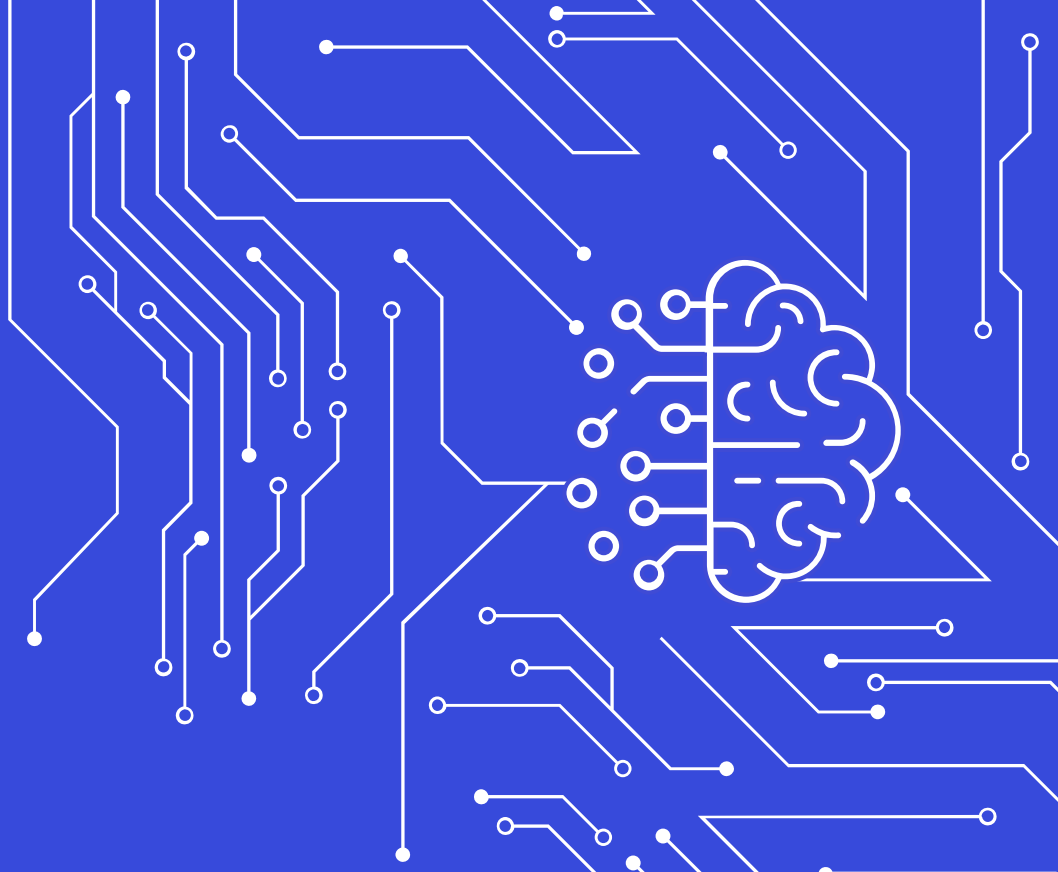


Uso responsável da IA para as políticas públicas:
Manual de Ciência de Dados

Felipe González
Teresa Ortiz
Roberto Sánchez

OECD BID

Uso responsável da IA para as políticas públicas:

Manual de Ciência de Dados

Felipe González, Teresa Ortiz y Roberto Sánchez Ávalos

https://www.iadb.org/

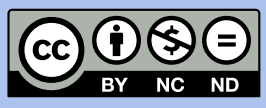
Copyright © 2020 Banco Interamericano de Desenvolvimento. Esta obra está licenciada sob uma licença Creative Commons IGO 3.0 Atribuição-NãoComercial-SemDerivações (CC BY-NC-ND 3.0 IGO) (http://creativecommons.org/licenses/by-nc-nd/3.0/igo/legalcode) e pode ser reproduzida com atribuição ao BID e para qualquer finalidade não comercial. Nenhum trabalho derivado é permitido.

Qualquer controvérsia relativa à utilização de obras do BID que não possa ser resolvida amigavelmente será submetida à arbitragem em conformidade com as regras da UNCITRAL. O uso do nome do BID para qualquer outra finalidade que não a atribuição, bem como a utilização do logotipo do BID serão objetos de um contrato por escrito de licença separado entre o BID e o usuário e não está autorizado como parte desta licença CC-IGO.

Note-se que o link fornecido acima inclui termos e condições adicionais da licença.

As opiniões expressas nesta publicação são de responsabilidade dos autores e não refletem necessariamente a posição do Banco Interamericano de Desenvolvimento, de sua Diretoria Executiva, ou dos países que eles representam.

Tradução de Mariana Fagundez, otraspalabras.com.



Banco Interamericano de Desenvolvimento (BID) - Setor Social

O Setor Social (SCL) é composto por uma equipe multidisciplinar que atua com a convicção de que investir nas pessoas permite melhorar suas vidas e superar os desafios de desenvolvimento na América Latina e no Caribe. Junto com os países da região, o Setor Social formula soluções de políticas públicas para reduzir a pobreza e melhorar a oferta de serviços de educação, trabalho, proteção social e saúde. O objetivo é construir uma região mais produtiva onde prevaleça a igualdade de oportunidades entre homens e mulheres, bem como uma maior inclusão dos grupos mais vulneráveis. www.iadb.org/en/about-us/departments/scl

Banco Interamericano de Desenvolvimento (BID) – BID Lab

O BID Lab é o laboratório de inovação do Grupo BID. Nele, financiamento, conhecimento e conexões são mobilizados para catalisar a inovação voltada para a inclusão na América Latina e no Caribe. Para o BID Lab, a inovação é uma ferramenta poderosa que pode transformar a região ao criar novas oportunidades para as populações em situação de vulnerabilidade devido às condições econômicas e socioambientais nas quais se encontram. https://bidlab.org/

Organização para a Cooperação e o Desenvolvimento Econômico

A OCDE é uma organização internacional que atua para construir políticas melhores para uma vida melhor. O nosso objetivo é desenvolver políticas que promovam a prosperidade, a igualdade, as oportunidades e o bem-estar para todos.

Juntamente com os governos, os responsáveis pelas políticas e os cidadãos, trabalhamos para estabelecer padrões internacionais baseados em evidências e encontrar soluções para uma série de desafios socioeconômicos e ambientais. Um exemplo de estabelecimento de normas são os Princípios da OCDE para a Inteligência Artificial (IA), que são os primeiros princípios desse tipo adotados pelos governos. Esses princípios promovem uma IA inovadora e confiável que respeita os direitos humanos e os valores democráticos. https://ia-latam.com/portfolio/principios-de-la-ocde-sobre-ia/

OECD.AI Policy Observatory

O Observatório OECD.AI é um hub de políticas públicas para a IA. Ajuda os países a promoverem, nutrirem e supervisionarem o desenvolvimento e o uso confiável da IA.

O OECD.AI é utilizado por responsáveis pelas políticas e outras partes interessadas em mais de 170 países. Tem-se tornado um renomado centro de evidências, debates e orientações voltadas para as políticas, sendo apoiado por fortes parcerias com atores de todos os grupos de interesse e outras organizações internacionais. Oferece análises da IA baseadas em evidências.

O OECD.AI é uma fonte única de dados e visualizações em tempo real sobre a evolução da IA. Também contém um banco de dados de políticas de IA de mais de 60 países que permite que os governos comparem respostas políticas e desenvolvam boas práticas. O OECD.AI mede o nosso progresso coletivo rumo a uma IA confiável; sua rede de especialistas e o blog AI Wonk facilitam os debates colaborativos sobre políticas de IA.

Outros trabalhos da OCDE relacionados à aplicação específica da IA no setor público podem ser acessados por meio do Observatório de Inovação no Setor Público (OPSI, na sigla em inglês) da OCDE (https://oecd-opsi.org/), que fornece uma visão geral das medidas de IA aplicadas no setor público, inclusive as relacionadas ao desenvolvimento de políticas de governança da IA e à formulação e prestação de serviços públicos.

Iniciativa fAIr LAC

O Banco Interamericano de Desenvolvimento (BID), em colaboração com parceiros e aliados estratégicos, lidera a iniciativa fAIr LAC, por meio da qual busca promover a adoção responsável da Inteligência Artificial (IA) e dos sistemas de suporte à decisão. O objetivo é melhorar a prestação de serviços sociais e criar oportunidades de desenvolvimento a fim de reduzir a desigualdade social. Este manual faz parte de uma série de ferramentas e documentos elaborados para orientar os responsáveis pela formulação de políticas públicas e suas equipes técnicas na mitigação dos desafios inerentes aos sistemas de suporte à decisão baseados em IA e na promoção de sua adoção responsável (Cabrol et al., 2020).

Agradecimentos

Por seu tempo e valiosas contribuições, expressamos os nossos agradecimentos especiais a Cristina Pombo, coordenadora da iniciativa fAIr LAC do BID, e ao Prof. Ricardo Baeza-Yates, Diretor de Ciência de Dados da Northeastern University, Silicon Valley Campus, e membro do Grupo de Especialistas da fAIr LAC. Os autores também agradecem as contribuições recebidas de Karine Perset, Administradora do OECD.AI, e Luis Aranda, Analista de Políticas do OECD.AI.

Também expressamos a nossa gratidão pelo apoio prestado e pelos comentários recebidos de Luis Tejerina, Elena Arias Ortiz, Natalia González Alarcón, Tetsuro Narita, Constanza Gómez-Mont, Daniel Korn, Ulises Cortés, José Antonio Guridi Bustos, César Rosales e Sofía Trejo.

ÍNDICE

SUMÁRIO EXECUTIVO

Das finanças e seguros à agricultura e transporte, a Inteligência Artificial (IA) está sendo difundida rapidamente por todos os setores, criando oportunidades, mas também gerando novos problemas de políticas públicas. No setor público, a IA promete aumentar a produtividade e aprimorar a qualidade dos serviços públicos. Ao analisar a atividade das mídias sociais em tempo real, os responsáveis pelas políticas públicas podem, por exemplo, aproveitar os sistemas de IA para obter uma avaliação mais precisa e baseada em evidências dos problemas e necessidades sociais mais urgentes. Os resultados e previsões dos sistemas de IA podem servir de base para a formulação, implementação e avaliação de políticas.

Nesse contexto, governos de todo o mundo estão adquirindo o conhecimento técnico necessário para aproveitar o poder da IA em prol do desenvolvimento de políticas públicas. No entanto, como essas políticas baseadas em IA podem ter um impacto significativo na vida e no bem-estar das pessoas, é necessária uma abordagem sistêmica para garantir que existam salvaguardas adequadas para aproveitar as oportunidades oferecidas pelo uso desses sistemas por parte das equipes de políticas públicas, bem como para enfrentar os desafios decorrentes dele.

Usando o ciclo de vida dos sistemas de IA como base de análise, este conjunto de ferramentas fornece orientações técnicas às equipes de políticas públicas que desejam utilizar tecnologias de IA para melhorar seus processos de tomada de decisão e seus resultados. Para cada fase do ciclo de vida do sistema de IA (“conceitualização e formulação”, “coleta e processamento de dados”, “desenvolvimento e validação de modelos” e “uso e monitoramento”), o conjunto de ferramentas identifica os desafios de uso da IA mais comuns em contextos de políticas públicas e delineia os mecanismos práticos para detectar e mitigar esses desafios.

Os responsáveis pelas políticas e suas equipes técnicas devem responsabilizar-se pelo bom funcionamento de um sistema de IA em cada fase de seu ciclo de vida. Nesse sentido, um dos capítulos da caixa de ferramentas foca-se em analisar os problemas relacionados à responsabilidade no uso da IA para as políticas públicas e a delinear mecanismos práticos para solucioná-los.

Para cumprir seu objetivo de promover o uso responsável da IA para a elaboração de políticas públicas, cada seção do conjunto de ferramentas inclui listas de verificação que ajudam a orientar a aplicação prática. Também são fornecidas uma ferramenta de “perfil de dados” e uma “ficha do modelo” para ajudar a avaliar os problemas nos dados e documentar as características de um sistema de IA, as suposições feitas e as medidas de mitigação de risco aplicadas ao longo do ciclo de vida. Além disso, o conjunto de ferramentas oferece uma seção com um caderno de trabalho que apresenta exemplos práticos de alguns desafios e estratégias de mitigação tratados no relatório, bem como o código relevante para implementá-los usando R ou outras linguagens de programação.

Por meio da iniciativa fAIr LAC e do Observatório de Políticas de IA da OCDE, o BID e a OCDE firmaram uma parceria para levar o debate sobre políticas de IA dos princípios de alto nível à prática e à implementação. Este conjunto de ferramentas é um passo concreto nessa direção.

Por que este manual?

Apesar de haver um número significativo de princípios que buscam uma IA ética, eles fornecem apenas orientações de alto nível sobre o que deve ou não ser feito em seu desenvolvimento, havendo muito pouca clareza sobre quais são as boas práticas para colocá-la em andamento (Vayena , 2019). O objetivo deste manual é fornecer essas recomendações e boas práticas técnicas para evitar resultados contrários (muitas vezes inesperados) aos objetivos dos tomadores de decisão. Esses propósitos são variados: podem referir-se a consequências indesejáveis do ponto de vista dos tomadores de decisão, desperdício de recursos devido a um direcionamento inadequado ou qualquer outro objetivo que o tomador de decisão esteja buscando atingir.[1]

A quem este manual se dirige?

Este manual é destinado a equipes técnicas que trabalham na aplicação de algoritmos de aprendizado de máquina para políticas públicas. No entanto, todos os desafios que ele abrange são comuns a qualquer aplicação dessa tecnologia. Presume-se que o leitor tenha conhecimentos básicos de estatística e programação, embora, quando os conceitos são mencionados, sejam incluídas breves descrições e compartilhada uma bibliografia adicional. O manual inclui cadernos de trabalho com vários exemplos dos desafios e as soluções explicadas. Diferentes tipos de modelo (linear, baseado em árvore e outros) e diferentes implementações (R, Keras, Xgboost) são usados para mostrar que esses problemas surgem independentemente da escolha de ferramentas específicas. Embora os códigos e exemplos tenham sido desenvolvidos em R, todos os tópicos e metodologias aplicados e descritos neste manual podem ser implementados em qualquer outra linguagem de programação.[2]

1 Este manual não pretende regular nem explicar quais devem ser as finalidades e objetivos das organizações e atores que tomam as decisões.

2 Todo o material deste documento pode ser reproduzido conforme as instruções encontradas no repositório https://github.com/EL-BID/Manual-IA-Responsable, que contém um Dockerfile que descreve as dependências de infraestrutura para sua replicação. A linguagem de programação R e os seguintes pacotes são usados: tidyverse, recipes, themis, rsample, parsnip, yardstick, workflows, tune, knitr, patchwork.

GLOSSÁRIO

- **Aprendizado de máquina:** conjunto de técnicas que permitem que um sistema aprenda comportamentos automaticamente por meio de padrões e inferências, não de instruções explícitas ou simbólicas inseridas por seres humanos (OCDE, 2019c).
- **Atributo protegido:** uma característica ou variável protegida é aquela em que queremos que certos critérios de equidade sejam cumpridos nas previsões. Em um conjunto de dados, podemos ter mais de uma variável protegida, como idade, gênero, raça etc.
- **Critério de justiça algorítmica**: representação matemática de uma definição específica de justiça que é incorporada ao processo de seleção e ajuste do modelo. É importante levar em consideração que essas definições podem ser excludentes, ou seja, satisfazer uma pode resultar em não satisfazer as outras (Verma & Rubin, 2018).
- **Estrutura preditiva:** é usada para falar em geral sobre os tipos de modelo usados para fazer previsões (lineares, florestas aleatórias, redes neurais), as características usadas e como o modelo as utiliza (interações, transformações não lineares).
- **Garantias probabilísticas:** em amostras randomizadas, é possível, em certas hipóteses, caracterizar o comportamento dos estimadores e procedimentos (com alta probabilidade). Por exemplo, um intervalo de confiança de 95% para as métricas de desempenho que contém o valor real a ser observado.
- **Inteligência artificial:** sistema computacional capaz de influenciar o ambiente e produzir um resultado (previsões, recomendações ou decisões) para um determinado conjunto de objetivos. Ela usa dados e contribuições de fontes humanas ou sensores para (i) perceber ambientes reais e/ou virtuais; (ii) abstrair essas percepções em modelos por meio de análises automatizadas (por exemplo, aprendizado de máquina) ou manuais; e (iii) usar as inferências do modelo para formular resultados. Os sistemas de IA são projetados para funcionar com diferentes níveis de autonomia (adaptado de OCDE 2019c).
- **Inequidade algorítmica:** falha técnica nos modelos que gera uma disparidade de resultados para grupos protegidos, que devem ser avaliados de acordo com a definição de justiça algorítmica determinada em um ponto anterior (pode ser mais de uma).
- **População-alvo:** conjunto de elementos nos quais se pretende intervir (pessoas, lares, áreas geográficas etc.). Os modelos são construídos com o objetivo de ser aplicados à população-alvo.
- **Sistemas de suporte à decisão:** relacionados ao conceito de inteligência assistida ou aumentada, são usados para descrever os sistemas nos quais as informações geradas pelos modelos de aprendizado de máquina são utilizadas como base para a tomada de decisão realizada por seres humanos.
- **Sistemas de tomada de decisão:** relacionados ao conceito de inteligência automatizada e autônoma. As decisões finais e as ações decorrentes delas são tomadas sem intervenção humana direta. Ou seja, o sistema passa a realizar tarefas que antes eram executadas por um ser humano.
- **Subpopulações de interesse ou subpopulações protegidas:** são subpopulações da população-alvo para as quais se deseja realizar avaliações específicas do desempenho de estimativas ou dos modelos.

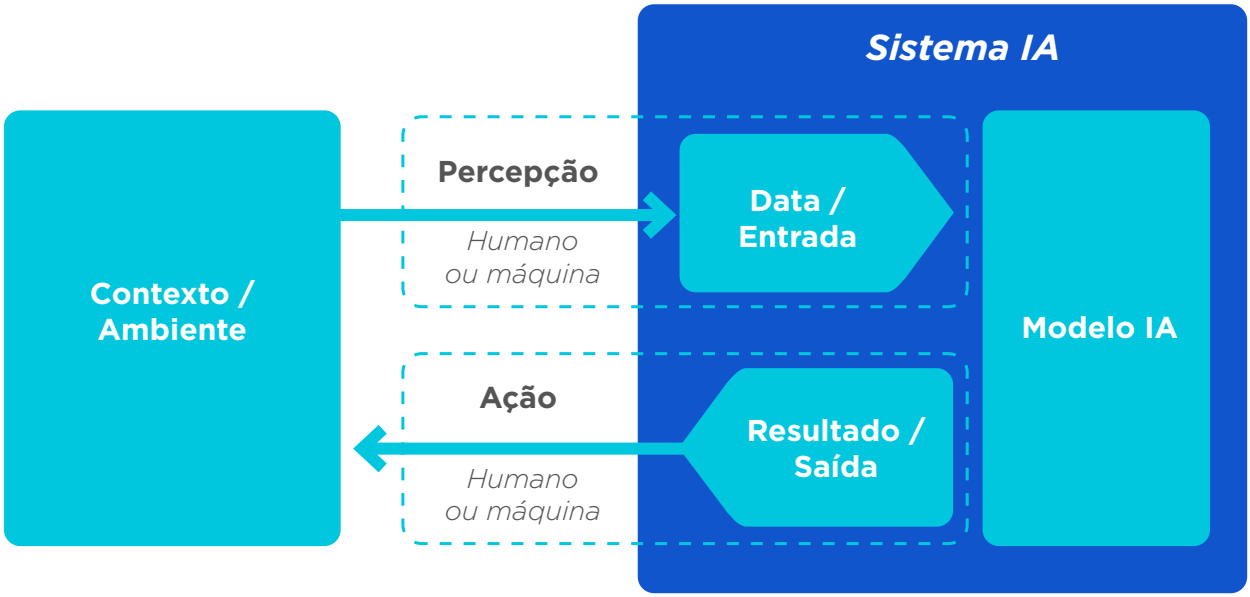
INTRODUÇÃO

Os métodos de aprendizado de máquina (que, para resumir neste documento, chamamos de ML, sigla em inglês de Machine Learning), como um subconjunto do que se conhece como inteligência artificial, fazem-se cada vez mais necessários e são usados pelos tomadores de decisão para fundamentar ações ou intervenções em vários contextos, de negócios a políticas públicas. Na prática, esses métodos têm sido usados com diferentes graus de sucesso, e, em decorrência disso, tem aumentado a preocupação sobre como entender o desempenho e a influência positiva ou negativa desses métodos na sociedade (Barocas and Selbst 2016; Suresh and Guttag 2019).

Machine Learning (ML) e sistemas de suporte à decisão/tomada de decisão

A Organização para a Cooperação e Desenvolvimento Econômico (OCDE) descreve a IA como um sistema computacional capaz de influenciar o ambiente e produzir um resultado (previsões, recomendações ou decisões) para um determinado conjunto de objetivos. Ela usa dados e contribuições de fontes humanas ou sensores para (i) perceber ambientes reais e/ou virtuais; (ii) abstrair essas percepções em modelos por meio de análises automatizadas (por exemplo, aprendizado de máquina) ou manuais; e (iii) usar as inferências do modelo para formular resultados. Os sistemas de IA são projetados para funcionar com diferentes níveis de autonomia (adaptado de OECD 2019c).

Figura 1. Visão conceitual de um sistema de IA



Embora os métodos de aprendizado de máquina não sejam o único tipo de algoritmo que pode ser usado pelos sistemas de IA, foram os que mais cresceram nos últimos anos. Trata-se de um conjunto de técnicas que permitem que um sistema aprenda comportamentos automaticamente por meio de padrões e inferências, não de instruções explícitas ou simbólicas inseridas por seres humanos (OECD 2019c).

Este manual analisa os desafios mais comuns no uso das tecnologias de aprendizado de máquina para o suporte à decisão e a tomada de decisão, entre eles a detecção e a mitigação de erros e vieses, e a avaliação de resultados indesejáveis por uma empresa, uma instituição do setor público ou a sociedade.

São considerados dois arquétipos de inclusão do aprendizado de máquina no processo de tomada de decisão:[3]

1. **Sistemas de suporte à decisão:** relacionados ao conceito de inteligência assistida ou aumentada, são usados para descrever os sistemas nos quais as informações geradas pelos modelos de aprendizado de máquina são utilizadas como base para a tomada de decisão realizada por seres humanos.

2. **Sistemas de tomada de decisão:** relacionados ao conceito de inteligência automatizada e autônoma. As decisões finais e as ações decorrentes delas são tomadas sem intervenção humana direta. Isso significa que o sistema passa a realizar tarefas que antes eram executadas por seres humanos. Em muitos contextos, usa-se a sigla ADM para denominar esses sistemas, uma abreviação do inglês *Automated Decision Making*.

Para o desenvolvimento de um sistema de suporte à decisão/tomada de decisão bem-sucedido baseado em aprendizado de máquina, deve-se considerar que existe uma grande variedade de técnicas, conhecimento especializado do assunto e modelagem em geral. Este manual não pretende abordar métodos específicos de aprendizado de máquina nem processos específicos de ajuste de hiperparâmetros (veja, por exemplo, Hastie, Tibshirani e Friedman, 2017; Kuhn e Johnson, 2013; Gelman e Hill, 2006), mas manter o foco em sua avaliação e nos desafios mais importantes que os sistemas compartilham, independentemente do tipo de algoritmo ou tecnologia utilizada.

Por outro lado, a avaliação de um sistema de aprendizado não faz sentido fora de seu contexto. Perguntas como "Qual é a taxa de erro apropriada?" ou "Quais vieses são inaceitáveis?", entre outras, só podem ser consideradas e respondidas dentro do contexto específico de sua aplicação, dos propósitos e motivações dos tomadores de decisão, bem como do risco que apresenta para os usuários finais. Ou seja, muitos dos critérios técnicos devem ser entendidos à luz do problema específico. Os sistemas de suporte à decisão/tomada de decisão nunca são perfeitos, mas, se seus vieses e limitações forem conhecidos, mesmo um sistema com baixa precisão pode ser útil e usado com responsabilidade. Caso contrário, se suas limitações não forem compreendidas, dispor de um sistema com altas métricas de avaliação não elimina o risco de uso irresponsável.

Objetivos

- Este manual foca-se no subconjunto de desafios relacionados aos processos técnicos ao longo do ciclo de vida dos sistemas de IA usados para o suporte à decisão e a tomada de decisão referentes a políticas públicas.

- Este manual descreve como diferentes vieses e deficiências podem ser causadas pelos dados de treinamento, seja por decisões tomadas no desenvolvimento do modelo ou durante o processo de validação e monitoramento.

3 Esses dois tipos de sistema são genéricos, ou seja, não necessariamente utilizam aprendizado de máquina. Além disso, esses sistemas podem ser interativos e aprender dinamicamente, usando técnicas de aprendizado por reforço, mas, neste manual, consideramos apenas sistemas não interativos.

Componentes de um sistema de IA para políticas públicas

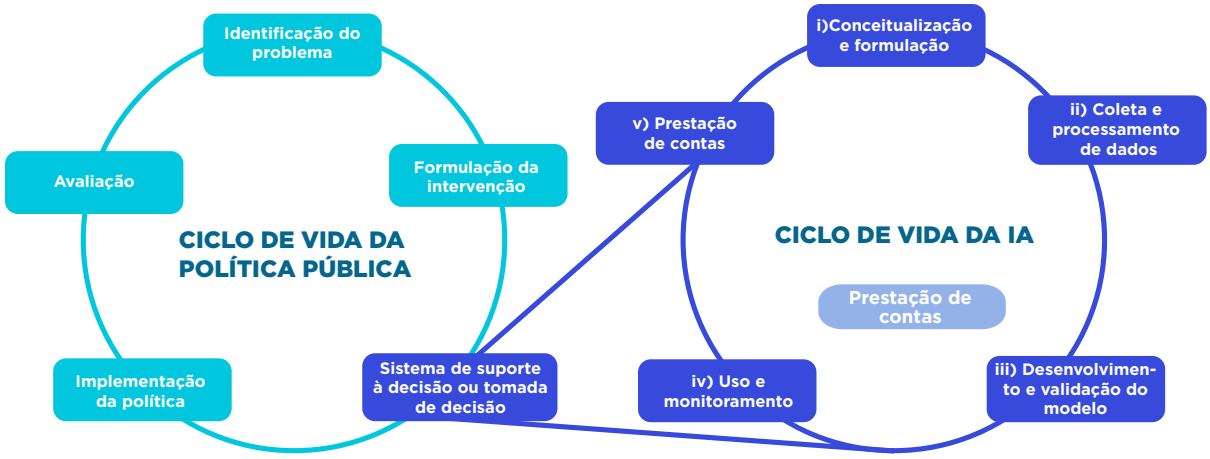
Ciclo de vida das políticas públicas com IA

A IA não substitui as políticas públicas, pois a IA por si só não resolve o problema social. Sua função é auxiliar, fornecendo informações para a tomada de decisão ou suporte à decisão. O ciclo de políticas públicas assistidas por IA é composto pelas seguintes etapas:

1. **Identificação do problema:** todo projeto de IA deve começar pela devida identificação do problema social que a política pública busca impactar, detalhando suas possíveis causas e consequências.
2. **Formulação da intervenção:** explicita-se a intervenção ou política que se considera aplicar a determinadas pessoas, unidades ou processos. Geralmente, pressupõe-se que haja evidências do benefício dessa política quando aplicada à população-alvo.
3. **Sistema de suporte à decisão/tomada de decisão:** uma vez definida a intervenção, inicia-se o ciclo de IA com a formulação e o desenvolvimento do sistema de suporte à decisão/tomada de decisão, cujo resultado será utilizado para focalizar ou orientar a intervenção escolhida no ponto anterior.4
4. **Implementação da política:** a política pública é colocada em operação, seja como projeto-piloto e/ou em maior escala.
5. **Avaliação da política:** são avaliadas a eficácia, a confiabilidade, o custo e as consequências previstas e não previstas, bem como outras características relevantes da medida de política em questão. Se seus resultados forem positivos, a intervenção é redimensionada ou continuada.

Paralelamente ao ciclo de elaboração de políticas públicas, o desenvolvimento de um sistema de IA tem seu próprio ciclo de vida, que inclui as seguintes etapas (OECD, 2019c): (i) conceitualização e formulação; (ii) coleta e processamento de dados; (iii) desenvolvimento do modelo e validação, e (iv) uso e monitoramento. Essas fases geralmente ocorrem de forma iterativa e não necessariamente são sequenciais (Figura 2).

Figura 2. Ciclo de vida das políticas públicas assistidas por um sistema de suporte à decisão/tomada de decisão



Fonte: preparada pelos autores.

4 A IA pode ser usada de diferentes maneiras. Algumas delas podem ser: i) sistemas de alerta precoce ou detecção de anomalias (previsão de abandono escolar ou alerta de fenômenos hidrometeorológicos); ii) sistemas de recomendação ou personalização (recomendação de vagas de emprego ou personalização de materiais educativos); e iii) sistemas de reconhecimento, diagnóstico de doenças, detecção de objetos ou reconhecimento biométrico.

A inter-relação desses dois ciclos gera importantes desafios que precisam ser avaliados e considerados durante o desenvolvimento e uso de sistemas de IA robustos e responsáveis.

Desafios do ciclo de vida de ML

Para construir sistemas de suporte à decisão/tomada de decisão robustos e responsáveis, é necessário realizar várias tarefas: considerar as possíveis fontes de vieses e deficiências que podem advir dos dados de treinamento e de problemas e decisões no desenvolvimento do modelo; definir claramente os objetivos dos sistemas e os critérios de justiça que serão buscados; entender as limitações e erros no contexto do projeto específico, e estabelecer medidas de monitoramento dos sistemas para evitar resultados indesejáveis e iniquidade na tomada de decisão.

Para isso, este manual apresenta os desafios e erros habituais na construção e aplicação de métodos de aprendizado de máquina durante o ciclo de vida da IA. Cinco seções descrevem os problemas mais comuns que podem ser encontrados, diagnósticos para detectá-los e sugestões para mitigá-los:

1. **Conceitualização e formulação:** refere-se às informações e critérios necessários que o tomador de decisão de políticas públicas deve obter para iniciar um projeto de IA.
2. **Coleta e processamento de dados:** concentra-se no processo de geração de dados, na seleção e controle das diferentes fontes, e na identificação e mitigação de deficiências e vieses.
3. **Desenvolvimento do modelo e validação:** refere-se a métodos e princípios importantes para a construção de modelos robustos e validados corretamente.
4. **Uso e monitoramento:** trata-se da avaliação do modelo em produção e do acompanhamento dos princípios fundamentais para evitar degradações inesperadas.

Além disso, um quinto aspecto, a prestação de contas, é uma dimensão transversal no ciclo de vida do sistema de IA que se refere a medidas de transparência e explicabilidade para promover a compreensão dos mecanismos pelos quais um sistema de IA produz um resultado, a reprodutibilidade do resultado e a capacidade do usuário de identificar e contestar erros ou resultados inesperados.

Em cada etapa do ciclo de vida, os atores da IA devem ser responsabilizados pelo funcionamento adequado de um sistema de IA com base em seus papéis no desenvolvimento e no contexto de uso do sistema.

Três ferramentas são propostas para acompanhar o desenvolvimento do sistema de IA:

- **Ferramenta 1: Lista de verificação para uma IA robusta e responsável.** Esta ferramenta consolida os principais receios da dimensão de risco do ciclo de vida da IA. A lista deve ser revisada continuamente pela equipe técnica, acompanhada pelo tomador de decisão.
- **Ferramenta 2: Perfil de dados.** O perfil é uma análise exploratória inicial durante a fase de Coleta e processamento de dados do ciclo de vida da IA. Fornece informações para avaliar a qualidade, integridade, pontualidade, consistência e possíveis vieses, danos em potencial e implicações de seu uso.
- **Ferramenta 3: Perfil do modelo.** É a descrição final de um sistema de IA; relata as principais premissas, as características mais importantes do sistema e as medidas de mitigação implementadas.

Caixa 1. Fontes de viés em um sistema de IA

Um dos conceitos mais importantes para os desafios que surgem no ciclo de vida da IA é o de viés, pois muitas das medidas de mitigação e desafios que precisam ser considerados durante o desenvolvimento dos modelos dependem de seus devidos entendimento e tratamento. Para resolver esse problema precocemente, é conveniente contar com revisões específicas nas diferentes fases do ciclo de vida. Em cada revisão, os especialistas e usuários finais do sistema correspondente devem ser convidados a verificar e defender as hipóteses lançadas durante cada etapa. Isso permite enriquecer os pontos de vista, encontrar suposições errôneas e adicionar aspectos não considerados.

O **erro do sistema** é a diferença entre o valor previsto, resultante do modelo, e o valor real da variável que está sendo estimada. Se o erro for sistemático em uma direção ou em um subconjunto específico dos dados, ele é chamado de **viés**5. Por exemplo, se uma balança pesa sempre um quilo a mais, ela está enviesada; ou, se um valor é sempre menor, como o salário da mulher para um trabalho equivalente ao do homem, a variável "salário" está enviesada. Por outro lado, quando o **erro** é aleatório, é chamado de **ruído**.

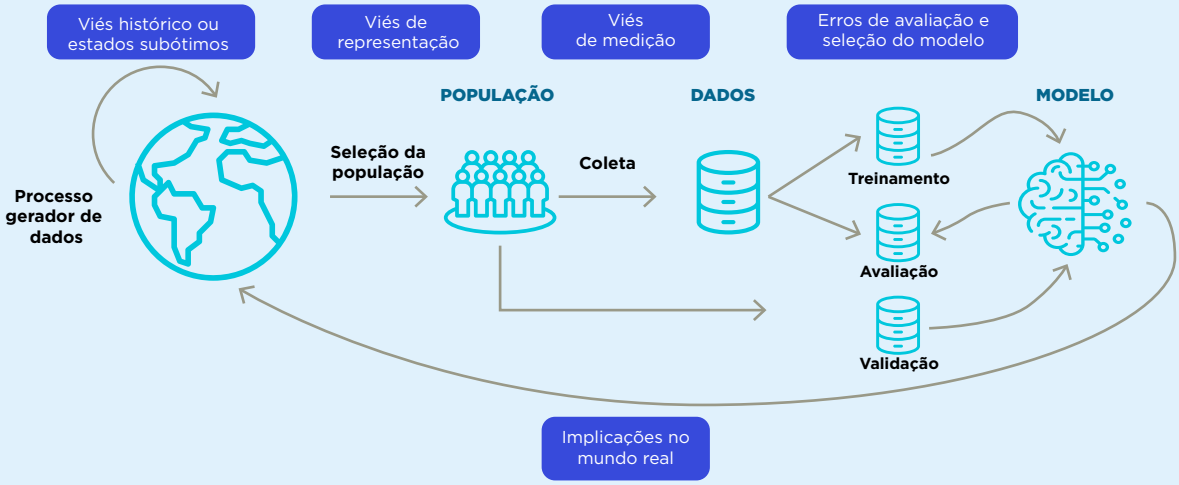
O viés de um sistema de IA pode ter implicações éticas quando seus resultados são usados para formular políticas públicas que podem ser consideradas injustas ou prejudiciais a determinados subgrupos da população. Essa avaliação de vieses está sujeita a uma definição específica de equidade algorítmica, a ser determinada pelos formuladores de políticas.

Uma definição de equidade algorítmica é uma representação matemática de um objetivo de política pública que é incorporada ao processo de seleção e ajuste do modelo. Por exemplo, em alguns casos, o objetivo de um sistema pode estar vinculado a critérios como paridade demográfica, igualdade de oportunidades e representação por cotas, entre vários outros critérios. Em algumas ocasiões, o cumprimento de uma definição de justiça algorítmica impossibilita o cumprimento de outra, ou seja, elas podem ser parcial ou totalmente excludentes. A definição de equidade algorítmica é uma tarefa para formuladores de políticas públicas, não para equipes técnicas. A equipe técnica é responsável apenas por realizar validações para garantir a conformidade. A Seção 3 deste manual analisa em profundidade as diferentes definições de equidade algorítmica e suas implicações.

Existem **diferentes fontes de viés**. Alguns vieses são intrínsecos aos dados, como os vieses históricos ou os estados indesejáveis, que são padrões pré-existentes na sociedade ou nos dados coletados que não é desejável reproduzir no modelo. O **viés de representação** ocorre quando há informações incompletas devido à falta de atributos, ao delineamento da amostra ou à ausência parcial ou total de dados de subpopulação. Os vieses de medição surgem da omissão (inclusão) de variáveis que (não) deveriam ser incluídas no modelo (Suresh e Guttag, 2019). Outros vieses aparecem devido a erros metodológicos: por exemplo, durante o treinamento, devido a erros nos processos de validação, definição de métricas e avaliação de resultados (**viés de avaliação**), ou **devido a suposições** errôneas sobre a população-alvo que podem afetar a definição do modelo. Também podem surgir devido a um uso e monitoramento

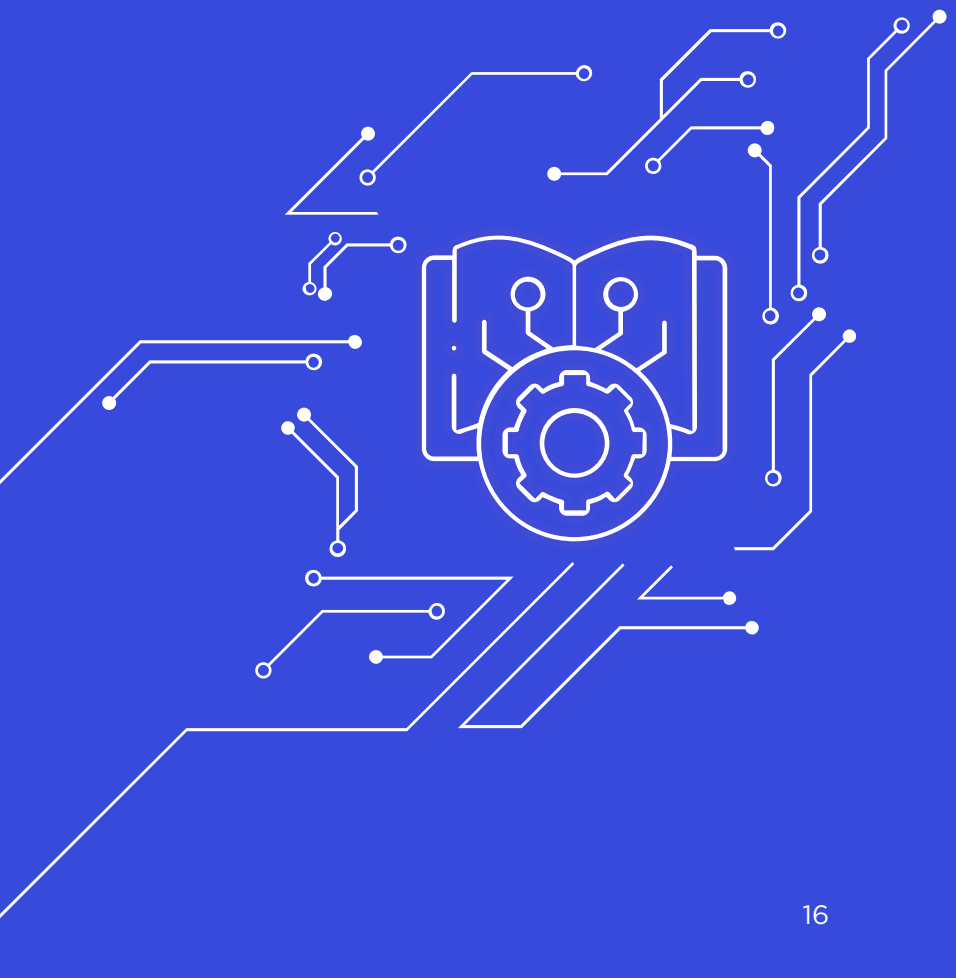
5 Nos modelos de previsão, há uma compensação entre a variância e o viés captado pelo modelo e seu objetivo de generalização de aprendizado. Por um lado, um modelo altamente enviesado pode criar sistemas que não se subajustam e aprendem muito pouco com os dados observados, mas modelos com uma variância alta podem ter o efeito oposto e sobreajustar, aprendendo perfeitamente os dados de treinamento. A seção "Desenvolvimento de modelos e validação" deste manual descreve esses fenômenos mais detalhadamente e oferece medidas para mitigar seus riscos.

incorretos dos modelos, seja em decorrência de interpretações inadequadas de seus resultados ou mudanças temporárias nos padrões do mundo real ou nos métodos de coleta de dados. Ao longo das diferentes seções deste manual, serão apresentadas as principais razões para esses vieses e propostas diferentes medidas para mitigá-los.



Fonte: baseado em Suresh Guttag (2019).

1. CONCEITUALIZAÇÃO E FORMULAÇÃO



1. Conceitualização e formulação

A implementação de uma solução de IA não pode ser dissociada do ciclo de vida das políticas públicas com IA[6]. A IA é uma ferramenta que deve estar condicionada a uma boa formulação da intervenção ou ação que será implementada com os resultados do sistema. A IA nunca é um substituto para políticas públicas. Isso significa que qualquer projeto de IA robusto e responsável deve partir do problema, não da tecnologia.

1.1 Definição correta do problema e da resposta por meio de uma política pública

Este manual considera que há pelo menos dois atores envolvidos no desenvolvimento dos sistemas: o tomador de decisão das políticas públicas e a equipe técnica que as implementará. A responsabilidade pela definição da intervenção deve ser sempre do tomador de decisão, que é quem tem conhecimento do problema social.

No entanto, a equipe técnica deve ser capaz de entender o problema para poder vincular os resultados do modelo à intervenção desejada. Da mesma forma, ela é responsável por guiar e orientar durante a formulação do sistema, explicando o que é viável e definindo claramente as limitações e riscos do sistema, sendo necessária, portanto, uma comunicação constante entre ambos os atores.

Um caso específico é a definição da população à qual o sistema será aplicado, a definição de grupos protegidos e as medidas de justiça algorítmica a serem aplicadas[7]. Essas definições têm um impacto direto em como se pode avaliar a qualidade e a cobertura dos dados ou o possível viés nos resultados do modelo.

1.2 Princípios de uma IA responsável

Embora a IA tenha um potencial significativo de agilizar processos e expandir a capacidade do Estado, também deve ser observado que não é uma panaceia. Uma vez definidos o problema e o tipo de intervenção, é necessário contextualizar e repensar a utilização da IA e do aprendizado de máquina de acordo com os Princípios da IA da Organização para a Cooperação e Desenvolvimento Econômico (OECD, na sigla em inglês; veja a Caixa 2).

É importante considerar a governança mais ampla que a aplicação de um sistema de IA engloba, incluindo as regras e leis da jurisdição onde o sistema será implementado. Também é importante estabelecer os requisitos adequados durante a conceitualização e formulação do sistema, pois eles podem definir ou limitar as opções de desenvolvimento para a equipe técnica. Por exemplo, requisitos de explicabilidade em previsões podem limitar o uso de alguns algoritmos cujos resultados seriam muito difíceis de interpretar.

6 Veja a seção “Componentes de um sistema de IA para políticas públicas”.

7 A Seção 3 deste manual aprofunda-se em diferentes definições de justiça algorítmica e suas implicações.

Caixa 2. Princípios de uma IA responsável segundo a OCDE

Os Princípios da IA da Organização para Cooperação e Desenvolvimento Econômico (OCDE) promovem o uso responsável da Inteligência Artificial (IA), respeitando os direitos humanos e os valores democráticos. Os princípios estabelecem padrões para a IA que são práticos e flexíveis o suficiente para resistir à passagem do tempo. Eles incluem cinco princípios baseados em valores para gerenciar uma IA responsável:

- **Crescimento inclusivo, desenvolvimento sustentável e bem-estar:** as partes interessadas devem comprometer-se a criar uma IA confiável que ajude a gerar resultados benéficos para as pessoas e para o planeta.
- **Valores centrados no ser humano e na equidade:** os valores dos direitos humanos, da democracia e do Estado de direito devem ser incorporados ao longo do ciclo de vida do sistema de IA, permitindo a intervenção humana por meio de mecanismos de salvaguarda.
- **Transparência e explicabilidade:** os atores que desenvolvem ou operam sistemas de IA devem fornecer informações para promover um entendimento geral dos sistemas entre as partes interessadas, permitindo que as pessoas afetadas pelos sistemas de IA entendam o resultado e contestem a decisão, quando necessário.
- **Robustez, segurança e proteção:** os sistemas de IA devem funcionar adequadamente durante todo o seu ciclo de vida. Os atores devem garantir a rastreabilidade e aplicar abordagens sistemáticas de gerenciamento de riscos para sua mitigação.
- **Responsabilidade:** os atores que desenvolvem, implantam ou operam sistemas de IA devem respeitar os princípios e responsabilizar-se pelo bom funcionamento desses sistemas.

Os Princípios IA da OCDE contêm cinco recomendações para políticas nacionais e para a cooperação internacional: (1) Investir em pesquisa e desenvolvimento de IA; (2) Promover um ecossistema digital para IA; (3) Criar um ambiente político propício para a IA; (4) Desenvolver a capacidade humana e se preparar para a transformação do mercado de trabalho e (5) Promover a cooperação internacional para uma IA confiável (OCDE 2019a, 2019b). Os princípios foram adotados em maio de 2019 pelos países membros da OCDE e são o primeiro padrão internacional sobre IA a ser assinado pelos governos. Além dos membros da OCDE, outros países como Argentina, Brasil, Costa Rica, Malta, Peru, Romênia, Ucrânia, Cingapura e Egito aderiram aos Princípios da IA e mais adesões são bem-vindas. Em junho de 2019, o G20 adotou os Princípios da IA centrados no ser humano, que se baseiam nos Princípios da IA da OCDE.

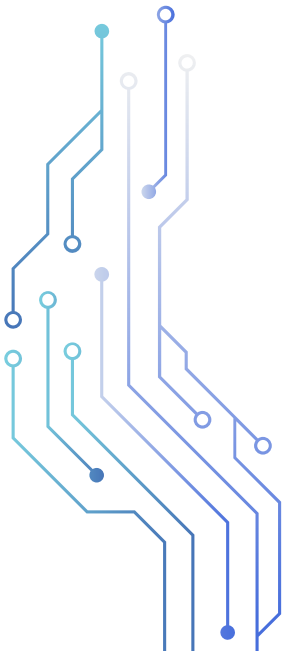
Caixa 3. Lista de verificação - Conceitualização e formulação

☑ Definição correta do problema e da resposta por meio de uma política pública:

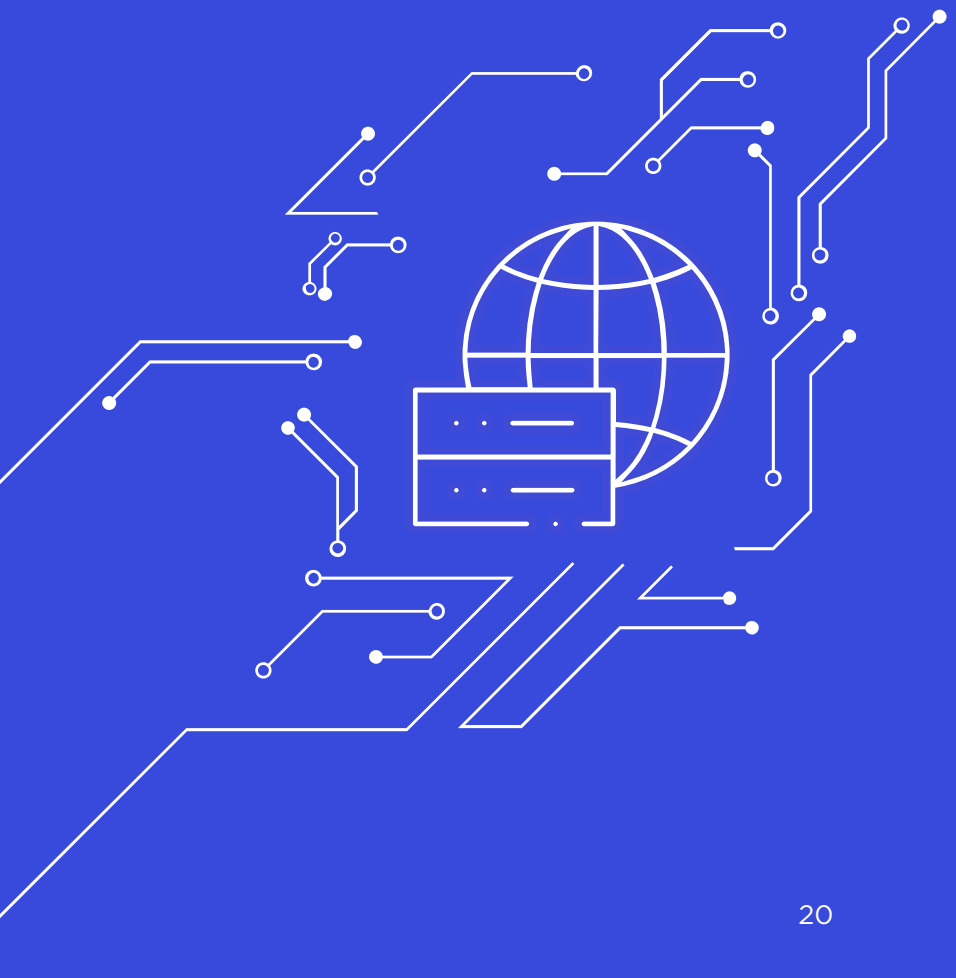
- (Qualitativo) O problema de política pública está claramente definido?
- (Qualitativo) Descrever como esse problema está sendo abordado atualmente, considerando-se as respostas de instituições relacionadas, e como o uso da IA melhoraria a resposta do governo a esse problema.
- (Qualitativo) Os grupos ou atributos protegidos foram identificados dentro do projeto? (Por exemplo, idade, gênero, escolaridade, raça, nível de marginalização etc.)
- (Qualitativo) Foram definidas as ações ou intervenções que serão realizadas com base no resultado do sistema de IA?

☑ Princípios da IA

- (Quantitativo) A necessidade de um sistema de IA foi justificada, levando-se em conta outras possíveis soluções que não requeiram a utilização de dados pessoais e decisões automatizadas?
- (Quantitativo) Há evidências de que tanto a ação das políticas públicas quanto a recomendação do sistema de IA beneficiarão as pessoas e o planeta, impulsionando o crescimento inclusivo, o desenvolvimento sustentável e o bem-estar?
- (Qualitativo) Projetos semelhantes foram identificados e analisados em busca de lições e erros comuns?
- (Quantitativo) Considerou-se a possibilidade de minimizar a exposição de informações pessoais identificáveis? (Por exemplo, anonimizando-se ou não se coletando informações não relevantes para a análise).



2. COLETA E PROCESSAMENTO DE DADOS



2. Coleta e processamento de dados

Não obstante, os dados coletados nem sempre têm uma frequência, desagregação ou cobertura que os torne relevantes, ou carecem da qualidade necessária para ser utilizados na tomada de decisão. Por exemplo, as pesquisas desenvolvidas com amostragem probabilística especificam o tipo de análise que se pode fazer com base nelas, mas esses tipos de ferramenta tendem a ser usados com pouca frequência e podem ser insuficientes para captar o movimento dos padrões que serão estudados. Por outro lado, informações de registros administrativos ou dados da internet (interações em redes sociais, visitas e outras medições em sites, etc.) e de telefonia (chamadas, localização por GPS etc.) tendem a ter uma frequência muito maior, mas, em poucos casos, abrangem a população como um todo; por isso, nem sempre podem ser utilizadas para tomar decisões que afetem toda a população.

No caso da implementação de um modelo supervisionado ou não supervisionado, os dados de treinamento são um ponto muito importante de qualquer sistema de ML. A qualidade dos dados pode ser analisada com base em critérios como volume, integridade, validade, relevância, precisão, pontualidade, acessibilidade, comparabilidade e interoperabilidade de diferentes fontes. Definir com precisão esses critérios em geral é difícil, pois o contexto de cada problema contém idiossincrasias sutis. Relevância e precisão referem-se à qualidade da mensuração e utilidade para informar a decisão, enquanto pontualidade refere-se ao fato de os dados ocorrerem com a temporalidade necessária para informar o problema que requer uma decisão. Acessibilidade, comparabilidade e interoperabilidade referem-se aos fatos de que os dados podem ser extraídos em tempo hábil e de que diferentes fontes de dados têm a consistência necessária para ser aplicadas em conjunto na análise.[8]

Dois problemas comuns para sistemas de aprendizado de máquina durante a fase de coleta e processamento de dados serão abordados nesta seção[9]:

1. Qualidade e relevância dos dados disponíveis; e

2. Qualificação e exaustividade dos dados para a população-alvo.

As Seções 2.1 e 2.2 abordam algumas das questões destacadas nos Princípios de Boas Práticas da OCDE relativos à ética de dados no setor público em relação à qualidade e qualificação dos dados. Os Princípios de Boas Práticas visam a apoiar os servidores públicos na aplicação da ética de dados em projetos, produtos e serviços de governo digital, de modo que: i) a confiança esteja no centro de sua formulação e entrega; e ii) a integridade pública seja mantida por meio de medidas específicas adotadas por governos, organizações públicas e, em uma camada mais granular, servidores públicos (OCDE, 2021).

2.1 Qualidade e relevância dos dados disponíveis

Os algoritmos de aprendizado de máquina captam padrões e relações observados com base nos dados com os quais foram treinados. Seu objetivo é identificar esses mesmos padrões para novos casos não observados durante o treinamento do modelo. Por esse motivo, os dados de treinamento determinam como o algoritmo se comportará. No entanto, os dados disponíveis nem sempre são ideais para todos os casos de uso. Dois dos principais problemas são:

8 Nessa fase, é recomendável preencher o Perfil de dados na seção de ferramentas deste manual.

9 Embora não sejam abordadas detalhadamente nesta seção, outras questões relacionadas aos dados, como o domínio e a estrutura dos dados, estão incluídas no perfil de dados.

1. Estados indesejáveis ou subótimos nos dados coletados;
2. Má correspondência entre variáveis disponíveis e variáveis ideais.

2.1.1 Estados indesejáveis ou subótimos nos dados coletados

O primeiro desafio é não levar em conta que os dados com os quais treinamos um modelo de ML podem ter captado estados indesejáveis do mundo real. Esses "estados indesejáveis" podem ser vieses e iniquidades prejudiciais para certos subgrupos, mas também podem ser qualquer outro padrão considerado subótimo ou indesejável do ponto de vista das políticas sociais.

Exemplo

Um caso que exemplifica esse desafio ocorreu em 2015, quando a Amazon experimentou um sistema de recomendação de recursos humanos baseado em técnicas de aprendizado supervisionado. O modelo foi treinado com um banco de dados dos processos de seleção de candidatos da empresa armazenados nos últimos dez anos. O banco de dados informava se um candidato havia sido aceito ou rejeitado para o cargo pelo departamento. O sistema partiu da hipótese de que o algoritmo poderia captar bons candidatos e diminuir o trabalho do departamento de recursos humanos na hora de fazer uma seleção inicial de candidatos. O que a equipe não levou em consideração é que o setor de tecnologia foi caracterizado como predominantemente masculino; assim, o sistema recomendou uma proporção maior de homens, já que mais homens foram aceitos nessas posições historicamente, criando um viés que parecia mostrar que os homens tiveram mais sucesso, quando, na verdade, estava captando uma iniquidade.

Caixa 4. Lista de verificação - Estados indesejáveis ou subótimos nos dados coletados

- (Qualitativo) Consultar especialistas no assunto sobre possíveis desigualdades sociais históricas no caso de uso.
- (Quantitativo) Realizar uma análise exploratória dos dados disponíveis com os quais o modelo será treinado, a fim de identificar possíveis vieses históricos ou estados indesejáveis.

2.1.2 Má correspondência entre variáveis disponíveis e variáveis ideais

As decisões relativas a políticas públicas são tomadas com base na definição de uma ou várias variáveis-alvo "ideais" que o tomador de decisão tem em mente. No entanto, as variáveis ideais podem ou não estar disponíveis nos dados acessados. Em muitas ocasiões, é necessário usar variáveis substitutas ou proxy para ajudar a aproximar-nos da variável ideal. Quando introduzimos esses tipos de variável em modelos de ML, podemos estar aprendendo vieses implícitos que podem não ser desejáveis. Por exemplo, uma bolsa de estudos que busca beneficiar os alunos mais inteligentes (variável ideal) encontrará o problema de definir esse conceito e encontrar uma variável capaz de descrevê-lo. Um teste de QI atribui um valor

baseado em um teste padronizado, que é descrito como uma variável proxy da inteligência. No entanto, o teste mede apenas algumas dimensões da inteligência e, por isso, subestimará a inteligência de algumas pessoas (Wilson, 2014).

As variáveis-alvo devem ser claramente definidas, mesmo que sejam ideais. As variáveis disponíveis devem ser analisadas para entender o quão adequadas são para ser usadas como proxy da variável ideal. Vieses sistemáticos devem ser identificados dentro de seu contexto de uso.

Exemplo

O sistema de saúde dos Estados Unidos implementou um algoritmo para prever as necessidades de cuidados médicos de diferentes pacientes. Nesse caso, o tomador de decisão da política pública queria uma ferramenta que indicasse preventivamente quais pacientes tinham um alto risco de precisar de mais cuidados médicos com base em informações históricas dos hospitais. Como a variável ideal para o risco de complicação não estava disponível, eles usaram os valores gastos pelos pacientes durante o período da doença como variável proxy, supondo que pessoas mais doentes acabariam gastando mais nos tratamentos médicos da doença. Pesquisadores (Obermeyer, Powers, Vogeli, & Mullainathan, 2019) mostraram que esse sistema tinha um viés racial porque subestimava o número de pacientes negros com necessidades de cuidados de saúde. O viés racial ocorreu porque essa subpopulação gasta, em média, menos dinheiro que os pacientes brancos. Usando a despesa como variável proxy para o risco de complicações, os pacientes brancos mais saudáveis pareciam exigir mais cuidados que os pacientes negros mais doentes. Nesse caso, foi inadequado usar o gasto com saúde como uma medida substituta da necessidade de cuidados médicos, uma vez que isso foi influenciado por uma variável omitida de desigualdade econômica.

Caixa 5. Má correspondência entre as variáveis disponíveis e as variáveis ideais

- (Qualitativo) As variáveis-alvo ideais devem ser claramente estabelecidas. As variáveis coletadas/disponíveis devem ser analisadas para entender até que ponto são adequadas para substituir a variável-alvo. Vieses sistemáticos ou a validade da métrica substituta devem ser identificados.
- (Qualitativo) O uso da variável de resposta selecionada para os propósitos da intervenção foi claramente justificado?

2.2 Qualificação e abrangência dos dados para a população-alvo

Os modelos de ML visam a gerar informações para a tomada de ações ou elaboração de políticas para uma população-alvo. Na maioria das vezes, as fontes de dados não incluem toda a população (como seria o caso de um censo); por isso, é comum que apenas um subconjunto ou amostra dela esteja disponível (uma pesquisa, um banco de dados administrativo etc.), a partir da qual se deve buscar fazer extrapolações, previsões ou estimativas para auxiliar na tomada de decisão.

2.2.1 Amostras probabilísticas e naturais

Em estatística, uma amostra é um subconjunto de casos ou indivíduos de uma população. Existem dois extremos possíveis:

1. **Amostragem probabilística:** é o nome dado a uma amostra na qual os casos são selecionados a partir de um delineamento probabilístico. Por exemplo: amostra aleatória simples, estratificada, por conglomerados etc. Nesse caso, todas as previsões e estimativas destinadas a ser aplicadas à população-alvo podem ser testadas quanto à precisão com garantias probabilísticas. Ou seja, podemos fornecer intervalos de erro para estimativas de quantidades associadas a toda a população-alvo.

   Por exemplo: uma pesquisa domiciliar nacional com delineamento probabilístico geralmente consiste em uma definição de estratificação, unidades de seleção aleatória em diferentes níveis (unidades primárias, unidades secundárias etc.). Cada domicílio é selecionado com uma probabilidade conhecida. Mesmo que a amostra seja delineada de forma não representativa (por exemplo, mais domicílios em áreas rurais ou de baixa renda), é possível fazer inferências para toda a população com certas garantias sobre o tamanho do erro de estimativa.

2. **Amostras naturais (não probabilísticas):** por outro lado, uma amostra natural ou não probabilística ocorre quando os casos não são selecionados aleatoriamente, mas por um processo natural pouco ou parcialmente conhecido. Nesse caso, não é possível saber o que acontecerá quando uma política resultante de um modelo for aplicada à população geral, não sendo possível definir intervalos de erro de previsões e estimativas com base em métodos estatísticos com garantias probabilísticas. Ou seja, as quantidades e previsões estimadas têm um erro desconhecido, os modelos e características úteis na amostragem podem não se aplicar à população-alvo, e a situação pode ser agravada para grupos protegidos sub-representados. A referência bibliográfica (Williams, 1981), por exemplo, mostra que os valores preditivos de anemia podem ser diferentes para diferentes grupos raciais e que as previsões desenvolvidas para um grupo podem ter um baixo desempenho em outro.

   Um caso comum desse tipo de amostra ocorre quando determinados subgrupos da população são excluídos pelo canal de captação de informações (viés de seleção). Por exemplo, em aplicações que excluirão a população que não tem acesso à internet ou smartphones. É o caso de informações de redes sociais, registros de ligações telefônicas etc.

   Amostras de dados naturais podem levar a:

   - Erros ou vieses de estimativa e/ou previsão.
   - Estruturas preditivas diferentes das que observaríamos na população-alvo (modelos inválidos).
   - Extrapolações que não são suportadas pelos dados.
   - Sub-representação ou sobrerrepresentação de subconjuntos da população.

   A amostragem probabilística seria a situação ideal para a maioria dos projetos de aprendizado de máquina. Nesse caso, pode-se entender exatamente quais subpopulações foram amostradas, a quais taxas e como essas taxas estão relacionadas às taxas populacionais. o delineamento da amostra determina o escopo inferencial. No entanto, nem sempre é possível contar com uma amostra probabilística.

Isso não quer dizer que as amostras naturais não sejam úteis; muitas vezes, são a única fonte de dados disponível para a tomada de decisão. No entanto, é importante entender de onde vêm os dados para levar em consideração suas limitações e identificar os riscos envolvidos na tomada de decisão para toda a população.

Um caso típico são amostras de dados provenientes de redes sociais nas quais a composição demográfica dos usuários difere substancialmente da população em geral. Um estudo realizado no Reino Unido descobriu que, em média, os usuários do Twitter e do Facebook são consideravelmente mais jovens que a população em geral, sendo mais provável que tenham níveis de escolaridade mais altos que os não usuários (Prosser & Mellon, 2016). Qualquer estudo com esses dados deve explicar como essas particularidades podem afetar os resultados.

É importante levar em conta que dispor de amostras balanceadas quanto às características populacionais não é condição necessária nem suficiente para qualificar o banco de dados como apropriado para a construção de modelos de ML. Por exemplo, no caso de informações coletadas pelas redes sociais, o fato de existir uma amostra que contém 50% de homens e 50% de mulheres não diz nada sobre as conclusões às quais se pode chegar com base nesses dados, pois a seleção dessas observações, por não se dar por meio de um processo probabilístico, pode apresentar um viés em alguma outra dimensão, não necessariamente sendo generalizada para a população total.

Caixa 6. Lista de verificação - Amostras probabilísticas e naturais

- (Qualitativo) Foram analisadas possíveis diferenças entre o banco de dados e a população para a qual o sistema de IA está sendo desenvolvido? (Usar a bibliografia relacionada ao tema e as informações dos especialistas. Estudar em particular os vieses de seleção não medidos).
- (Quantitativo) Embora os modelos possam ser elaborados com diferentes fontes de dados, artificiais ou naturais, o ideal é que a validação seja realizada com uma amostra que permita uma inferência estatística para a população. A amostra de validação deve abranger adequadamente a população-alvo e as subpopulações de interesse.

2.2.2 Atributos ausentes ou incompletos

Muitos projetos de aprendizado de máquina estão fadados ao fracasso devido à baixa qualidade dos dados disponíveis. Quando dados do mundo real são coletados por meio de amostras não probabilísticas, é muito comum que algumas observações tenham dados ausentes, ou seja, observações para as quais nem todos os atributos estão disponíveis.

Atributos ausentes ou incompletos são um fenômeno que pode ter um efeito significativo nas conclusões extraíveis dos dados. Por um lado, quando informações cruciais sobre as unidades são totalmente desconhecidas, isso pode resultar em modelos de baixo desempenho, com pouca utilidade para a tomada de decisão, e, por outro lado, a ausência de informações também pode estar associada a características relevantes das unidades para as quais se pretende fazer previsões.

Quando há observações ausentes, é possível implementar diferentes métodos de imputação, mas é importante explorar as razões ou o "mecanismo de censura" pelo qual uma observação pode ter valores ausentes. Na literatura, existem três premissas principais (Rubin, 2002):

- **Valores ausentes de forma completamente aleatória (Missing Completely at Random - MCAR):** ocorre quando a probabilidade de ausência é a mesma para todas as observações. Ou seja, a censura ou ausência de valores ocorre totalmente ao acaso.

- **Valores ausentes de forma aleatória (Missing at Random - MAR):** ocorre quando os valores ausentes não dependem dos valores adotados pela variável em questão, mas há uma relação entre os valores ausentes e outros dados observados do indivíduo.

- **Valores ausentes não aleatórios (Missing Not at Random - MNAR):** ocorre quando os valores ausentes dependem dos valores que a variável em questão adota ou de dados não observados. Por exemplo, é um fenômeno conhecido que, quando são realizadas pesquisas de renda autodeclarada, as pessoas com renda mais alta tendem a não revelar sua renda.

Caixa 7. Atributos ausentes ou incompletos

- (Qualitativo) Realizou-se uma análise de valores ausentes e variáveis omitidas?
- (Qualitativo) Identificou-se a existência de variáveis importantes omitidas para as quais não existem medidas associadas (se for o caso)?
- (Qualitativo) Os motivos para a falta de observações foram identificados (se for o caso)?
- (Quantitativo) A sensibilidade a suposições e dados dos processos de imputação devem ser avaliados. De preferência, métodos de imputação múltipla devem ser usados para avaliar a incerteza de imputação (Little & Rubin, 2002), (Buuren & Groothuis-Oudshoorn, 2011).

2.3 Comparação causal

Quando os humanos racionalizam o mundo, tentam entendê-lo em termos de causa e efeito: se entendermos por que algo aconteceu, podemos alterar o nosso comportamento para mudar os resultados futuros.

Um modelo de ML pode fornecer-nos resultados que parecem descrever relações causais sem que esse necessariamente seja o caso. Se a política for aplicada com base nos resultados das variáveis inclusas no modelo, a derivação de políticas desses modelos pode levar a decisões erradas.
Técnicas econométricas, como ensaios controlados e randomizados ou RCTs (Randomized Controlled Trials), experimentos naturais, diferenças em diferenças, e variáveis instrumentais, são utilizadas para esses propósitos, a fim de controlar fenômenos como viés de seleção ou endogênese de variável omitida, entre outros. Nos últimos anos, trabalhos como o de Athey (2018) começaram a introduzir essas técnicas e processos experimentais do tipo A/B testing nos algoritmos de ML, que passaram a ser usados em massa em contextos digitais devido à facilidade de criar experimentos em massa na internet. No entanto, na maioria dos casos, os algoritmos de ML não buscam descrever relações causais, sendo necessário ter muito cuidado com esse tipo de uso (Stuart, 2008).

Caixa 8. Lista de Verificação - Comparação causal

- (Qualitativo) Compreender e descrever as razões pelas quais a variável de resposta está correlacionada com variáveis conhecidas e desconhecidas. Descrever possíveis vieses com base no conhecimento e na análise de especialistas.
- (Qualitativo) Caso nenhum trabalho tenha sido feito para garantir a causalidade nos resultados, as limitações dos resultados foram explicitamente comunicadas ao responsável pela política pública?
- (Quantitativo) No caso de uma tentativa de inferência causal com modelos, as hipóteses, considerações ou métodos usados para apoiar uma interpretação causal devem ser descritos. Verificações de robustez devem ser realizadas e documentadas.

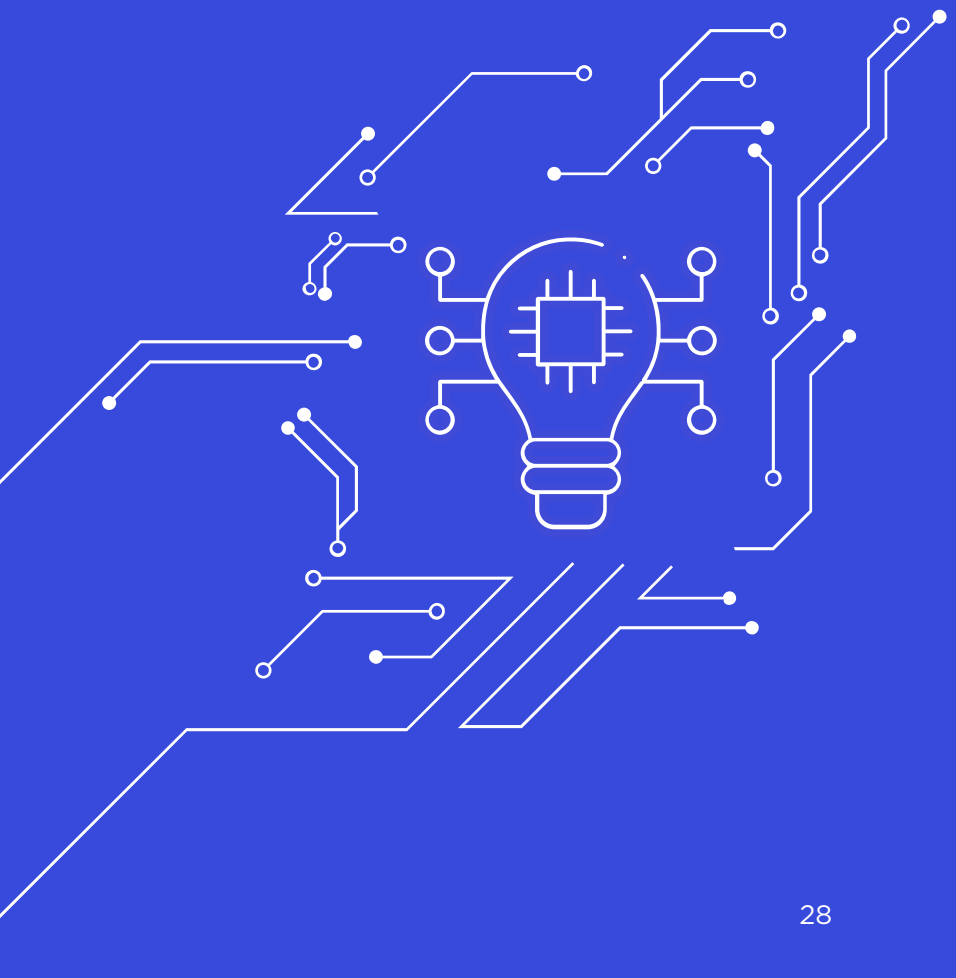
Atividade:

O preenchimento do Perfil de dados é recomendado durante a fase de conscientização e preparação de dados do ciclo de vida da IA (consulte a Ferramenta 2). No fim desta fase, recomenda-se preencher a seção de fonte e gestão de dados do Perfil do modelo e realizar uma discussão com o tomador de decisão de políticas públicas.

VER FERRAMENTA 2

VER FERRAMENTA 3

3. DESENVOLVIMENTO DO MODELO E VALIDAÇÃO



3. Desenvolvimento do modelo e validação

O processo de desenvolvimento do modelo envolve muitas decisões que têm implicações em seus resultados. Algumas decisões podem levar a erros metodológicos que geram vieses ou impedem que o sistema generalize adequadamente. Entre eles, encontramos vazamentos de informações, sobreajuste e subajuste.

Além disso, existe um outro grupo de decisões que não necessariamente são problemas metodológicos que podem alterar substancialmente o comportamento do sistema: como escolher entre dois modelos? Que tipos de erros relatar? Que definição de justiça algorítmica será escolhida? Como mencionado no início do manual, nenhuma dessas perguntas faz sentido fora do contexto da aplicação específica. O que se pode fazer é criar um quadro de compreensão desses erros para que possam ser discutidos entre equipes técnicas e tomadores de decisão de políticas públicas.

Nesta seção do manual, serão expostos os desafios que surgem durante os processos de treinamento e validação dos sistemas de suporte à decisão e tomada de decisão. Neste caso, a maioria dos erros deve-se a falhas metodológicas na avaliação e à não explicitação correta do objetivo de ajustar o sistema ou das métricas que se busca otimizar.

3.1 Ausência ou uso inadequado de amostras de validação

Os modelos de aprendizado de máquina são treinados principalmente para criar previsões em casos não observados. É inútil avaliar o desempenho de um sistema com base na previsão das observações com as quais foi treinado, pois o sistema pode simplesmente memorizar cada resposta.[10] Sua utilidade está em até que ponto o sistema consegue generalizar o aprendizado para fazer previsões com dados fora do conjunto de treinamento (out-of-sample).

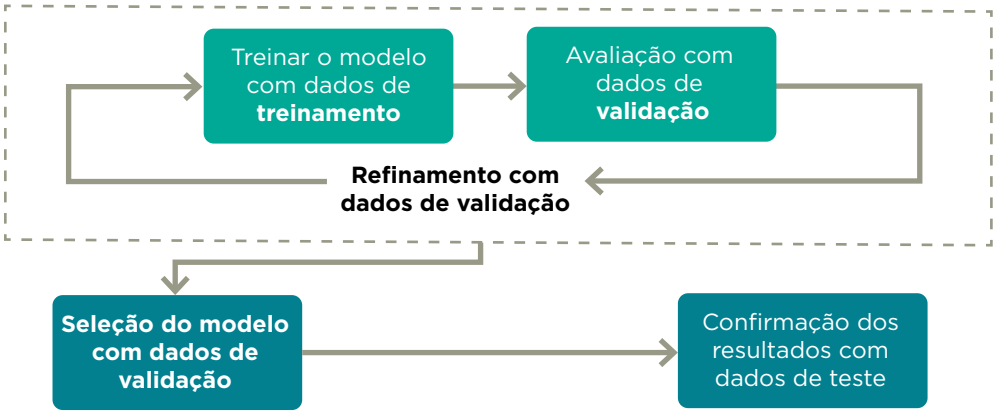
A validação geralmente envolve pelo menos duas amostras (1 e 2), preferencialmente três:

1. **Dados de treinamento:** subconjunto dos dados usados para treinar o modelo.
2. **Dados de validação:** subconjunto dos dados com os quais o treinamento é avaliado iterativamente.
3. **Dados de teste:** subconjunto dos dados que devem ficar ocultos até que o modelo seja selecionado e usado para confirmar os resultados.

Para evitar que uma porção aleatória nos dados de treinamento e validação favoreça ou prejudique a avaliação, geralmente se realiza uma validação cruzada. Ela consiste em dividir os dados em k pedaços, calculando uma média de k avaliações, onde os dados de validação são cada um dos pedaços, e os k-1 restantes são os dados de treinamento. Em inglês, isso é chamado de k-fold evaluation, e geralmente escolhemos k=5 ou k=10.

10 Esse fenômeno está relacionado ao sobreajuste que será visto mais adiante.

Figura 3. Fases de avaliação



Fonte: Construção própria

O primeiro desafio é contar com um processo de validação inadequado ou mesmo inexistente. Nesse caso, os resultados do modelo seriam apresentados apenas com o desempenho do conjunto de dados de treinamento. As métricas de desempenho desse conjunto não devem ser utilizadas como indicador do comportamento em potencial do modelo para novos casos, pois seu desempenho pode ser superestimado.
Uma validação bem-sucedida também está relacionada a critérios de qualidade, como a abrangência e a representatividade das informações vistas no capítulo anterior, pois, se a população-alvo for diferente da representada pelos dados utilizados durante o treinamento, mesmo que o processo de avaliação tenha sido realizado corretamente, é possível que haja um comportamento completamente diferente.

Caixa 9. Lista de verificação - Ausência ou uso inadequado de amostras de validação

- (Quantitativo) As amostras de validação e teste foram geradas adequadamente, considerando-se um tamanho adequado, abrangendo-se subgrupos de interesse e protegidos, e evitando-se vazamentos de informações durante sua implementação?
  - A amostra de validação deve ser construída sobre uma base de amostragem que permita inferências sobre a população-alvo (Lohr, 2009).
  - A amostra de validação deve abranger subgrupos de interesse e protegidos, para que seja possível fazer inferências sobre suas subpopulações. Isso inclui tamanhos de amostras apropriados, de acordo com a metodologia de amostragem (Lohr, 2009).
  - Se essa amostra não estiver disponível, uma análise dos riscos e limitações da amostra natural, realizada por especialistas e indivíduos familiarizados com o processo que gerou os dados da amostra, é essencial.

3.2 Vazamentos de informações

O vazamento de dados ocorre quando o modelo observa informações adicionais ao conjunto de treinamento (Kaufman, Rosset, & Perlich, 2011). Essas informações adicionais modificam o processo de aprendizado e colocam em questão a validação do modelo como forma de estimar o desempenho produtivo do sistema.

Isso ocorre de duas formas:

- **Contaminação treinamento-validação:** a amostra de treinamento recebe vazamentos dos dados de validação, o que implica o uso de dados de validação no treinamento e invalida a estimativa do erro de previsão.

- **Vazamentos de dados indisponíveis na previsão:** as amostras de treinamento e validação têm agrupamentos temporais ou de outros tipos que não são preservados no processo de treinamento e validação. Nesse caso, o treinamento e a validação recebem vazamentos de informações que estarão indisponíveis no momento de fazer as previsões.

3.2.1 Contaminação treinamento-validação

A contaminação treinamento-validação ocorre quando todas ou algumas das amostras de validação ou teste são usadas para construir os modelos durante o treinamento. Esse erro geralmente resulta em níveis de desempenho pouco realistas no conjunto de validação, porque o modelo fará previsões com base em observações feitas antes.

Esse erro costuma ocorrer quando se aplicam metodologias de pré-processamento que adicionam e compartilham informações de composição entre o banco de dados e as observações individuais: por exemplo, redimensionando uma variável, criando covariáveis com médias ou recontagens, metodologias de sobreamostragem ou subamostragem, etc. Esses processos devem ser feitos após a divisão do conjunto de dados de treinamento e validação.

Caixa 10. Lista de verificação - Contaminação treinamento-validação

- (Quantitativo) O processamento e a preparação dos dados de treinamento devem evitar ao máximo o uso de dados de validação ou teste. Uma barreira sólida deve ser mantida entre o treinamento em contraste com a validação e os testes. Isso inclui recodificação de dados, normalizações, seleção de variáveis, identificação de dados atípicos e qualquer outro tipo de preparação de qualquer variável a ser incluída nos modelos, o que também envolve ponderações ou balanceamento de amostras com base em sobre/subamostragem.

3.2.2 Vazamentos de dados indisponíveis na previsão

Esse erro ocorre quando um modelo é treinado com informações que não estarão disponíveis da mesma forma ou com a mesma qualidade quando o modelo for colocado em produção. Geralmente, refere-se à temporalidade dos dados ou dos agrupamentos. Em casos mais sutis, esse erro pode ser difícil de detectar, pois a variável está presente, mas as informações são atualizadas retroativamente.

Um exemplo disso pode ser visto nas estatísticas de criminalidade e mortalidade. As denúncias de furto podem demorar para aparecer nos bancos de dados das autoridades devido a processos burocráticos ou administrativos, e a incidência observada em um período pode aumentar sistematicamente com o tempo. Nesse exemplo, a variável-alvo está disponível na produção, mas pode não estar completa devido a certas defasagens inerentes à notificação. Se esse fenômeno não for considerado durante o treinamento e se dados que já estão completos forem usados, a avaliação do modelo pode parecer precisa, mas, na produção, a precisão dos dados diminuirá consideravelmente.

Caixa 11. Vazamentos de dados indisponíveis na previsão

O esquema de validação **deve replicar com a maior precisão possível** o esquema por meio do qual as previsões serão aplicadas. Isso inclui a necessidade de replicar:

- Janelas temporárias para observação e registro de variáveis, bem como janelas de previsão.
- Se existirem grupos nos dados, considerar se as informações de cada grupo estarão disponíveis quando a previsão for feita ou se será necessário prever para novos grupos.

3.3 Modelos de classificação: probabilidades e classes

No aprendizado de máquina, os algoritmos de classificação supervisionada são sistemas cujo objetivo é atribuir uma categoria ou rótulo de classe às novas observações. Denomina-se classificação binária quando a variável-alvo possui duas classes (por exemplo, classificar um e-mail como spam ou não spam) e classificação multiclasses quando há mais de duas classes (por exemplo, algoritmo de identificação de espécies vegetais).

3.3.1 Dados desbalanceados

Em um problema de classificação, um conjunto de dados desbalanceados ocorre quando a distribuição de observações entre as classes conhecidas não está distribuída uniformemente. Esses tipos de conjunto de dados terão uma ou mais classes com muitos exemplos, chamadas de classes majoritárias, e uma ou mais classes com menos observações, chamadas de classes minoritárias (por exemplo, grupos com menos de 1% do total de observações). Estes últimos grupos apresentam dificuldades consideráveis para os modelos de previsão, pois pode haver poucas informações sobre eles.

Em dados altamente desbalanceados, os preditores de classe podem ter um desempenho baixo (por exemplo, nunca preveem a classe minoritária), mesmo que as medidas de desempenho sejam boas. Em particular, se sempre previrmos a classe majoritária, a precisão será igual à porcentagem de elementos dessa classe.

Exemplos

- **Considere que há um milhão de dados: 999.000 negativos e 1.000 positivos. Pode ser uma boa ideia fazer uma subamostragem dos negativos por uma determinada fração (por exemplo, 10%), ponderando-se cada caso amostrado por 10 no ajuste e pós-processamento.**
- **Considere que há um milhão de dados: 999.950 negativos e 50 positivos. Pode ser impossível discriminar adequadamente os 50 dados positivos. Construir conjuntos de validação piora a situação: não é possível validar o desempenho preditivo nem construir um modelo de alto desempenho.**

Caixa 12. Lista de verificação - Dados desbalanceados

- (Quantitativo) Fazer **previsões de probabilidade** em vez de classe. Essas probabilidades podem ser incorporadas ao processo de decisão subsequente. Evitar pontos de corte padrão de probabilidade, como 0,5, ou fazer previsões conforme a probabilidade máxima.
- (Quantitativo) Quando o número absoluto de casos minoritários é muito pequeno, pode ser muito difícil encontrar informações apropriadas para discriminar essa classe. Mais **dados da classe minoritária precisam ser coletados.**
- (Quantitativo) A subamostragem da classe dominante (ponderação para cima dos casos para evitar a perda de calibração) pode ser uma estratégia bem-sucedida para reduzir o tamanho dos dados e o tempo de treinamento sem afetar o desempenho preditivo.
- (Quantitativo) Replicar a classe minoritária para balancear melhor as classes (sobreamostragem).
- (Quantitativo) Algumas técnicas de aprendizado de máquina permitem que cada classe seja ponderada com um peso diferente, de modo que o peso total de cada classe esteja balanceado. Se isso for possível, é preferível à sub ou sobreamostragem.

3.3.2 Pontos de corte arbitrários

Em problemas de classificação para a tomada de decisão, recomenda-se utilizar algoritmos de classificação de probabilidade em vez de classificar a observação apenas com sua classe mais provável. O resultado de um algoritmo de classificação de probabilidade é uma distribuição da probabilidade pelo conjunto de classes. Esses métodos podem fornecer informações ao tomador de decisão sobre a incerteza relacionada à classificação.

Para decidir se a observação deve ser classificada como positiva ou negativa, a equipe técnica deve escolher o limiar a partir do qual a observação é classificada como pertencente a cada classe. Um corte de 0,5 para classificações binárias geralmente é incorretamente aceito, pois é o valor padrão para muitos modelos de aprendizado de máquina. Essa decisão pode ter implicações importantes se tomada fora do contexto do problema em questão, sendo importante, portanto, que seja debatida e selecionada levando-se em conta os tipos de erro e suas implicações.

Caixa 13. Lista de verificação - Pontos de corte arbitrários

- (Quantitativo) É mais adequado usar **algoritmos de classificação probabilística** para que a tomada de decisão incorpore a incerteza em relação à classificação.
- (Quantitativo) Evitar pontos de corte de probabilidade padrão, como 0,5. Escolher uma interpretação ideal das probabilidades previstas, analisando-se as métricas de erro.

3.3.3 Adequação das métricas de avaliação

Em problemas de classificação, os pontos de corte são adotados conforme critérios relacionados ao contexto da decisão. A maioria é construída por meio de uma análise da matriz de confusão de classificação.

Tabela 1. Matriz de confusão

		Real	
		Positivo	**Negativo**
Previsto	**Positivo**	**Verdadeiro Positivo (VP)**	**Falso Positivo (FP)**
	Negativo	**Falso Negativo (FN)**	**Verdadeiro Negativo (TN)**

Os erros de um modelo de classificação podem ser divididos em falsos positivos e falsos negativos. Um falso positivo é uma observação para a qual o modelo prevê incorretamente a classe positiva, enquanto um falso negativo é uma observação para a qual o modelo prevê incorretamente a classe negativa. Essas métricas podem ser combinadas de diferentes maneiras, dependendo do caso de uso e do objetivo da política social. As métricas mais utilizadas são:

1. **Exatidão (Accuracy):** uma das métricas mais comumente usadas para avaliar modelos de classificação é a fração das previsões que o modelo acertou:

$$\text{Exactitud} = \frac{VP + VN}{VP + VN + FP + FN}$$

2. **Precisão:** fração dos classificados como positivos pelo modelo, mas que, na realidade, eram positivos:

$$\text{Precisión} = \frac{VP}{VP + FP}$$

3. **Sensibilidade (Recall):** fração de positivos que o modelo classificou corretamente:

$$\text{Sensibilidad} = \frac{VP}{VP + FN}$$

4. **Especificidade:** fração de negativos que o modelo classificou corretamente:

$$\text{Especificidad} = \frac{VN}{VN + FP}$$

É necessário levar em conta o contexto do problema ao definir os critérios de avaliação dos modelos de classificação. Por exemplo, se o modelo classificar a prevalência de uma doença letal, o custo de não diagnosticar (falso negativo) é muito maior que o custo de encaminhar uma pessoa saudável para testes adicionais (falso positivo). Ou seja, dependendo do caso, o custo dos falsos negativos pode ser muito diferente do custo dos falsos positivos. Por isso, recomenda-se utilizar a análise de custo-benefício, que compara o resultado do modelo no contexto da tomada de decisão.

Esses critérios também podem enganar, dependendo da composição do banco de dados de treinamento e avaliação. Por exemplo, ao usar dados não balanceados, uma precisão de 95% pode, na verdade, significar um desempenho inferior do modelo. Soluções parciais para esse problema incluem o uso de medidas que combinam precisão e sensibilidade, como a métrica F1 ou a curva Precision-Recall, que podem ajudar a analisar o equilíbrio entre verdadeiros positivos e falsos positivos no caso em questão.

Caixa 14. Lista de verificação - Adequação das métricas de avaliação

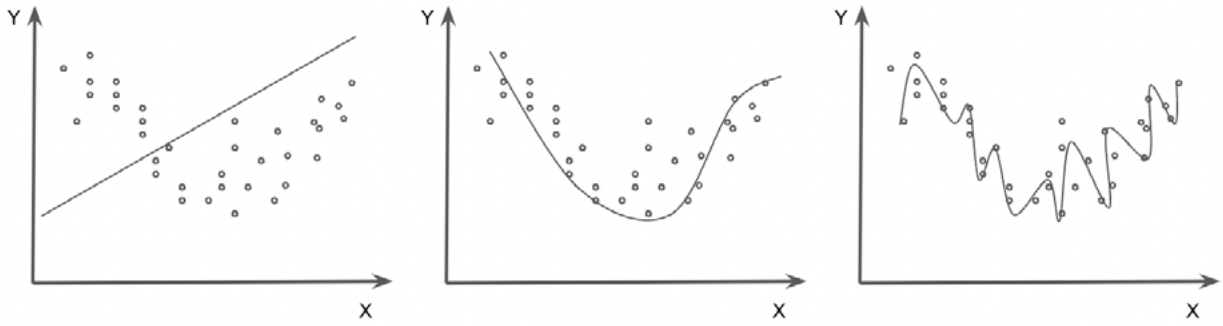
- (Qualitativo) As implicações dos diferentes tipos de erro foram questionadas para o caso de uso específico, bem como a forma correta de avaliá-los?
- (Qualitativo) As limitações do modelo foram claramente explicadas, identificando-se falsos positivos e falsos negativos, bem como as implicações que uma decisão do sistema teria na vida da população beneficiária?
- (Quantitativo) Implementou-se uma análise de custo-benefício do sistema em contraste com o status quo ou outras estratégias de tomada de decisão/suporte à decisão (quando possível)?

3.4 Sub e sobreajuste

Sub e sobreajuste ocorrem quando as informações preditivas dos dados são usadas de forma inadequada para o objetivo final do aprendizado de máquina, que é a generalização do aprendizado e seu uso em conjuntos de dados não observados no treinamento.

- O **sobreajuste** ocorre quando o modelo memoriza as particularidades dos dados de treinamento, mas é incapaz de generalizar para exemplos não vistos. Um modelo complexo demais para os dados disponíveis tende a captar características não informativas como parte da estrutura preditiva. Isso geralmente se reflete em um modelo que funciona muito bem nos dados de treinamento, mas tem um desempenho baixo no conjunto de dados de validação.

- O **subajuste** ocorre quando o modelo não consegue ter um bom desempenho com os dados de treinamento ou generalizar para os novos dados. Isso ocorre quando as características individuais das observações são superaglomeradas e recebem pouca importância. Um modelo pouco ajustado tende a ignorar os padrões da estrutura preditiva. Isso se reflete em erros sistemáticos e identificáveis, como, por exemplo, sub/sobreprevisão sistemática para determinados grupos ou valores das variáveis de entrada.



Subajuste | Ajuste robusto | Sobreajuste

Fonte: preparada pelos autores

Caixa 15. Lista de verificação - Sub e sobreajuste

- (Quantitativo) Sobreajuste: modelos com uma lacuna considerável entre validação e treinamento (indícios de sobreajuste) devem ser evitados. Se necessário, os métodos devem ser refinados para moderar o sobreajuste, como regularização, restrição do espaço funcional de modelos possíveis, uso de mais dados de treinamento ou perturbação dos dados de treinamento, entre outros (Hastie, Tibshirani, & Friedman, 2017).
- (Quantitativo) Subajuste: subconjuntos importantes de casos (por exemplo, grupos protegidos) devem ser revisados para verificar se há erros sistemáticos indesejáveis.

3.5 Erros não quantificados e avaliação humana

Em muitos casos, haverá aspectos do modelo que não são medidos pelas métricas de desempenho escolhidas.

Por exemplo, em um sistema de busca de documentos que, embora tenha um bom desempenho de validação nas métricas, seleciona documentos que tendem a ser muito curtos, produz resultados pouco úteis ou imparciais para determinadas pesquisas, ou prefere documentos do tipo promocional ou publicitário. Os motivos podem variar desde erros de pré-processamento (alguns atributos mal calculados) até a seleção de atributos para fazer as previsões que consideram apenas parte do problema.

3.5.1 Falhas não medidas pelo modelo

Alguns algoritmos produzem resultados de baixa qualidade que escapam à avaliação das métricas de validação. Esses modelos podem apresentar um baixo desempenho quando colocados em produção. Os motivos por trás disso são, entre outros, os seguintes:

- Erros de pré-processamento no momento do cálculo das previsões.
- Tratamento de dados que excluem métricas importantes para fazer previsões de qualidade ou não injustas.
- Ausência de métricas que medem certos tipos de erro particulares graves.

Isso pode ser um problema difícil, pois, por sua natureza, são erros não visíveis ou mensuráveis diretamente. É necessário descobrir esses vieses ou erros fora do contexto técnico de avaliação e, se possível, incluir métricas de avaliação adicionais que considerem esses problemas.

Caixa 16. Lista de verificação - Falhas não medidas pelo modelo

- (Qualitativo) Uma avaliação com especialistas do caso de uso foi realizada para procurar vieses ou erros conhecidos? (Por exemplo, painéis de revisores podem ser usados para examinar previsões específicas e considerar se são razoáveis ou não. Esses painéis precisam ser balanceados em relação aos tipos de usuário esperados, inclusive tomadores de decisão, se necessário).

3.6 Equidade e desempenho diferencial de preditores

Métodos baseados em aprendizado de máquina podem gerar resultados injustos ou discriminatórios para certos subgrupos (Buolamwini & Gebru, 2018) (Barocas & Selbst, 2014) (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016). Esses resultados podem ser causados por todos os desafios mencionados, tanto da obtenção e manuseio dos dados quanto de erros na formulação do modelo.

Exemplos de desempenho diferencial e iniquidade incluem diferentes taxas de aceitação para receber benefícios em diferentes grupos ou erros de detecção de rostos humanos que diferem de acordo com a raça.

É importante lembrar que a avaliação dos resultados de um sistema de tomada de decisão/suporte à decisão é feita levando-se em conta os objetivos do tomador de decisão, que podem ser diferentes dos objetivos e até contrários a eles do ponto de vista do problema de aprendizado de máquina. Por exemplo, um tomador de decisão pode sacrificar o desempenho geral de um modelo para melhorar o desempenho do modelo em um subgrupo, mesmo que esse subgrupo seja pequeno em comparação com a população como um todo (por exemplo, ação afirmativa para corrigir alguma discriminação social existente).

Embora a análise das implicações éticas dos modelos de aprendizado de máquina e da relação que elas têm com uma definição de justiça ainda seja um campo de estudo em aberto, existe uma literatura considerável que busca implementar definições matemáticas de equidade nos modelos para descrever sua imparcialidade ou discriminação entre subgrupos e tomar decisões que mitiguem os resultados indesejáveis.

3.6.1 Definição de justiça e equidade algorítmicas

O que se entende por "justiça" pode variar de acordo com a cultura ou tradição, e também pode ser específico para uma questão ou projeto de política pública. Por exemplo, em alguns casos, as políticas buscam a inclusão social por meio de ações afirmativas, como cotas de diversidade e políticas de reparação, enquanto, em outros casos, essas políticas baseiam-se simplesmente em argumentos regionais ou territoriais. Esses critérios devem ser integrados durante o processo de formulação, a análise dos dados de treinamento, o processo de avaliação de erros e a saída do sistema. Esse processo pode ser dividido em duas etapas importantes:

- **Definição de justiça algorítmica:** uma representação matemática de uma definição de justiça específica que é incorporada ao processo de ajuste e seleção do modelo. É importante levar em consideração que essas definições podem ser excludentes, ou seja, satisfazer uma pode significar não satisfazer as demais (Verma & Rubin, 2018).

- **Iniquidade algorítmica:** falhas técnicas nos modelos que produzem disparidade de resultados para grupos protegidos, devendo ser avaliadas de acordo com a definição de justiça algorítmica determinada no ponto anterior (pode ser mais de uma).

O objetivo do modelador é estabelecer diretrizes para evitar que deficiências nos modelos produzam disparidades indesejáveis de acordo com os diferentes subgrupos associados a uma variável protegida (por exemplo, gênero, raça ou nível de marginalização). Para isso, é necessário selecionar previamente uma definição de justiça algorítmica. Algumas das mais utilizadas são as seguintes (embora outras possam ser definidas, dependendo do problema específico e dos objetivos dos tomadores de decisão):

1) Omissão de variáveis protegidas

Uma estratégia amplamente contestada para evitar disparidades entre grupos de uma variável protegida é ignorar a variável . Nesse processo, pretende-se eliminar a possibilidade de disparidade ao não incluir a variável no processo de construção de preditores. Essa abordagem NÃO resolve o problema porque:

- Normalmente, existem outros atributos associados a que podem produzir resultados semelhantes, embora não seja considerada (por exemplo, área geográfica ou código postal e nível socioeconômico).
- Pode haver razões importantes para incluir nos modelos preditivos. Por exemplo, no caso da pressão arterial, há variações entre os grupos raciais () quanto à predisposição à hipertensão (Lackland, 2014); portanto, um modelo que avalie o risco seria mais preciso e adequado se incluísse a variável .

(2) Paridade demográfica

Já a paridade demográfica estabelece que a proporção de cada segmento de uma classe protegida (por exemplo, gênero ou determinadas faixas etárias) deve obter um resultado positivo na mesma proporção (por exemplo, a concessão de bolsas escolares). Isso é indesejável por si só: por exemplo, se quiséssemos construir um classificador para uma determinada doença, precisaríamos considerar que é possível que mulheres e homens sejam afetados de formas diferentes. No entanto, a paridade demográfica pode ser um objetivo para os tomadores de decisão, e isso deve ser levado em consideração ao tomar a decisão associada à previsão.

(3) Equidade de possibilidades

O conceito de equidade de possibilidades (Hardt, 2016) depende menos dos objetivos dos tomadores de decisão; refere-se ao desempenho preditivo em diferentes grupos definidos por uma variável protegida (Verma & Rubin, 2018). Se é a variável que queremos prever, e é a nossa previsão, dizemos que a nossa previsão satisfaz a equidade de possibilidades quando e são independentes, dado o valor verdadeiro .

Isso significa que não deve influenciar a previsão quando sabemos o valor verdadeiro ou, em outras palavras, o pertencimento ou não pertencimento ao grupo protegido A não deve influenciar o resultado da classificação.

Considera-se, então, que os preditores que estiverem muito distantes desse critério provavelmente incluirão disparidades associadas à variável protegida A. Uma implicação desse critério é que, de acordo com a hipótese de equidade de possibilidades, as taxas de erro preditivo em cada subgrupo de A são semelhantes, e, para a classificação binária, as taxas de falsos positivos e falsos negativos são semelhantes.

Por exemplo, suponhamos que o objetivo seja criar um sistema para selecionar bolsistas de uma universidade reconhecida. A instituição define o pertencimento a uma comunidade indígena como uma variável protegida (que presumiremos que, neste caso, adota dois va-

lores: autodenomina-se indígena ou não se autodenomina indígena). O preditor satisfaz a equidade de possibilidades quando tanto a taxa de falsos positivos quanto a taxa de falsos negativos são as mesmas para pessoas indígenas e pessoas não indígenas.

(4) Justiça contrafactual

Esta medida considera um preditor como "justo" se seu resultado continuar igual quando o valor do atributo protegido for obtido e alterado para outro valor possível do atributo protegido (como, por exemplo, a introdução de uma mudança de raça, gênero ou outra condição).

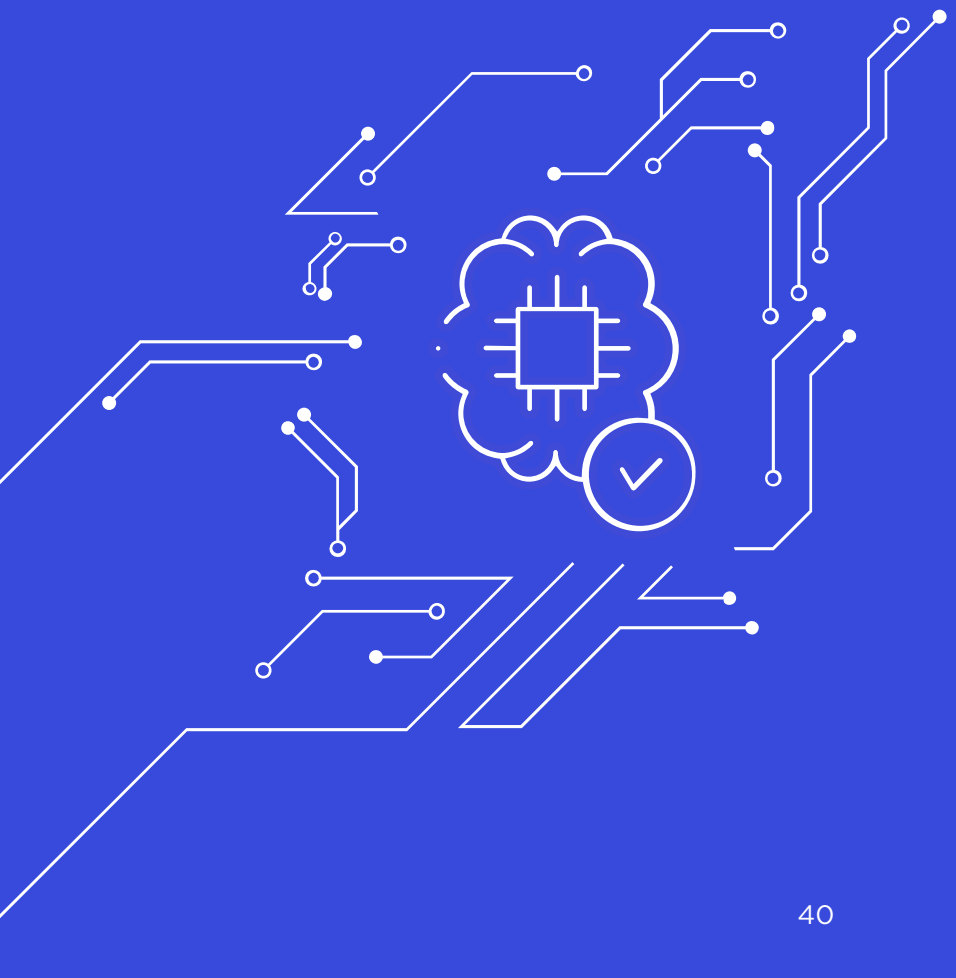
Na prática, não existe uma resposta única nem uma medida de justiça algorítmica que funcione para todos os problemas. Na maioria dos casos, buscar o cumprimento de uma implica não cumprir integralmente as demais; portanto, sua escolha deve ser feita no contexto do problema, e suas razões devem ser justificadas. A equidade de oportunidades muitas vezes é um critério aceitável que introduz critérios de justiça algorítmica, permitindo, também, a otimização de outros resultados desejáveis.

Caixa 17. Lista de verificação - Definição de justiça e equidade algorítmicas

Atividade:  No fim desta fase, recomenda-se preencher as seções de desenvolvimento do modelo do Perfil do modelo e debater com o tomador de decisão de políticas públicas (veja a Ferramenta 3).

VER FERRAMENTA 3

4. USO E MONITORAMENTO



- (Qualitativo) Identificar grupos ou atributos protegidos. Por exemplo: idade, gênero, raça, nível de marginalização etc.
- (Qualitativo) A medida de justiça algorítmica a ser utilizada no projeto foi definida com especialistas e tomadores de decisão?
- (Quantitativo) Quando há atributos protegidos, deve-se avaliar até que ponto as previsões se desviam da definição de justiça algorítmica escolhida.
- (Quantitativo) Pós-processar as previsões devidamente, se necessário, para satisfazer os critérios de justiça algorítmica escolhida (por exemplo, equidade de possibilidades, oportunidades).
- (Quantitativo) No caso da classificação, os pontos de corte podem ser ajustados para diferentes subgrupos, a fim de alcançar a equidade de oportunidades.
- (Quantitativo) Coletar informações mais relevantes sobre subgrupos protegidos (casos e características) para melhorar o desempenho preditivo em grupos minoritários.

4. Uso e monitoramento

Logo depois que os métodos de aprendizado de máquina começarem a ser usados para tomar decisões, será necessário:

- Monitorar, em geral, o desempenho e as características usadas ao longo do tempo.
- Monitorar, em particular, os resultados indesejáveis que podem advir da interação do usuário com os sistemas de tomada de decisão/suporte à decisão.
- Avaliar a coleta e o processamento de dados para melhorar o desempenho ou avaliar os resultados.

4.1 Degradação do desempenho

O desempenho de um modelo pode degradar-se com o passar do tempo por vários motivos:

- Os modelos de ML que estabelecem uma relação estática entre variáveis de entrada e saída podem degradar a qualidade de suas previsões devido a mudanças nas relações subjacentes do contexto de estudo.

- Também pode advir de uma mudança na qualidade dos dados devida ao método de coleta ou mesmo redefinições metodológicas usadas para coletar informações. Por exemplo, em registros administrativos, um ministério ou secretaria pode alterar processos de coleta de dados, digitalizar sistemas, sistematizar limpezas ou processamentos, tornando o aprendizado de um sistema irrelevante.

- Isso também ocorre em sistemas interativos, nos quais o sistema e seus usuários formam um ciclo fechado de feedback e, assim, causam uma degradação do sistema, uma vez que os usuários só podem interagir com elementos decididos pelo sistema.

- Para mitigar esses possíveis erros, é necessário monitorar o comportamento das variáveis de entrada e atualizar as premissas com os tomadores de decisão e com base em conhecimentos especializados.

Também é necessário monitorar o comportamento das métricas de erro ao longo do tempo: quantidades com a taxa total de positivos e negativos (inclusive desagregações por outras variáveis protegidas ou de interesse), distribuição de previsões e atributos.

Caixa 18. Lista de verificação - Degradação do desempenho

Degradação do desempenho:

- (Qualitativo) Existe um plano para monitorar o desempenho do modelo e a coleta de informações ao longo do tempo?
- (Quantitativo) Monitorar várias métricas associadas às previsões em subgrupos previamente definidos (inclusive variáveis protegidas).
- (Quantitativo) Monitorar a deriva (drift) nas distribuições de características em relação ao conjunto de treinamento.
- (Quantitativo) Monitorar mudanças na metodologia de coleta e processamento de dados que possam reduzir a qualidade das previsões.
- (Quantitativo) Idealmente, planejar a coleta de dados sobre a variável não observada, a fim de reajustar os modelos e manter o desempenho.
- (Qualitativo) Quando aplicável e viável, uma fração das previsões deve ser examinada por um grupo de seres humanos e classificada de acordo com algum critério ou medição das variáveis a serem previstas.

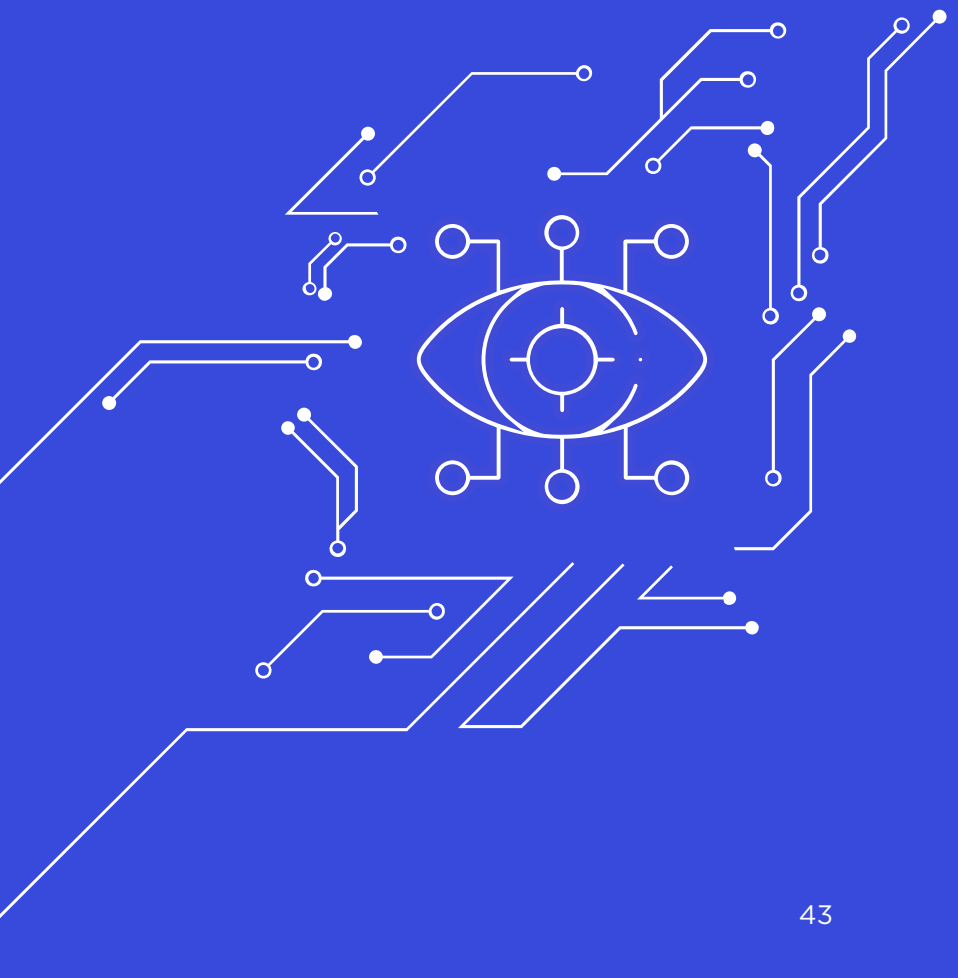
4.2 Experimentos e avaliação do modelo

A forma e os dados coletados para manutenção dos algoritmos de previsão devem ser planejados de modo a melhorar, sempre que possível, e compreender melhor as consequências da utilização dos modelos.

Atividade: No fim desta fase, recomenda-se preencher a seção de uso e monitoramento do Perfil do modelo (veja a Ferramenta 3) e realizar uma discussão com o tomador de decisão de políticas públicas.

VER FERRAMENTA 3

5. PRESTAÇÃO DE CONTAS



As melhorias esperadas nos processos podem ser difíceis de avaliar sem contrafactuais sólidos.

Podem ser planejados testes com delineamento experimental, como, por exemplo, do tipo A/B ou outros (Vaver and Koehler, 2011), quando possível, para entender quais mudanças em particular, desejáveis ou indesejáveis, seriam introduzidas pelo uso dos modelos.

Caixa 19. Lista de verificação - Experimentos e avaliação do modelo

- (Quantitativo) Sempre que possível, planejar atribuir, sob delineamentos experimentais, tratamentos aleatórios (ou conforme o status quo) a algumas unidades. Traçar comparações de desempenho e comportamento entre essa amostra e os resultados de acordo com o regime algorítmico.
- (Quantitativo) Identificar variáveis não observadas e encontrar uma maneira de medi-las. Se possível, reajustar o modelo e avaliar seu desempenho com base nessas novas informações.

5. Prestação de contas

Regulamentos como, por exemplo, o Regulamento Geral de Proteção de Dados (RGPD) da União Europeia, definem responsabilidade como requisito para que as organizações implementem medidas técnicas e organizacionais apropriadas e sejam capazes de demonstrar o que fizeram e sua eficácia, quando solicitado.

Embora o desenvolvimento de padrões e normas técnicas para sistemas de IA ainda seja um trabalho em andamento para a comunidade de IA, este manual descreve os principais aspectos técnicos e medidas para evitar e mitigar vieses durante o ciclo de vida da IA. No entanto, restam vários desafios relacionados aos requisitos sociais e jurídicos que acompanham o uso desses sistemas em aplicações do mundo real.

Esta seção revisará os conceitos de interpretabilidade, explicabilidade e rastreabilidade de sistemas de IA.

5.1 Interpretabilidade e explicação das previsões

5.1.1 Interpretabilidade

É difícil propor uma definição técnica de interpretabilidade ou explicabilidade, termos que geralmente se referem a tornar o funcionamento de um algoritmo e seus resultados inteligíveis para o ser humano (Molnar, 2019). Quanto mais interpretável for um modelo, mais fácil será para um indivíduo entender o processo que o levou a uma determinada decisão (Miller, 2019). Um modelo com alta interpretabilidade é desejável em uma aplicação de política social de alto risco, na qual o critério de responsabilidade se torna essencial.

Existem várias razões pelas quais é importante contar com um certo grau de interpretabilidade nos modelos usados para tomar decisões (Molnar, 2019):

1. Aprendizagem sobre o domínio do problema.
2. Obtenção da aceitação social em relação ao uso dos sistemas.
3. Detecção de vieses em potencial nos algoritmos.

4. Depuração e aprimoramento de modelos.

Algoritmos complexos, como redes neurais profundas, podem ter milhões de relações entre seus parâmetros; portanto, obter a interpretabilidade do modelo nesses algoritmos continua sendo um campo em aberto no aprendizado de máquina. Quando uma alta interpretabilidade é necessária, recomenda-se o uso de métodos intrinsecamente interpretáveis, como regressão linear, regressão logística e árvores de decisão.

5.1.2 Explicabilidade de previsões individuais

Em muitos casos, há uma necessidade jurídica e/ou ética de fornecer explicações individuais sobre como determinadas decisões foram tomadas (por exemplo, por que uma pessoa não obteve crédito ou por que alguém não se qualificou para um programa social).[11]

Em áreas de pesquisa, como visão artificial e processamento de linguagem natural, as implementações mais bem-sucedidas tendem a ser desenvolvidas com modelos altamente complexos, como redes neurais profundas, que inicialmente não são muito transparentes em relação a como as previsões subjacentes são feitas (Carrillo, Cantu , & Noriega, 2020).

Embora seja uma área de pesquisa em andamento, já existem vários métodos para aumentar a explicabilidade das previsões (Molnar, 2019). Podem ser usados métodos como o método de explicações contrafactuais (Wachter, Mittelstadt, & Russell, 2017), valores de Shapley (Lundberg & Lee, 2017) ou gradientes integrados para redes profundas (Sundararajan, Taly, & Yan, 2017).

Caixa 20. Explicabilidade de previsões individuais

- (Qualitativo) Foram analisados os requisitos jurídicos e éticos de explicabilidade e interpretabilidade necessários para o caso de uso?
- (Qualitativo) Caso um usuário seja prejudicado pelos resultados, um plano de resposta foi definido?
- (Qualitativo) Existe um processo para explicar a um determinado indivíduo por que uma decisão foi tomada?
- (Qualitativo) Foram debatidos os prós e contras dos algoritmos de acordo com seu nível de interpretabilidade e explicabilidade, a fim de decidir qual é o mais adequado?
- (Quantitativo) Para modelos mais simples (por exemplo, lineares ou árvores de decisão), explicações ad hoc podem ser elaboradas.
- (Quantitativo) Usar métodos como explicações contrafactuais, valores de Shapley ou gradientes integrados para redes profundas.

5.1.3 Modelos parcimoniosos

Há uma ideia generalizada de que um modelo de ML é sempre melhor quando mais covariáveis são usadas; isso está parcialmente correto, porque o modelo pode encontrar

11 Na União Europeia, por exemplo, o artigo 22 do GDPR descreve o direito de uma pessoa de contestar a decisão de um sistema, principalmente se for uma decisão automática.

padrões entre as inter-relações das variáveis. No entanto, quando se leva em consideração a interpretabilidade, métodos mais parcimoniosos que usam menos características, mas características que são relevantes, são preferíveis a modelos que usam muitas, mas que talvez sejam menos relevantes.

Vieses em potencial podem ocorrer quando se utilizam características ou variáveis de dados que, embora válidas para um período e um conjunto de dados determinados, podem facilmente ser alterada, à medida que o processo de geração de dados evolui. Algoritmos ou métodos de previsão que usam muitos atributos irrelevantes correm um risco maior de falhar, tanto explícita quanto silenciosamente, quando as fontes de dados ou os processos de geração de dados mudam.

Exemplos disso podem ser o uso de variáveis ativamente influenciadas por alguma política que não continuará no futuro ou o aprendizado de características com base em um conjunto de treinamento não exaustivo (por exemplo, no reconhecimento de imagens, reconhecimento de espécies animais conforme o contexto no qual as informações foram coletadas, como um zoológico, uma armadilha fotográfica, uma paisagem etc.).
Esse tipo de viés prejudica a explicabilidade de um sistema e pode ser difícil de detectar, mas métodos parcimoniosos e conhecimentos especializados podem mitigar o risco.

Caixa 21. Lista de verificação - Modelos parcimoniosos

- (Qualitativo) Incluir todas as características disponíveis ao elaborar modelos aumenta o risco de viés. As variáveis a serem incluídas no processo de aprendizagem devem ter algum embasamento teórico ou explicação de por que podem ajudar na tarefa de previsão.
- (Quantitativo) Métodos mais parcimoniosos que usam menos características são preferíveis a modelos que usam muitas características.
- (Quantitativo) Métodos como gráficos de dependência parcial (Friedman, 2001) ou importância baseada em permutação (Breiman, 2001) (Molnar, 2019) podem sinalizar variáveis problemáticas que recebem muito peso na previsão, em contraste com observações anteriores ou os conhecimentos especializados.

5.2 Rastreabilidade

Um processo de dados pouco rastreável para tomada de decisão é aquele que contém etapas com documentação insuficiente sobre sua execução. Essas etapas incluem processos manuais ou decisões do operador pouco especificadas, extraem dados de fontes não documentadas ou inacessíveis, omitem códigos ou materiais necessários, ou não explicam os ambientes de computação para garantir resultados reprodutíveis.

A rastreabilidade permite que o usuário entenda os processos seguidos por um sistema de IA para chegar a um resultado, incluindo quaisquer deficiências e limitações do sistema. Quando há pouca rastreabilidade em um modelo, os riscos observados ao longo deste documento podem ser difíceis de identificar, podendo até ser exacerbados. Por outro lado, todas as etapas, da coleta de dados à tomada de decisão, são claramente documentadas e especificadas de forma inequívoca em um projeto rastreável.

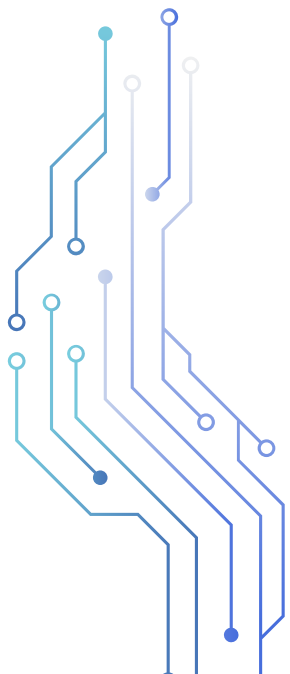
Caixa 22. Lista de verificação - Rastreabilidade

- (Quantitativo) O processo de ingestão, transformação, modelagem e tomada de decisão está bem documentado? (Incluindo fontes de dados, infraestrutura e dependências, código, métricas e interpretação de resultados.)

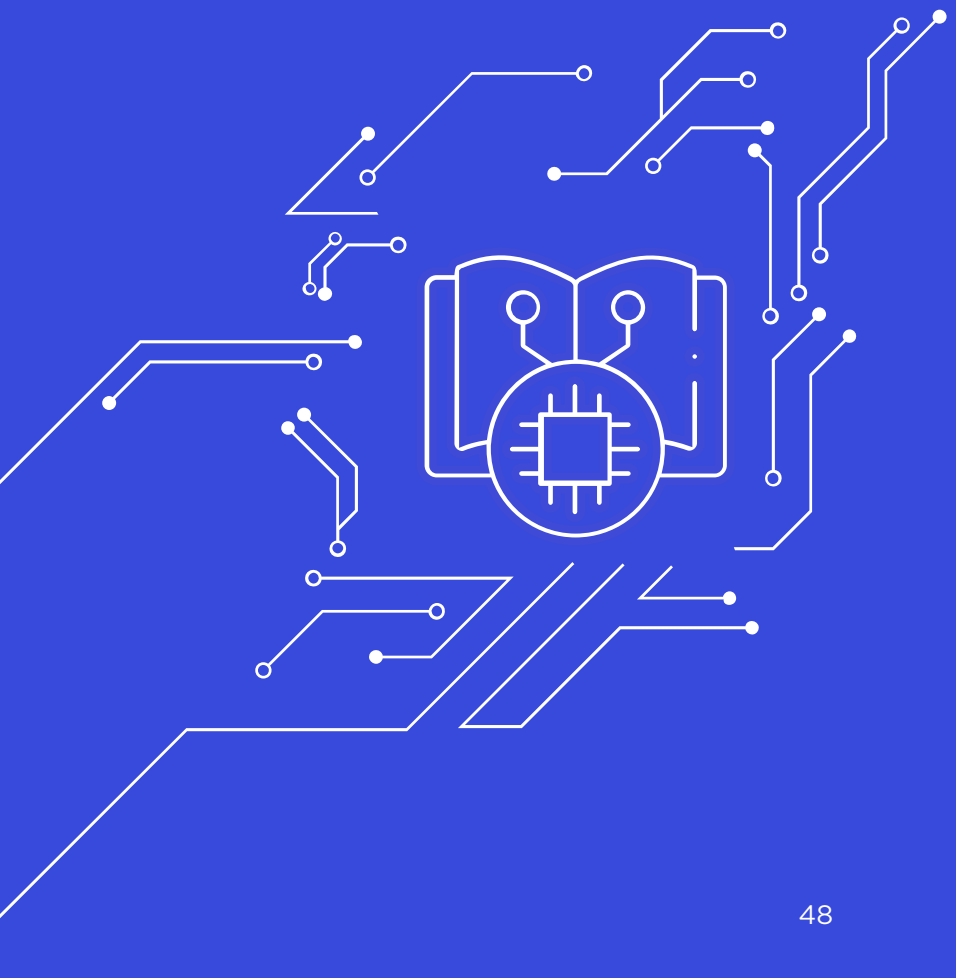
1. **Fontes de dados**, incluindo metadados de conjuntos de dados, processos de coleta de dados e etapas de processamento desses dados (veja a Ferramenta 2).
2. **Código completo e devidamente documentado** que define as bibliotecas necessárias e suas versões apropriadas, permitindo que terceiros entendam o propósito de cada parte do código.
3. **Informações sobre como o código deve ser executado**, incluindo documentação detalhada dos parâmetros e requisitos de computação. Essas informações precisam garantir a reprodutibilidade dos resultados originais por terceiros.
4. **Informações sobre como os resultados do sistema foram utilizados e incluídos** no processo de tomada de decisão de políticas públicas.
5. **Informações sobre a estratégia de monitoramento**, incluindo detalhes sobre métricas de desempenho e limiares, bem como o comportamento esperado do modelo e as ações de mitigação.

Idealmente, um terceiro pode replicar as etapas acima com pouca ou nenhuma intervenção dos criadores e operadores do sistema original.

- (Qualitativo) As deficiências, limitações e vieses do modelo foram comunicados às partes interessadas para consideração na tomada de decisão e suporte à decisão?
- (Qualitativo) A equipe técnica preencheu o Perfil de dados (veja a Ferramenta 2) e o Perfil do modelo (veja a Ferramenta 3), e definiu-se um processo de atualização contínua dessas ferramentas?



FERRAMENTAS



Ferramenta 1: Lista de verificação de uma IA robusta e responsável

Esta ferramenta consolida os principais receios relativos à dimensão de risco do ciclo de vida da IA. A lista de verificação deve ser continuamente revisada pela equipe técnica, acompanhada pelo tomador de decisão (Fritzler 2015; drivendata 2019).

Conceitualização e formulação

☑ ***Definição correta do problema e da resposta por meio de uma política pública:***

- (Qualitativo) O problema de política pública foi claramente definido?
- (Qualitativo) Descrever como esse problema está sendo abordado atualmente, considerando-se as respostas de instituições relacionadas, e como o uso da IA melhoraria a resposta do governo a esse problema.
- (Qualitativo) Os grupos ou atributos protegidos foram identificados dentro do projeto (por exemplo, idade, gênero, escolaridade, raça, nível de marginalização etc.)?
- (Qualitativo) Foram definidas as ações ou intervenções que serão realizadas com base no resultado do sistema de IA?

☑ ***Princípios da IA***

- (Quantitativo) A necessidade de um sistema de IA foi justificada, levando-se em conta outras possíveis soluções que não requeiram a utilização de dados pessoais e decisões automatizadas?
- (Quantitativo) Há evidências de que tanto a ação das políticas públicas quanto a recomendação do sistema de IA beneficiarão as pessoas e o planeta, impulsionando o crescimento inclusivo, o desenvolvimento sustentável e o bem-estar?
- (Qualitativo) Projetos semelhantes foram identificados e analisados em busca de lições e erros comuns?
- (Quantitativo) Considerou-se a possibilidade de minimizar a exposição de informações pessoais identificáveis (por exemplo, anonimizando-se ou não se coletando informações não relevantes para a análise)?

Ciclo da vida

☑ ***Leitura e processamento de dados***
Estados indesejáveis ou subótimos nos dados coletados

- (Qualitativo) Consultar especialistas no assunto sobre possíveis desigualdades sociais históricas no caso de uso .
- (Quantitativo) Realizar uma análise exploratória dos dados disponíveis com os quais o modelo será treinado para identificar possíveis vieses históricos ou estados indesejáveis.

☑ ***Correspondência inadequada entre as variáveis disponíveis e as variáveis ideais***

- (Qualitativo) As variáveis-alvo ideais devem ser claramente estabelecidas. As variáveis coletadas/disponíveis devem ser analisadas para entender até que ponto são adequadas para substituir a variável-alvo. Vieses sistemáticos ou a validade da métrica substituta devem ser identificados.
- (Qualitativo) O uso da variável de resposta selecionada para os propósitos da intervenção foi claramente justificado?

☑ ***Amostras probabilísticas e naturais***

- (Qualitativo) Foram analisadas possíveis diferenças entre o banco de dados e a população para a qual o sistema de IA está sendo desenvolvido? (Usar a bibliografia relacionada ao tema e as informações dos especialistas. Estudar em particular os vieses de seleção não medidos).
- (Quantitativo) Embora os modelos possam ser elaborados com diferentes fontes de dados, artificiais ou naturais, o ideal é que a validação seja realizada com uma amostra que permita uma inferência estatística para a população. A amostra de validação deve abranger adequadamente a população-alvo e as subpopulações de interesse.

☑ ***Atributos ausentes ou incompletos***

- (Qualitativa) Realizou-se uma análise de valores ausentes e variáveis omitidas?
- (Qualitativa) Identificou-se a existência de variáveis importantes omitidas para as quais não existem medidas associadas (se for o caso)?
- (Qualitativa) Os motivos para a falta de observações foram identificados (se for o caso)?
- (Quantitativa) A sensibilidade a suposições e dados dos processos de imputação devem ser avaliados. De preferência, métodos de imputação múltipla devem ser usados para avaliar a incerteza de imputação (Little & Rubin, 2002) (Buuren & Groothuis-Oudshoorn, 2011).

☑ ***Comparação causal***

- (Qualitativo) Compreender e descrever as razões pelas quais a variável de resposta está correlacionada com variáveis conhecidas e desconhecidas. Descrever possíveis vieses com base no conhecimento e na análise de especialistas.
- (Qualitativo) Caso nenhum trabalho tenha sido feito para garantir a causalidade nos resultados, as limitações dos resultados foram explicitamente comunicadas ao responsável pela política pública?
- (Quantitativo) No caso de uma tentativa de inferência causal com modelos, as hipóteses, considerações ou métodos usados para apoiar uma interpretação causal devem ser descritos. Verificações de robustez devem ser realizadas e documentadas.

CoDesenvolvimento do modelo e validação

☑ ***Ausência ou uso inadequado de amostras de validação***

- (Quantitativa) As amostras de validação e teste foram geradas adequadamente, considerando-se um tamanho adequado, abrangendo-se subgrupos de interesse e protegidos, e evitando-se vazamentos de informações durante sua implementação?

☑ ***Contaminação treinamento-validação***

- (Quantitativa) O processamento e a preparação dos dados de treinamento devem evitar ao máximo o uso de dados de validação ou teste. Uma barreira sólida deve ser mantida entre o treinamento em contraste com a validação e os testes. Isso inclui recodificação de dados, normalizações, seleção de variáveis, identificação de dados atípicos e qualquer outro tipo de preparação de qualquer variável a ser incluída nos modelos, o que também envolve ponderações ou balanceamento de amostras com base em sobre/subamostragem.

☑ ***Vazamentos de dados indisponíveis na previsão***

- O esquema de validação deve replicar com a maior precisão possível o esquema por meio do qual as previsões serão aplicadas. Isso inclui a necessidade de replicar:
- Janelas temporárias para observação e registro de variáveis, bem como janelas de previsão.
- Se existirem grupos nos dados, considerar se as informações de cada grupo estarão disponíveis quando a previsão for feita ou se será necessário prever para novos grupos.

☑ ***Dados desbalanceados***

- (Quantitativa) Fazer previsões de probabilidade em vez de classe. Essas probabilidades podem ser incorporadas ao processo de decisão subsequente. Evitar pontos de corte padrão de probabilidade, como 0,5, ou fazer previsões conforme a probabilidade máxima.
- (Quantitativa) Quando o número absoluto de casos minoritários é muito pequeno, pode ser muito difícil encontrar informações apropriadas para discriminar essa classe. Mais dados da classe minoritária precisam ser coletados.
- (Quantitativa) A subamostragem da classe dominante (ponderação para cima dos casos para evitar a perda de calibração) pode ser uma estratégia bem-sucedida para reduzir o tamanho dos dados e o tempo de treinamento sem afetar o desempenho preditivo.
- (Quantitativa) Replicar a classe minoritária para balancear melhor as classes (sobreamostragem).
- (Quantitativa) Algumas técnicas de aprendizado de máquina permitem que cada classe seja ponderada com um peso diferente, de modo que o peso total de cada classe esteja balanceado. Se isso for possível, é preferível à sub ou sobreamostragem.

☑ ***Pontos de corte arbitrários***

- (Quantitativo) É mais adequado usar algoritmos de classificação probabilística, para que a tomada de decisão incorpore a incerteza em relação à classificação.
- (Quantitativo) Evitar pontos de corte de probabilidade padrão, como 0,5. Escolher uma interpretação ideal das probabilidades previstas, analisando-se as métricas de erro.

☑ ***Adequação das métricas de avaliação***

- (Qualitativa) As implicações dos diferentes tipos de erro foram questionadas para o caso de uso específico, bem como a forma correta de avaliá-los?
- (Qualitativa) As limitações do modelo foram claramente explicadas, identificando-se falsos positivos e falsos negativos, bem como as implicações que uma decisão do sistema teria na vida da população beneficiária?
- (Quantitativa) Implementou-se uma análise de custo-benefício do sistema em contraste com o status quo ou outras estratégias de tomada de decisão/suporte à decisão (quando possível)?

☑ ***Sub e sobreajuste***

- (Quantitativa) Sobreajuste: modelos com uma lacuna considerável entre validação e treinamento (indícios de sobreajuste) devem ser evitados. Se necessário, os métodos devem ser refinados para moderar o sobreajuste, como regularização, restrição do espaço funcional de modelos possíveis, uso de mais dados de treinamento ou perturbação dos dados de treinamento, entre outros (Hastie, Tibshirani, & Friedman, 2017).
- (Quantitativa) Subajuste: subconjuntos importantes de casos (por exemplo, grupos protegidos) devem ser revisados para verificar se há erros sistemáticos indesejáveis.

☑ ***Falhas não medidas pelo modelo***

- (Qualitativa) Uma avaliação com especialistas do caso de uso foi realizada para procurar vieses ou erros conhecidos? (Por exemplo, painéis de revisores podem ser usados para examinar previsões específicas e considerar se são razoáveis ou não. Esses painéis precisam ser balanceados em relação aos tipos de usuário esperados, inclusive tomadores de decisão, se necessário).

☑ ***Definição de justiça e equidade algorítmicas***

- (Qualitativa) Identificar grupos ou atributos protegidos. Por exemplo, idade, gênero, raça, nível de marginalização etc.
- (Qualitativa) A medida de justiça algorítmica a ser utilizada no projeto foi definida com especialistas e tomadores de decisão?
- (Quantitativa) Quando há atributos protegidos, deve-se avaliar até que ponto as previsões se desviam da definição de justiça algorítmica escolhida.
- (Quantitativa) Pós-processar as previsões devidamente, se necessário, para satisfazer os critérios de justiça algorítmica escolhida (por exemplo, equidade de possibilidades, oportunidades).

- (Quantitativa) No caso da classificação, os pontos de corte podem ser ajustados para diferentes subgrupos, a fim de alcançar a equidade de oportunidades.
- (Quantitativa) Coletar informações mais relevantes sobre subgrupos protegidos (casos e características) para melhorar o desempenho preditivo em grupos minoritários.

Uso e monitoramento

☑ ***Degradação do desempenho***

- (Qualitativo) Existe um plano para monitorar o desempenho do modelo e a coleta de informações ao longo do tempo?
- (Quantitativa) Monitorar várias métricas associadas às previsões em subgrupos previamente definidos (inclusive variáveis protegidas).
- (Quantitativa) Monitorar a deriva (drift) nas distribuições de características em relação ao conjunto de treinamento.
- (Quantitativa) Monitorar mudanças na metodologia de coleta e processamento de dados que possam reduzir a qualidade das previsões.
- (Quantitativa) Idealmente, planejar-se para coletar dados sobre a variável não observada a fim de reajustar os modelos e manter o desempenho.
- (Qualitativa) Quando aplicável e viável, uma fração das previsões deve ser examinada por um grupo de seres humanos e classificada de acordo com algum critério ou medição das variáveis a serem previstas.

☑ ***Experimentos e avaliação do modelo***

- (Quantitativa) Sempre que possível, planejar atribuir sob delineamentos experimentais, tratamentos aleatórios (ou conforme o status quo) a algumas unidades. Traçar comparações de desempenho e comportamento entre essa amostra e os resultados de acordo com o regime algorítmico.
- (Quantitativo) Identificar variáveis não observadas e encontrar uma maneira de medi-las. Se possível, reajustar o modelo e avaliar seu desempenho com base nessas novas informações.

Prestação de contas

☑ ***Explicabilidade de previsões individuais***

- (Qualitativo) Os requisitos jurídicos e éticos de explicabilidade e interpretabilidade necessários foram analisados para o caso de uso?
- (Qualitativo) Caso um usuário seja prejudicado pelos resultados, um plano de resposta foi definido?
- (Qualitativo) Existe um processo para explicar a um determinado indivíduo por que uma decisão foi tomada?
- (Qualitativo) Os prós e contras dos algoritmos de acordo com seu nível de interpretabilidade e explicabilidade foram debatidos para decidir qual é o mais adequado?

- (Quantitativo) Para modelos mais simples (por exemplo, lineares ou árvores de decisão), explicações ad hoc podem ser elaboradas.
- (Quantitativo) Usar métodos como explicações contrafactuais, valores de Shapley ou gradientes integrados para redes profundas.

☑ ***Modelos parcimoniosos***

- (Qualitativo) Incluir todas as características disponíveis ao elaborar modelos aumenta o risco de viés. As variáveis a serem incluídas no processo de aprendizado devem ter algum embasamento teórico ou explicação de por que podem ajudar na tarefa de previsão.
- (Quantitativa) Métodos mais parcimoniosos que usam menos características são preferíveis a modelos que usam muitas características.
- (Quantitativa) Métodos como gráficos de dependência parcial (Friedman, 2001) ou importância baseada em permutação (Breiman, 2001) (Molnar, 2019) podem sinalizar variáveis problemáticas que recebem muito peso na previsão, em contraste com observações anteriores ou os conhecimentos especializados.

☑ ***Rastreabilidade***

- (Quantitativa) O processo de ingestão, transformação, modelagem e tomada de decisão está bem documentado (incluindo fontes de dados, infraestrutura e dependências, código, métricas e interpretação de resultados)?
  - Fontes de dados, incluindo metadados de conjuntos de dados, processos de coleta de dados e etapas de processamento desses dados (veja a Ferramenta 2).
  - Código completo e adequadamente documentado que define as bibliotecas necessárias e suas versões apropriadas, permitindo que terceiros entendam o propósito de cada parte do código.
  - Informações sobre como o código deve ser executado, incluindo documentação detalhada dos parâmetros e requisitos de computação. Essas informações precisam garantir a reprodutibilidade dos resultados originais por terceiros.
  - Informações sobre como os resultados do sistema foram utilizados e incluídos no processo de tomada de decisão de políticas públicas.
  - Informações sobre a estratégia de monitoramento, incluindo detalhes sobre métricas de desempenho e limiares, bem como o comportamento esperado do modelo e as ações de mitigação.
- (Qualitativo) As deficiências, limitações e vieses do modelo foram comunicados às partes interessadas para consideração na tomada de decisão e suporte à decisão?
- (Qualitativo) A equipe técnica preencheu o Perfil de dados (veja a Ferramenta 2) e o Perfil do modelo (veja a Ferramenta 3), e definiu-se um processo de atualização contínua dessas ferramentas?

Ferramenta 2: Perfil de dados

O perfil de dados é uma análise exploratória que fornece informações para avaliar a qualidade, integridade, temporalidade, consistência e possíveis vieses de um conjunto de dados que será usado para treinar um modelo de aprendizado de máquina (Gebru et al., 2018).

Fonte de coleta e origem dos dados

- Nome do conjunto de dados usado.
- Que instituição criou o conjunto de dados?
- Com que finalidade a instituição criou o conjunto de dados utilizado?
- Que mecanismos ou procedimentos foram usados para coletar os dados (por exemplo, pesquisa domiciliar, sensor, software, API)? Eles cumprem as normas vigentes relativas à proteção de dados?
- Qual é a escala do conjunto de dados?
- Obter a documentação de cada variável do conjunto de dados. Fornecer uma breve descrição, incluindo seu nome e tipo, o que representa, como é medido etc.

Governança de dados

- Qual é o domínio dos dados? (Por exemplo, proprietário, público, pessoal.)
- Se pessoais, os dados são identificados, contêm pseudônimos ou pseudônimos não vinculados, são anônimos ou agregados?
- Se privados, os direitos de propriedade intelectual e proteção de dados pessoais foram levados em consideração?

Estrutura dos dados

- Os dados são estáticos ou dinâmicos? Se dinâmicos, com que frequência serão atualizados?
- Captar a frequência (diária, semanal, mensal) ou o número médio de observações por indivíduo. Que versão do conjunto de dados está sendo usada?
- O conjunto de dados é o mais apropriado disponível, considerando-se o problema em questão?

Qualidade dos dados

- Como os dados foram obtidos? (Observados, derivados, sintéticos ou fornecidos por indivíduos ou organizações).
- Os dados são representativos da população de interesse?
- Descrever o tipo de amostragem usado para obter os dados.
- Analisar a cobertura espacial e temporal dos dados.
- Analisar a cobertura dos grupos protegidos (gênero, raça, idade etc.).

- Descrever as dimensões importantes nas quais a amostra de dados pode diferir da população, particularmente vieses de seleção não medidos. Usar literatura relacionada ao assunto e informações de especialistas.
- Identificar possíveis "estados indesejáveis" nos dados que possam levar a vieses e inequidades prejudiciais a um determinado subgrupo, ou qualquer outro padrão considerado subótimo ou indesejável do ponto de vista da política social.
- Há valores ausentes? Em caso afirmativo, explicar por que essas informações não estão disponíveis (inclusive informações removidas intencionalmente). Identificar os motivos da ausência dos dados e considerar se estão associados à variável a ser prevista. Documentar qualquer processo de imputação usado para substituir dados ausentes.

Ferramenta 3: Perfil do modelo (Model Card)

O guia apresentado aqui é um cartão de acompanhamento que resume as principais características de um sistema de tomada de decisão/suporte à decisão baseado em ML e destaca as principais premissas, as características mais importantes do sistema e as medidas de mitigação implementadas (Mitchell et al., 2019).

Conceitualização e formulação de políticas públicas

1. Informações básicas
   - Pessoas que desenvolveram o modelo, data, versão, tipo.
2. Casos de uso
   - Antecedentes.
   - População-alvo e horizonte de previsões.
   - Atores e componentes que interagirão com os resultados.
   - Casos de uso considerados durante o desenvolvimento.
   - Usos não considerados e advertências relacionadas.
   - Definição de grupos protegidos.

Fonte e gerenciamento de dados

3. Dados de treinamento
   - Conjunto de dados usado e sua rotulagem.
   - Etapas de pré-processamento ou preparação de dados.
   - Possíveis vieses e deficiências, dependendo do caso de uso.

Desenvolvimento do modelo

4. Modelagem
   - Algoritmos usados para treinar, parâmetros assumidos ou restrições.
5. Métricas de desempenho
   - Métricas técnicas usadas para selecionar e avaliar modelos.
   - Análise de custo-benefício do modelo para o caso de uso em questão.
   - Definição de grupos protegidos e medidas de equidade selecionadas.
6. Dados de validação
   - Conjuntos de dados usados e sua rotulagem.
   - Etapas de pré-processamento.
   - Avaliação da adaptação dos dados de validação de acordo com o caso de uso.

- Possíveis vieses e deficiências, dependendo do caso de uso.

7. Resumo da análise quantitativa
   - Erro de validação relatado.
   - Resumo da análise custo-benefício.
   - Relatório de medidas de equidade para grupos protegidos.

Uso e monitoramento

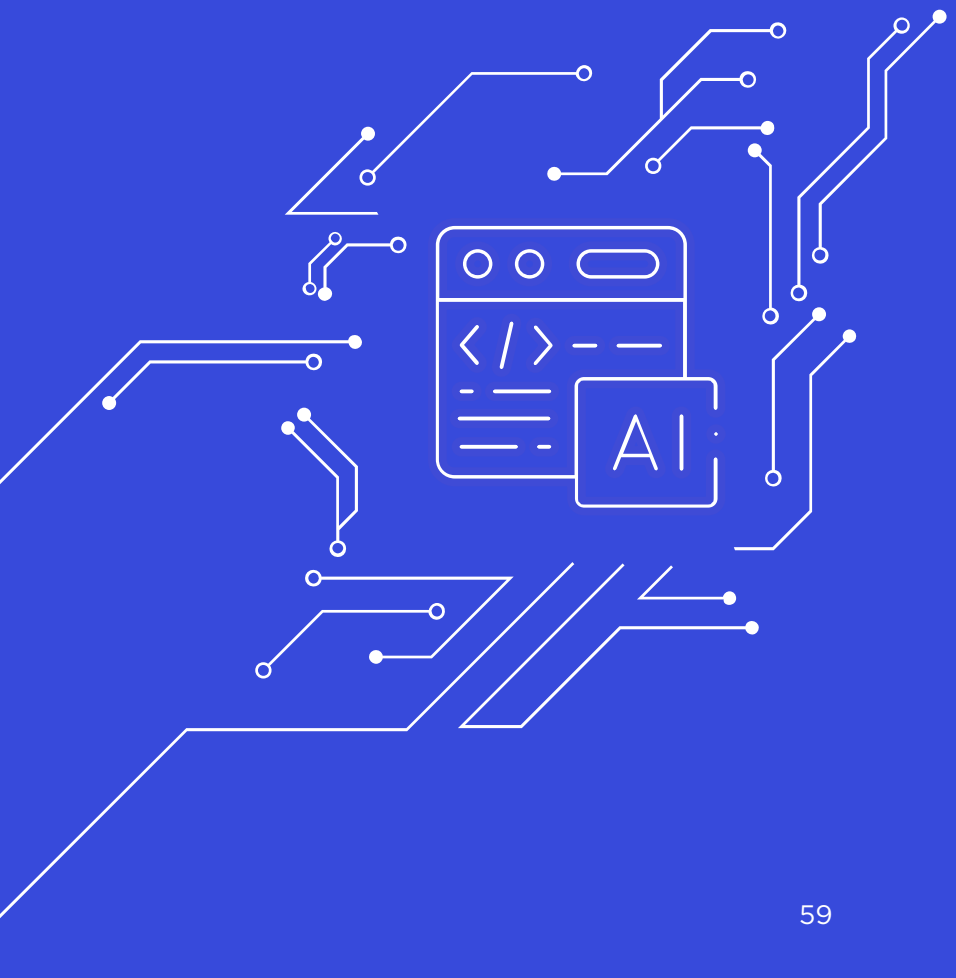
8. Recomendações de monitoramento
   - Estratégia de monitoramento e aprimoramento da produção
   - Estratégias de monitoramento humano das previsões (se for o caso).

Prestação de contas

9. (Opcional) Explicabilidade das previsões.
   - Estratégia para explicar previsões específicas (se necessário).
   - Estratégia para entender a importância de diferentes atributos.

10. Outras considerações éticas, recomendações e advertências.

CADERNOS DE TRABALHO



Cadernos de trabalho

Esta seção oferece vários exemplos dos desafios e soluções explicados no documento principal. Diferentes tipos de modelo (linear, baseados em árvore e outros) e diferentes implementações (R, keras, xgboost) são usados para mostrar que esses problemas ocorrem independentemente da escolha de ferramentas específicas.

Os cadernos usam a notação de ponto decimal para manter a consistência com os pacotes que a utilizam. A linguagem de programação R e os seguintes pacotes são usados: tidyverse, recipes, themis, rsample, parsnip, yardstick, workflows, tune, knitr, patchwork.

Todo o material é reproduzível conforme as instruções contidas neste repositório, que contém um arquivo Dockerfile que descreve as dependências de infraestrutura para replicação.

Coleta e processamento de dados

Correspondência inadequada entre as variáveis disponíveis e as variáveis ideais

O uso de modelos que preveem a métrica incorreta pode levar a decisões erradas. Às vezes, quando a métrica substituta apresenta deficiências óbvias, o problema fica claro; em outras, pode ser mais sutil.

No exemplo a seguir, buscamos prever a demanda por um determinado produto (como vacinas ou algum medicamento) para tomar decisões de abastecimento.

Existem dados históricos de inventário (80 semanas), vendas e uma variável predictor associada às vendas (no caso das vacinas, pode ser a temperatura) e outra para o esgotamento do inventário. Separamos os dados de treinamento e teste, auxiliando o modelo com o subconjunto de dados de treinamento. Neste caso, um modelo linear é usado com a variável dependente “vendas” e as covariáveis “semana” e “predictor”.

```
entrena <- ventas %>% filter(semana < 60)

prueba <- ventas %>% filter(semana >= 60, semana <= 80)

entrena %>% select(-demanda) %>%  head() %>% kable()
```

Semana	inventário	vendas	predictor	esgotamento
1	153	110	-27.7014124	0
2	170	148	0.7664636	0
3	158	130	-15.2606032	0
4	162	142	4.2461227	0
5	159	159	28.5107593	1
6	162	162	14.8895964	1

```
mod_lineal <- lm(ventas ~ semana + predictor, data = ventas)

mod_lineal
```

```
##

## Call:

## lm(formula = ventas ~ semana + predictor, data = ventas)

##

## Coefficients:

## (Intercept)       semana    predictor

##    140.9935       0.8166       0.5535
```

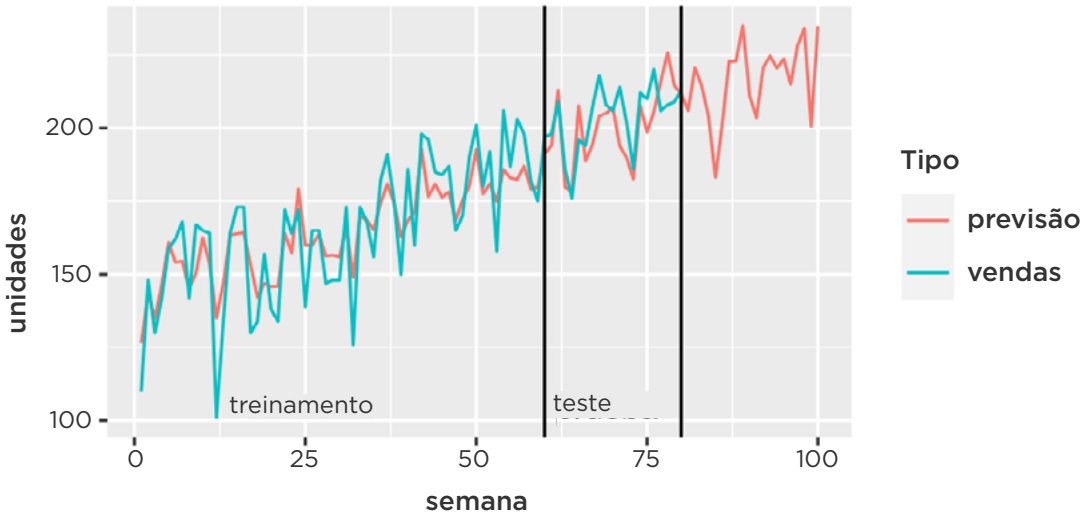
Avaliamos o erro de previsão.

```
preds <- predict(mod_lineal, newdata = prueba)

round(mean(abs(preds - prueba$ventas))/mean(prueba$ventas), 3)
```

```
## [1] 0.04
```

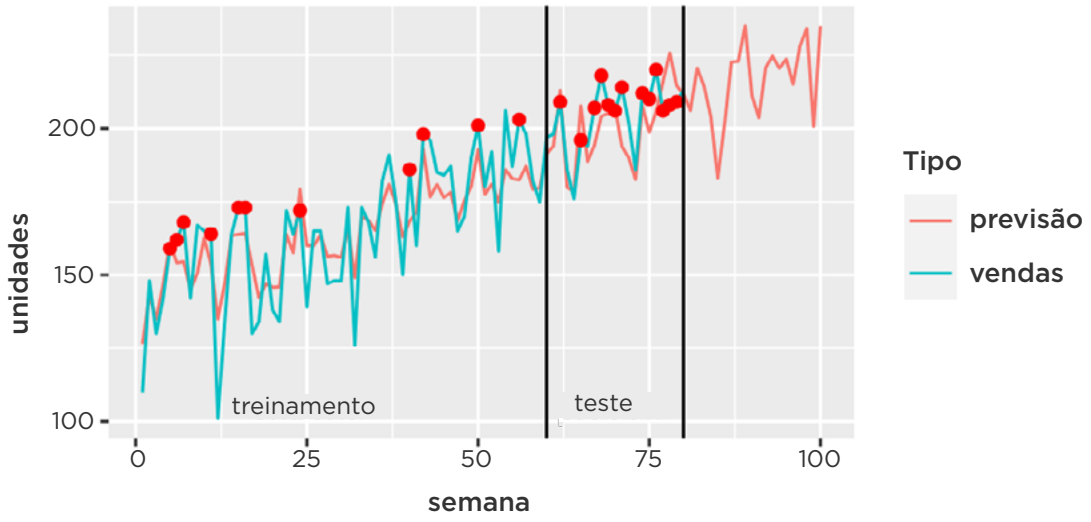
O erro percentual é baixo. Os dados ajustados e as previsões são os seguintes:

```
preds <- predict(mod_lineal, newdata = ventas)

ventas_larga <- ventas %>% mutate(pred = preds) %>%

  pivot_longer(cols = all_of(c("ventas","pred")), names_to = "tipo", values_to
= "unidades")

ggplot(ventas_larga %>% mutate(unidades = ifelse(tipo=="ventas" & semana > 80,
NA, unidades)),

       aes(x = semana, y = unidades, group = tipo, colour = tipo)) +

  geom_line() +

  geom_vline(xintercept = 80) +

  geom_vline(xintercept = 60) +

  annotate("text", x = 25, y=105, label = "entrena") +

  annotate("text", x = 69, y=105, label = "prueba")
```



No entanto, tomar decisões de demçanda ou estoque é errado, pois existe uma diferença entre a variável ideal (demanda real por medicamentos) e a variável observada (venda de medicamentos). A diferença é que existem esgotamentos de estoque, ou seja, períodos durante os quais, embora houvesse demanda, não havia estoque suficiente para todos os compradores. Isso está marcado em vermelho no gráfico a seguir.

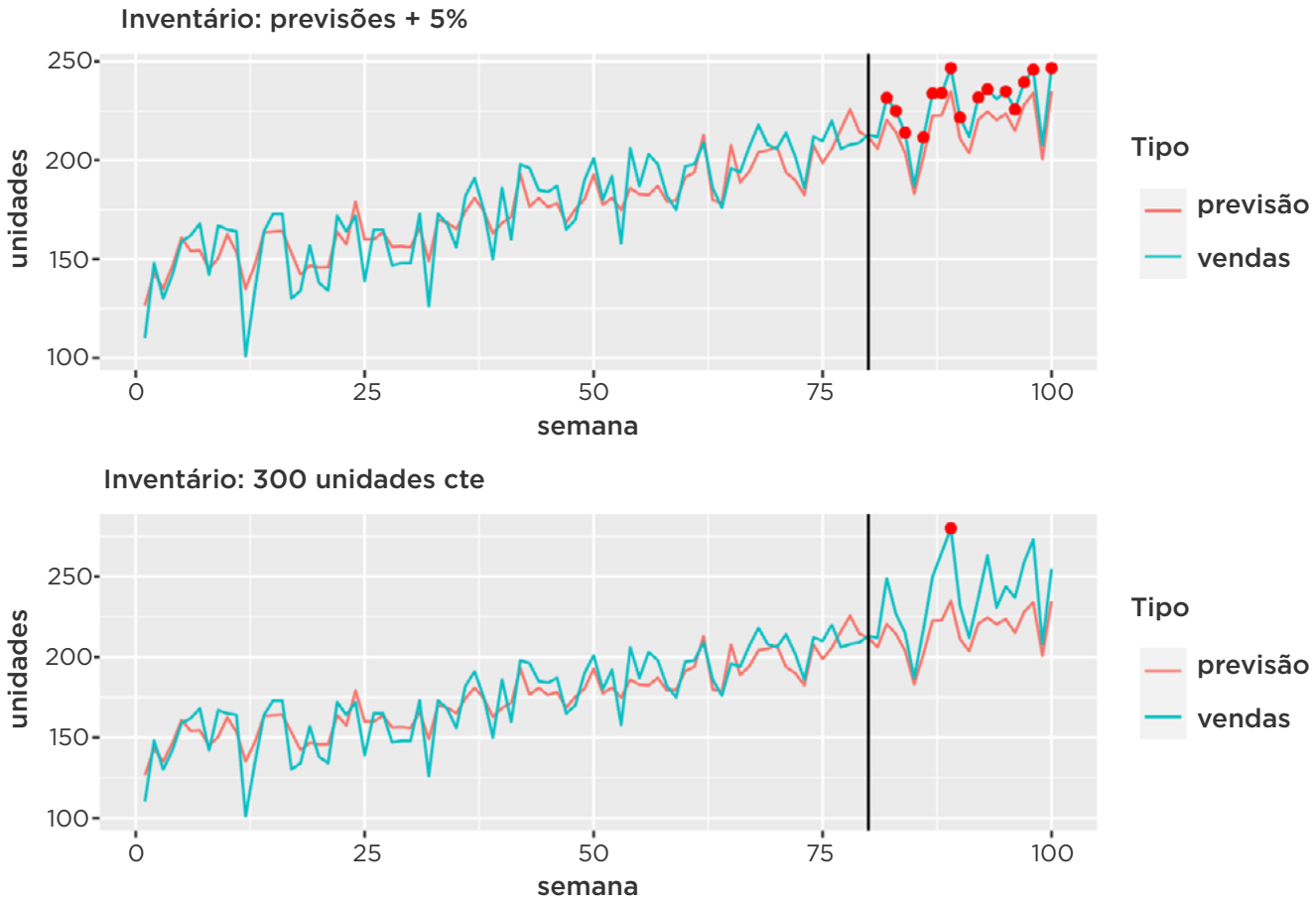
```
preds <- predict(mod_lineal, newdata = ventas)

ventas_larga <- ventas %>% mutate(pred = preds) %>%

  pivot_longer(cols = all_of(c("ventas","pred")), names_to = "tipo", values_to
= "unidades")

ggplot(ventas_larga %>% mutate(unidades = ifelse(tipo=="ventas" & semana > 80,
NA, unidades)), aes(x = semana)) +

  geom_line(aes(group = tipo, colour = tipo, y = unidades)) +

  geom_point(data = filter(ventas, agotamiento==1, semana < 80), aes(y = ventas),
colour = "red") +

  geom_vline(xintercept = 80) +

  geom_vline(xintercept = 60) +

  annotate("text", x = 25, y=105, label = "entrena") +

  annotate("text", x = 69, y=105, label = "prueba")
```



Se usássemos a política sugerida pelas previsões (por exemplo, 5% a mais), veríamos as vendas do primeiro gráfico abaixo. Porém, se adotássemos uma política de estoque com 280 unidades, observaríamos:

```
preds <- predict(mod_lineal, newdata = ventas)

ventas_obs <- ventas %>% mutate(pred = preds) %>%

  mutate(inventario = 1.05 *  pred) %>%

  mutate(ventas = ifelse(semana > 80, pmin(inventario, demanda), ventas))

ventas_larga <- ventas_obs %>%

  pivot_longer(cols = all_of(c("ventas","pred")), names_to = "tipo", values_to
= "unidades")

g1 <- ggplot(ventas_larga, aes(x = semana)) +

  geom_line(aes(group = tipo, colour = tipo, y = unidades)) +

  geom_point(data = filter(ventas_obs, ventas == inventario, semana > 80), aes(y
= ventas), colour = "red") +

  geom_vline(xintercept = 80) + labs(subtitle = "Inventario: Predicciones + 5%")

preds <- predict(mod_lineal, newdata = ventas)

ventas_obs <- ventas %>% mutate(pred = preds) %>%

  mutate(inventario = 280) %>%

  mutate(ventas = ifelse(semana > 80, pmin(inventario, demanda), ventas))
```

```
ventas_larga <- ventas_obs %>%

pivot_longer(cols = all_of(c("ventas","pred")), names_to = "tipo", values_to = "unidades")

g2 <- ggplot(ventas_larga, aes(x = semana)) +

  geom_line(aes(group = tipo, colour = tipo, y = unidades)) +

  geom_point(data = filter(ventas_obs, ventas == inventario, semana > 80), aes(y = ventas), colour = "red") +

  geom_vline(xintercept = 80)+ labs(subtitle = "Inventario: 300 unidades cte")

g1 / g2
```



Então,

- A política baseada em previsões **exacerba** o problema dos esgotamentos.
- Um uso não intencional dos dados, sem considerar o processo que os gera, pode produzir grandes erros nas decisões.
- Nesse caso, a confusão deve-se à não separação entre os conceitos de demanda e vendas. Outros indicadores de demanda ou modelos mais apropriados ajudariam a resolver o problema.
- Soluções simplistas, como apenas coletar dados onde não ocorrem esgotamentos, podem piorar ainda mais a situação: aumentam o viés (selecionamos semanas durante as quais as vendas tendem a ser baixas) e reduzem a precisão.

Amostras probabilísticas e naturais

Quando as amostras de treinamento são diferentes das populações às quais os modelos serão aplicados, é difícil validar corretamente as previsões.

Neste exemplo, os dados da pesquisa nacional de renda e despesa familiar do México (INEGI, 2014) serão usados para simular o cenário que queremos exemplificar:

```
set.seed(128)

encuesta_ingreso <- read_csv("datos/enigh-ejemplo.csv")

datos_ingreso <- encuesta_ingreso %>%

  mutate(num_focos = FOCOS) %>%

  mutate(ingreso_miles = (INGCOR / 1000)) %>%

  mutate(tel_celular = ifelse(SERV_2 == 1, "Sí", "No")) %>%

  mutate(piso_firme = ifelse(PISOS != 1 | is.na(PISOS), "Sí", "No")) %>%

  mutate(lavadora = ifelse(LAVAD != 1 | is.na(LAVAD), "Sí", "No")) %>%

  mutate(automovil = VEHI1_N > 0) %>%

  mutate(marginacion = fct_reorder(marginación, ingreso_miles, median)) %>%

  rename(ocupadas = PEROCU) %>%

  rename(educacion_jef = NIVELAPROB) %>%

  select(ingreso_miles, num_focos, tel_celular,

         marginacion, ocupadas, piso_firme, lavadora, automovil, educacion_jef)

ingreso_split <- initial_split(datos_ingreso, prop = 0.7)

entrena <- training(ingreso_split)

prueba <- testing(ingreso_split)
```

Suponhamos que estamos interessados em estimar a renda domiciliar. Para isso, é utilizada uma pesquisa por celular; além disso, suponhamos que apenas as áreas que não possuem uma marginalização muito alta são acessadas.

```
muestra_sesgada <- filter(entrena,

                          tel_celular == "Sí",

                          marginacion=="Muy bajo")

sesgados_split <- initial_split(muestra_sesgada)

entrena_sesgo <- training(sesgados_split)

validacion_sesgo <- testing(sesgados_split)
```

Um modelo linear é construído para o logaritmo da renda com os dados disponíveis.

```
library(splines)

formula <- as.formula("log(ingreso_miles) ~ ns(num_focos, 3) +

              ns(ocupadas, 3) +  lavadora + automovil + piso_firme +

              ns(educacion_jef, 3)")

mod_sesgo <- lm(formula, data = entrena_sesgo)

# pegamos uma amostra representativa para comparação, do mesmo tamanho da amostra enviesada

mod_representativa <- lm(formula, data = sample_n(entrena, nrow(entrena_sesgo)))
```

E o erro é avaliado em uma amostra de teste construída com dados que têm as mesmas características enviesadas dos dados de treinamento (domicílios com celular e baixíssimo grau de marginalização).

```
preds_val <- predict(mod_sesgo, newdata = validacion_sesgo)

mean(abs(preds_val - log(1 + validacion_sesgo$ingreso_miles))) %>% round(2)
```

```
## [1] 0.37
```

O erro é maior em uma amostra mais semelhante à população à qual se pretende aplicar o algoritmo:

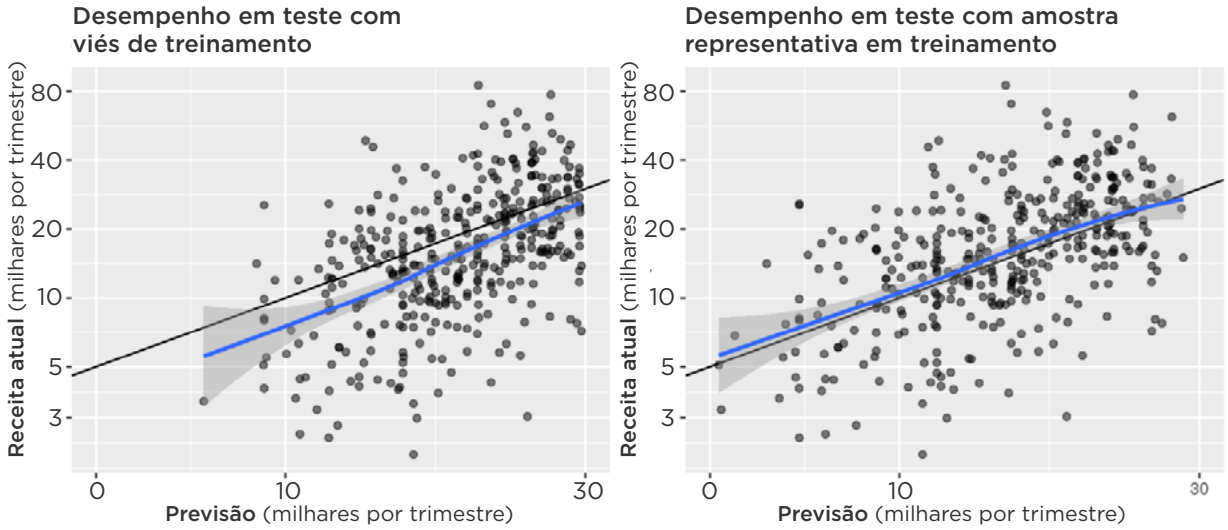
```
preds_prueba_sesgo <- predict(mod_sesgo, newdata = prueba)
```

```
preds_prueba <- predict(mod_representativa, newdata = prueba)

prueba$pred_sesgada <- preds_prueba_sesgo

prueba$pred_rep <- preds_prueba

mean(abs(preds_prueba_sesgo - log(1 + prueba$ingreso_miles))) %>% round(2)
```

```
## [1] 0.42
```

No entanto, o principal problema está refletido no gráfico a seguir, no qual escalas logarítmicas são usadas para traçar comparações multiplicativas, que são interessantes devido à natureza da renda. Cada ponto representa um domicílio. A amostra é mais semelhante à população à qual a metodologia será aplicada. O eixo horizontal mostra a previsão dos domicílios, usando o modelo, enquanto o eixo vertical corresponde à renda de cada domicílio. Como referência, adicione a linha $y = x$ e um suavizador (veja aqui, por exemplo). O foco está no desempenho para os domicílios de renda relativamente baixa (menos de 10 mil pesos por mês):

```
breaks_y <- c(3, 5, 10, 20, 40, 80)

g_sesgo <- ggplot(prueba %>% filter(pred_sesgada < log(30)),

  aes(x = exp(pred_sesgada), y = ingreso_miles)) +

  geom_point(alpha = 0.5) +

  geom_abline() + geom_smooth(method = "loess", span = 1) +

  scale_x_log10(limits=c(5, 30)) + scale_y_log10(breaks = breaks_y) +

  xlab("Predicción (miles al trimestre)") +

  ylab("Ingreso corriente (miles al trimestre)") +

  labs(subtitle = "Desempeño en prueba \ncon sesgo en entrenamiento")

g_representativa <- ggplot(prueba %>% filter(pred_sesgada < log(30)),

  aes(x = exp(pred_rep), y = ingreso_miles)) +

  geom_point(alpha = 0.5) +

  geom_abline() + geom_smooth(method = "loess", span = 1) +

  scale_x_log10(limits = c(5, 30)) + scale_y_log10(breaks = breaks_y) +

  xlab("Predicción (miles al trimestre)") +

  ylab("Ingreso corriente (miles al trimestre)") +

  labs(subtitle = "Desempeño en prueba \ncon muestra representativa en entrenamiento")

g_sesgo + g_representativa
```



Embora normalmente se espere superestimar valores observados relativamente baixos e o contrário para valores relativamente altos, para aqueles com renda inferior a 10 mil pesos por mês, o modelo enviesado superestima a renda real em cerca de 40%:

```
prueba_bajo <- prueba %>% filter(ingreso_miles < 3*10)

sesgo <- mean(exp(prueba_bajo$pred_sesgada))/mean(prueba_bajo$ingreso_miles) -
1

round(sesgo,3)
```

```
## [1] 0.412
```

Quando comparado ao mesmo modelo treinado em uma amostra representativa, na qual o efeito é consideravelmente menor:

```
prueba_bajo <- prueba %>% filter(ingreso_miles < 3*10)

sesgo <- mean(exp(prueba_bajo$pred_rep))/mean(prueba_bajo$ingreso_miles) - 1

round(sesgo,3)
```

```
## [1] 0.152
```

Então, existem dois problemas:

1. O viés produz um erro consideravelmente maior na implementação que na validação.
2. Pior ainda, o viés é maior para as famílias de baixa renda (as previsões são altas), o que pode levar a uma focalização inadequada ao tentar identificar os domicílios de baixa renda.

Amostras naturais: comparações causais

Este exemplo foi retirado de (Hastie, Tibshirani, & Friedman, 2017) e (Rossouw, 1983). São considerados os seguintes dados, com o objetivo de prever doenças cardíacas (chd)[12]:

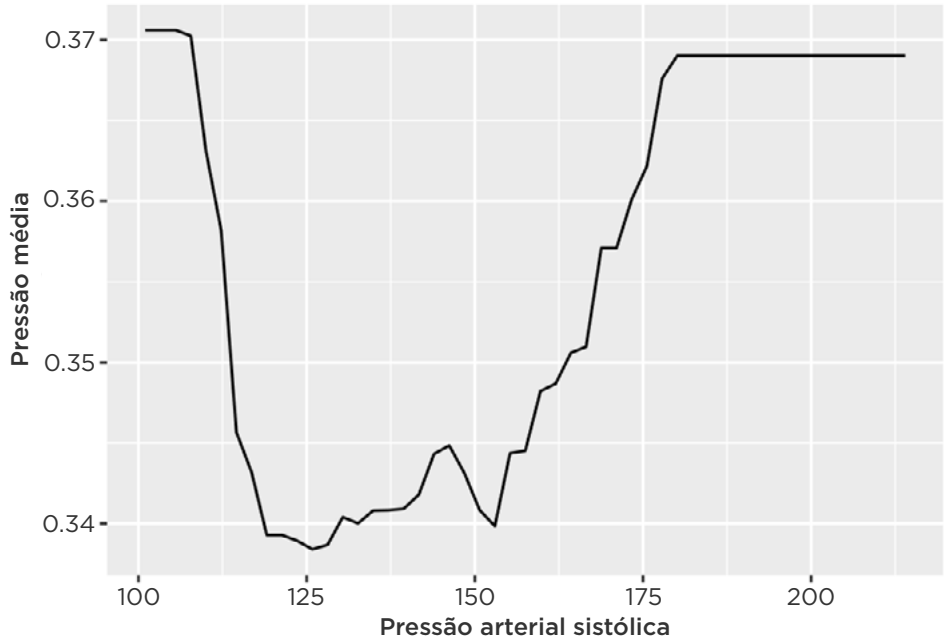
```
sa_heart <- read_csv("datos/sa-heart.csv")

sa_heart <- sa_heart %>%

  rename(presion_arterial = sbp, tabaco = tobacco, colesterol_ldl = ldl,

         adiposidad = adiposity, historia_fam = famhist, tipo_a = typea, obesi-
dad = obesity,

         edad = age, enf_coronaria = chd)

sa_heart

## # A tibble: 462 x 10

##    presion_arterial tabaco colesterol_ldl adiposidad historia_fam tipo_a

##               <dbl>  <dbl>          <dbl>      <dbl> <chr>         <dbl>

##  1              160  12              5.73       23.1 Present          49

##  2              144   0.01           4.41       28.6 Absent           55

##  3              118   0.08           3.48       32.3 Present          52

##  4              170   7.5            6.41       38.0 Present          51

##  5              134  13.6            3.5        27.8 Present          60

##  6              132   6.2            6.47       36.2 Present          62

##  7              142   4.05           3.38       16.2 Absent           59

##  8              114   4.08           4.59       14.6 Present          62

##  9              114   0              3.83       19.4 Present          49

## 10              132   0              5.8        31.0 Present          69

## # … with 452 more rows, and 4 more variables: obesidad <dbl>, alcohol <dbl>,

## #   edad <dbl>, enf_coronaria <dbl>

library(recipes)

set.seed(125)

sa_split <- rsample::initial_split(sa_heart, prop = 0.75)

sa_split

## <Training/Validation/Total>
```

12 Dados acessíveis em http://archive.ics.uci.edu/ml/datasets/heart+Disease

```
## <347/115/462>

receta_sa <- training(sa_split) %>%

  recipe(enf_coronaria ~ .) %>%

  step_dummy(historia_fam) %>%

  step_mutate(enf_coronaria = factor(enf_coronaria)) %>%

  prep()

sa_entrena <- receta_sa %>% juice

sa_boosted <- boost_tree(trees = 3000, mode = "classification",

                         learn_rate = 0.001, tree_depth = 2,

                         sample_size = 0.5) %>%

  set_engine("xgboost") %>%

  fit(enf_coronaria ~ ., data = sa_entrena)
```

É possível avaliar o modelo e refinar os parâmetros. Aqui, é interessante interpretar o efeito das variáveis nesse modelo. Para isso, considera-se o gráfico de dependência parcial da prevalência de doença cardíaca e da variável "obesidade":

```
library(pdp)

pdp_ob <- pdp::partial(sa_boosted$fit,  pred.var = "presion_arterial",

  plot = TRUE, plot.engine = "ggplot2", prob = TRUE,

  train = sa_entrena %>% dplyr::select(-enf_coronaria))

pdp_ob + xlab("Presión arterial sistólica") + ylab("Predicción promedio")
```



A interpretação correta desse gráfico de **dependência parcial** (Hastie, Tibshirani, & Friedman, 2017) depende do fato de que se trata de um estudo retrospectivo, no **qual alguns pacientes com risco de doença cardíaca tiveram intervenções para reduzir seu risco**, incluindo medicamentos para reduzir a pressão. Uma interpretação causal da redução da pressão arterial como promotora de doenças cardíacas é incorreta e potencialmente perigosa.

Desenvolvimento do modelo e validação

Contaminação treinamento-validação

A seguir, serão apresentados vários exemplos de como os vazamentos do treinamento para a validação geram estimativas enviesadas do desempenho dos preditores.

Seleção de variáveis antes de dividir os dados

Todas as etapas de pré-processamento devem ser realizadas sem o uso de dados de validação. Isso inclui os casos nos quais métodos como validação cruzada são usados.

Este exemplo é oriundo de (Hastie, Tibshirani, & Friedman, 2017). Serão utilizados dados sintéticos, gerados com o seguinte processo:

1. Simulando-se as variáveis de resposta y com distribuição binomial.

1. Simulando-se 1.000 covariáveis independentes, cada uma com uma distribuição normal padrão.

```
simular <- function(n = 100, p = 500, prob = 0.5){

  datos <- map(1:p, ~ rnorm(n)) %>%

    bind_cols()

  datos$y <- rbinom(n, 1, prob)

  datos

}

set.seed(8234)

datos_entrena <- simular(n = 200, p = 1000)

datos_prueba <- simular(n = 2000, p = 1000)

dim(datos_entrena)
```

```
## [1]  200 1001
```

```
datos_entrena %>% group_by(y) %>% tally() %>% kable()
```

y	n
0	113
1	87

A seleção das variáveis é dada pela função abaixo. Esta função seleciona as variáveis mais correlacionadas com a variável-alvo:

```
seleccionar <- function(datos, num_var = 10){

  correlaciones <- datos %>%

   pivot_longer(cols = matches("V"), names_to = "variable", values_to = "x") %>%
```

```
    group_by(variable) %>%
    summarise(corr = abs(cor(y, x))) %>%
    arrange(desc(corr))
  # seleccionar
  seleccionadas <- correlaciones %>%
    top_n(num_var, wt = corr) %>%
    pull(variable) datos %>% select(one_of(c("y", seleccionadas)))
}
```

Método incorreto

Aqui serão apresentadas as 10 variáveis selecionadas. Por si só, este método não está incorreto, mas, quando executado com base nos dados a serem usados na validação (validação cruzada), a estimativa de desempenho é otimista:

```
datos_filtrados <- seleccionar(datos_entrena)

datos_filtrados %>% head %>%

  mutate_if(is.numeric, round, 3) %>% kable()
```

y	V337	V464	V984	V461	V525	V732	V39	V774	V491	V682
0	1.592	-0.587	1.763	-0.847	0.452	-0.604	-0.400	-1.146	-0.938	0.136
0	1.782	0.604	0.739	-0.533	1.752	0.945	1.142	-0.638	-0.342	-1.308
0	1.528	0.635	-0.326	0.734	-0.207	-0.974	1.574	2.401	0.428	0.176
0	0.799	-1.436	0.724	0.366	1.680	0.476	0.376	-1.673	-0.683	0.161
0	0.759	-0.208	-0.373	0.208	-1.009	-0.028	-1.209	0.759	2.038	1.402
1	-0.377	-1.044	1.358	-0.223	0.469	1.221	0.582	0.378	-0.116	0.173

Para qualquer corte de validação feito (seja dividindo-se um conjunto de dados ou realizando-se uma validação cruzada), a porcentagem de acertos parece ser maior que 0,5:

```
corte_validacion <- datos_filtrados %>% sample_frac(0.7)

valida <- anti_join(datos_filtrados, corte_validacion)
```

```
modelo_1 <- glm(y ~ ., corte_validacion, family = "binomial")

mean(as.numeric(predict(modelo_1, valida) > 0) == valida$y) %>% round(2)
```

```
## [1] 0.73
```

No entanto, o desempenho real do modelo será:

```
mean(as.numeric(predict(modelo_1, datos_prueba) > 0) == datos_prueba$y) %>%
round(2)
```

```
## [1] 0.49
```

Método correto

A seleção de variáveis deve ser feita em cada rodada de validação cruzada:

```
corte_validacion <- datos_entrena %>% sample_frac(0.7)

datos_filtrados_corte <- seleccionar(corte_validacion)

valida <- anti_join(datos_entrena, corte_validacion)

modelo_1 <- glm(y ~ ., datos_filtrados_corte, family = "binomial")

mean(as.numeric(predict(modelo_1, valida) > 0) == valida$y) %>% round(2)
```

```
## [1] 0.52
```

Sobreamostragem antes de particionar

Uma das formas de resolver problemas de desbalanceamento de classes é a técnica de sobreamostragem. No entanto, deve-se ter muito cuidado para evitar erros de vazamento de informações ao aplicar estas técnicas.

Neste exemplo, veremos que a sobreamostragem de uma classe reduzida antes de separar os dados de validação ou fazer a validação cruzada pode produzir estimativas excessivamente otimistas do erro de previsão.

Suponhamos que temos um desbalanceamento grave entre as nossas duas classes:

```
set.seed(99134)

datos_desbalance <- simular(n = 500, p = 20, prob = 0.1) %>%

  mutate(y = factor(y, levels = c(1, 0)))

datos_desbalance %>% group_by(y) %>% tally() %>% kable()
```

y	n
1	41
0	459

Modo incorreto

Suponhamos que seja aplicado primeiro (SMOTE) (Chawla, 2002) para tentar balancear os dados:

```
receta_balance <- recipe(y ~ ., datos_desbalance) %>%

  step_smote(y) %>%

  prep()

datos_smote <- juice(receta_balance)
```

Obtendo-se, assim:

```
datos_smote %>% group_by(y) %>% tally() %>% kable()
```

y	n
1	459
0	459

Agora, "treinamento" e "validação" estão separados:

```
sep_datos_smote <- initial_split(datos_smote)

entrena_smote <- training(sep_datos_smote)

prueba_smote <- testing(sep_datos_smote)
```

E gera-se um método de classificação baseado em uma floresta aleatória de árvores de decisão:

```
metricas <- metric_set(accuracy, recall, precision)
bosque <- rand_forest(trees = 500, mtry = 20, mode = "classification") %>%
  set_engine("ranger") %>%
  fit(y ~ ., data = entrena_smote)
bosque %>%
  predict(prueba_smote) %>%
  bind_cols(prueba_smote) %>%
  metricas(truth = y, estimate = .pred_class) %>%
  mutate_if(is.numeric, round, 3) %>% kable
```

.metric	.estimator	.estimate
accuracy	binary	0.926
recall	binary	0.973
precision	binary	0.887

A princípio, parece que o desempenho é muito bom. Sabe-se que isso é fictício, pois não há nenhuma relação entre y e as outras covariáveis.

Modo correto

Antes de fazer o rebalanceamento de classes, "treinamento" e "validação" são separados. Se desejado, esta parte pode ser feita com base em uma amostragem estratificada, por exemplo, mas aqui ela é construída com uma amostragem aleatória simples:

```
sep_datos <- initial_split(datos_desbalance, prop = 0.5)

entrena <- training(sep_datos)

prueba <- testing(sep_datos)

receta_balance <- recipe(y ~ ., data = entrena) %>%

  step_smote(y) %>%

  prep()

entrena_balanceado <- juice(receta_balance)

bosque_1 <- rand_forest(trees = 500, mtry = 20, mode = "classification") %>%

  set_engine("ranger") %>%

  fit(y ~ ., data = entrena_balanceado)

bosque_1 %>%

  predict(prueba) %>%

  bind_cols(prueba) %>%

  metricas(truth = y, estimate=.pred_class) %>%

  mutate_if(is.numeric, round, 3) %>%

  kable()
```

.metric	.estimator	.estimate
accuracy	binary	0.828
recall	binary	0.000
precision	binary	0.000

Embora a *accuracy* pareça alta, a precisão e a sensibilidade são zero. Um classificador trivial que sempre prevê a classe dominante pode ter uma precisão melhor que o que construímos.

Variáveis indisponíveis no momento da previsão

Neste caso, mostramos um exemplo em que uma variável que estará indisponível no momento de fazer as previsões é utilizada incorretamente (dados de (Greene, 2003)).

```
credito <- read_csv("datos/AER_credit_card_data.csv") %>%
  rename(gasto = expenditure, dependientes = dependents, ingreso = income,
         edad = age, propietario = owner) %>%
  mutate(propietario = fct_recode(propietario, c(si = "yes")))
credito %>% head %>%
  mutate_if(is.numeric, round, 1) %>% kable()
```

card	reports	idade	renda	share	gasto	propietário	selfemp	dependentes	months	majorcards	active
yes	0	37.7	4.5	0.0	125.0	si	no	3	54	1	12
yes	0	33.2	2.4	0.0	9.9	no	no	3	34	1	13
yes	0	33.7	4.5	0.0	15.0	si	no	4	58	1	5
yes	0	30.5	2.5	0.1	137.9	no	no	0	25	1	7
yes	0	32.2	9.8	0.1	546.5	si	no	2	64	1	5
yes	0	23.2	2.5	0.0	92.0	no	no	0	54	1	1

Pretende-se construir um modelo para prever quais solicitações foram aceitas e automatizar o processo de seleção. Utiliza-se uma regressão logística com Keras e penalidade L2:

```
set.seed(823)
credito_split <- initial_split(credito)
entrena <- training(credito_split)
prueba <- testing(credito_split)
# preparacion de datos
credito_receta <- recipe(card ~ ., credito) %>%
  step_normalize(all_numeric()) %>%
```

```
  step_dummy(all_nominal(), -card)

# modelo

modelo_regularizado <-

  logistic_reg(penalty = 1) %>%

  set_engine("keras", epochs = 500, verbose = FALSE)  %>%

  set_mode("classification")

# ajustar parametros de preprocesamiento

receta_prep <- credito_receta %>% prep(entrena)

# preprocesar datos

entrena_prep <- bake(receta_prep, entrena)

prueba_prep <- bake(receta_prep, prueba)

# ajustar modelo

ajuste <- modelo_regularizado

  fit(card~ gasto + dependientes + ingreso + edad + propietario_si, data = en-
trena_prep)

# evaluar

metricas <- metric_set(accuracy, recall, precision)

ajuste %>% predict(prueba_prep) %>%

  bind_cols(prueba) %>%

  metricas(truth = factor(card), estimate = .pred_class) %>%

  mutate_if(is.numeric, round, 3) %>%

  kable()
```

.metric	.estimator	.estimate
accuracy	binary	0.833
recall	binary	0.393
precision	binary	0.892

E parece ter um desempenho razoável. Se retirarmos a variável "despesas" (expenditure), o desempenho do modelo fica totalmente degradado:

```
 ajuste_2 <- modelo_regularizado %>%

  fit(card~ dependientes + ingreso + edad + propietario_si, data = entrena_prep)

ajuste_2 %>% predict(prueba_prep) %>%

  bind_cols(prueba) %>%

  metricas(truth = factor(card), estimate = .pred_class) %>%

  mutate_if(is.numeric, round, 3) %>%

  kable()
```

.metric	.estimator	.estimate
accuracy	binary	0.745
recall	binary	0.000
precision	binary	NA

A sensibilidade é muito baixa, e a precisão não pode ser calculada, pois o modelo não faz previsões positivas para o conjunto de teste.

O motivo dessa degradação do desempenho é que a variável *despesas* refere-se ao uso de cartões de crédito. Isso inclui o cartão para o qual se deseja fazer uma previsão de aceitação:

```
entrena %>%

  mutate(algun_gasto = gasto > 0) %>%

  group_by(algun_gasto, card) %>%

  tally() %>%

  kable()
```

algun_gasto	card	n
FALSE	no	212
FALSE	yes	19
TRUE	yes	759

O que indica que alguma despesa provavelmente inclui a despesa no cartão atual. A variável despesas é mensurada após a entrega do cartão:

- O desempenho desse modelo para novas aplicações será muito baixo, pois a variável *despesas*, no momento da aplicação, evidentemente não contabiliza quanto cada cliente gastará no futuro.

Pontos de corte arbitrários

As melhores decisões de ponto de corte podem ser tomadas com base em uma análise de custo-benefício e com curvas de elevação (*lift*), como as do exemplo acima, baseadas nos ganhos e perdas de cada decisão. Embora essas informações muitas vezes não estejam disponíveis, é a situação ideal para avaliar como o modelo ajuda e quanto valem as ações que pretendemos realizar. É possível fazer essa análise com valores incertos de custo-benefício.

Suponhamos que estamos concebendo uma intervenção para reter estudantes em um programa de treinamento ou aperfeiçoamento.

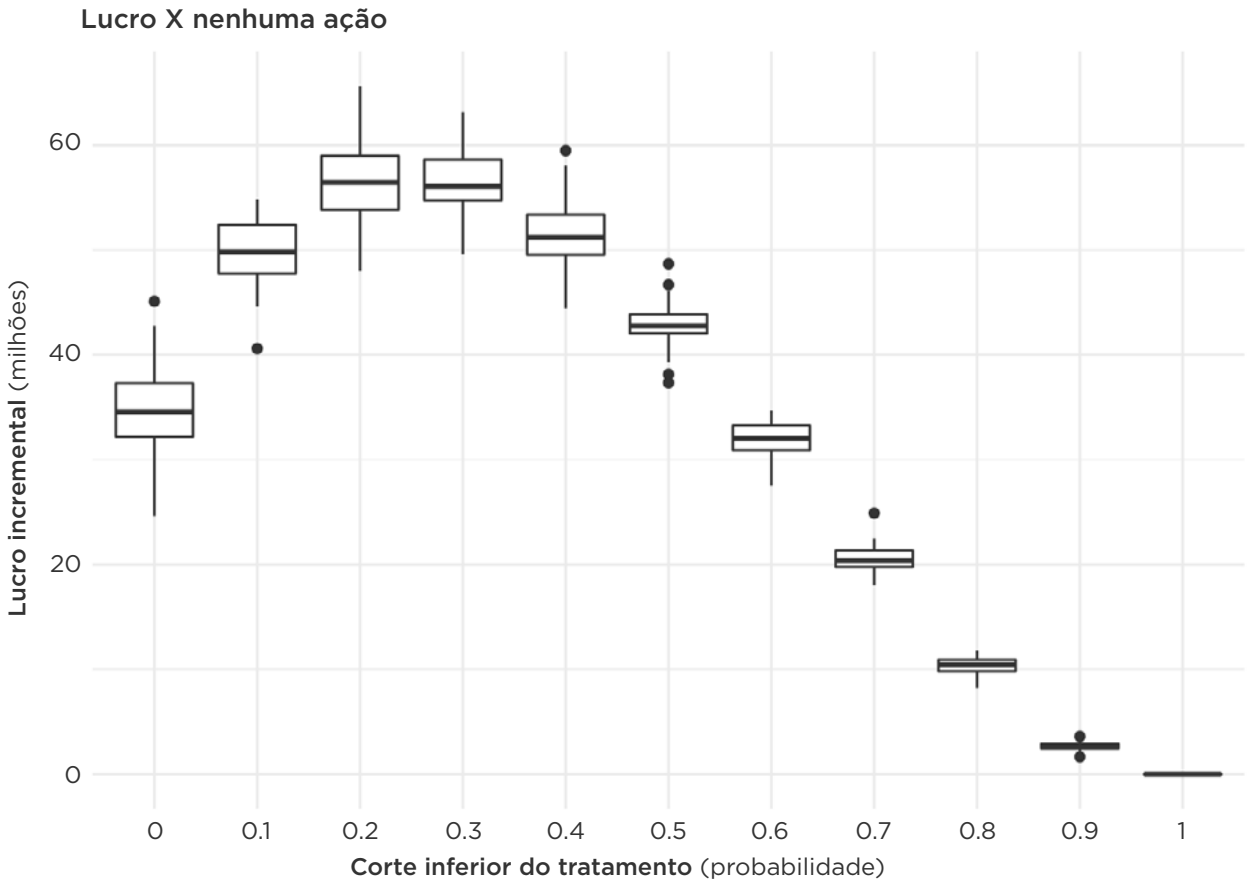
- A intervenção de retenção custa 5 mil pesos por estudante.
- Com experimentos ou alguma análise externa, estima-se que a intervenção reduza a probabilidade de abandono em 60%.
- Existe algum tipo de estimação do valor social associado à permanência de um estudante no programa.

O modelo pode ser avaliado no contexto do problema da seguinte maneira:

- Supondo-se que a intervenção afetará uma porcentagem dos estudantes com maior probabilidade de rotatividade.
- Calculando-se o custo esperado, se a intervenção afetar uma porcentagem dos estudantes: ela é simulada, reduzindo-se a probabilidade de abandono graças à intervenção e adicionando-se os custos associados a ela.
- Comparando-se o modelo com o que ocorreria se nenhuma intervenção fosse realizada.

Não é necessário usar medidas muito técnicas para resumir como a intervenção e o modelo podem ajudar a manter o valor do portfólio:

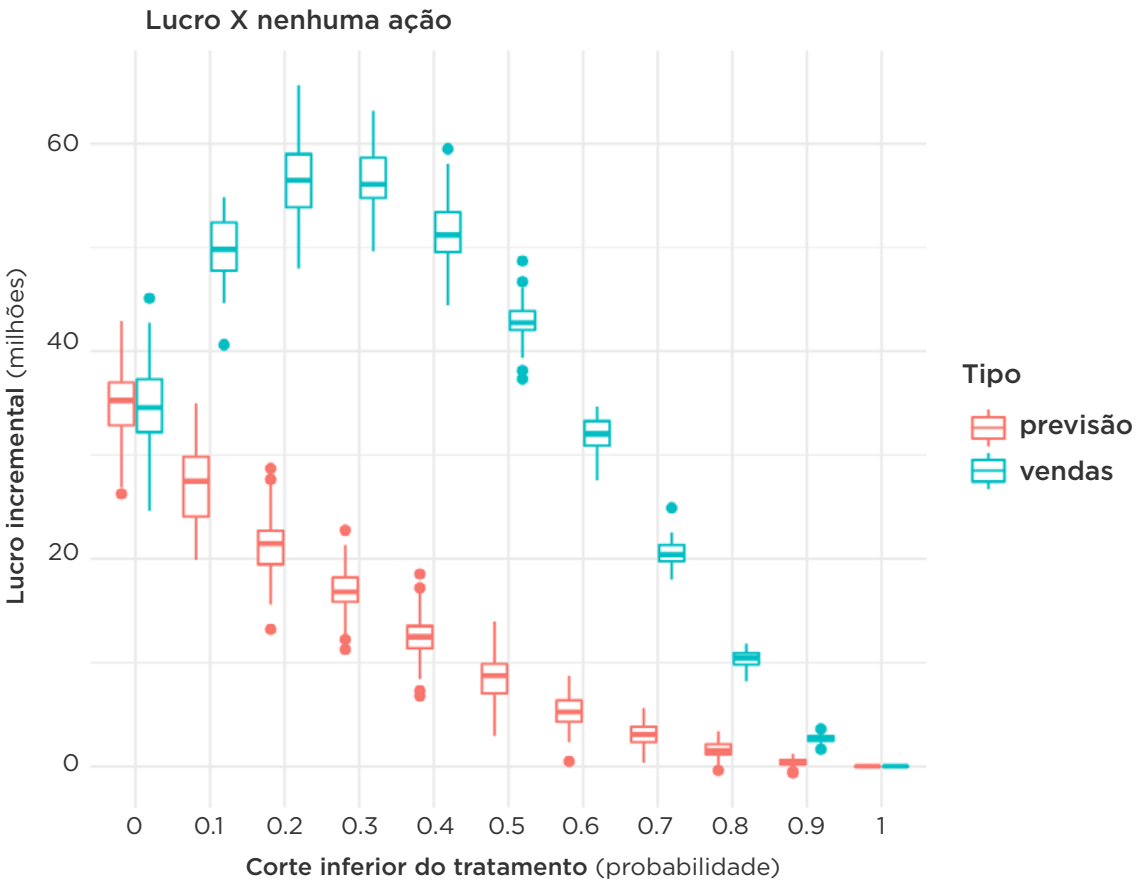
```
ggplot(filter(perdidas_sim, tipo==”Tratamiento modelo”),
       aes(x = factor(corte), y = - perdida / 1e6)) +
  geom_boxplot() + ylab(“Ganancia incremental (millones)”) +
  xlab(“Corte inferior de tratamiento (probabilidad)”) +
  labs(subtitle = “Ganancia vs ninguna acción”) + theme_minimal()
```



É possível escolher um ponto de corte entre 0,2 e 0,3, por exemplo, ou realizar mais simulações para refinar a escolha.

A fim de separar o efeito da intervenção do efeito da intervenção aplicada de acordo com o modelo, pode-se traçar uma comparação com a ação que consiste em realizar uma intervenção que afete os estudantes aleatoriamente:

```
ggplot(perdidas_sim, aes(x = factor(corte), y = - perdida / 1e6,

                         group = interaction(tipo, corte), colour = tipo)) +

  geom_boxplot() + ylab("Ganancia incremental (millones)") +

  xlab("Corte inferior de tratamiento (probabilidad)") +

  labs(subtitle = "Ganancia vs ninguna acción") + theme_minimal()
```



- A conclusão é que o modelo **auxilia consideravelmente no direcionamento do programa** (área entre as duas curvas mostradas acima).

Desbalanceamento de classes

Quando existe um grave desbalanceamento de classes, dois problemas podem surgir: 1) em termos absolutos, há poucos elementos de uma classe para uma discriminação efetiva (mesmo se os atributos ou *features* corretos estiverem presentes); e 2) os métodos usuais de avaliação de previsões são deficientes para avaliar o desempenho das previsões.

Os seguintes dados devem ser considerados:

> "Os dados contêm 5.822 registros de clientes reais. Cada registro é composto por 86 variáveis, que contêm dados sociodemográficos (variáveis 1-43) e de propriedade do produto (variáveis 44-86). Os dados sociodemográficos são derivados dos códigos postais. Todos os clientes que residem em áreas com o mesmo CEP possuem os mesmos atributos sociodemográficos. A variável 86 (compra) indica se o cliente adquiriu uma apólice de seguro de trailer" (***James***, 2017)13.

13 Dados e mais informações em http://www.liacs.nl/~putten/library/cc2000/data.html

Pretende-se prever a variável *purchase*:

```
caravan <- read_csv("datos/caravan.csv") %>%
  mutate(MOSTYPE = factor(MOSTYPE),
         MOSHOOFD = factor(MOSHOOFD)) %>%
  mutate(Compra = fct_recode(Purchase, si = "Yes", no = "No")) %>%
  mutate(Compra = fct_rev(Compra)) %>%
  select(-Purchase)
nrow(caravan)
```

```
## [1] 5822
```

```
caravan %>% count(Compra) %>%
  mutate(pct = 100 * n / sum(n)) %>%
  mutate(pct = round(pct, 2))
```

```
## # A tibble: 2 x 3
##   Compra     n   pct
##   <fct>  <int> <dbl>
## 1 si       348  5.98
## 2 no      5474 94.0
```

Essa é a distribuição natural de resposta observada nos dados, havendo relativamente poucos dados na categoria "Sim".

Utilizar-se-á uma amostragem estratificada para obter proporções semelhantes nos conjuntos de treinamento e teste:

```
set.seed(823)
caravan_split = initial_split(caravan, strata = Compra, prop = 0.9)
caravan_split
```

```
## <Training/Validation/Total>
## <5240/582/5822>
```

```
entrena <- training(caravan_split)
prueba <- testing(caravan_split)
```

E utilizar-se-á a regressão logística (o mesmo vale para outros métodos que produzem probabilidades de classe, como *boosting*, árvores aleatórias ou redes neurais):

```
library(tune)

# preparacion de datos

caravan_receta <- recipe(Compra ~ ., entrena) %>%

  step_dummy(all_nominal(), -Compra)

caravan_receta_prep <- caravan_receta %>% prep

# modelo

modelo_log <-

  logistic_reg() %>%

  set_engine("glm")  %>%

  set_mode("classification") %>%

  fit(Compra ~ ., data = caravan_receta_prep %>% juice)

set.seed(823)
```

Análise incorreta

A matriz de confusão dos dados de treinamento é:

```
predictions_ent_glm <- modelo_log %>%

  predict(new_data = juice(caravan_receta_prep)) %>%

  bind_cols(juice(caravan_receta_prep) %>% select(Compra))

predictions_ent_glm %>%

  conf_mat(Compra, .pred_class)
```

```
##            Truth

## Prediction   si   no

##         si    6    9

##         no  299 4926
```

E os de teste:

```
prueba_procesado <- bake(caravan_receta_prep, prueba)

predictions_glm <- modelo_log %>%

  predict(new_data = prueba_procesado) %>%

  bind_cols(prueba_procesado %>% select(Compra))

predictions_glm %>%
```

```
  conf_mat(Compra, .pred_class)
```

```
##           Truth
## Prediction  si  no
##         si   0   4
##         no  43 535
```

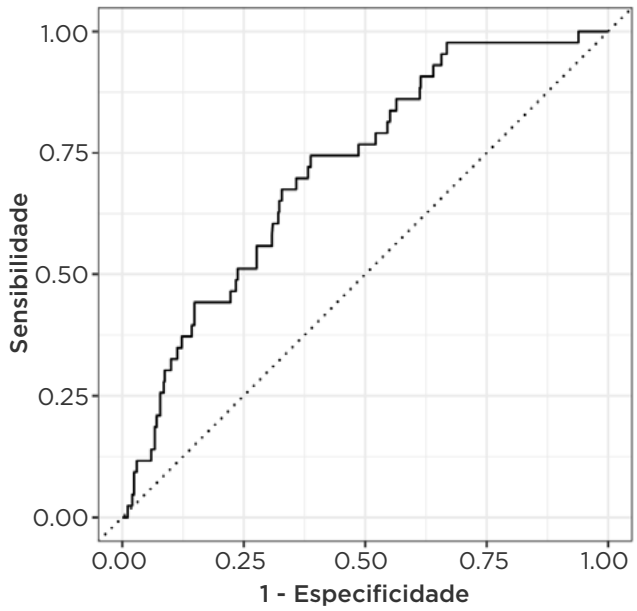
De acordo com essa matriz de confusão (teste e treinamento), o desempenho obtido é baixo. A sensibilidade é muito baixa, embora a especificidade (taxa de negativos corretos) seja alta. Uma conclusão típica é que o *modelo não tem valor preditivo ou que é necessário sobreamostrar a classe de ocorrência baixa*.

Análise correta

Em vez de começar com a sub/sobreamostragem, que altera as proporções naturais das categorias nos dados, é possível trabalhar com probabilidades em vez de previsões de classe com ponto de corte de 0,5.

Por exemplo, isso pode ser visualizado com uma curva ROC (ou curva de elevação *[lift]*, *precision-recall* ou qualquer outra curva semelhante que leve em consideração as probabilidades):

```
predictions_prob <- modelo_log %>%
  predict(new_data = prueba_procesado, type = "prob") %>%
  bind_cols(prueba_procesado %>% select(Compra)) %>%
  select(.pred_si, Compra)
datos_roc <- roc_curve(predictions_prob, Compra, .pred_si)
autoplot(datos_roc) +
  xlab("1 - especificidad") + ylab("sensibilidad")
```



Onde se vê que é possível atingir bons níveis de sensibilidade, se for aceita alguma degradação da especificidade, que originalmente é muito alta. Por exemplo, cortando-se em 0,05, é possível obter uma especificidade e uma sensibilidade que podem ser adequadas para o problema:

```
datos_roc %>% filter(abs(.threshold - 0.04) < 1e-4) %>% round(4)
```

```
## # A tibble: 2 x 3
##   .threshold specificity sensitivity
##        <dbl>       <dbl>       <dbl>
## 1     0.0399       0.553       0.744
## 2     0.0399       0.555       0.744
```

O que acontece se for feita uma sub/sobreamostragem?

Com a sobreamostragem:

```
caravan_receta_smote <- recipe(Compra ~ ., entrena) %>%
  step_dummy(MOSTYPE, MOSHOOFD) %>%
  step_smote(Compra)
smote_prep <- prep(caravan_receta_smote)
# modelo
entrena_1 <- juice(smote_prep)
entrena_1 %>% count(Compra)
```

```
## # A tibble: 2 x 2
##   Compra     n
##   <fct>  <int>
## 1 si      4935
## 2 no      4935
```

```
modelo_log_smote <-
  logistic_reg() %>%
  set_engine("glm")  %>%
  set_mode("classification") %>%
  fit(Compra ~ ., data = entrena_1)
```

No treinamento, a matriz de confusão é *aparentemente* melhor:

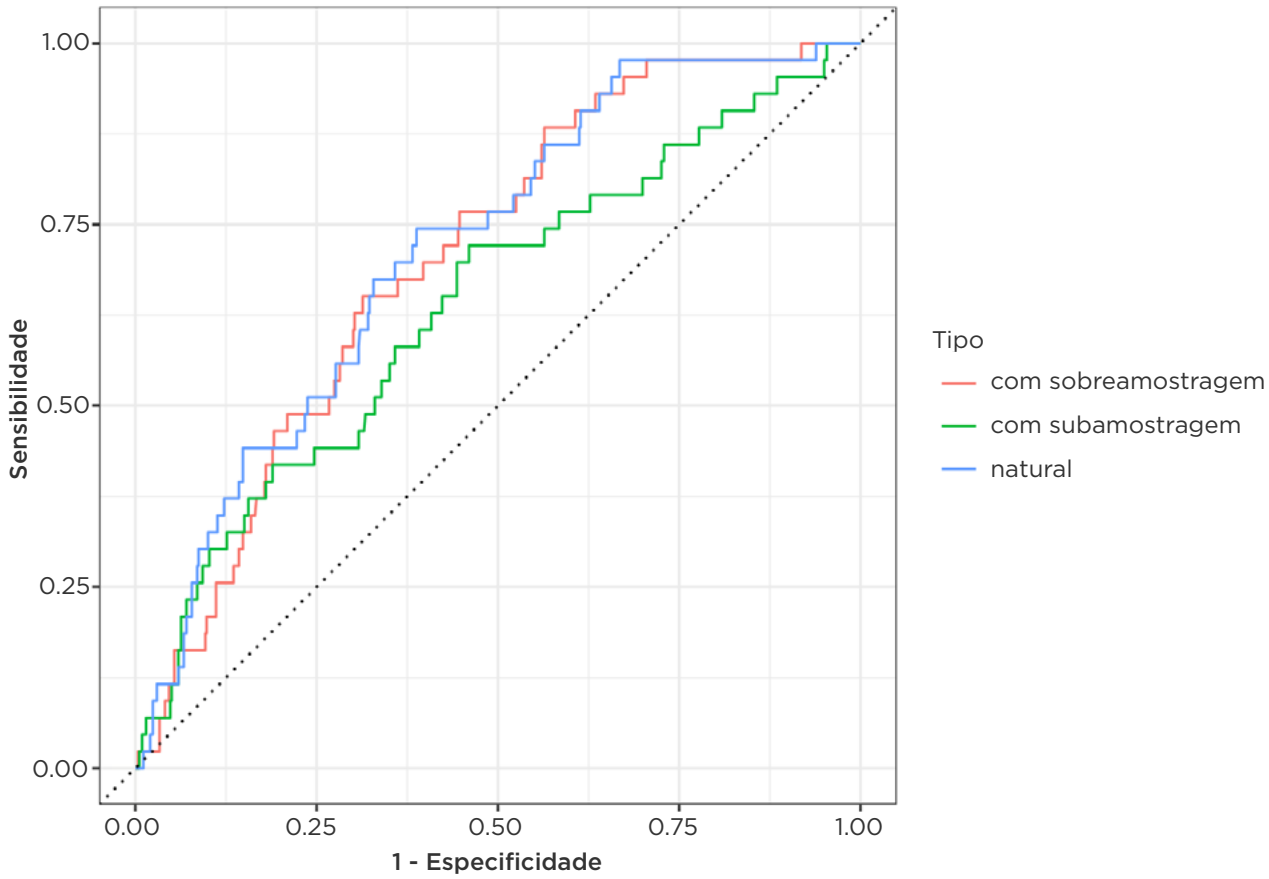
```
predictions_ent_glm <- modelo_log_smote %>%
  predict(new_data = entrena_1) %>%
  bind_cols(entrena_1 %>% select(Compra))
predictions_ent_glm %>%
  conf_mat(Compra, .pred_class)
```

```
##            Truth
## Prediction   si   no
##         si 3854 1271
##         no 1081 3664
```

Porém, no teste, os resultados são muito semelhantes. O modelo construído com a subamostragem da classe dominante também é adicionado:

```
entrena_sub <- caravan_receta %>% step_downsample(Compra) %>% prep() %>% juice
modelo_log_sub <-
 logistic_reg() %>%
  set_engine("glm")  %>%
  set_mode("classification") %>%
  fit(Compra ~ ., data = entrena_sub)
predictions_prob <- modelo_log_smote %>%
  predict(new_data = prueba_procesado, type = "prob") %>%
  bind_cols(prueba_procesado %>% select(Compra)) %>%
  select(.pred_si, Compra)
predictions_prob_sub <- modelo_log_sub %>%
  predict(new_data = prueba_procesado, type = "prob") %>%
  bind_cols(prueba_procesado %>% select(Compra)) %>%
  select(.pred_si, Compra)
datos_roc_smote <- roc_curve(predictions_prob, Compra, .pred_si)
datos_roc_sub <- roc_curve(predictions_prob_sub, Compra, .pred_si)
datos_roc_comp <- bind_rows(datos_roc %>% mutate(tipo = "natural"),
                        datos_roc_smote %>% mutate(tipo = "con sobre muestreo"),
                        datos_roc_sub %>% mutate(tipo = "con sub muestreo")
)
```

```
ggplot(datos_roc_comp,
       aes(x = 1 - specificity, y = sensitivity, colour = tipo)) +
  geom_path() +
  geom_abline(lty = 3) +
  coord_equal() +
  theme_bw()
```



- **O problema original não era o não funcionamento do ajuste, mas a avaliação de um ponto de corte errado.** Um ponto de corte de 0,5 com SMOTE é equivalente a um ponto de corte muito mais reduzido sem SMOTE.

- **Pior ainda, as probabilidades do modelo construído com sobreamostragem não refletem as taxas de ocorrência da resposta que interessa,** o que pode produzir resumos enganosos das taxas de resposta que se espera observar na produção.

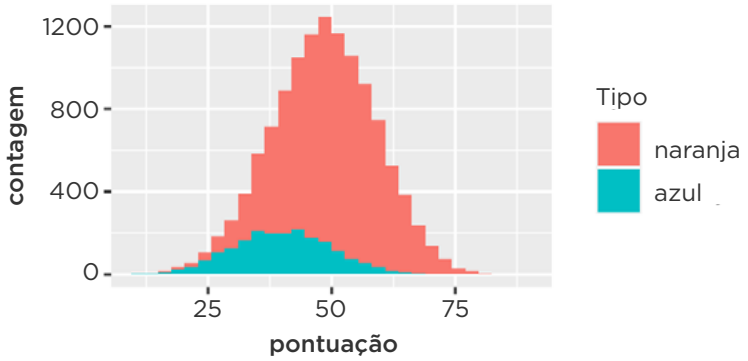
Equidade com atributos protegidos

O exemplo a seguir é derivado de (Hardt, 2016). Suponhamos que existe um atributo protegido A com dois valores: azul e laranja. Laranja é o grupo minoritário desfavorecido. Utilizamos dados simulados da seguinte forma: o atributo *score* é associado ao atributo protegido.

```
inv_logit <- function(x){
  1 / (1 + exp(-x))
}
simular_datos <- function(n = c(10000, 2000)){
  score_azul <- pmax(rnorm(n[1], 50, 10), 0)
  score_naranja <- pmax(rnorm(n[2], 40, 10), 0)
  azul <- tibble(tipo = "azul", score = score_azul)
  naranja <- tibble(tipo = "naranja", score = score_naranja)
  datos <- bind_rows(azul, naranja) %>%
    mutate(coef_0 = ifelse(tipo == "azul", 0.0, 0),
           prob_real_pos = inv_logit(-1 + coef_0 + 0.1 * (score-40))) %>%
 mutate(atr_1 = rpois(nrow(.), 3))
  datos %>% select(-coef_0) %>%
    mutate(paga = map_dbl(prob_real_pos, ~ rbinom(1, 1, .x))) %>%
    select(-prob_real_pos)
}
set.seed(1221)
tbl_datos <- simular_datos()
```

Utilizando-se um histograma referente ao *score*, obtém-se um grupo minoritário com valores da variável *score* mais baixa:

```
ggplot(tbl_datos, aes(x = score, fill = tipo)) + geom_histogram()
```



Um modelo de regressão logística simples é ajustado:

```
reg_log <- glm(paga ~  score + atr_1 + tipo, tbl_datos, family = "binomial")

tbl_datos <- tbl_datos %>% mutate(prob_pos = predict(reg_log, type = "response"))
```

As taxas reais de conformidade são as mesmas para os dois grupos. Primeiramente, considera-se uma estratégia segundo a qual o mesmo ponto de corte é aplicado a todos os grupos:

```
resultado_cortes <- function(tbl_datos, cortes){

  resultado <- tbl_datos %>%

     mutate(recibe = ifelse(tipo == "azul", prob_pos > cortes[1], prob_pos >
cortes[2]),

            decision = ifelse(recibe, "Aceptado", "Rechazado"))

  resultado %>% group_by(tipo, decision, paga) %>% count() %>%

    ungroup()

}

resultados_conteo <- resultado_cortes(tbl_datos, c(0.6, 0.6))

resultados_conteo
```

```
## # A tibble: 8 x 4

##   tipo    decision   paga     n

##   <chr>   <chr>     <dbl> <int>

## 1 azul    Aceptado      0   905

## 2 azul    Aceptado      1  2400

## 3 azul    Rechazado     0  4149

## 4 azul    Rechazado     1  2546
```

```
## 5 naranja Aceptado      0    47
## 6 naranja Aceptado      1   101
## 7 naranja Rechazado     0  1353
## 8 naranja Rechazado     1   499
```

```
resultados_conteo %>%
  group_by(tipo, decision) %>%
  summarise(n = sum(n)) %>%
  mutate(total = sum(n)) %>%
  mutate(prop = n / total) %>%
  filter(decision == "Aceptado")
```

```
## # A tibble: 2 x 5
## # Groups:   tipo [2]
##   tipo    decision     n total  prop
##   <chr>   <chr>    <int> <int> <dbl>
## 1 azul    Aceptado  3305 10000 0.330
## 2 naranja Aceptado   148  2000 0.074
```

Observe que o grupo laranja recebeu consideravelmente menos aceitações que o grupo azul, tanto no total quanto proporcionalmente. Além disso, com a precisão ou taxa de verdadeiros positivos, é possível avaliar a proporção dos que foram aceitos entre aqueles que cumpriri-am se fossem aceitos, de acordo com o ponto de corte:

```
resultados_conteo %>%
  filter(paga == 1) %>%
  group_by(tipo) %>%
mutate(tvp = n / sum(n)) %>%
  filter(decision == "Aceptado")
```

```
## # A tibble: 2 x 5
## # Groups:   tipo [2]
##   tipo    decision  paga     n   tvp
##   <chr>   <chr>    <dbl> <int> <dbl>
## 1 azul    Aceptado     1  2400 0.485
## 2 naranja Aceptado     1   101 0.168
```

E percebe-se que o grupo laranja também está em desvantagem, pois há menos decisões de aceitação entre os que cumprem.

O próximo passo é considerar a **paridade demográfica**. Neste caso, decide-se conceder o mesmo número de empréstimos a cada grupo, dependendo do tamanho:

```
calcular_puntos_paridad <- function(tbl_datos, prop){

  tbl_datos %>% group_by(tipo) %>%

    summarise(corte = quantile(prob_pos, 1 - prop))

}

cortes_paridad_tbl <- calcular_puntos_paridad(tbl_datos, 0.45)

cortes_paridad_tbl
```

```
## # A tibble: 2 x 2

##   tipo    corte

##   <chr>   <dbl>

## 1 azul    0.521

## 2 naranja 0.297
```

O corte para o azul é mais exigente que para o laranja. Isso por si só não é um problema, mas observa-se:

```
cortes_paridad <- cortes_paridad_tbl %>% pull(corte)

resultados_conteo <- resultado_cortes(tbl_datos, cortes_paridad)

resultados_conteo %>%

  filter(paga == 1) %>%

  group_by(tipo) %>%

  mutate(tvp = n / sum(n)) %>%

  filter(decision == "Aceptado")
```

```
## # A tibble: 2 x 5

## # Groups:   tipo [2]

##   tipo    decision  paga     n   tvp

##   <chr>   <chr>    <dbl> <int> <dbl>

## 1 azul    Aceptado     1  3094 0.626

## 2 naranja Aceptado     1   410 0.683
```

E, assim, além de serem mais exigentes com o grupo azul, aqueles do grupo azul que cumprem os requisitos também recebem menos decisões de aceitação. Além disso, são aceitas consideravelmente menos pessoas da população.

A solução de **igualdade de oportunidades** é cortar de forma que a taxa de aceitação dentro do grupo dos pagantes seja semelhante para ambas as populações, o que ocorre aproximadamente a 0,35:

```
calcular_cortes_oportunidad <- function(tbl_datos, prop){

tbl_datos %>%

    filter(paga==1) %>%

    group_by(tipo) %>%

    mutate(rank_p = rank(prob_pos) / length(prob_pos) ) %>%

    filter(rank_p < prop) %>%

    top_n(1, rank_p) %>%

    select(tipo, corte = prob_pos)

}

cortes_op <- calcular_cortes_oportunidad(tbl_datos, 0.35)

resultados_conteo <- resultado_cortes(tbl_datos, cortes_op %>% pull(corte))

resultados_conteo %>%

  filter(paga == 1) %>%

  group_by(tipo) %>%

  mutate(tvp = n / sum(n)) %>%

  filter(decision == "Aceptado")
```

```
## # A tibble: 2 x 5

## # Groups:   tipo [2]

##   tipo    decision  paga     n   tvp

##   <chr>   <chr>    <dbl> <int> <dbl>

 ## 1 azul    Aceptado     1  3215 0.650

## 2 naranja Aceptado     1   391 0.652
```

Observação: é importante observar que, se a variável de resultado positivo for atribuída injustamente, esse método não resolverá o problema. Nesse caso, é relevante perceber quais são os critérios com os quais se considera um resultado positivo em função do grupo do atributo protegido (por exemplo, se um determinado segmento puder atrasar pagamentos mais que outro, ou se os membros de um grupo forem considerados infratores reincidentes por uma infração muito menor que os de outros).

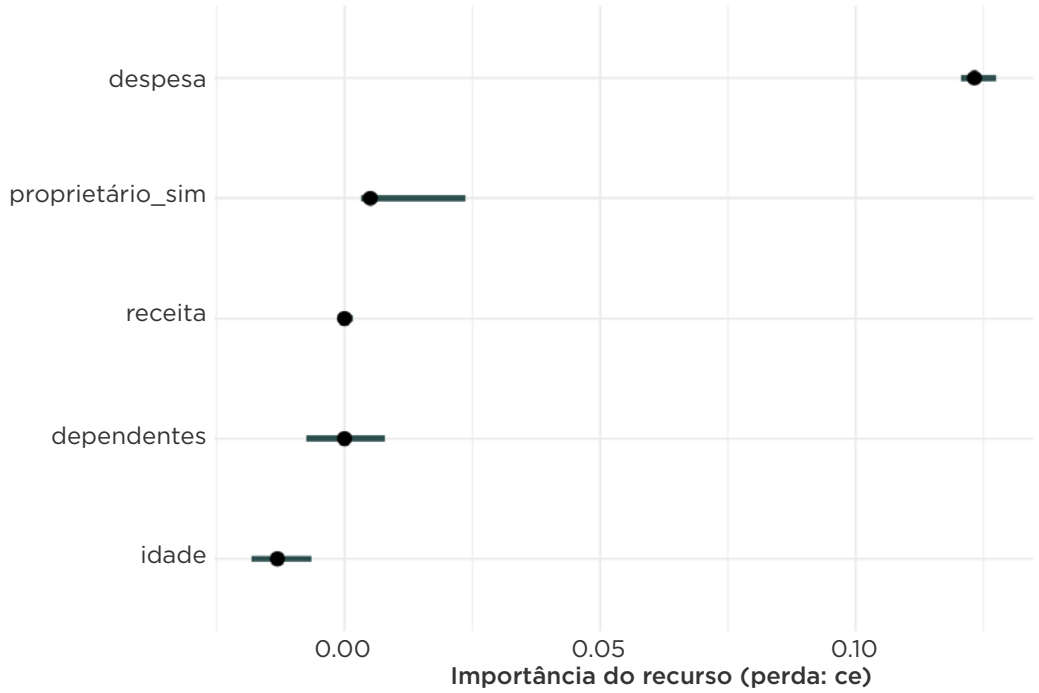
Prestação de contas

Interpretabilidade

Medidas como a importância das permutações podem ser usadas para examinar os modelos. Neste exemplo, voltamos ao exercício de previsão de aceitação do pedido de crédito e consideramos a importância com base em permutações (Molnar, 2019):

```
set.seed(823)

credito_split <- initial_split(credito)

entrena <- training(credito_split)

prueba <- testing(credito_split)

# preparacion de datos

credito_receta <- recipe(card ~ ., credito) %>%

  step_normalize(all_numeric()) %>%

  step_dummy(all_nominal(), -card)

# modelo

modelo_regularizado <-

  logistic_reg(penalty = 1) %>%

  set_engine("keras", epochs = 500, verbose = FALSE)  %>%

  set_mode("classification")

# ajustar parametros de preprocesamiento

receta_prep <- credito_receta %>% prep(entrena)

# preprocesar datos

entrena_prep <- bake(receta_prep, entrena)

prueba_prep <- bake(receta_prep, prueba)

# ajustar modelo

ajuste <- modelo_regularizado %>%

  fit(card~ gasto + dependientes + ingreso + edad + propietario_si, data = en-
trena_prep)

library(iml)

modelo <- ajuste$fit

entrena_x <- entrena_prep %>% dplyr::select(gasto, dependientes, ingreso, edad,
propietario_si)
```

```
predictor <- Predictor$new(modelo, data = entrena_x, y = ifelse(entrena_prep$-
card == "yes",2,1) ,

                           type = "prob")

imp <- FeatureImp$new(predictor, loss = "ce", compare = "difference")

plot(imp) + theme_minimal()
```



- Percebe-se que, para essa rede sem camadas ocultas, a importância está concentrada em um único preditor, *despesas*, que, como se vê, representa um vazamento de informações. Esse diagnóstico é útil em geral e, embora não seja tão drástico quanto o exemplo, pode apontar quais variáveis devem ser consideradas com cuidado.

- É essencial considerar, também, o efeito de variáveis associadas a grupos protegidos e, se necessário, examinar cuidadosamente como elas afetam as previsões.

- Modelos parcimoniosos que usam menos atributos facilitam a análise, mantêm o fluxo de dados e reduzem a exposição a problemas de vazamento ou efeitos indesejáveis.

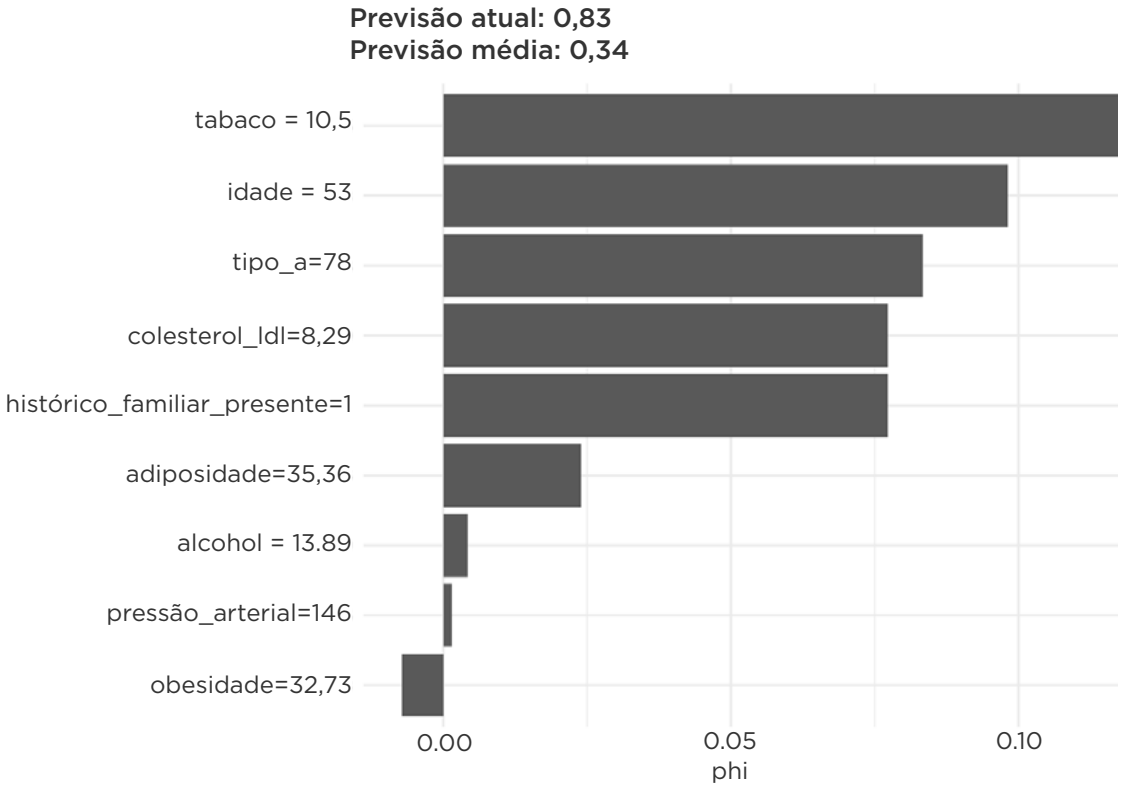
Explicação das previsões

Os valores de Shapley (Molnar, 2019), (Lundberg & Lee, 2017) podem ser usados para explicar previsões individuais. Esses gráficos indicam a contribuição de cada atributo designada a uma previsão individual, com a ideia de considerar efeitos marginais na previsão de acordo com a presença ou ausência de outros atributos. **As contribuições obtidas** somam a diferença existente entre a previsão particular e a previsão média.

As médias entre os grupos de interesse também podem ser examinadas.

Considere o exemplo de fatores para detectar doenças cardíacas (Rossouw, 1983):

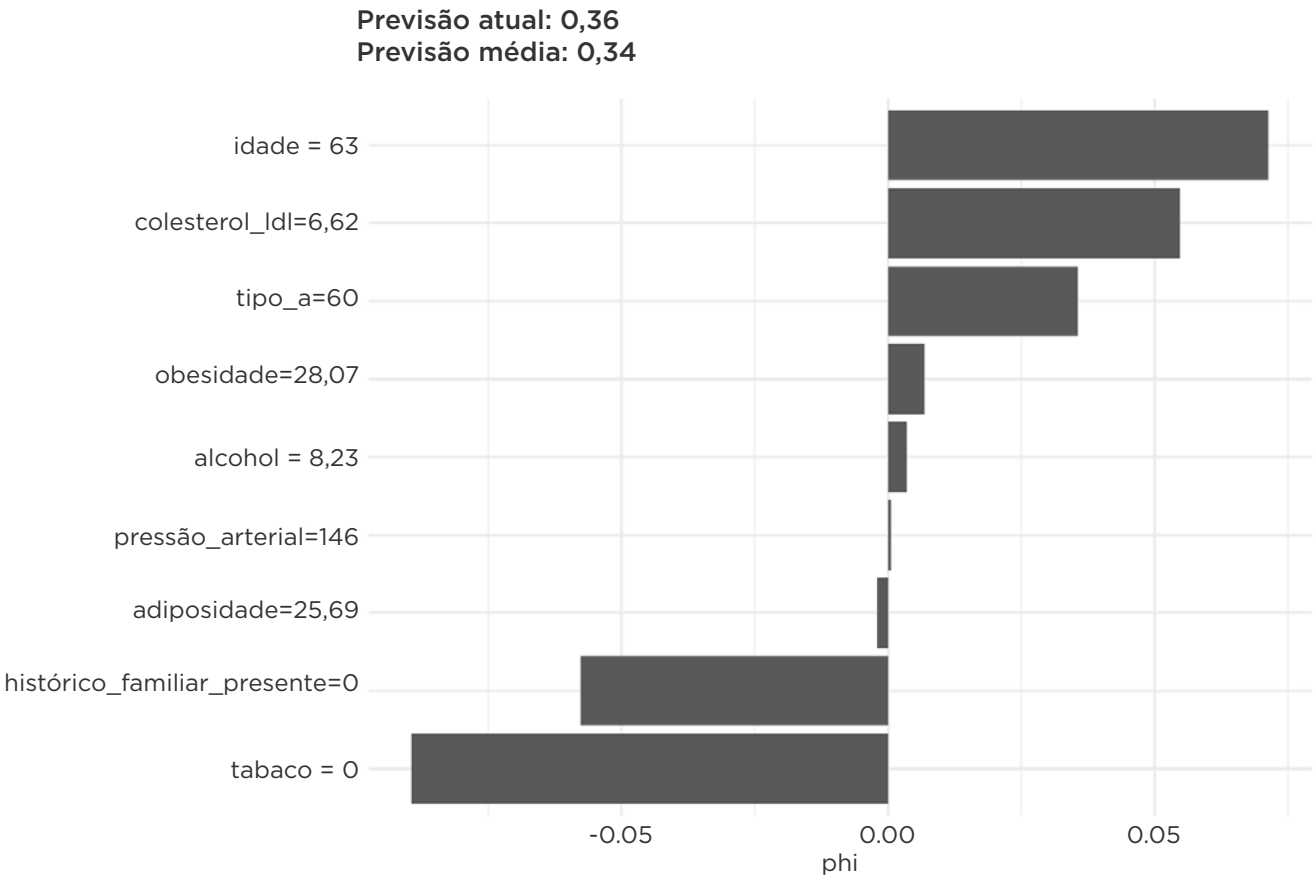
```
modelo_sa <- sa_boosted$fit

sa_entrena_x <- sa_entrena %>% dplyr::select(-enf_coronaria)

predict_fun <- function(object, newdata){

  new_data_x = xgb.DMatrix(data.matrix(newdata), missing = NA)

  results<-predict(modelo_sa, new_data_x)

  return(results)

}

predictor <- Predictor$new(modelo_sa, data = sa_entrena_x, y = sa_entrena$chd ,

                           type = "prob", predict.function = predict_fun)


# o caso de interesse é o caso 15

valores_shapley <- Shapley$new(predictor, x.interest = (sa_entrena_x[15, ]))

valores_shapley$plot() + theme_minimal()
```



Nesse caso, várias medidas contribuem positivamente para a probabilidade de doenças cardíacas, como o tabagismo, a idade e os valores de colesterol. Essas contribuições explicam a alta probabilidade desse indivíduo em particular.

Em contraste, a pessoa a seguir aproxima-se da média. A idade e os valores de colesterol aumentam positivamente a probabilidade, mas o não consumo de tabaco e a inexistência de histórico de diabetes na família têm um impacto no sentido negativo:

```
# o caso de interesse é o caso 24

valores_shapley <- Shapley$new(predictor, x.interest = (sa_entrena_x[24, ]))

valores_shapley$plot() + theme_minimal()
```



Observação: como no modelo e nos gráficos de dependência parcial discutidos acima, esses coeficientes **não** devem ser interpretados de forma causal (por exemplo, o colesterol precisa ser reduzido para esses dois indivíduos). Essas são as informações que o modelo usa para construir a previsão a partir da previsão média para toda a população.

Os valores de Shapley podem ser calculados para dois grupos de idade, por exemplo.

Referências

Washingtonpost. (abril de 2019). *21 more studies showing racial disparities in the criminal justice system*. Obtido em https://www.washingtonpost.com/opinions/2019/04/09/more-studies--showing-racial-disparities-criminal-justice-system/

Rubin, R. J. (2002). *Statistical Analysis with Missing Data, Second Edition.* John Wiley & Sons, Inc.

Vayena, A. J. (2019). *The global landscape of AI ethics guidelines.* Springer Science and Business Media LLC.

Williams, D. M. (09 de 1981). Racial differences of hemoglobin concentration: measurements of iron, copper, and zinc. *The American Journal of Clinical Nutrition, 34*(9), pp. 1694-1700.

Wachter, S., Mittelstadt, B. D. e Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *ArXiv, abs/1711.00399*.

Verma, S. e Rubin, J. (2018). Fairness Definitions Explained. *ConferenceProceedings of the International Workshop on Software Fairness* (pp. 1-7). Nova York, NY, EUA: Association for Computing Machinery.

Vaver, J. e Koehler, J. (2011). *Measuring Ad Effectiveness Using Geo Experiments.* Google Inc.

Suresh, H. e Guttag, J. V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. *ArXiv, abs/1901.10002*.

Sundararajan, M., Taly, A. e Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *ArXiv, abs/1703.01365*.

Obermeyer, Z., Powers, B., Vogeli, C. e Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), pp. 447-453.

Molnar, C. (2019). *Interpretable Machine Learning.*

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artif. Intell., 267*, pp. 1-38.

Lundberg, S. e Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *ArXiv, abs/1705.07874*.

Lohr, S. L. (2009). *Sampling: Design and Analysis.* Cengage Learning.

Little, R. J. e Rubin, D. B. (2002). *Statistical analysis with missing data.* Wiley.

Lackland, D. (junho de 2014). Racial Differences in Hypertension: Implications for High Blood Pressure Management. *The American journal of the medical sciences, 348*.

Kuhn, M. e Johnson, K. (2013). *Applied Predictive Modeling.* Springer New York.

Kaufman, S., Rosset, S. e Perlich, C. (janeiro de 2011). Leakage in Data Mining: Formulation, Detection, and Avoidance., *6*, pp. 556-563.

Hastie, T., Tibshirani, R. e Friedman, J. (2017). *The Elements of Statistical Learning.* Springer New York Inc.

Gelman, A. e Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models* (1 Ed.). Cambridge University Press.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics, 29*(5), pp. 1189-1232.

Buuren, S. V. e Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations. In: R. *Journal of Statistical Software, 45*(3), pp. 1-67.

Buolamwini, J. e Gebru, T. (23–24 de fevereiro de 2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. En S. A. Friedler, & C. Wilson (Ed.), *ConferenceProceedings of the 1st Conference on Fairness, Accountability and Transparency. 81*, pp. 77-91. Nova York, NY, EUA: PMLR.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), pp. 5-32.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. e Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *ArXiv, abs/1607.06520*.

Barocas, S. e Selbst, A. D. (2014). Big Data's Disparate Impact. *SSRN eLibrary*.

Stuart, K. I. (2008). Misunderstandings Among Experimentalists and Observationalists about Causal Inference. *Journal of the Royal Statistical Society, Series A, 171, part 2*, pp. 481-502.

Hardt, M. A. (2016). Equality of Opportunity in Supervised Learning. *CoRR, abs/1610.02413*.

INEGI. (2014). *Encuesta Nacional de Ingresos Y Gastos de Los Hogares (Enigh-2014). Diseño Muestral.* Obtido em http://internet.contenidos.inegi.org.mx/contenidos/Productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825070359.pdf

Rossouw, J. E. (1983). *Coronary Risk Factor Screening in Three Rural Communities. The Coris Baseline Study.* South African Medical Journal = Suid-Afrikaanse Tydskrif Vir Geneeskunde.

Chawla, N. V. (2002). *SMOTE: Synthetic Minority over-Sampling Technique.* Journal of Artificial Intelligence Research 16: 321–57.

Greene, W. (2003). *Econometric Analysis.* Pearson Education. Obtido em https://books.google.com.mx/books?id=njAcXDlR5U8C.

James, G. D. (2017). *Data for an Introduction to Statistical Learning with Applications in R.* Obtido em https://CRAN.R-project.org/package=ISLR.

Harini Suresh, J. V. (2019). *A Framework for Understanding Unintended Consequences of Machine Learning.* MIT. Obtido em https://arxiv.org/pdf/1901.10002.pdf

Kuhn, M. F. (2020). *Rsample: General Resampling Infrastructure.* Obtido em https://CRAN.R-project.org/package=rsample.

Pombo, C., Cabrol, M., Alarcón, N. G. e Ávalos, R. S. (2020). *fAIr LAC: Adopción ética y responsable de la inteligencia artificial en América Latina y el Caribe.* doi:http://dx.doi.org/10.18235/0002169

Wilson, J. (2014). *What your IQ score doesn't tell you.* CNN.

drivendata. (2019). *An ethics checklist for data scientists.* Obtido em https://deon.drivendata.org/

Fritzler, A. (2015). *An ethical checklist for data science.* Obtido em http://www.dssgfellowship.org/2015/09/18/an-ethical-checklist-for-data-science/

Carrillo, A., Cantú, L. e Noriega, A. (2020). *Individual Explanations in Machine.* IADB.

Athey, S. W. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*.

Gebru, T., Morgenstern, J., Vecchione, B., Wortman, J., Wallach, H., Daumé, H. e Crawford, K. (2018). *Datasheets for Datasets.* Obtido em https://arxiv.org/pdf/1803.09010.pdf

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B. e Gebru, T. (2019). *Model Cards for Model Reporting.* Obtido em https://arxiv.org/abs/1810.03993

Prosser, C. e Mellon, J. (2016). *Twitter and Facebook are Not Representative of the General Population: Political Attitudes and Demographics of Social Media Users.* Disponível em SSRN: https://ssrn.com/abstract=2791625 ou http://dx.doi.org/10.2139/ssrn.2791625.

Kim, B. R. (2016). *Examples are not enough, learn to criticize! Criticism for interpretability.* Advances in Neural Information Processing Systems.

OECD. (2019c). *Artificial Intelligence in Society.* Paris: OECD Publishing.

OECD (Em breve). (s.f.). *Framework for the Classification of AI Systems.* Paris: OECD Publishing.

OECD. (2021). Good Practice Principles for Data Ethics in the Public Sector.

Pombo, C., Cabrol, M., González, N. e Sánchez, R. (2020). *fAIr LAC: Adopción ética y responsable de la inteligencia artificial en América Latina y el Caribe.* doi:http://dx.doi.org/10.18235/0002169

Athey, S. W. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*.

OECD. (2019). *Artificial Intelligence in Society.* Paris: OECD Publishing.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B. e Gebru, T. (2019). *Model Cards for Model Reporting.* Obtido em https://arxiv.org/abs/1810.03993

Gebru, T., Morgenstern, J., Vecchione, B., Wortman, J., Wallach, H., Daumé, H. e Crawford, K. (2018). *Datasheets for Datasets.* Obtido em https://arxiv.org/pdf/1803.09010.pdf

Anna Jobin, M. I. (2019). *The global landscape of AI ethics guidelines.* Springer Science and Business Media LLC.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A. e Ghani, R. (2019). *Aequitas: A Bias and Fairness Audit Toolkit.* Center for Data Science and Public Policy.


BID
Melhorando vidas