recursos de aprendizado em sala de aula, a promoção da colaboração entre alunos, entre outros desafios educacionais. É uma área que resulta da interseção de três grandes campos: ciência da computação, estatística e educação. A Figura 2.1 ilustra o ecossistema de criação e utilização dos modelos resultantes da análise de dados educacionais. Neste ecossistema, alunos, professores e gestores acadêmicos geram os dados a serem processados, normalmente através da utilização de ambientes virtuais educacionais (como o Moodle, Google Classroom e Sistemas Acadêmicos internos). Esses dados passam pelo processo de análise, a ser detalhado na Seção 2.2.2. Por fim, o conhecimento adquirido nesse processo é reincorporado ao sistema através de gráficos, dashboards e alertas exibidos aos usuários do ecossistema, que se traduzem em ações e intervenções.

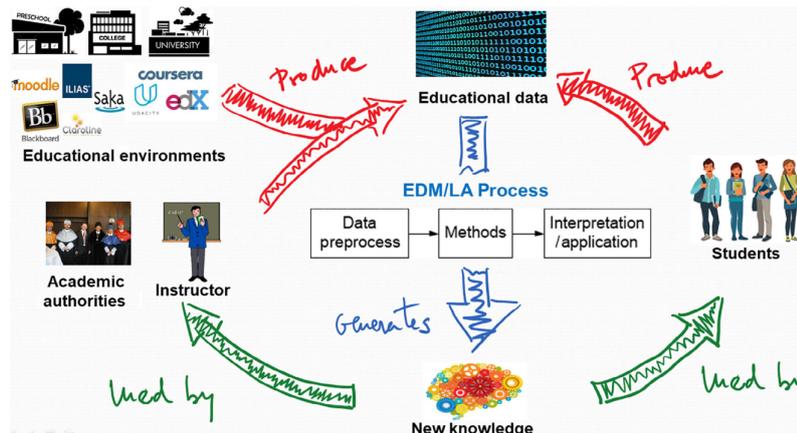


Figura 2.1. Ecossistema da análise de dados educacional. Retirado de: [Romero and Ventura 2020].

A Tabela 2.1 descreve as principais aplicações encontradas nesta área, seguidas de uma breve descrição e as ações e conhecimentos que podem vir a resultar das análises [Romero and Ventura 2020, Ahmad et al. 2024].

Existem duas grandes categorias [Ahmad et al. 2024] de sistemas de apoio à decisão educacional: 1) representacionais ou baseados em conhecimento e 2) baseados em dados. Os métodos baseados em conhecimento utilizam o conhecimento de especialistas para sugerir ações e apoiar a tomada de decisão, sendo mais comuns anteriormente, quando os dados provenientes de sistemas educacionais não estavam tão amplamente disponíveis. Atualmente, os métodos mais utilizados para apoio à decisão envolvem a análise de dados por meio de técnicas de inteligência artificial.

Dentro da área de inteligência artificial, e mais especificamente no campo do aprendizado de máquina (em que modelos são desenvolvidos a partir de dados), os tipos de aprendizado são tradicionalmente classificados em três categorias: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço [Murphy 2012, Sutton and Barto 2018]. Devido às características distintas da inteligência artificial generativa, suas aplicações são, muitas vezes, agrupadas em uma categoria à parte, ressaltando seu caráter disruptivo [Ahmad et al. 2024, Vasconcelos and Clua 2017].

1. **Aprendizado supervisionado:** este tipo de aprendizado predomina na maioria dos trabalhos publicados na área. Os algoritmos dessa categoria necessitam de pares de entrada associados a uma ou mais saídas e têm como objetivo aproximar uma função desconhecida, aprendendo o mapeamento entre as entradas e suas respectivas saídas. Uma vez que a função aproximada é obtida, torna-se possível prever as saídas para novos valores de entrada. No entanto, no contexto educacional, o acesso aos dados são limitados, sendo uma dificuldade consegui-los fora da instituição ao qual o pesquisador está vinculado. Exemplos de aplicações nessa categoria incluem a previsão de notas, a identificação de reprovações e a detecção de evasão escolar.
2. **Aprendizado não supervisionado:** neste tipo de aprendizado, não há pares correspondentes entre entrada e saída, ou seja, os dados utilizados não possuem rótulos

Tabela 2.1. Aplicações, ações e conhecimentos adquiridos em contextos educacionais

Aplicação	Ações e conhecimento adquirido
Análise de teorias educacionais ou estratégias pedagógicas: Entender como determinada teoria educacional ou estratégia pedagógica impacta o aprendizado.	Dinâmicas diferenciadas em sala de aula, reestruturação de currículo e reforma de material pedagógico.
Aprendizado colaborativo e trabalhos em grupos: Entender como o aprendizado colaborativo ou os trabalhos em grupos impactam o aprendizado.	Dinâmicas diferenciadas em sala de aula, estruturação da sala de aula, tipos de exercícios, orquestração de tecnologia em sala de aula.
Análise de currículos acadêmicos: Analisar a estrutura dos programas e os dados curriculares administrativos, verificando a dependência entre as ementas e conhecimentos necessários para melhor aproveitamento das disciplinas.	Aprimoramento do currículo, da qualidade dos cursos e na criação de um panorama do que é ensinado na área.
Sistema de alerta de abandono ou reprovação precoce: Prever o desempenho dos alunos ou alunos em risco de evasão.	Intervenções precoces que permitam a permanência e/ou sucesso do aluno, fatores e comportamentos de risco, feedback personalizado.
Análise de sentimentos e emoções: Entender os sentimentos dos alunos em relação às aulas ou algum recurso didático.	Reforma de material pedagógico, dinâmicas diferenciadas em sala de aula, itens de aprendizado personalizados, intervenção psicológica.
Análise de dados de gamificação: Aplicar técnicas de mineração de dados e visualização às interações dos jogadores em jogos sérios.	Avaliar o conhecimento adquirido dos alunos através de métodos de avaliação não convencionais, identificar dificuldades de aprendizado.
Análise de escrita: Aplicar ferramentas de mineração de texto e análise a dados textuais provenientes de fóruns, chats, redes sociais, avaliações, redações, etc.	Entender o sentimento e dificuldades dos alunos, feedback.
Análise de dados de orquestração: Estudar o efeito da adoção de tecnologias em sala de aula.	Maneiras de inserir a tecnologia em sala de aula de modo a maximizar o aprendizado dos alunos e otimização dos recursos.
Sistemas de recomendação: Recomendação de material de estudo ou material similar.	Estudo personalizado a partir de materiais adaptados à necessidade do aluno, pareamento de alunos que podem se ajudar.
Transfer learning (aprendizado por transferência): Desenvolver modelos que possam ser transferíveis ou aplicados a outros cursos, instituições, etc.	Desenvolver modelos mais robustos a variações, generalizações de teoria de aprendizagem e práticas pedagógicas e permitir o compartilhamento do conhecimento adquirido entre instituições, mitigando problemas de poucos dados ou dados de somente uma instituição, comumente encontrados no contexto educacional.

indicando os resultados esperados. O objetivo principal é identificar padrões em um conjunto de dados, normalmente agrupando observações semelhantes (clustering) ou extraindo os atributos (ou combinações de atributos) mais relevantes para representar esses padrões (redução de dimensionalidade). Exemplos de aplicações nessa categoria incluem a descoberta de perfis de alunos e a recomendação de materiais educacionais.

3. **Aprendizado por reforço:** frequentemente confundido com o aprendizado supervisionado, o aprendizado por reforço se diferencia por não utilizar pares de entrada e saída rotulados. Nesse paradigma, um agente autônomo aprende a tomar decisões ao interagir com um ambiente, buscando maximizar uma função de recompensa ao longo do tempo. Exemplos de aplicações nessa categoria incluem sistemas de aprendizado adaptativo e modelagem de estilos de aprendizagem.
4. **Inteligência artificial generativa:** A IA generativa é um campo recente da inteligência artificial que se dedica ao desenvolvimento de modelos capazes de criar novos conteúdos, como textos, imagens e até músicas a partir de padrões aprendidos nos dados de treinamento. Utiliza arquiteturas de redes neurais avançadas para criação dos modelos. Exemplos de aplicações nessa categoria incluem *feedbacks* personalizados, revisões de texto e *chatbots* para elucidar dúvidas.

As próximas subseções abordam conceitos, definições e aplicações do aprendizado supervisionado no contexto educacional. Os tópicos de aprendizado não supervisionado e aprendizado por reforço não serão discutidos. Devido às suas características inovadoras e ao seu impacto nas aplicações, a inteligência artificial generativa será tratada separadamente em uma seção específica (Seção 2.3).

2.2.2. Metodologia para análise de dados

Nessa seção apresentaremos a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*, do inglês, Padrão de Process Inter-Indústrias em Mineração de Dados) [Provost and Fawcett 2016] para desenvolvimento de projetos em Ciência de Dados. A Figura 2.2 detalha cada etapa da metodologia.

1. **Entendimento do problema:** identificação, definição e motivação do problema a ser resolvido. Responde às perguntas:
 - (a) Como uma solução de mineração de dados resolverá este problema?
 - (b) O que se visa alcançar ao realizar essa análise?
 - (c) Quais métricas conseguirão medir uma solução satisfatória para o problema?

Nesta etapa, pesquisa-se na literatura existente, extraindo dos artigos relacionados os modelos existentes e como seu impacto foi medido. É preciso pesquisar métricas quantitativas e é preciso pesquisar também as métricas qualitativas, principalmente o impacto causado pelos modelos em situações autênticas (por exemplo, quando utilizado durante um curso).

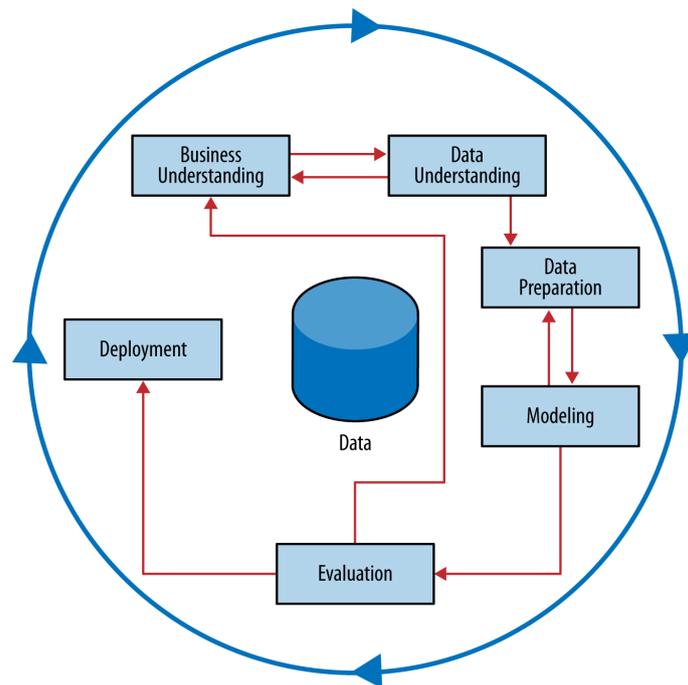


Figura 2.2. Etapas da metodologia CRISP-DM para projetos de Ciência de Dados.
Retirado de [Provost and Fawcett 2016].

2. **Entendimento dos dados:** identificação dos dados necessários para resolução do problema proposto. Essa etapa consiste em avaliar o banco de dados, entender o que existe nele, suas limitações e indicar outros dados de interesse que complementem a análise. Responde às perguntas:
- (a) Há dados faltantes? Se sim, é preciso remover os dados faltantes, utilizar alguma técnica de imputação ou optar por deixá-los no banco de dados, utilizando um modelo que saiba lidar com esse tipo de dado.
 - (b) Há dados duplicados? Se sim, é preciso removê-los caso suspeite que os dados sejam errôneos.
 - (c) Os dados se encontram dentro dos limites esperados? É possível verificar para cada atributo numérico os valores mínimos e máximos, verificando se se encontram dentro dos limites conceituais esperados. Por exemplo, se uma variável é a nota, levanta-se uma suspeita se essa nota for maior que 10 vindo de instituto educacional brasileiro.
 - (d) Qual a distribuição dos dados? Para verificar essa pergunta, utiliza-se um histograma para cada variável de interesse. Por exemplo, é possível entender se há uma desigualdade de gênero em determinado curso.
 - (e) Há *outliers*? Outliers são chamados os casos anômalos ou atípicos que podem ser acontecimentos reais no banco de dados ou resultado de erros de medição, entrada ou processamento. Há a possibilidade de remover essas observações uma vez que podem não representar padrão nenhum, distorcer estatísticas e mascarar tendências gerais. No entanto, é preciso ter cuidado ao lidar com esses casos atípicos: essas observações podem representar o comportamento

legítimo de uma minoria (e por serem minoria não constar um volume de dados significativo no banco a ponto de ser considerado um padrão) ou serem relevantes se a análise se refere a encontrar comportamentos fora do comum. Portanto, a decisão de remover, transformar ou analisar separadamente os *outliers* deve ser baseada no entendimento do contexto dos dados e nos objetivos da análise, para evitar a perda de informações relevantes ou a introdução de vieses nos resultados.

- (f) Há dependência entre os atributos? Essa é uma das análises mais importantes no entendimento dos dados. É importante verificar as variáveis com maior poder de predição da variável-alvo (variável que se deseja prever) e quais variáveis estão relacionadas entre si. Isso influencia a escolha dos modelos e a interpretação dos resultados. Por exemplo, se há duas variáveis muito correlacionadas e deseja-se prever uma terceira, um modelo de regressão pode apresentar somente uma das variáveis como relevantes para a previsão. No entanto, no momento de análise dos resultados, é preciso lembrar que ambas as variáveis estão correlacionadas e, portanto, é preciso entender mais a fundo o que está influenciando a previsão. Essa análise é feita com um gráfico de dispersão e métricas de correlação.
3. **Preparação dos dados:** nesta etapa os dados são avaliados através da análise exploratória e formatados para serem utilizados nos modelos. Aqui é verificada a completude dos dados, possíveis inconsistências, além de realizar algumas transformações como limpeza, imputação, normalização, seleção e agrupamento de atributos, redução de dimensionalidade, entre outras. Um exemplo prático será disponibilizado na Seção 2.2.4.
4. **Modelagem:** nesta etapa são criados e aplicados os modelos preditivos que auxiliam as partes interessadas a alcançarem os objetivos determinados na etapa 1 (entendimento do problema). Por isso, é importante a escolha de modelos alinhados com a pergunta que se deseja responder. Como diferentes modelos possuem diferentes hiperparâmetros que influenciam diretamente nos resultados, é necessário identificar os hiperparâmetros existentes em cada modelo e avaliar como sua alteração afeta o resultado. Responde às perguntas:
- (a) Quais modelos testar? A escolha dos métodos e algoritmos deve ser feita de acordo com a pergunta que se deseja responder. Por exemplo, se o objetivo é prever um valor contínuo, como uma nota, é necessário selecionar um modelo adequado para tarefas de regressão.
 - (b) Como testar os modelos? Para avaliar a eficiência e mitigar riscos, a preferência inicial é pela construção de modelos preditivos de forma *offline*, ou seja, utilizando dados históricos, sem interação em tempo real com alunos ou outros participantes. O método *offline* não envolve intervenções, ou seja, não altera os resultados já ocorridos, servindo apenas para simular e avaliar o desempenho dos modelos. Somente após a seleção de um modelo *offline* adequado (definida na etapa de avaliação), ele passa a ser utilizado de forma *online* (em tempo real) na fase de implantação.

5. **Avaliação:** após a criação dos modelos, é preciso avaliar se os resultados estão conforme o esperado. Neste ponto, as métricas e seus valores aceitáveis já foram definidos (etapa 1, entendimento do problema). Portanto, nesta etapa, avaliamos se os modelos definidos na etapa anterior cumprem os requisitos e onde é possível melhorá-los. É também possível reavaliar as métricas definidas anteriormente. Responde às perguntas:

- (a) Quais métricas são apropriadas para o problema?
- (b) Como simular um conjunto de dados não vistos para avaliar a generalização do modelo?

6. **Implantação:** nesta etapa final, o modelo é colocado em produção para ser utilizado em tempo real, coletando novos dados e validando o modelo criado.

2.2.3. Exemplo ilustrativo de aplicação da metodologia: Prevenção da evasão escolar

1. **Entendimento do problema:** Uma escola pública busca reduzir os índices de evasão escolar no ensino médio. A gestão escolar suspeita que alguns fatores registrados em sua base de dados podem antecipar comportamentos de risco. O objetivo da análise é prever quais alunos têm maior probabilidade de abandonar os estudos ainda no primeiro semestre, de modo a permitir intervenções pedagógicas antecipadas. *Métricas definidas:* acurácia geral, *recall* para a classe de risco (evadidos) e precisão.

2. **Entendimento dos dados:** A base de dados da escola contém informações históricas dos estudantes: frequência mensal (em %), notas por disciplina, histórico de reprovações, dados socioeconômicos (renda familiar, ocupação dos responsáveis), participação em programas sociais (ex: Bolsa Família), abandonou ou não a escola.

Análises exploratórias indicaram:

- Ausência de dados em cerca de 10% das entradas na variável “renda familiar”.
- Notas com outliers (alunos com notas consistentemente muito baixas).
- Correlação moderada entre alta frequência e permanência na escola (scatter plot).
- Distribuição das faltas com concentração no segundo semestre (histograma).

3. **Preparação dos dados:**

Foram realizadas as seguintes etapas de preparação: remoção de registros duplicados, imputação da “renda familiar” com a mediana segmentada por bairro, normalização das variáveis contínuas (frequência, notas) e codificação de variáveis categóricas (ocupação dos responsáveis).

4. **Modelagem:** Modelos preditivos testados: regressão logística (*baseline*), Random Forest e XGBoost.

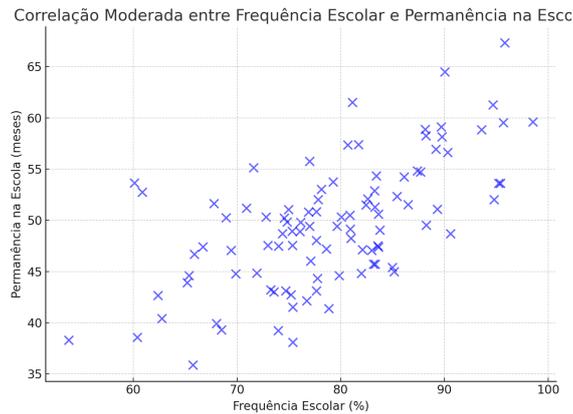


Figura 2.3. Exemplo fictício de correlação moderada entre alta frequência e permanência na escola.

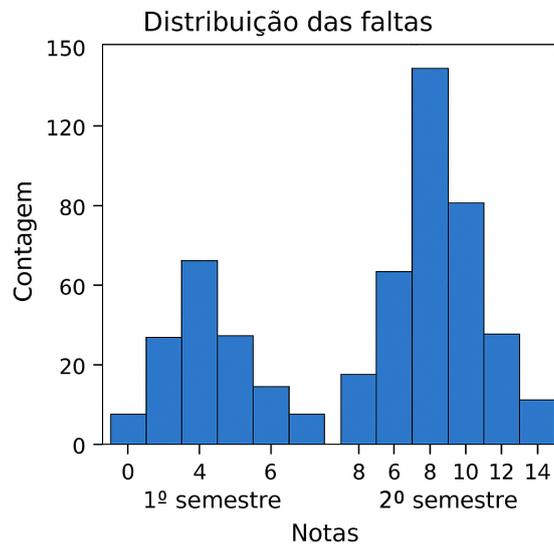


Figura 2.4. Distribuição das faltas com concentração no segundo semestre.

Hiperparâmetros como a regularização da regressão logística, profundidade de árvore e número de estimadores foram ajustados por busca em grade. A avaliação dos modelos foi feita por validação cruzada com *k-fold* ($k=5$).

5. **Avaliação:** Os resultados foram:

- Random Forest: recall de 87% para evasão, acurácia de 83%.
- Regressão logística: menor recall (74%), porém fácil interpretabilidade.
- XGBoost: melhor acurácia (85%), mas maior tempo de execução.

Com base nos objetivos da escola, o modelo Random Forest foi selecionado.

6. **Implantação:** A escola optou por adotar o modelo como ferramenta de alerta precoce, incorporando-o ao planejamento pedagógico semestral.

2.2.4. Análise exploratória de dados

Esta seção apresenta os conceitos de estatística básica necessários para o entendimento dos modelos preditivos e conduz uma análise exploratória, mostrando ao leitor as principais análises a serem realizadas e os pontos de atenção ao realizá-las.

Terminologia Os dados coletados em sistemas virtuais de aprendizado, acadêmicos ou entrevistas são, em sua maioria, organizados em formato de tabela. Essa organização pode ocorrer tanto em bancos de dados relacionais quanto em arquivos de texto estruturados, como aqueles com extensão .CSV (*Comma-Separated Values*).

Nessa estrutura tabular, cada linha representa uma observação individual, que pode ser um caso, um evento ou qualquer outra unidade de análise relevante para o estudo. Por exemplo, em uma base de dados escolar, cada linha pode corresponder a um estudante específico. As colunas da tabela, por sua vez, representam variáveis ou atributos que descrevem diferentes características das observações. No exemplo da base escolar, as colunas poderiam conter informações como frequência, nota em matemática, número de reprovações, renda familiar, entre outros. Assim, a tabela pode ser vista como uma matriz de dados, onde as linhas são chamadas de vetores. Cada vetor contém os valores dos atributos para uma única observação.

Em modelos preditivos, geralmente existe um atributo principal chamado de atributo-alvo (ou variável-alvo), embora em alguns casos possa haver mais de um. O atributo-alvo é aquele que desejamos prever utilizando as demais informações disponíveis no conjunto de dados. Por exemplo, considere o modelo que busca prever o abandono escolar, exemplificado na Seção 2.2.3. Nesse caso, o atributo-alvo é “Abandono”, e o objetivo é estimar se um aluno irá ou não abandonar a escola com base em outras variáveis, como suas notas e o número de reprovações.

Tabela 2.2. Exemplo resumido de dados estudantis

Aluno	Freq. (%)	Nota Mat.	Reprov.	Renda (R\$)	Bolsa Fam.	Abandono
Ana	95	8,5	0	1800	Sim	Não
Bruno	87	6,0	1	2500	Não	Não
Carla	92	7,0	0	1200	Sim	Não
Diego	78	5,5	2	900	Sim	Sim
Eduarda	99	9,5	0	3200	Não	Não

Medidas descritivas:

Medidas de tendência/posição central - média, mediana e moda: representam o meio do conjunto de dados. Como se fosse a escolha de um representante para aquele conjunto.

Medidas de dispersão - desvio padrão, coeficiente de variação, quartis, percentis: indica o quanto os valores variam entre eles. O quão dispersos eles estão ao redor de uma medida central?

Outras sumarizações: mínimo e máximo.

Começaremos essa explicação apresentando um problema. Anotamos a idade de um grupo de pessoas e obtemos: 4, 1 e 10. Qual valor melhor representa esse grupo? O que a gente gostaria é que o erro/diferença entre valor representante e os dados fosse o menor possível, correto? Como calcular? Definindo o representante como \bar{x} , podemos utilizar dois tipos de erros para calcular essa diferença. Qual o melhor representante?

$$Erro = |4 - \bar{x}| + |1 - \bar{x}| + |10 - \bar{x}| \quad \text{ou} \quad Erro = (4 - \bar{x})^2 + (1 - \bar{x})^2 + (10 - \bar{x})^2$$

Medidas clássicas: média, desvio padrão e variação Se escolhermos minimizar o erro quadrático, o melhor representante \bar{x} é a média aritmética, definida por:

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N}$$

Considere o exemplo da Tabela 2.3, retirado de [Barbetta et al. 2010], em que temos as notas finais relativas a alunos de três turmas. Nos três casos, a média 6,00 é o valor que minimiza o somatório do erro quadrático entre o valor do representante e os valores de cada nota. No entanto, apesar de apresentarem a mesma média, a distribuição das notas em cada turma é bastante diferente. Ao comparar as turmas A e B, por exemplo, observa-se que a turma B possui notas mais extremas, ou seja, há alunos com notas muito baixas e outros com notas muito altas, distantes da média. Isso indica que a turma B é mais heterogênea, com maior dispersão dos dados em relação à média, enquanto a turma A apresenta uma distribuição mais concentrada em torno desse valor.

Já na turma C, a situação é diferente: a maioria dos alunos obteve notas próximas a 7, mas há um único aluno com uma nota extremamente baixa (zero). Esse valor discrepante, chamado de *outlier*, exerce grande influência sobre a média, puxando-a para baixo. Como resultado, a média de 6,00 não representa adequadamente a maioria dos alunos dessa turma, pois a maior parte das notas está acima desse valor. Dessa forma, podemos concluir que a presença de valores muito discrepantes em um conjunto de dados pode distorcer a média, tornando-a um representante inadequado da tendência central dos dados. Por isso, é importante analisar não apenas a média, mas também a dispersão e a existência de possíveis *outliers* ao interpretar conjuntos de dados.

Tabela 2.3. Notas e média para três turmas hipotéticas. Retirado de: [Barbetta et al. 2010]

Turma	Notas dos alunos								Média da turma
A	4	5	5	6	6	7	7	8	6,00
B	1	2	4	6	6	9	10	10	6,00
C	0	6	7	7	8	8			6,00

Além da média, é importante apresentar uma medida complementar que indique o grau de dispersão dos dados em relação a esse representante. Essa medida, chamada

de desvio padrão, mostra o quanto os dados variam em torno da média e, portanto, o quão representativa a média é para o conjunto analisado. O desvio padrão é calculado em alguns passos: primeiro, determina-se a média aritmética dos dados. Em seguida, calcula-se a diferença entre cada valor e a média, elevando cada uma dessas diferenças ao quadrado. Depois, soma-se todos esses valores quadrados e divide-se o resultado pelo número de dados. Por fim, tira-se a raiz quadrada desse resultado para se obter o desvio padrão. Em resumo, o desvio padrão é definido por:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}}$$

Ao calcularmos a dispersão dos dados em relação à média, utilizamos o somatório dos erros ao quadrado, ou seja, elevamos ao quadrado a diferença entre cada valor e a média. O resultado desse cálculo é chamado de variância. No entanto, como estamos somando erros quadráticos, a unidade da variância é o quadrado da unidade original dos dados. Por exemplo, se os dados estão em metros, a variância estará em metros ao quadrado. Para que a medida de dispersão volte à mesma unidade dos dados originais e seja mais facilmente interpretada, definimos o desvio padrão como a raiz quadrada da variância. Assim, o desvio padrão expressa a variabilidade dos dados na mesma unidade de medida utilizada inicialmente, facilitando a comparação e a compreensão dos resultados.

Medidas baseadas na ordenação dos dados: mediana, quartis e percentis Se escolhermos minimizar o erro com valor absoluto, o melhor representante \bar{x} é a mediana. Para calculá-la, é necessário, primeiramente, organizar os dados em ordem crescente (ou decrescente). Em seguida, identifica-se o ponto central dessa sequência, de modo que o conjunto seja dividido em duas partes com a mesma quantidade de elementos em cada lado. Se o número de dados for ímpar, a mediana será o valor que ocupa exatamente a posição central da lista ordenada. Nesse caso, haverá o mesmo número de elementos à esquerda e à direita desse valor. Por outro lado, se o número de dados for par, não existe um único valor central. Nessa situação, a mediana é calculada fazendo-se a média aritmética dos dois valores que ocupam as posições centrais da sequência ordenada. Assim, garante-se que o conjunto de dados seja dividido em duas partes iguais, cada uma contendo metade dos elementos. Ainda utilizando como exemplo a turma A da Tabela 2.3, por essa turma ter um número par de elementos (oito), a mediana será a média aritmética dos dois elementos centrais (no caso, 6). Calculando para as três turmas, o valor entre a média aritmética e a mediana serão diferentes somente na turma C, onde a mediana é 7.

De forma semelhante ao cálculo da mediana, os quartis e os percentis também servem para dividir o conjunto de dados em partes iguais, facilitando a análise da distribuição dos valores. Os quartis segmentam os dados em quatro partes iguais, cada uma correspondendo a 25% do total. O primeiro quartil (Q1) indica o valor abaixo do qual estão 25% dos dados, o segundo quartil (Q2) corresponde à mediana (50%), e o terceiro quartil (Q3) marca o ponto abaixo do qual estão 75% dos dados. Já os percentis dividem o conjunto de dados em cem partes iguais, cada uma representando 1% da amostra. Por exemplo, o 30º percentil indica o valor abaixo do qual estão 30% dos dados. Para avaliar o grau de dispersão dos dados em relação à mediana, costuma-se calcular os quartis. Dessa

forma, é possível visualizar como os dados estão distribuídos em intervalos de 25%, identificando concentrações, dispersões e possíveis assimetrias. Essas medidas de posição são frequentemente representadas de forma gráfica por meio do boxplot (ou diagrama de caixa), que mostra, de maneira visual, a distribuição dos dados, destacando os quartis, a mediana e possíveis valores discrepantes (*outliers*).

2.2.4.1. Entendimento e preparação dos dados

De acordo com [Ahmad et al. 2024], as principais atividades desenvolvidas na área de inteligência artificial aplicada à educação envolvem a avaliação do desempenho dos estudantes, a previsão de retenção e a identificação de possíveis casos de evasão escolar. Nesse contexto, os modelos de previsão utilizam, em sua maioria, dados relacionados a fatores socioeconômicos (como gênero, idade e situação profissional), desempenho acadêmico prévio (por exemplo, média geral das notas e resultados em exames de admissão), atividades na internet (como o uso de ambientes virtuais de aprendizagem - AVA) e fatores psicológicos ou de saúde (incluindo qualidade do sono e vida social). Além desses, alguns estudos também consideram avaliações específicas em disciplinas, hábitos de estudo e indicadores de saúde mental dos estudantes [Rodrigues et al. 2025].

Como explicado no item 2 da Metodologia (Seção 2.2.2), a análise de dependência dos atributos entre si e dos atributos com a variável-alvo é uma das mais importantes, pois já pode fornecer *insights* sobre os dados e guiar as próximas etapas da metodologia. Nesta etapa, o método mais utilizado é o índice de correlação com o apoio visual de um gráfico de dispersão.

Correlação Dizemos que duas variáveis estão positivamente correlacionadas quando o aumento em uma delas tende a ser acompanhado pelo aumento na outra. Por outro lado, a correlação negativa ocorre quando o aumento em uma variável está associado à diminuição da outra. É importante ressaltar que o conceito de correlação refere-se apenas à existência de uma relação numérica entre duas variáveis, sem implicar, necessariamente, em uma relação de causa e efeito ou na existência de alguma estrutura subjacente [Barbetta et al. 2010]. Cabe ao pesquisador analisar os fundamentos teóricos que poderiam explicar a correlação observada e investigá-los por meio de hipóteses. Assim, a correlação funciona como uma ferramenta auxiliar na formulação e verificação de hipóteses baseadas em fundamentos teóricos, sendo normalmente utilizada em estudos exploratórios.

Existem diferentes maneiras de calcular a correlação, sendo as três mais utilizadas as de Pearson, Spearman e Kendall [Okoye and Hosseini 2024]. A correlação de Pearson caracteriza-se principalmente por medir a relação linear entre duas variáveis. Isso significa que, se um aumento x em uma variável corresponde a um aumento y em outra, espera-se que esses incrementos sejam proporcionais e constantes. Por exemplo, se a primeira variável aumenta em $2x$, a segunda deveria aumentar em $2y$. A correlação de Pearson pode ser calculada pela equação 1 e sua relação linear pode ser visualizada em um gráfico de dispersão, onde os pontos tendem a se alinhar formando uma reta. Detalhes sobre a construção da equação 1 podem ser encontrados em [Barbetta et al. 2010].

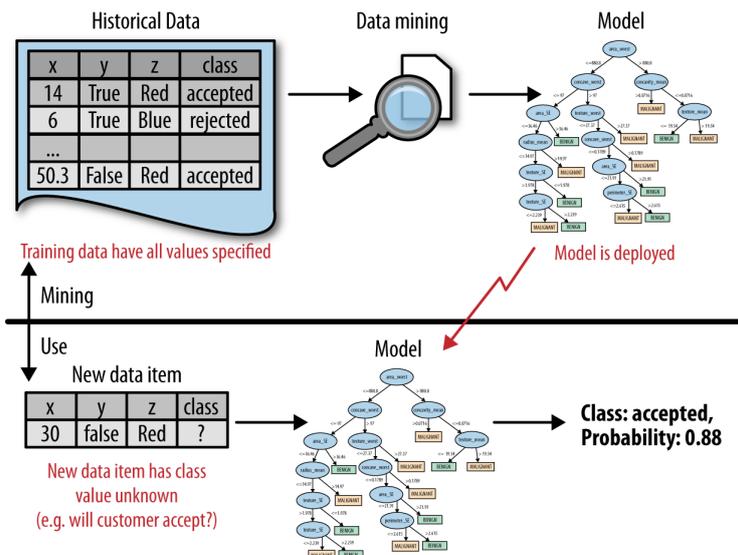


Figura 2.5. Comparação entre o processo de criação de um modelo preditivo e utilização de um modelo implantado. Retirado de: [Provost and Fawcett 2016]

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

Durante o minicurso haverá uma parte prática em que a análise exploratória será conduzida, compartilhada com os participantes através do repositório do GitHub: <https://github.com/MachineTeachingEdu/JAI2025-IA-Educacao.git>.

2.2.5. Aprendizagem estatística e modelagem

Após entendimento dos dados nas seções anteriores, essa seção introduz o leitor à criação de modelos preditivos. Para isso, são apresentados e definidos os conceitos principais para a sua criação e utilização correta.

2.2.5.1. Conjunto de treino, teste, sobreajuste e generalização

A criação de um modelo preditivo normalmente visa sua utilização no futuro, conforme novos dados são obtidos. Por exemplo, ao criar um modelo preditivo para prever a nota ou a evasão de alunos com dados históricos, visa-se utilizar esse modelo com novos alunos, que ainda estão cursando e, portanto, não temos o resultado final ainda. O processo de criação e utilização de um modelo preditivo está ilustrado na Figura 2.5. Na parte superior da figura, dados históricos são explorados para produzir o modelo. Note que os dados históricos possuem um atributo-alvo chamado “class” que será utilizado para criar o modelo que melhor se adapta aos dados (menor erro fora da amostra). Na parte inferior da figura, o modelo é aplicado um dado novo que ainda não possui o atributo-alvo e, portanto, gostaríamos de prevê-lo. O modelo prevê o valor da classe e a sua probabilidade de ocorrência.

No entanto, se criarmos um modelo e verificarmos seu funcionamento utilizando

somente um conjunto de dados, muito provavelmente criaremos um modelo que não tenha aprendido um padrão sobre o comportamento dos alunos e sim aprendido sobre aqueles dados. Uma analogia comum é pensar em um aluno estudando para uma prova nova a partir de uma prova antiga. O aluno pode aprender o conteúdo ou somente aprender as respostas da prova. No primeiro caso, em uma prova nova, espera-se que ele acerte as questões. No segundo, como não terá aprendido o conteúdo, ele provavelmente não saberá responder caso a pergunta mude. O mesmo acontece nos modelos preditivos de aprendizado de máquina. O modelo pode aprender o padrão (conteúdo) ou aprender somente o que acontece naqueles dados.

Idealmente, usaríamos um conjunto de dados só para o treinamento e criação do modelo e outro diferente para verificar seu funcionamento, similar a utilização de uma prova antiga e uma prova nova para estudo. No entanto, frequentemente só temos um conjunto de dados, como podemos então maximizar nossas chances de criarmos um modelo que aprenda o padrão e não somente o que acontece nos dados? Separaremos o conjunto único que temos em dois: o primeiro fará o papel da prova antiga, sendo utilizado na treinamento e ajuste do modelo. Chamaremos esse de conjunto de treino. O segundo, por mais que tenha vindo da mesma prova, será escondido e não será utilizado em momento algum do processo, somente no final, onde seu propósito será somente para avaliar se o modelo é capaz de aprender o padrão, simulando um prova nova. Este será chamado de conjunto de teste. A Tabela 2.4 apresenta as definições dos conceitos fundamentais em aprendizado de máquina.

Mas como saberemos que o modelo criado não generaliza e possui sobreajuste? Existem técnicas para identificar e evitar o sobreajuste. A Tabela 2.5 resume de maneira simplificada os principais sinais de alerta para identificação do sobreajuste e do subajuste a partir dos resultados de erro dentro e fora da amostra e da característica do modelo. Sobre as técnicas para evitar, elas, em sua maioria, dependem do tipo de modelo escolhido.

No contexto educacional, frequentemente lidamos com dados que possuem uma dependência temporal. Nesses casos, a divisão entre treino e teste deve respeitar a ordem cronológica dos eventos, evitando o vazamento de informações do “futuro” para o “passado”. Por exemplo, suponha uma coleta de dados realizada entre 2020 e 2022 ao longo de cinco semestres. O modelo pode ser treinado e testado utilizando cinco conjuntos de dados distintos, onde a cada rodada o conjunto de treinamento é incrementado com mais um semestre e o semestre subsequente é utilizado como conjunto de teste, conforme ilustrado na Figura 2.6. Esse procedimento, conhecido como validação cruzada para séries temporais, garante que, em cada iteração, o conjunto de teste sempre contenha dados posteriores aos do conjunto de treinamento, respeitando a sequência temporal e evitando avaliações tendenciosas do desempenho do modelo. Dessa forma, o modelo que apresentar melhor desempenho nas quatro situações é considerado o mais adequado, pois demonstra robustez e capacidade de generalização para diferentes cenários ao longo do tempo.

Tabela 2.4. Definições de conceitos fundamentais em aprendizado de máquina.

Termo	Definição
Conjunto de treino	Conjunto de dados que será utilizado para realizar o treino e ajuste do modelo.
Conjunto de teste	Conjunto de dados que não será utilizado durante o processo de construção do modelo, somente no final, para avaliar se o modelo aprendeu o padrão.
Erro dentro da amostra	Erro medido no conjunto de treino. Utilizaremos essa métrica para avaliar se o modelo conseguiu aprender dentro dos próprios dados. Continuando nossa analogia, seria verificar se o aluno consegue acertar as questões na prova antiga, na mesma prova que ele estudou.
Erro fora da amostra	Erro medido no conjunto de teste. Essa métrica é a mais importante, visto que verifica se o modelo é capaz de aprender um padrão (conteúdo) e não somente o que aconteceu em um conjunto particular de dados. Continuando a analogia, seria verificar na simulação de uma prova nova, se o aluno é capaz de acertar as novas questões.
Generalização	Capacidade do modelo de identificar o padrão dos dados e acertar a resposta em dados não vistos no treinamento do modelo. Só conseguimos ter certeza da generalização do modelo na prática, mas utilizamos o erro fora da amostra para fazer essa estimativa.
Sobreajuste	Um ditado popular nesta área diz que “se você torturar os dados por tempo suficiente, eles vão confessar” [Abu-Mostafa et al. 2012, Provost and Fawcett 2016]. O sobreajuste acontece quando “torturamos” demais os dados, ou seja, encontramos não somente um padrão geral, mas padrões falsos (que não generalizam) que se aplicam somente ao conjunto de dados em particular.
Subajuste	O subajuste é quando não encontramos padrão nenhum nos dados. Normalmente indica que não estamos usando um modelo apropriado (depende da complexidade do padrão a ser encontrado e da quantidade de dados que temos para encontrar esse padrão) ou não o configuramos corretamente. É mais fácil de identificar e contornar que o sobreajuste.

2.2.5.2. Modelos

Nesta subseção, explicaremos superficialmente três diferentes modelos comumente utilizados no contexto educacional. Detalhes sobre o funcionamento desses modelos podem ser encontrados em [James et al. 2023, Abu-Mostafa et al. 2012]:

1. Regressão linear: modelos lineares convencionais assumem que a variável dependente segue uma distribuição Normal e é modelada diretamente como uma combi-

Tabela 2.5. Sinais de alerta para identificação de sobreajuste e subajuste

Característica	Sobreajuste	Subajuste
Erro no treinamento (dentro da amostra)	Baixo	Alto
Erro no teste (fora da amostra)	Alto	Alto
Diferença do erro no treino e no teste	Grande	Pequena
Complexidade do modelo	Alta (modelo muito complexo, normalmente com muitos parâmetros)	Baixa (modelo muito simples, normalmente com poucos parâmetros)
Capacidade de generalização	Ruim	Ruim

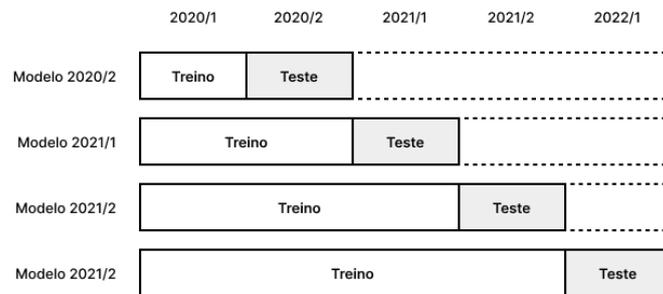


Figura 2.6. Divisões entre dados de treinamento e teste em cada modelo

nação das variáveis preditoras. Essa relação pode ser escrita na forma:

$$E(Y_i) = \mu_i = x_i^T \beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in} \quad (2)$$

Para estimar os valores de β , normalmente utiliza-se o método dos mínimos quadrados. Este método consiste em achar os valores de β que formam a reta (quando trabalha-se com muitas variáveis, essa reta é chamada de hiperplano) cuja diferença entre os pontos da reta (hiperplano) e os pontos observados possua o menor erro quadrático possível. Por ser um modelo simples, as principais maneiras de se evitar o sobreajuste são através da regularização e da seleção de variáveis [James et al. 2023].

2. Regressão logística: em situações em que a variável resposta é binária e, portanto, não podemos assumir a normalidade dos dados — como no caso de abandono ou não de um aluno em uma disciplina — é recomendada a utilização de uma importante extensão dos modelos lineares tradicionais: os modelos lineares generalizados (GLMs). Os GLMs permitem modelar variáveis resposta que seguem distribuições diferentes da normal, como distribuições binomial, de contagem e de proporções, ampliando consideravelmente as possibilidades de análise estatística [Dobson and Barnett 2008]. No caso específico de uma variável resposta binária

ria, o modelo mais adequado dentro do *framework* dos GLMs é a regressão logística [James et al. 2023], que utiliza a função de ligação *logit* para relacionar o preditor linear à probabilidade do evento de interesse. No caso da regressão logística, em vez de modelarmos diretamente a variável resposta, modelamos a probabilidade de que a resposta pertença a uma determinada categoria (por exemplo, a categoria “abandono”). Tal relação pode ser escrita como:

$$Pr(Y = abandono|X) = p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad (3)$$

onde β representa os coeficientes da função a serem estimados e X as variáveis preditoras. Após alguma manipulação, temos que a função logística é dada por

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X \quad (4)$$

e tem uma interpretação natural como o logaritmo da razão entre a probabilidade de um evento ocorrer e a probabilidade de não ocorrer. Assim como na regressão linear, as principais maneiras de se evitar o sobreajuste são através da regularização e da seleção de variáveis [James et al. 2023].

3. **Árvore de Decisão:** é um método sistemático para segmentar dados, particionando o espaço em regiões menores de forma sequencial, utilizando um atributo por vez. A escolha do atributo em cada divisão prioriza aquele que proporciona a melhor separação, tornando as regiões resultantes mais homogêneas em relação à variável-alvo (ou seja, com menor variância em relação a essa variável) [James et al. 2023, Provost and Fawcett 2016]. Esse método pode ser aplicado tanto em tarefas de classificação quanto de regressão. Por sua natureza, a árvore de decisão pode continuar dividindo o espaço até que não restem regiões com erro, o que a torna suscetível ao sobreajuste. Para evitar esse problema, os principais mecanismos utilizados são a poda e a parada precoce. Na poda, a árvore é inicialmente construída até sua máxima profundidade (ficando altamente especializada e, muitas vezes, sem erros nos dados de treinamento). Em seguida, são removidos ramos que não contribuem significativamente para a capacidade de generalização do modelo, ou seja, cuja remoção não aumenta substancialmente o erro do modelo. Já a parada precoce consiste em definir critérios restritivos durante a construção da árvore—como um número mínimo de amostras em cada nó ou um limite máximo de erro tolerado—fazendo com que o crescimento da árvore seja interrompido antes que ela se torne excessivamente especializada [de Oliveira F. Moraes 2016]. No entanto, a construção de uma árvore de decisão é altamente sensível à forma como os dados são divididos em conjuntos de treino e teste. Pequenas alterações no conjunto de treino podem resultar em árvores significativamente diferentes [James et al. 2023]. Para contornar esse problema e aumentar a robustez dos modelos, foram propostas algumas variações do algoritmo de árvore de decisão, como Florestas Aleatórias, Bagging e Boosting. Embora não entremos em detalhes sobre as diferenças entre esses métodos, todos eles se baseiam no princípio do aprendizado em conjunto (*ensemble learning* em inglês). Em vez de construir apenas uma única árvore, esses métodos

constroem múltiplas árvores simples, seja por meio da aleatorização das observações, dos atributos ou pela construção sequencial de árvores que corrigem os erros das anteriores. Assim, o resultado final se torna mais robusto, pois não depende das decisões de uma única árvore, reduzindo a sensibilidade a pequenas variações nos dados de treino.

2.2.5.3. A complexidade dos modelos e sua interpretabilidade

Modelos que possuem uma grande variedade de parâmetros a serem ajustados são considerados mais flexíveis ou complexos e têm a tendência a produzir melhores resultados em aplicações gerais. No entanto, em educação, onde os conjuntos de dados são mais escassos, diversos e segregados entre instituições [Hellas et al. 2018], é difícil coletar um conjunto de dados com um número suficientemente grande de observações que justifique a utilização de métodos complexos (como redes neurais profundas, por exemplo) sem causar sobreajuste. Além disso, há uma discussão vigente sobre a troca entre interpretabilidade e complexidade do modelo [James et al. 2023]. Modelos complexos tendem a ser menos interpretáveis, ou seja, é mais difícil entender (ou interpretar) como as variáveis influenciam o resultado final. No contexto educacional, é fundamental que os tomadores de decisão (normalmente professores e gestores, mas podem ser também os próprios alunos) não recebam somente a previsão como resultado, mas consigam entender o processo decisório do modelo para tomar tal decisão. Por exemplo, não é informativo a um aluno ou professor simplesmente informar que um aluno está propenso a abandonar um curso ou tirar uma nota baixa, mas sim quais dos fatores observados indicam que o aluno possa ter tal desempenho para que o tomador de decisão possa tomar ações que remediem o resultado previsto.

[Rodrigues et al. 2025] apresenta os algoritmos mais utilizados e com melhores resultados para previsão de desempenho em universidades reportados por diversos artigos. A pesquisa mostrou que o algoritmo de Floresta Aleatória foi o mais utilizado, mostrando a preferência por algoritmos complexos mas que ainda oferecem interpretabilidade.

Técnicas mais recentes chamadas de Inteligência Artificial Explicável (XAI ou eXplainable Artificial Intelligence, em inglês) como LIME [Ribeiro et al. 2016] e SHAP [Lundberg and Lee 2017] estão permitindo quebrar essa troca, oferecendo explicações e interpretabilidade mesmo em modelos complexos [Belle and Papantonis 2021]. No entanto, a dificuldade no contexto educacional continua ainda na coleta de um conjunto grande de dados para justificar a utilização dessas técnicas, o que só é conseguido normalmente através de cursos online abertos e massivos (MOOC), coletas longitudinais (ao longo de muitos anos) e colaborações institucionais.

2.2.6. Avaliação e interpretação dos resultados

Após a criação dos modelos preditivos, é preciso tirar *insights* dos resultados e transformá-los em ações, apresentando-os aos diferentes interessados: gestores acadêmicos, professores e alunos. Cada interessado utilizará e interpretará o resultado de uma maneira diferente, tirando proveitos distintos. Essa seção apresenta as principais técnicas de apresentação de resultados de acordo com o tipo de interessado.

2.2.6.1. Matriz de confusão e suas métricas derivadas

A matriz de confusão, ou tabela de contingência, é o instrumento mais completo para avaliar o resultado dos modelos preditivos quando estamos trabalhando com um conjunto pequeno de classes, como por exemplo, no caso da previsão de evasão. A matriz de confusão mostra a contagem ou a proporção de previsões corretas e incorretas feitas pelo modelo preditivo, comparando-as com as classes verdadeiras dos dados. Cada coluna representa a classe verdadeira (rótulo real) de uma instância de dado e cada linha representa a classe prevista pelo classificador, conforme apresentado na Tabela 2.6.

Tabela 2.6. Matriz de confusão para o caso de classificação binária

Valor predito	Valor real	
	$Y = 0$	$Y = 1$
$c(X) = 0$	VN (verdadeiro negativo)	FN (falso negativo)
$c(X) = 1$	FP (falso positivo)	VP (verdadeiro positivo)

Note que a matriz de confusão deve ser construída a partir das classificações no conjunto de teste para que as estimativas das medidas de desempenho não sejam subestimadas. Algumas das métricas definidas com base em uma matriz de confusão são apresentadas na Tabela 2.7. Pode-se utilizar tais métricas para a escolha do melhor classificador. Em muitos casos, busca-se maximizar simultaneamente a sensibilidade e a especificidade, ou seja, um equilíbrio entre os erros de classificação entre as duas classes. Em outros casos, pode-se ter interesse em maximizar a estatística F1 ou outra métrica obtida a partir da matriz de confusão [Izbicki and dos Santos 2020].

Outra forma de analisar o desempenho das classificações de modelos é usando a curva ROC (do inglês, *Receiver Operating Characteristic*), uma representação gráfica que permite avaliar o desempenho de um classificador binário em diferentes valores de K. A curva ROC é construída plotando a taxa de verdadeiros positivos (Sensibilidade) no eixo vertical e o complementar da taxa de falsos positivos (1 - Especificidade) no eixo horizontal. Cada ponto na curva ROC representa um determinado limiar de classificação aplicado ao modelo. A curva ROC permite a comparação visual da capacidade preditiva de diferentes modelos ajustados. Curvas com maior área abaixo dela indicam melhor desempenho do modelo, o valor da área nos leva a mais uma métrica, conhecida por AUC (Area Under the Curve, em inglês), que reflete um desempenho mais elevado à medida que se aproxima do valor 1. Além disso, a curva ROC é usada para definir o ponto de corte do classificador que maximiza a sensibilidade e a especificidade [James et al. 2023]. A fim de definir um modelo preditivo para classificação, precisamos escolher um valor para o limiar de decisão K, acima do qual classificamos uma observação com a classe de interesse. Para isso, podemos testar diferentes cortes K entre 0 e 1 e selecionar aquele que tem melhor desempenho, segundo algum critério estabelecido usando os valores obtidos de uma matriz de confusão.

Tabela 2.7. Métricas de avaliação: descrição e fórmulas

Métrica	Fórmula
Acurácia: do total de observações, quantas foram classificadas corretamente.	$A = \frac{VP+VN}{VP+VN+FP+FN}$
Sensibilidade/Recall: das observações em que o valor real é 1 ($Y = 1$), quantas foram classificadas corretamente ($c(X) = 1$).	$S = \hat{P}(c(X) = 1 Y = 1) = \frac{VP}{VP+FN}$
Especificidade: das observações em que o valor real é 0 ($Y = 0$), quantas foram classificadas corretamente ($c(X) = 0$).	$E = \hat{P}(c(X) = 0 Y = 0) = \frac{VN}{VN+FP}$
Valor preditivo positivo/Precision: das observações classificadas como 1 ($c(X) = 1$), quais são de fato da classe 1 ($Y = 1$).	$VPP = \hat{P}(Y = 1 c(X) = 1) = \frac{VP}{VP+FP}$
Valor preditivo negativo: das observações classificadas como 0 ($c(X) = 0$), quais são de fato da classe 0 ($Y = 0$).	$VPN = \hat{P}(Y = 0 c(X) = 0) = \frac{VN}{VN+FN}$
Estatística F1: média harmônica entre o Recall e a Precision.	$F1 = \frac{2}{\frac{1}{S} + \frac{1}{VPP}}$

O contexto educacional, assim como outros contextos sensíveis que lidam com dados pessoais, como saúde e justiça, apresenta um desbalanceamento natural não apenas na quantidade de dados por classe, mas também em relação a quais classes são de maior interesse. Por exemplo, no caso da previsão de evasão ou reprovação, há maior interesse em identificar quem está em maior risco de evasão ou reprovação do que em identificar quem está em menor risco. Isso nos leva a refletir sobre quais tipos de erros são mais prejudiciais, pois, ao utilizar métricas simples de classificação não se faz distinção entre erros falsos positivos e falsos negativos, assumindo tacitamente que ambos os erros são igualmente importantes, quando, na prática, esse raramente é o caso [Provost and Fawcett 2016]. Por exemplo, é mais grave errar ao afirmar que um aluno com baixo risco de abandonar o curso irá abandoná-lo, ou ao afirmar que um aluno com alto risco de abandono não irá abandonar? Esses são, normalmente, tipos de erros com custos muito diferentes, porque as classificações incorretas têm consequências de diversas gravidades.

Considerando que os resultados dos modelos podem ser utilizados para orientar ações e intervenções por parte de professores e gestores, no primeiro caso, gastaríamos recursos extras com um aluno que provavelmente não precisaria de intervenção, sem causar grandes prejuízos a ele. No entanto, no segundo caso, deixaríamos de intervir em um aluno que provavelmente se beneficiaria da ação, podendo comprometer seu desempenho ou permanência. Portanto, é fundamental refletir sobre essas questões ao trabalhar com modelos preditivos na educação, ajustando as métricas e os pontos de corte de modo que reflitam esse desequilíbrio e priorizem intervenções mais eficazes.

O desafio de escrever sobre este tema é que cada situação é diferente, não havendo uma solução única ou “correta” que possa ser ditada [Provost and Fawcett 2016].

2.2.6.2. Apresentação dos resultados

As métricas são indicadores importantes do poder preditivo de um modelo. No entanto, principalmente no contexto educacional, os interessados (alunos, professores e gestores) estão interessados não somente na previsão (por exemplo, o risco de abandono ou reprovação) mas também em como mitigar esse risco. E, inspecionando alguns outros fatores dos modelos, conseguimos algumas informações sobre os comportamentos preditores do desfecho de interesse. Por exemplo, em uma regressão linear ou logística, é comum inspecionarmos os coeficientes ajustados dos atributos para entender aqueles de maior relevância, através de sua magnitude e p-valor. Já em uma árvore de decisão, a interpretação é facilitada, pois o modelo pode ser representado visualmente por um diagrama, permitindo que suas decisões sejam compreendidas como um conjunto de regras lógicas.

Ainda, para a utilização apropriada dos modelos e promover a auto-avaliação, auto-reflexão e consciência situacional nos contextos educacionais, *dashboards* (ou painéis, em português) podem ser utilizados, organizando a informação em gráficos informativos direcionados às pessoas de interesse.

2.3. Inteligência artificial generativa para alunos e professores

A inteligência artificial generativa, mais conhecida através de seus modelos de Linguagem de Larga Escala (*Large Language Models* em inglês ou LLMs) ganharam destaque nos últimos anos por sua facilidade de interação e respostas coerentes [Brown et al. 2020]. Na educação em computação, modelos como ChatGPT, Llama, Gemini, Sabiá, entre outros, têm sido alvos de pesquisa sobre a sua capacidade de responder corretamente questões [Denny et al. 2024], prover explicações de código para alunos [Denny et al. 2024, Nam et al. 2024] e professores [Hassany et al. 2024]; apoiar os alunos na escrita de código [Kazemitabaar et al. 2024, Nam et al. 2024] e atuar como um monitor, elicitando dúvidas [Liffiton et al. 2024]. Sendo um campo de pesquisa incipiente, a literatura tem pesquisado e perguntado que tipo de interações são percebidas como mais úteis pelos alunos [Denny et al. 2024, Liffiton et al. 2024], quais produzem resultados mais acurados [Liffiton et al. 2024, Nam et al. 2024] e quais conseguem uma retenção maior do conhecimento [Kazemitabaar et al. 2024]. Essas pesquisas apontam que a utilização de ferramentas baseadas em LLMs na sala de aula pode impulsionar a criação de experiências personalizadas para os alunos [Kasneci et al. 2023], atuando como um agente capaz de identificar a dificuldade de cada aluno e prover sugestões de melhorias e comentários sobre os códigos escritos.

Esse cenário torna promissora a ideia de utilizar LLMs para oferecer suporte aos alunos e professores. No entanto, há preocupações sobre a possível dependência excessiva dos estudantes em relação a esses modelos, especialmente quando são usados para gerar respostas completas [Becker et al. 2022]. Uma descrição simplificada da ameaça percebida é a seguintes: até recentemente, a tarefa de articular ideias com base em fontes selecionadas — seja para produzir textos, seja para resolver problemas matemático-científicos — era amplamente valorizada como estratégia eficaz para o desenvolvimento de competências fundamentais da educação básica. Ao interagir com diferentes fontes, os estudantes aprendiam a identificar informações relevantes, distinguir fatos de opiniões, avaliar a confiabilidade dos conteúdos e organizar o conhecimento de forma coerente.

Essa prática fomenta habilidades como pensamento crítico, argumentação e letramento informacional — competências essenciais para a formação de cidadãos autônomos e conscientes no século XXI. A possibilidade de que um estudante delegue essas tarefas a uma LLM, abrindo mão de todo esse aprendizado, levanta preocupações legítimas. No entanto, o cerne dessa questão, evidenciado pelas novas tecnologias, já fazia parte dos debates educacionais: a corresponsabilidade do estudante por sua própria trajetória de aprendizagem e do valor percebido pelo estudante acerca da formação escolar. Apesar de percebidas como uma ameaça, as novas tecnologias e as transformações que elas impulsionam podem enriquecer esse debate e contribuir para a construção de uma experiência escolar mais significativa e alinhada aos propósitos dos estudantes, respeitando o espaço de atuação e a responsabilidade do professor.

Esta seção se propõe a apresentar como utilizar LLMs incentivando o aprendizado dos alunos e apoiando o trabalho do professor. Mais especificamente, auxiliaremos os professores a responder às seguintes perguntas:

1. Como estimular o pensamento crítico dos alunos através da avaliação das alucinações de IAs generativas?
2. Quais exercícios os alunos podem fazer com o auxílio de IA generativa?
3. Em quais tarefas a IA generativa pode ajudar o professor?
4. Como escrever *prompts* (comandos para a IA generativa) que forneçam melhores resultados?

2.3.1. Conceitos fundamentais de LLMs

Esta seção apresenta a fundamentação teórica do funcionamento das LLMs, dando uma visão geral e definindo conceitos como *redes neurais*, *embedding* e *tokens* de modo que o leitor entenda a natureza probabilísticas das respostas fornecidas por um modelo de IA generativa.

Redes neurais são modelos computacionais inspirados no funcionamento do cérebro humano, compostos por unidades chamadas neurônios artificiais ou perceptrons. O perceptron é a unidade fundamental dessas redes: ele recebe um conjunto de sinais de entrada, cada um multiplicado por um peso, soma esses valores e processa o resultado por meio de uma função de ativação, geralmente não linear [Vasconcelos and Clua 2017]. As redes neurais são formadas por vários perceptrons organizados em camadas: uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. A arquitetura da rede refere-se à forma como essas camadas e seus neurônios estão conectados, podendo variar conforme a aplicação e o tipo de problema a ser resolvido.

Antes de serem utilizados no treinamento do modelo, os dados passam por várias etapas de preparação, como limpeza, normalização e, principalmente, a *tokenização*. Nesse processo, a entrada é dividida em pequenas partes chamadas *tokens*. No caso de textos, esses *tokens* geralmente correspondem a palavras, subpalavras ou combinações frequentes de caracteres. Em outros tipos de dados, como imagens ou áudios, os *tokens* podem ser pequenos trechos da imagem ou fragmentos do som.

Cada um desses *tokens* é então associado a um vetor — uma lista de números — que busca capturar o significado daquele fragmento. Esses vetores, conhecidos como *embeddings*, posicionam cada *token* em um espaço de alta dimensionalidade, de forma que *tokens* com significados semelhantes fiquem próximos uns dos outros nesse espaço. Assim, o modelo pode representar o texto (ou outros dados) como uma sequência de vetores, facilitando a identificação de relações complexas entre palavras, frases ou até mesmo elementos visuais e sonoros. O Exemplo 1 ilustra como uma frase pode ser segmentada em *tokens*, e como cada *token* recebe um identificador numérico correspondente ao seu *embedding*. Os IDs são fictícios, apenas para ilustrar como cada *token* é associado a um índice numérico no vocabulário do modelo.

Exemplo 1, adaptado de [de Sousa 2025]

Texto: “Estou com dificuldades para entender estruturas de repetição”

Tokens: “Est”, “ou”, “com”, “di”, “fic”, “uldades”, “para”, “en”, “tender”, “es”, “tru”, “turas”, “de”, “re”, “pe”, “ti”, “ção”]

IDs: [501, 208, 305, 612, 713, 499, 402, 311, 718, 256, 337, 805, 109, 410, 527, 634, 999]

Após o pré-processamento, os dados — agora representados como uma sequência de vetores — são processados por arquiteturas baseadas em *transformers*. Esses modelos combinam camadas que realizam diferentes operações para analisar os *tokens* e aprender as relações entre eles. Entre as camadas mais utilizadas estão os mecanismos de atenção, redes neurais *feed forward* e *encoders* e *decoders*.

As camadas de atenção revolucionaram a IA Generativa, pois ela é que permite que os vetores “conversem” entre si, trocando informações para atualizar seus valores de acordo com o contexto. Por exemplo, a palavra “modelo” em “modelo de aprendizado de máquina” tem um significado diferente de “modelo de moda”. As camadas de atenção são responsáveis por identificar quais palavras, dentro do contexto, são relevantes para atualizar o significado das demais, ajustando as representações de acordo com a relação entre os *tokens*. Assim, mesmo que dois *tokens* estejam distantes na sequência, o mecanismo de atenção pode estabelecer conexões importantes entre eles, refinando a compreensão contextual do texto [de Sousa 2025].

A geração do texto final ocorre por meio de um processo sequencial, no qual o modelo prevê os *tokens* um a um. A cada etapa, a escolha do próximo *token* é baseada no histórico dos *tokens* já gerados e nas representações contextuais (vetores ou *embeddings*) construídas. Esse processo envolve repetidas passagens pelos blocos de atenção e camadas de rede *feed forward*, permitindo que o modelo refine continuamente a compreensão do contexto [Barros et al. 2025].

Ao final dessas iterações, espera-se que todo o significado essencial do trecho esteja incorporado no último vetor da sequência. Sobre esse vetor final, o modelo realiza uma operação que gera uma distribuição de probabilidade sobre todos os possíveis *tokens* seguintes — ou seja, sobre todas as pequenas unidades de texto que podem ser geradas na sequência. Dessa forma, o modelo seleciona o próximo *token* mais provável e repete o processo até concluir a geração do texto.

2.3.2. Técnicas de criação de *prompts*

Os *prompts* são instruções fornecidas aos LLMs com o objetivo de orientar a geração de texto e obter respostas específicas, sem a necessidade de alterar os parâmetros do modelo. Eles constituem o principal meio de interação com o sistema, estabelecendo o contexto, a tarefa a ser realizada e até mesmo o tom desejado para a resposta. Por isso, a elaboração cuidadosa do *prompt* é fundamental para garantir que o modelo forneça respostas adequadas à tarefa proposta [Sahoo et al. 2024]. Quando os modelos recebem instruções vagas ou pouco detalhadas, acabam considerando uma ampla variedade de possibilidades. Isso geralmente resulta em respostas genéricas e superficiais, semelhantes ao que se encontraria em uma simples pesquisa na internet. Embora essas respostas não sejam erradas ou inúteis, muitas vezes não atendem às necessidades específicas de determinadas tarefas [Chen et al. 2024]. No Exemplo 2, observa-se um caso em que um *prompt* genérico levou o modelo a gerar uma resposta ampla. Como o tema não foi claramente especificado, o modelo apresentou uma lista que pode não contemplar exatamente o interesse do usuário.

Uma das estratégias mais eficazes ao utilizar LLMs é elaborar instruções claras e específicas, evitando qualquer tipo de ambiguidade. Esses modelos, treinados com grandes volumes de dados textuais provenientes de diversas fontes, conseguem sintetizar informações variadas. No entanto, quando recebem *prompts* pouco detalhados, tendem a gerar respostas genéricas, o que pode comprometer a relevância em situações que exigem precisão. Em contrapartida, *prompts* bem estruturados e detalhados ajudam a reduzir a incerteza do modelo, direcionando suas respostas para atender exatamente às necessidades da tarefa. Isso resulta em conteúdos mais precisos e relevantes, aprimorando o desempenho do modelo [Sahoo et al. 2024].

Exemplo 2, retirado de [de Sousa 2025]

Entrada: Eu gostaria de saber sobre o assunto do momento no campo da tecnologia.

Saída:

As principais tendências tecnológicas de 2025 incluem:

- **Computação Quântica:** Oferece capacidades de cálculo revolucionárias para aplicações em ciência dos materiais e descoberta de drogas
- **IA Generativa:** Transforma setores através da criação automática de conteúdo complexo, melhorando a produtividade e o engajamento do cliente
- **Geminação Digital:** Utiliza IoT para replicar e analisar sistemas e processos do mundo real em ambientes virtuais, beneficiando saúde e manufatura
- **Bioprinting:** Avança na medicina com a impressão de tecidos e órgãos para transplantes, promovendo a medicina personalizada

Outra prática recomendada é atribuir ao modelo um papel específico, como o de professor ou especialista em determinado tema. Essa estratégia orienta a forma como o modelo responde, garantindo maior alinhamento com o contexto desejado. Ao definir

um papel claro, não só se direciona a produção do conteúdo, mas também se fortalece a coerência entre o *prompt* e a resposta gerada. No Exemplo 3, por exemplo, o modelo foi instruído a agir como professor de um grupo de alunos de determinada faixa etária, o que resultou em uma resposta mais adequada ao interesse do usuário.

Essas estratégias apresentam como benefícios a diminuição da necessidade de reunir grandes quantidades de dados específicos para cada tarefa, além de reduzirem o risco de o modelo limitar suas respostas apenas a perguntas previamente conhecidas. Dessa forma, o modelo se torna mais flexível e capaz de lidar com uma variedade maior de solicitações.

Exemplo 3, retirado de [de Sousa 2025]

Entrada: Você é um professor de introdução a programação para alunos do ensino médio com idades entre 15 e 18 anos. Quais projetos você poderia criar com os seus alunos usando Python?

Saída: Como professor de programação para alunos do ensino médio, você pode desenvolver projetos interessantes usando Python para ajudar a ensinar conceitos básicos e estimular o interesse dos alunos. Algumas ideias incluem:

- Calculadora Simples: Para praticar operações matemáticas básicas.
- Jogos Simples: Como “Adivinhe o Número” ou “Tic-Tac-Toe”.
- Gerador de Histórias Aleatórias: Combinando partes de frases para criar histórias.

2.3.2.1. Prompts básicos: *zero-shot*, *few-shot* e *chain-of-thought*

Entre as principais estratégias para a criação de *prompts* estão o *zero-shot*, *one-shot* e *few-shot prompting*. Essas abordagens se diferenciam pela quantidade de exemplos incluídos no *prompt*, o que impacta diretamente o desempenho do modelo em diferentes tipos de tarefas. No *zero-shot prompting*, por exemplo, o modelo é solicitado a executar uma tarefa sem receber nenhum exemplo concreto. Nesse caso, o *prompt* apenas descreve a tarefa e o contexto, sem fornecer modelos de resposta. Embora seja uma estratégia rápida e simples, ela pode não garantir alta precisão em situações em que o contexto ou o formato esperado da resposta não estejam claramente definidos [Brown et al. 2020]. No Exemplo 4 o usuário pediu que o modelo criasse um código para dividir dois números de forma genérica. Contudo, se fosse necessário um formato específico, como a inclusão de uma descrição detalhada (*docstring*), essa informação não seria contemplada automaticamente pelo modelo em sua resposta.

No *one-shot prompting*, é fornecido ao modelo um exemplo de resposta para a tarefa, o que facilita a compreensão do formato e do tipo de resposta esperado. Normalmente, o *prompt* apresenta uma instrução seguida de um exemplo, e então pede-se ao modelo que produza uma resposta semelhante. No Exemplo 5, por exemplo, o usuário

Exemplo 4, retirado de [de Sousa 2025]

Entrada: Faça um código python que calcule a divisão de dois números.

Saída:

```
def divisao( Numerador, divisor):  
    try:  
        resultado = Numerador / divisor  
        return resultado  
    except ZeroDivisionError:  
        return "Não é possível dividir por zero"
```

incluiu um código de referência que continha tanto uma *docstring* quanto um comando de impressão (*print*). O modelo, por sua vez, utilizou esse exemplo como guia para gerar o novo código solicitado, seguindo o formato especificado pelo usuário.

Visando quebrar o problema proposto em problemas menores e diminuir a chance de alucinações e respostas incorretas, as técnicas *Chain-of-Thought* [Wei et al. 2022] e *Plan-and-Solve* [Wang et al. 2023] buscam inserir etapas intermediárias de processamento nos modelos. No entanto, não há garantia de que as etapas intermediárias de processamento assegurem a correção das respostas, podendo resultar tanto em respostas incorretas quanto em respostas certas baseadas em processos falhos [Peres 2023]. A maior parte da LLMs disponíveis em formato de *chatbot* atualmente utiliza uma mistura dessas técnicas, já incorporando as etapas intermediárias em seu raciocínio. O Exemplo 6 apresenta a resposta fornecida pelo ChatGPT-4o a uma pergunta de raciocínio matemático, indicando o processamento por etapas.

Por fim, existem os *prompts* iterativos (ou de múltiplos estágios), nos quais o processamento é dividido em várias entradas enviadas ao LLM, de modo que cada nova entrada depende da resposta obtida à etapa anterior [Qiao et al. 2023]. Combinando essa abordagem com a especialização do modelo para contextos ou bases de dados específicas, essas técnicas têm demonstrado resultados superiores [Sun et al. 2024, Gao et al. 2024]. Isso ocorre porque o usuário ganha o poder de avaliar e direcionar as respostas do modelo a cada etapa, promovendo maior precisão e adequação ao objetivo desejado.

2.3.3. Exemplos práticos no contexto educacional

Esta seção traz exemplos de utilização de IA generativa no contexto educacional. São apresentados casos de uso para alunos e professores, além de fomentada a discussão sobre a dependência excessiva, as alucinações e o pensamento crítico na utilização desta tecnologia.

Para os professores da educação básica, compreender tanto as potencialidades quanto os riscos associados às tecnologias baseadas em Inteligência Artificial é essencial para integrá-las de forma crítica, ética e pedagógica ao cotidiano escolar. A construção desse entendimento requer processos formativos que desenvolvam o senso crítico e favoreçam a apropriação consciente dos mecanismos, benefícios e limitações dessas

Exemplo 5, retirado de [de Sousa 2025]

Entrada: Com base no código de exemplo abaixo, faça um código python que calcule a divisão de dois números.

```
def somar(num_a, num_b):  
    """Soma dois números e retorna o resultado.  
    Entradas:  
        num_a -> int, float  
        num_b -> int, float  
    Retorna:  
        int, float: A soma de num_a e num_b."""  
    resultado = num_a + num_b  
    print("O resultado da soma dos números é",  
          resultado)  
    return resultado
```

Saída:

```
def dividir(num_a, num_b):  
    """Divide o primeiro número pelo segundo e retorna  
    o resultado.  
    Entradas:  
        num_a -> int, float  
        num_b -> int, float  
    Retorna:  
        float: O resultado da divisão de num_a por  
        num_b."""  
    if num_b == 0:  
        print("Não é possível dividir por zero.")  
        return None  
    resultado = num_a / num_b  
    print("O resultado da divisão dos números é",  
          resultado)  
    return resultado
```

Exemplo 6, adaptado de [Kojima et al. 2022]

Entrada: Um malabarista consegue fazer malabarismo com 16 bolas. Metade das bolas são bolas de golfe e metade das bolas de golfe são azuis. Quantas bolas de golfe azuis há?

Saída:

Vamos analisar passo a passo:

1. O malabarista faz malabarismo com 16 bolas.
2. Metade das bolas são bolas de golfe $\rightarrow \frac{16}{2} = 8$ bolas de golfe.
3. Metade das bolas de golfe são azuis $\rightarrow \frac{8}{2} = 4$ bolas de golfe azuis.

Resposta: há 4 bolas de golfe azuis.

ferramentas — com destaque para sistemas generativos como o ChatGPT.

Uma introdução reflexiva e prática para docentes pode ser feita em oficinas de curta duração, abordando inicialmente os principais riscos associados ao uso da IA, como vieses algorítmicos, desinformação e questões éticas. Após essas primeiras considerações, é indicada a organização de discussões sobre os riscos e ameaças percebidos e a contraposição destes com outros riscos e ameaças já presentes no cotidiano escolar. É importante lembrar que o contexto escolar é parte da vida social e que toda implicação à vida social é uma implicação também para o contexto escolar. A IA chega na escola e cabe ao professor analisar criticamente a bagagem que ela traz.

2.3.4. Identificação e reflexão sobre os riscos da IA

A partir de uma pesquisa simples na internet sobre os riscos do uso do ChatGPT, os professores são convidados a identificar e analisar criticamente informações provenientes de diferentes fontes. Essa etapa busca revelar questões pouco percebidas previamente, promovendo o desenvolvimento do pensamento crítico e a sensibilização para a necessidade da mitigação responsável desses riscos em contextos educacionais.

Também é interessante que os professores utilizem experimentalmente o ChatGPT para gerar descrições visuais de diferentes perfis humanos (pessoas bem-sucedidas, boas, simpáticas) e avaliem se existem indícios de vieses culturais nas respostas. Tal exercício favorece a compreensão de que sistemas de IA refletem valores e limitações de seus dados de treinamento, estimulando a reflexão sobre os desafios éticos inerentes à adoção dessas tecnologias.

2.3.5. Utilizando IA para desenvolver atividades didáticas e avaliações

A proposta prática de elaboração de atividades didáticas mediadas por IA oferece aos professores a oportunidade de experimentar o potencial das ferramentas na construção de planos de aula contextualizados, alinhados a demandas curriculares (como a BNCC) e metodologias ativas. A avaliação crítica da satisfação com os resultados gerados ajuda a balizar a efetividade e os limites da tecnologia no cotidiano escolar.

É importante ressaltar a importância de fornecer detalhes como segmento/ano, matéria, e qualquer outra informação que achar pertinente. Por exemplo: Sou professora do quinto ano do Ensino Fundamental em uma escola pública brasileira e preciso planejar uma aula de 50 minutos sobre o sistema solar. Utilizando metodologias ativas e se baseando nas habilidades da Base Nacional Comum Curricular, elabore três propostas de planos de aula detalhando cada etapa. Inclua informações sobre evidências de aprendizagem e uma proposta de avaliação.

2.3.5.1. Propondo atividades didáticas que envolvam o uso de LLMs pelos alunos

Para desenhar atividades que tragam o uso de LLMs para os estudantes, é importante pensar nas vantagens do uso dessa tecnologia. A maior delas é a possibilidade de interagir, criando seus próprios prompts. Como em qualquer atividade didática, tudo começa com a definição do objetivo de aprendizagem. Esse objetivo pode estar relacionado a habilidades como a produção escrita, a análise crítica, ou a resolução de problemas, entre outros.

Por exemplo, ao pedir que os alunos dialoguem com o ChatGPT para explorar ideias, revisar textos ou simular situações, o professor estimula a autonomia e o pensamento crítico. Essas interações devem ser acompanhadas de momentos de reflexão: “O que você aprendeu?”, “Você concorda com a resposta da IA?”, “O que mudaria?”.

Exemplo: Língua Portuguesa. Competências da BNCC: Desenvolver a oralidade e a escuta ativa; Identificar elementos estruturais da história (personagem, espaço, tempo e enredo): Recontar histórias com criatividade e autonomia. Relacionar textos ouvidos com experiências pessoais.

Com a ajuda da IA, criar uma versão de contos de fadas clássicos. A IA pode ser usada para fazer imagens ou sugerir textos, porém os alunos têm que partir de ideias próprias para criar o prompt. Há um amplo universo de questões que podem ser trabalhadas, além da construção de narrativas. Alguns exemplos mais diretos são questões de gênero, representatividade e regionalidades.

Outro tipo de atividade que atende a qualquer área do conhecimento é o uso da LLM para questionamento Socrático. A técnica do questionamento socrático, originada na prática filosófica de Sócrates, baseia-se em provocar o pensamento crítico por meio de perguntas que estimulam a reflexão, o raciocínio lógico e a construção do conhecimento pelo próprio sujeito. Em vez de fornecer respostas prontas, o método busca orientar o estudante a descobrir por si mesmo os fundamentos, contradições e implicações de suas crenças, fornecendo, ao invés de respostas, outras perguntas.

Como a IA é capaz de sustentar diálogos e responder a perguntas em linguagem natural, o professor ou o próprio aluno pode utilizar o ChatGPT para simular uma conversa socrática, formulando questões abertas, exploratórias e reflexivas.

Exemplos de comando:

- “Acredito que as redes sociais fazem mais mal do que bem. Faça 5 perguntas que desafiem meu ponto de vista.”

- “Estou com dificuldades para entender o Teorema de Pitágoras. Faça perguntas que me ajudem a descobrir sozinho como ele funciona.”
- “Aqui está um problema: ‘Um trem sai de uma cidade a 60 km/h...’ — Me ajude a resolver esse problema fazendo apenas perguntas.”
- “Estou tentando entender por que um objeto em queda livre acelera. Faça perguntas que me ajudem a entender isso sozinho.”
- “Aqui está um problema: Um carro de 1000 kg é puxado com uma força de 2000 N. Me ajude a entender o que está acontecendo fazendo apenas perguntas.”
- “Por que vemos um objeto de uma certa cor? Me guie por meio de perguntas.”

2.4. Questões éticas relacionadas ao uso de IA na educação

Os avanços tecnológicos, a Inteligência Artificial (IA) e o uso crescente de dados em ambientes educacionais ampliaram as possibilidades de personalização da aprendizagem e de aprimoramento dos processos pedagógicos. No entanto, essa transformação também traz desafios significativos relacionados ao uso ético, seguro e transparente dessas informações, bem como os impactos que essas novas no ambiente educacional.

É bastante perceptível a aceleração do uso de IA no setor educacional ocorrida nos últimos anos, especialmente após a pandemia de COVID-19, que impulsionou a digitalização do ensino. Como essas tecnologias têm alto potencial para interferir e até mesmo transformar o ambiente educacional, organizações internacionais que se ocupam de questões de impacto global produziram documentos que discutem a IA e até especificamente o uso de IA na educação.

A consideração dos riscos associados à aplicação da Inteligência Artificial (IA) na educação é fundamental para garantir que essas tecnologias sejam utilizadas de forma alinhada aos princípios educacionais democráticos, contribuindo para o objetivo primordial da educação – a formação do indivíduo livre, crítico, autônomo e apto a promover o bem-estar próprio e do seu entorno.

Sistemas de IA aplicados à educação processam grandes volumes de dados pessoais e sensíveis de estudantes, como desempenho acadêmico, histórico de comportamento, de condições de saúde física, emocional e psicológica, e algumas vezes até mesmo dados biométricos. O uso indevido de informações dessa natureza tão pessoal pode violar legislações como a Lei Geral de Proteção de Dados (LGPD) no Brasil e o Regulamento Geral sobre a Proteção de Dados (GDPR) na União Europeia, comprometendo a privacidade e a segurança dos alunos [UNESCO 2021a].

Algoritmos treinados com dados enviesados, muitas vezes por questões históricas e/ou estruturais, podem reproduzir e até intensificar desigualdades já existentes no sistema educacional, impactando estudantes de grupos vulneráveis em processos de recomendação de conteúdos e ações pedagógicas, e reforçando as situações estruturais de fracasso escolar.

Ferramentas de IA frequentemente operam como “caixas-pretas” para os usuários finais, dificultando a compreensão sobre o resultado produzido. A falta de transparência

compromete a possibilidade de responsabilização dos tomadores de decisão e impede que professores, estudantes e famílias entendam, ponderem ou contestem decisões automatizadas [UNESCO 2023].

A delegação de tarefas pedagógicas complexas à IA pode reduzir a autonomia dos educadores, ocultar nuances dos processos pessoais de aprendizagem dos alunos, coibir o diálogo e as oportunidades de acolhimento e suporte, comprometendo o papel da mediação humana no ensino-aprendizagem. A utilização de sistemas de monitoramento e personalização baseados em IA pode gerar estresse e sensação de vigilância constante, afetando o ambiente escolar e a saúde mental dos estudantes.

Há ainda a dependência de soluções tecnológicas desenvolvidas por grandes corporações estrangeiras, que pode comprometer a soberania educacional e tecnológica dos países. Isso levanta preocupações quanto à autonomia dos sistemas educacionais e à governança dos dados coletados.

2.4.1. Mapeamento dos riscos do uso de IA na educação

O relatório *Governing AI for Humanity* [High-Level Advisory Body on AI 2024], elaborado pelo Alto Comitê Consultivo das Nações Unidas (ONU) e publicado em 2024, propõe princípios e recomendações globais para a governança ética e segura da IA, com foco na promoção dos direitos humanos, na equidade e na sustentabilidade. Neste relatório, é apresentada uma lista de riscos relacionados à IA, categorizados com base na vulnerabilidade existente ou potencial. Esses riscos abrangem aspectos éticos, sociais, econômicos e ambientais, incluindo:

1. **Viés algorítmico e discriminação:** Sistemas de IA podem perpetuar ou amplificar preconceitos existentes, resultando em decisões injustas, especialmente em áreas como saúde, justiça e educação.
2. **Privacidade e segurança de dados:** O uso extensivo de dados pessoais por sistemas de IA levanta preocupações sobre a proteção da privacidade e a segurança das informações sensíveis.
3. **Desinformação e manipulação:** A IA pode ser utilizada para criar e disseminar informações falsas ou enganosas, impactando negativamente a opinião pública e os processos democráticos.
4. **Armas autônomas e uso malicioso:** O desenvolvimento de sistemas de IA para fins militares ou por atores mal-intencionados representa uma ameaça significativa à segurança global.
5. **Impactos socioeconômicos:** A automação impulsionada pela IA pode levar à perda de empregos e ao aumento das desigualdades econômicas, afetando especialmente comunidades vulneráveis.
6. **Consumo energético e sustentabilidade:** O treinamento e operação de modelos de IA de grande escala demandam altos níveis de energia, contribuindo para impactos ambientais adversos.

7. **Concentração de poder:** O domínio do desenvolvimento de IA por um número limitado de empresas e países pode resultar em desequilíbrios de poder e acesso desigual aos benefícios da tecnologia.

O relatório enfatiza a necessidade de uma governança global inclusiva e colaborativa para mitigar esses riscos e garantir que a IA seja desenvolvida e utilizada de maneira ética, segura e benéfica para toda a humanidade.

A Organização das Nações Unidas para a Educação, a Ciência e a Cultura – UNESCO publicou diversos relatórios que abordam os riscos associados ao uso da Inteligência Artificial (IA). Em 2021, a UNESCO publicou a Recomendação sobre a Ética da Inteligência Artificial [UNESCO 2021b], na qual reconhece que a IA pode trazer benefícios significativos para a sociedade; contudo, alcançar esses benefícios pode intensificar tensões já existentes, relacionadas à transformação das relações sociais e ao acesso desigual ao conhecimento e às tecnologias. Entre os desafios estão: Déficit de **alfabetização digital e cívica**, que limita a capacidade da população de se envolver criticamente com os temas ligados à IA; Lacunas em capacidades institucionais para mitigar questões como **transferência de responsabilidades**; Deficiências em **infraestrutura física e digital**; Ausência de **marcos regulatórios** adequados no que se refere à governança de dados, podendo ameaçar o respeito e a proteção à privacidade e à diversidade cultural e agravar situações de vulnerabilidade.

A publicação *Artificial Intelligence and Education: Guidance for Policy-makers* [UNESCO 2021a] oferece orientações para formuladores de políticas sobre como melhor aproveitar as oportunidades e enfrentar os riscos apresentados pela crescente conexão entre IA e educação. Os riscos já apontados nas referências anteriores foram analisados à luz das especificidades do contexto educacional. De forma resumida: **Ampliação das desigualdades educacionais:** a rápida adoção de tecnologias de IA pode acentuar disparidades existentes, especialmente em contextos com acesso limitado a recursos tecnológicos, infraestrutura e conectividade. **Viés algorítmico e discriminação:** sistemas de IA podem perpetuar ou amplificar preconceitos existentes, resultando em decisões injustas, especialmente em áreas como saúde, justiça e educação. **Privacidade e segurança de dados:** o uso extensivo de dados pessoais por sistemas de IA levanta preocupações sobre a proteção da privacidade e a segurança das informações sensíveis. **Dependência excessiva da IA:** a confiança excessiva em sistemas de IA pode levar à redução do papel dos educadores e ao enfraquecimento das interações humanas essenciais para o processo de ensino-aprendizagem. **Desinformação e manipulação:** a IA pode ser utilizada para criar e disseminar informações falsas ou enganosas, impactando negativamente a opinião pública e os processos democráticos.

Especificamente sobre o uso de IA generativa na educação, a UNESCO alerta para riscos como a distorção de fatos históricos e a propagação de discursos de ódio, conforme destacado no relatório *Generation AI: Navigating the opportunities and risks of artificial intelligence in education* [UNESCO 2024]. Para mitigar esses riscos, a UNESCO recomenda o desenvolvimento de estruturas normativas robustas, políticas educacionais inclusivas e estratégias de capacitação que promovam o uso ético e equitativo da IA na educação. Profissionais da computação e especialistas em informática na educação tem um lugar de fala importante para essa discussão.

Além dos documentos produzidos por estas organizações mundiais, voltadas primordialmente ao desenvolvimento sustentável e a propósitos humanitários e socioambientais, vale mencionar também o *AI Risk Repository*² do Instituto de Tecnologia de Massachusetts (MIT), lançado em 2024 e atualizado em abril de 2025. O *AI Risk Repository* é uma base de dados pública e dinâmica que reúne mais de 1.600 riscos associados à IA, extraídos de 65 estruturas e classificações existentes. Desenvolvido pelo *MIT FutureTech*, o repositório visa consolidar e esclarecer como os riscos da IA são categorizados nas literaturas acadêmica, industrial e política, oferecendo uma estrutura comum para monitoramento e supervisão desses riscos [Slattery et al. 2025].

2.4.2. Gestão de Dados Educacionais

O contexto educacional gera uma ampla gama de informações que se qualificam como dados educacionais. Isso inclui avaliações, informações pessoais dos estudantes, como endereço, idade, histórico médico e dados de contato, bem como registros administrativos, como horários de entrada e saída, frequência, relatórios descritivos sobre comportamento ou desempenho em atividades não avaliativas, e interações gerais no âmbito da comunidade escolar. Esses dados desempenham um papel crucial na gestão educacional, contribuindo para o monitoramento do bem-estar dos estudantes, a organização administrativa e o planejamento pedagógico.

Informações detalhadas sobre o desempenho acadêmico dos estudantes permitem que professores e gestores desenvolvam estratégias pedagógicas individualizadas, atendendo às necessidades específicas de cada aluno e promovendo um progresso educacional mais eficaz [Hernández-de Menéndez et al. 2022]. O uso desses dados é comum em aplicações de IA [Ahmad et al. 2024]. Eles podem apoiar a identificação precoce de dificuldades de aprendizagem, permitindo intervenções direcionadas que ajudam a reduzir as taxas de evasão escolar e a melhorar os indicadores de permanência dos alunos. Dados consolidados podem orientar políticas educacionais e permitir uma alocação mais eficiente de recursos, contribuindo para a redução das desigualdades no acesso à educação. Porém, o uso indevido de dados educacionais pode gerar consequências negativas. Um dos riscos mais significativos é a violação da privacidade dos estudantes, especialmente em contextos onde medidas de proteção ou anonimização são insuficientes. Tal exposição pode comprometer a segurança dos alunos e aumentar sua vulnerabilidade a abusos, discriminação ou práticas de vigilância excessiva [Gašević et al. 2015]. Outro impacto negativo potencial é o reforço das desigualdades educacionais preexistentes. Quando o uso de algoritmos ou ferramentas de análise de dados não é acompanhado de supervisão ética e técnica, pode-se perpetuar vieses sistêmicos, categorizando ou limitando oportunidades com base em padrões históricos ou distorcidos, e desprezando o poder que ações educativas têm de transformar indivíduos.

A gestão de dados educacionais exige uma análise cuidadosa dos parâmetros legais e regulatórios que regem a geração, o armazenamento e a utilização dessas informações por instituições educacionais no Brasil. No entanto, os marcos legais frequentemente não acompanham a velocidade dos avanços tecnológicos, o que gera lacunas que dificultam sua aplicação prática [Almeida and Soares 2022]. A proteção de dados

²<https://airisk.mit.edu>, acessado em maio de 2025

sensíveis representa um dos principais desafios da LGPD, especialmente devido ao seu potencial discriminatório e aos riscos associados ao manuseio inadequado dessas informações [Sarlet and Ruaro 2021]. No contexto educacional, essa preocupação se torna ainda mais premente, dada a vulnerabilidade dos estudantes e a necessidade de resguardar sua privacidade e dignidade.

Embora a LGPD forneça uma estrutura inicial para a proteção de dados pessoais, ainda há ausência de regulamentações específicas voltadas para o setor educacional. Além disso, a Autoridade Nacional de Proteção de Dados (ANPD) ainda não emitiu diretrizes específicas para instituições educacionais, contribuindo para um cenário de incerteza quanto ao uso adequado dos dados educacionais. Essa lacuna regulatória cria um ambiente de insegurança jurídica, no qual as instituições educacionais enfrentam dificuldades para alinhar suas práticas às disposições legais existentes.

A implementação da LGPD nas instituições de ensino exige a adoção de medidas de governança, segurança e transparência de dados, em conformidade com recomendações de adaptação já identificadas em outros contextos institucionais. Estudos sobre a aplicação dessa legislação em diferentes tipos de instituições, como repositórios institucionais [Nascimento and Silva 2023] e escolas de educação infantil [Wendling et al. 2023], demonstram que o cumprimento da LGPD demanda não apenas ajustes técnicos, mas também transformações organizacionais. Esse processo envolve a sensibilização dos atores institucionais, o estabelecimento de protocolos claros de gestão de dados e a oferta de formação contínua aos responsáveis pelo tratamento das informações.

Entre os princípios gerais da LGPD aplicáveis ao contexto educacional, destaca-se o princípio da *transparência*, que exige das instituições a comunicação clara sobre como os dados são coletados, processados e armazenados. Além disso, é exigida a publicação de políticas de privacidade que detalhem os procedimentos e finalidades relacionados ao uso dos dados. A legislação também estabelece a obrigatoriedade da indicação de um Encarregado pelo Tratamento de Dados Pessoais (DPO), responsável por assegurar o cumprimento legal, bem como pela formação contínua dos envolvidos na aplicação das normas de proteção de dados. A constituição de equipes de conformidade também é essencial para a identificação de riscos e o cumprimento dos requisitos legais. Por fim, a LGPD enfatiza a importância da *anonimização* dos dados sempre que possível, como medida de mitigação de riscos e prevenção à exposição de informações pessoais. No entanto, permanecem indefinições, como a classificação de dados sensíveis no contexto educacional e os métodos eficazes de anonimização, especialmente em avaliações de estudantes.

Em contrapartida, países como o Reino Unido e os Estados Unidos possuem estruturas regulatórias mais detalhadas e específicas, que podem servir de referência para o aprimoramento da legislação brasileira. No Reino Unido, o tratamento de dados educacionais é regido pelo UK GDPR e pela *Data Protection Act 2018* [Kingdom 2018], que adaptam o Regulamento Geral de Proteção de Dados da União Europeia (GDPR) [Parliament and Council 2016] ao contexto britânico. Nesse modelo, as escolas devem coletar dados exclusivamente para finalidades educacionais claras e justificáveis, minimizar a coleta de dados e garantir a transparência no uso dessas informações. Além disso, o consentimento é requisito obrigatório para o compartilhamento de dados com terceiros, exceto em casos com base legal adequada.

Nos Estados Unidos, o arcabouço regulatório é definido pelo *Family Educational Rights and Privacy Act* (FERPA) [of Education 1974], que protege a privacidade dos dados educacionais e garante aos pais o direito de acesso e correção dos registros escolares de seus filhos. A FERPA também estabelece limitações rigorosas quanto ao compartilhamento de informações, permitindo acesso apenas a indivíduos autorizados e para finalidades legítimas, como auditorias ou análises pedagógicas. Em casos de violação, a legislação prevê sanções federais às instituições educacionais, reforçando a importância da conformidade e da proteção da privacidade dos dados.

No entanto, mesmo em países com legislações robustas de proteção de dados, como os Estados Unidos e nações europeias, persistem desafios relacionados ao uso indevido dessas informações. Estudos apontam que, nos Estados Unidos, a implementação de políticas de responsabilização escolar baseadas em dados nem sempre resultou em melhoria do desempenho estudantil. Em alguns casos, essas políticas levaram a práticas indesejáveis, como a manipulação de resultados para alcançar metas predefinidas, sem, de fato, melhorar a qualidade da educação [Núñez et al. 2019].

É importante ressaltar que dados produzidos por ferramentas de IA a partir de dados educacionais — como em previsões de desempenho ou recomendações — também devem ser tratados com o mesmo rigor ético e em conformidade com a legislação que se aplica a outros tipos de dados educacionais, garantindo o alinhamento com os marcos regulatórios e as boas práticas de governança de dados. A falta deste conhecimento expõe não apenas as instituições e os profissionais a riscos, mas também os próprios estudantes e suas famílias. Inseguranças desses atores quanto à privacidade e proteção de dados fazem com que instituições relutem em adotar novas tecnologias educacionais [van Haastrecht et al. 2024]. Quanto menor a confiança das instituições em sua capacidade de coletar, armazenar e usar informações pessoais de forma segura, respeitando a legislação vigente, menor será a probabilidade de investimento nas tecnologias que dependem dessas informações.

2.4.2.1. Métodos e técnicas para mitigação de vies

A discussão sobre justiça e responsabilidades no uso da inteligência artificial (IA) tem se intensificado à medida que os sistemas algorítmicos assumem papéis cada vez mais decisivos em processos sociais, educacionais e governamentais. O reconhecimento dos potenciais vieses algorítmicos e de sua reprodução de desigualdades históricas tem sido alvo de críticas e análises profundas por diversos pesquisadores.

A mitigação de injustiças e vieses no uso de IA é um dos grandes desafios contemporâneos. Algoritmos de IA são treinados com dados históricos que muitas vezes refletem desigualdades sociais, raciais e econômicas, podendo reproduzi-las ou mesmo agravá-las [Noble 2018, Eubanks 2018]. Por esse motivo, diversas abordagens técnicas, institucionais e regulatórias têm sido propostas para promover maior justiça e equidade.

Do ponto de vista técnico, estratégias podem ser adotadas para tentar garantir dados diversos e representativos para alimentar os algoritmos de Inteligência Artificial. Técnicas como *pré-processamento* (*pre-processing*), *in-processamento* e *post-processamento* são amplamente utilizadas para mitigar vieses algorítmicos.

O pré-processamento atua diretamente na preparação dos dados, antes do treinamento dos modelos, com o objetivo de reduzir ou eliminar padrões discriminatórios que possam levar a decisões enviesadas [Barocas et al. 2019]. Pode envolver diferentes estratégias, como a *remoção de atributos sensíveis* (por exemplo, raça, gênero ou religião), um *balanceamento de classes* para igualar distribuições de grupos sub-representados, e a *transformação de variáveis correlacionadas* a atributos sensíveis, conhecidas como *proxies discriminatórios*. A exclusão direta de atributos sensíveis nem sempre é suficiente, uma vez que outras variáveis — como código postal, histórico educacional, profissão (própria ou dos pais) — podem indiretamente refletir essas características e perpetuar preconceitos ou situações desvantajosas [Zemel et al. 2013].

A técnica *Learning Fair Representations*, proposta por Zemel et al. (2013), transforma os dados em representações que retêm a informação necessária para a predição, mas minimizam a dependência de atributos sensíveis. Outra abordagem é o *Reweighting*, que ajusta os pesos das instâncias durante o treinamento para compensar disparidades entre diferentes grupos [Kamiran and Calders 2012]. Além disso, algoritmos de pré-processamento podem utilizar métodos estatísticos para verificar a presença de viés nos dados, como *disparate impact* e *equalized odds*, permitindo a identificação de fontes de discriminação antes do uso do modelo preditivo. O uso dessas técnicas contribui para o desenvolvimento de sistemas mais éticos e inclusivos, alinhando-se a princípios como justiça, responsabilidade e respeito aos direitos humanos [Russell and et al. 2022]. Contudo, a aplicação do pré-processamento requer cautela, pois alterações excessivas podem comprometer a acurácia e a utilidade dos modelos.

O *in-processing* inclui modificações no próprio algoritmo de aprendizado de máquina [Barocas et al. 2019], no intuito de promover equidade durante o treinamento dos modelos. Diferentemente das abordagens de pré-processamento, que modificam os dados antes do treinamento, o *in-processing* altera a função de otimização ou o próprio algoritmo de aprendizado para que ele produza previsões mais justas. Isso geralmente é feito por meio da introdução de restrições ou penalidades associadas a medidas de injustiça na função de custo do modelo [Hardt et al. 2016], ou seja, ajustes que penalizam decisões injustas.

Um exemplo é o método *Adversarial Debiasing*, no qual um modelo preditivo principal é treinado em conjunto com um adversário que tenta inferir atributos sensíveis a partir das previsões do modelo [Zhang et al. 2018]. O objetivo é minimizar simultaneamente o erro de predição e a capacidade do adversário de identificar os atributos sensíveis, forçando o modelo a fazer previsões menos enviesadas.

Outra técnica é a incorporação de métricas de equidade diretamente na função de penalidade do modelo, como a *Equalized Odds*, que busca garantir que a taxa de verdadeiros positivos e a taxa de falsos positivos sejam semelhantes entre diferentes grupos [Hardt et al. 2016]. Tais abordagens são especialmente úteis quando há interesse em promover justiça de resultados, e não apenas a justiça no tratamento de cada caso.

O *in-processing* tem como vantagem não requerer alterações nos dados nem etapas de correção posterior do resultado, o que pode preservar melhor a acurácia do modelo. No entanto, pode acarretar aumento da complexidade computacional e a necessidade de algoritmos de treinamento customizados.

O *post-processing* consiste na modificação das saídas do modelo após o processo de treinamento, com o objetivo de torná-las mais justas segundo critérios pré-estabelecidos de equidade [Barocas et al. 2019]. As técnicas de *post-processing* atuam exclusivamente sobre resultados do modelo já treinado. Por isso, são particularmente atrativas em contextos nos quais o modelo original não pode ser modificado, como em sistemas de terceiros.

O método *Equalized Odds Postprocessing* [Hardt et al. 2016], de forma análoga ao *Equalized Odds* do *in-processing*, ajusta as previsões positivas e negativas de um classificador de forma a garantir que a taxa de verdadeiros positivos e a taxa de falsos positivos sejam semelhantes entre diferentes grupos (por exemplo, raça ou gênero). O método utiliza programação linear para redistribuir as saídas sem comprometer significativamente a acurácia. Na técnica *Reject Option Classification* as decisões mais próximas ao limiar de decisão são reavaliadas com base em critérios de equidade. Assim, em casos onde a decisão já teria um alto grau de incerteza, o sistema pode favorecer o grupo historicamente desfavorecido, contribuindo para a redução de discriminação indireta [Kamiran and Calders 2012].

Embora sejam menos invasivas e mais fáceis de implementar, as técnicas de *post-processing* podem gerar inconsistências entre a previsão e a pontuação original do modelo, o que pode ser problemático em contextos sensíveis à rastreabilidade das decisões. Essa abordagem é uma opção valiosa quando não é possível reestruturar os dados ou o algoritmo.

2.4.2.2. Ferramentas para Avaliação e Mitigação de Viés

Ferramentas como o *AI Fairness 360* (AIF360), da IBM, e os *Fairness Indicators*, do Google, têm sido reconhecidas como recursos interessantes para identificar e mitigar vieses em sistemas de IA.

O *AI Fairness 360*³ é uma biblioteca de código aberto desenvolvida pela IBM, projetada para detectar, entender e mitigar vieses em modelos de aprendizado de máquina. Disponível em Python e R, esse *toolkit* oferece: mais de 70 métricas de equidade, como *Statistical Parity Difference*, *Equal Opportunity Difference* e *Disparate Impact*; dez algoritmos de mitigação de viés, abrangendo técnicas de pré-processamento; tutoriais interativos e documentação, aplicáveis a diversos domínios como saúde, educação e finanças.

Os *Fairness Indicators*⁴ desenvolvidos pelo Google, são utilizados para avaliar a equidade de modelos de aprendizado de máquina. Integrados ao TensorFlow, permitem: avaliar métricas de desempenho como *precision* e *recall* em diferentes subgrupos definidos por atributos sensíveis; visualizar disparidades entre grupos e facilitar a identificação de vieses; integração com outras ferramentas.

Embora eficazes para identificação de desigualdades, os *Fairness Indicators* não fornecem mecanismos de mitigação, sendo frequentemente usados em conjunto com ou-

³<https://ai-fairness-360.org/>, acessado em maio de 2025.

⁴<https://github.com/tensorflow/fairness-indicators>, acessado em maio de 2025.

tras bibliotecas como o AIF360.

A *What-If Tool*⁵ é uma ferramenta interativa de visualização desenvolvida pela equipe PAIR (People + AI Research) do Google [Wexler et al. 2020], projetada para facilitar a análise de modelos de aprendizado de máquina sem a necessidade de programação adicional. Integrada ao *TensorBoard* e compatível com o TensorFlow e outros frameworks via API, a ferramenta oferece suporte direto à avaliação de justiça, explicabilidade e desempenho dos modelos.

A *What-If Tool* permite aos usuários explorar o comportamento de modelos preditivos de forma interativa, destacando-se pelas seguintes funcionalidades: **Análise Contra-factual:** Permite modificar manualmente atributos de uma instância de entrada e observar como a previsão do modelo muda, ajudando a identificar variáveis sensíveis. **Comparação entre modelos:** Torna possível comparar o desempenho e as previsões de múltiplos modelos aplicados ao mesmo conjunto de dados, identificando inconsistências. **Visualização de Métricas de Justiça:** Oferece métricas como *Statistical Parity*, *Equal Opportunity* e *Disparate Impact*, facilitando a avaliação de justiça algorítmica entre subgrupos. **Agrupamento por atributos sensíveis:** O desempenho do modelo pode ser analisado por subgrupos definidos por atributos como raça, gênero ou idade, promovendo uma visão granular do comportamento preditivo.

Apesar de ser uma excelente ferramenta interativa de visualização, o que ajuda a perceber vieses e analisar suas possíveis causas, a *What-If Tool* não implementa algoritmos de mitigação. Uma vantagem dessa ferramenta é a possibilidade de uso das funcionalidades interativas e visuais por profissionais de diversas áreas, contribuindo para a democratização da análise crítica de modelos de aprendizado de máquina.

As ferramentas apresentadas, bem como outras já existentes ou ainda em construção, são importantes peças para a formação de um acervo de recursos para o desenvolvimento ético de sistemas inteligentes, possibilitando uma abordagem mais sistemática à mitigação de vieses. Contudo, é imprescindível ressaltar que ferramentas técnicas não enfrentam sozinhas as causas sociais e históricas dos vieses [Birhane 2021]. Decisões sobre o que é justo dependem de contextos socioculturais, e exigem participação ativa de grupos afetados. Uma abordagem sociotécnica é indicada, combinando análises técnicas com consultas públicas, auditorias participativas e governança inclusiva [Cath 2018].

A necessidade de atribuição de responsabilidades legais em casos de falhas ou discriminação algorítmica é um ponto central. Documentos internacionais como as diretrizes da [UNESCO 2021b] e o *White Paper* da [European Commission 2020] delineiam princípios éticos fundamentais, incluindo justiça, responsabilidade, transparência e não discriminação.

A implementação de mecanismos de responsabilização, como a nomeação de agentes de proteção de dados (DPOs), auditorias independentes e canais para denúncias de danos algorítmicos, reforça a responsabilidade das instituições pelo uso ético da IA. É importante também garantir diversidade nas equipes de desenvolvimento, incluindo diferentes origens sociais, de gênero e culturais, como forma de reduzir cegueiras epistêmicas e ampliar a compreensão dos impactos sociais dos sistemas de IA

⁵<https://pair-code.github.io/what-if-tool/>, acessado em maio de 2025.

[Buolamwini and Gebru 2018]. Portanto, a mitigação de injustiças no uso da IA não depende apenas de soluções técnicas, mas de um ecossistema que promova responsabilidade, participação democrática e justiça social.

Referências

- [Abu-Mostafa et al. 2012] Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning From Data: A Short Course*. AMLBook.
- [Ahmad et al. 2024] Ahmad, K., Iqbal, W., El-Hassan, A., Qadir, J., Benhaddou, D., Ayyash, M., and Al-Fuqaha, A. (2024). Data-Driven Artificial Intelligence in Education: A Comprehensive Review. *IEEE Transactions on Learning Technologies*, 17:12–31.
- [Almeida and Soares 2022] Almeida, S. D. C. D. D. and Soares, T. A. (2022). Os impactos da lei geral de proteção de dados-igpd no cenário digital. *Perspectivas em Ciência da Informação*, 27:26–45.
- [Barbetta et al. 2010] Barbetta, P. A., Reis, M. M., and Bornia, A. C. (2010). *Estatística Para Cursos de Engenharia e Informática*. Atlas, São Paulo, 3 edition.
- [Barocas et al. 2019] Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. Disponível em: <https://fairmlbook.org>.
- [Barros et al. 2025] Barros, J., Moraes, L. O., Oliveira, F., and Delgado, C. A. D. M. (2025). Large language models generating feedback for students of introductory programming courses. In *Proceedings of the Artificial Intelligence in Education Conference*, Palermo, Italy. Springer. A ser publicado.
- [Becker et al. 2022] Becker, B. A., Denny, P., Finnie-Ansley, J., Luxton-Reilly, A., Prather, J., and Santos, E. A. (2022). Programming is hard - or at least it used to be: Educational opportunities and challenges of ai code generation. *Arxiv*. Available at <https://arxiv.org/abs/2212.01020>.
- [Belle and Papantonis 2021] Belle, V. and Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data*, 4.
- [Birhane 2021] Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2).
- [Brown et al. 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *arXiv*. Available at <https://arxiv.org/abs/2005.14165v4>.

- [Buolamwini and Gebru 2018] Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT* 2018)*, pages 77–91. PMLR.
- [Campbell et al. 2000] Campbell, M., Jr., A. J. H., and Hsiung Hsu, F. (2000). Deep blue. *IBM Journal of Research and Development*, 44(3):213–228.
- [Cath 2018] Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180080.
- [Chen et al. 2024] Chen, B., Zhang, Z., Langrené, N., and Zhu, S. (2024). Unleashing the potential of prompt engineering in large language models: A comprehensive review. *arXiv*. Available at <https://arxiv.org/abs/2310.14735v5>.
- [Cunha 1980] Cunha, L. A. (1980). *Educação e desenvolvimento social no Brasil*. Francisco Alves, Rio de Janeiro.
- [de Carvalho et al. 2024] de Carvalho, A. C. P. L. F., Menezes, A. G., and Bonidia, R. P. (2024). *Ciência de Dados - Fundamentos e Aplicações*. LTC.
- [de Oliveira F. Moraes 2016] de Oliveira F. Moraes, L. (2016). Classificação de linfomas utilizando uma abordagem baseada em Árvores de decisão. Mestrado, Universidade Federal do Rio de Janeiro (UFRJ).
- [de Sousa 2025] de Sousa, J. B. (2025). Modelos de linguagem em larga escala na geração de feedbacks para alunos de cursos introdutórios de programação da ufrj. Trabalho de conclusão de curso, Universidade Federal do Rio de Janeiro (UFRJ), Escola Politécnica, Rio de Janeiro.
- [Denny et al. 2024] Denny, P., Prather, J., Becker, B. A., Finnie-Ansley, J., Hellas, A., Leinonen, J., Luxton-Reilly, A., Reeves, B. N., Santos, E. A., and Sarsa, S. (2024). Computing education in the era of generative ai. *Communications of the ACM*, 67(2):56–67.
- [Dobson and Barnett 2008] Dobson, A. and Barnett, A. (2008). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC Texts in Statistical Science. Chapman and Hall/CRC.
- [Eubanks 2018] Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin’s Press.
- [European Commission 2020] European Commission (2020). White paper on artificial intelligence: A european approach to excellence and trust.
- [Freire 1970] Freire, P. (1970). *Pedagogia do oprimido*. Paz e Terra, Rio de Janeiro.
- [Gao et al. 2024] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2024). Retrieval-augmented generation for large language models: A survey.

- [Gašević et al. 2015] Gašević, D., Dawson, S., and Siemens, G. (2015). Let’s not forget: Learning analytics are about learning. *TechTrends*, 59(1):64–71.
- [Hardt et al. 2016] Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3315–3323.
- [Hassany et al. 2024] Hassany, M., Brusilovsky, P., Jiaze, K., Akhsueyinoglu, K., and Narayanan, A. B. L. (2024). Human-ai co-creation of worked examples for programming classes. In *Workshops at the International Conference on Intelligent User Interfaces (IUI) 2024*.
- [Hellas et al. 2018] Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., and Liao, S. N. (2018). Predicting academic performance: a systematic literature review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, ITiCSE 2018 Companion*, page 175–199, New York, NY, USA. ACM.
- [Hernández-de Menéndez et al. 2022] Hernández-de Menéndez, M., Morales-Menendez, R., Escobar, C. A., and Ramírez Mendoza, R. A. (2022). Learning analytics: state of the art. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 16(3):1209–1230.
- [High-Level Advisory Body on AI 2024] High-Level Advisory Body on AI (2024). Governing ai for humanity: Report of the high-level advisory body on artificial intelligence. Acesso em: 22 maio 2025.
- [INEP 2023] INEP (2023). Sistema de avaliação da educação básica (saeb). <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb>. Acesso em: 2 jun. 2025.
- [Izbicki and dos Santos 2020] Izbicki, R. and dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*.
- [James et al. 2023] James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer, 1st edition.
- [Kamiran and Calders 2012] Kamiran, F. and Calders, T. (2012). Decision theory for discrimination-aware classification. *IEEE Transactions on Knowledge and Data Engineering*, 25(5):833–846.
- [Kasneci et al. 2023] Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., and Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

- [Kazemitabaar et al. 2024] Kazemitabaar, M., Hou, X., Henley, A., Ericson, B. J., Weintrop, D., and Grossman, T. (2024). How novices use llm-based code generators to solve cs1 coding tasks in a self-paced learning environment. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*, Koli Calling '23, New York, NY, USA. Association for Computing Machinery.
- [Kingdom 2018] Kingdom, U. (2018). Data protection act 2018. Accessed: 2025-01-20.
- [Kojima et al. 2022] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.
- [Libâneo 1994] Libâneo, J. C. (1994). *Pedagogia e pedagogos, para quê?* Cortez, São Paulo.
- [Liffiton et al. 2024] Liffiton, M., Sheese, B. E., Savelka, J., and Denny, P. (2024). Co-dehelp: Using large language models with guardrails for scalable support in programming classes. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*, Koli Calling '23, New York, NY, USA. ACM.
- [Luckesi 2011] Luckesi, C. C. (2011). *Avaliação da aprendizagem escolar: estudos e proposições*. Cortez, São Paulo, 22 edition. Capítulos sobre políticas públicas de avaliação, incluindo o SAEB.
- [Lundberg and Lee 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- [McCarthy et al. 1955] McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (1955). A proposal for the dartmouth summer research project on artificial intelligence. Memorandum, Dartmouth College. Disponível em: <https://www.jmc.stanford.edu/articles/dartmouth/dartmouth.pdf>.
- [Ministério da Educação do Brasil 2017] Ministério da Educação do Brasil (2017). Base nacional comum curricular. Acesso em: 22 maio 2025.
- [Murphy 2012] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, Massachusetts, illustrated edition.
- [Nam et al. 2024] Nam, D., Macvean, A., Hellendoorn, V., Vasilescu, B., and Myers, B. (2024). Using an llm to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, ICSE '24, New York, NY, USA. Association for Computing Machinery.
- [Nascimento and Silva 2023] Nascimento, B. L. C. and Silva, E. M. (2023). Lei geral de proteção de dados (lgpd) e repositórios institucionais: reflexões e adequações. *Em Questão*, 29:e-127314.

- [Newell et al. 1956] Newell, A., Shaw, J. C., and Simon, H. A. (1956). The logic theory machine – a complex information processing system. In *Proceedings of the Western Joint Computer Conference*, pages 218–230. IRE-AIEE-ACM.
- [Noble 2018] Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.
- [Núñez et al. 2019] Núñez, C. P., Koslinski, M. C., and Fernández, S. J. (2019). Políticas de incentivo ao uso de dados educacionais: experiências no contexto internacional e brasileiro. *Jornal de Políticas Educacionais*, 13.
- [of Education 1974] of Education, U. D. (1974). Family educational rights and privacy act (ferpa). 20 U.S.C. 1232g; 34 CFR Part 99.
- [Okoye and Hosseini 2024] Okoye, K. and Hosseini, S. (2024). *Correlation Tests in R: Pearson Cor, Kendall’s Tau, and Spearman’s Rho*, pages 247–277. Springer Nature Singapore, Singapore.
- [Parliament and Council 2016] Parliament, E. and Council (2016). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 – general data protection regulation. Accessed: 2025-01-20.
- [Peres 2023] Peres, R. S. (2023). Grandes modelos de linguagem na resolução de questões de vestibular: O caso dos institutos militares brasileiros. Mestrado, Universidade Federal do Estado do Rio de Janeiro (UNIRIO).
- [Piaget 1973] Piaget, J. (1973). *A epistemologia genética*. Bertrand Brasil, Rio de Janeiro. Tradução de E. F. Souza. Obra original: *L’*épistémologie génétique**, 1950.
- [Provost and Fawcett 2016] Provost, F. and Fawcett, T. (2016). *Data Science Para Negócios*. Editora Alta Books, edição: 1 edition.
- [Qiao et al. 2023] Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., and Chen, H. (2023). Reasoning with language model prompting: A survey. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.
- [Ribeiro et al. 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.
- [Rodrigues et al. 2025] Rodrigues, H. S., Moraes, L. O., Garcia, A. C. B., Alvares, R. V., dos Santos, R. P., and Delgado, C. (2025). Investigating artificial intelligence algorithms to predict college students’ academic performance: A systematic mapping study. In *Anais do Workshop de Educação em Computação*, Maceió, Brasil. A ser publicado.

- [Romero and Ventura 2020] Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3):e1355.
- [Russell and et al. 2022] Russell, S. and et al. (2022). Human rights in the age of ai. *UNESCO Report*.
- [Sahoo et al. 2024] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv*. Available at <https://arxiv.org/abs/2402.07927v1>.
- [Sarlet and Ruaro 2021] Sarlet, G. B. S. and Ruaro, R. L. (2021). A proteção de dados sensíveis no sistema normativo brasileiro sob o enfoque da lei geral de proteção de dados (lgpd) – l. 13.709/2018. *Revista de Direitos Fundamentais & Democracia*, 26(2):81–106.
- [Skinner 1953] Skinner, B. F. (1953). *Science and Human Behavior*. Macmillan, New York.
- [Slattery et al. 2025] Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., and Thompson, N. (2025). The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence.
- [Sun et al. 2024] Sun, S., Yuan, R., Cao, Z., Li, W., and Liu, P. (2024). Prompt chaining or stepwise prompt? refinement in text summarization. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7551–7558, Bangkok, Thailand. Association for Computational Linguistics.
- [Sutton and Barto 2018] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, Massachusetts, 2nd edition.
- [UNESCO 2021a] UNESCO (2021a). *Artificial Intelligence and Education: Guidance for Policy-makers*. United Nations Educational, Scientific and Cultural Organization, Paris.
- [UNESCO 2021b] UNESCO (2021b). Recommendation on the ethics of artificial intelligence. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.
- [UNESCO 2023] UNESCO (2023). *Governing Artificial Intelligence in Education: From Principles to Practice*. United Nations Educational, Scientific and Cultural Organization, Paris.
- [UNESCO 2024] UNESCO (2024). Generation ai: Navigating the opportunities and risks of artificial intelligence in education.

- [van Haastrecht et al. 2024] van Haastrecht, M., Brinkhuis, M., and Spruit, M. (2024). Federated learning analytics: Investigating the privacy-performance trade-off in machine learning for educational analytics. In Olney, A. M., Chounta, I.-A., Liu, Z., Santos, O. C., and Bittencourt, I. I., editors, *Artificial Intelligence in Education*, pages 62–74, Cham. Springer Nature Switzerland.
- [Vasconcelos and Clua 2017] Vasconcelos, C. N. and Clua, E. W. G. (2017). Deep learning - teoria e prática. In Delicato, F. C., Pires, P. F., and Silveira, I. F., editors, *Jornadas de Atualização em Informática 2017*, chapter 6. Sociedade Brasileira de Computação.
- [Vygotsky 1984] Vygotsky, L. S. (1984). *A formação social da mente: o desenvolvimento dos processos psicológicos superiores*. Livros Técnicos e Científicos, São Paulo. Tradução de José Cipolla Neto, Luiz Silveira Menna Barreto e Solange Castro Afeche. Obra original de 1930-1934.
- [Wang et al. 2023] Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., and Lim, E.-P. (2023). Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.
- [Wei et al. 2022] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- [Weizenbaum 1966] Weizenbaum, J. (1966). Eliza – a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.
- [Wendling et al. 2023] Wendling, M. I. et al. (2023). Diagnóstico do nível de maturidade da aplicação da IgpD nas escolas de educação infantil da rede municipal de passo fundo. *Revista Brasileira de Políticas Públicas e Direito da Educação*, 15(3):45–62.
- [Wexler et al. 2020] Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., and Wilson, J. (2020). The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65.
- [Zemel et al. 2013] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, pages 325–333. PMLR.
- [Zhang et al. 2018] Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unfairness in machine learning with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.