UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

CARLA PIAZZON RAMOS VIEIRA

Uso de agrupamento para alcançar explicabilidade global de modelos de aprendizado de máquina

São Paulo
2023

CARLA PIAZZON RAMOS VIEIRA

Uso de agrupamento para alcançar explicabilidade global de modelos de aprendizado de máquina

Versão corrigida

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Orientador: Prof. Dr. Luciano Antonio Digiampietri

São Paulo

2023

Ficha catalográfica elaborada pela Biblioteca da Escola de Artes, Ciências e Humanidades,
com os dados inseridos pelo(a) autor(a)
Brenda Fontes Malheiros de Castro CRB 8-7012; Sandra Tokarevicz CRB 8-4936

Vieira, Carla Piazzon Ramos
Uso de agrupamento para alcançar explicabilidade global de modelos de aprendizado de máquina / Carla Piazzon Ramos Vieira; orientador, Luciano Antonio Digiampietri. -- São Paulo, 2023.
68 p: il.

Dissertacao (Mestrado em Ciencias) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2023.
Versão corrigida

1. Interpretabilidade. 2. Explicabilidade. 3. Aprendizado de Máquina. I. Digiampietri, Luciano Antonio, orient. II. Título.

Dissertação de autoria de Carla Piazzon Ramos Vieira, sob o título **"Uso de agrupamento para alcançar explicabilidade global de modelos de aprendizado de máquina"**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 24 de março de 2023 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Luciano Antonio Digiampietri
Universidade de São Paulo
Presidente

Prof. Dr. Márcio Moretto Ribeiro
Universidade de São Paulo

Profa. Dra. Sandra Eliza Fontes de Avila
Universidade Estadual de Campinas

Agradecimentos

Primeiramente gostaria de agradecer a meus pais por nunca me deixarem desistir e terem me ensinado a depositar minha confiança em Deus sempre acreditando que no final tudo dá certo e que uma grande caminhada começa com o primeiro passo.

Ao meu orientador Luciano Antonio Digiampietri por ter aceitado me orientar durante os últimos quatro anos, pela paciência e compreensão durante tempos pandêmicos e pela confiança no meu potencial.

A meu colega pesquisador Tarcizio Silva que sempre deu todo apoio à minha carreira na pesquisa e me incentivou a seguir em frente com uma pesquisa crítica sobre computação.

As minhas queridas professoras Gisele Craveiro e Renata Araújo que me ensinaram a importância de uma pesquisa implicada.

A dissertação que você lerá a seguir foi escrita ao longo de três anos de muita pesquisa e empenho em criar um futuro tecnológico sob uma das perspectivas poucas representadas na tecnologia: a de uma mulher negra periférica.

Durante a graduação, eu me apaixonei pela área de Inteligência Artificial. No entanto, nunca me foi apresentado o debate sobre como a tecnologia poderia impactar de forma desproporcional certos grupos. A primeira vez que ouvi sobre o assunto foi através da minha mentora Raissa Kullian, uma das pessoas a quem dedico esse texto. Na minha jornada de definir uma proposta de pesquisa para me candidatar ao mestrado, ela disse que eu deveria pesquisar sobre os desafios da área de Inteligência Artificial. Essa pesquisa me levou até os posts de investigação algorítmica da Karen Hao no MIT Review que me revelaram como vieses humanos se tornavam parte da inteligência artificial e como, grande parte dos sistemas algorítmicos são caixas-pretas.

Após a leitura de dezenas de artigos e seguir alguns pesquisadores no twitter, descobri um fenômeno chamado "XAI" (Explicabilidade de Inteligência Artificial). Pesquisadores do mundo todo buscando abrir a caixa-preta e tornar a Inteligência Artificial confiável e explicável. Ali, eu vi uma possibilidade de construir um futuro melhor por meio da tecnologia. A partir desse dia, eu decidi que queria fazer parte do debate que critica sistemas que concentram poder e constroem sistemas que geram impacto social e positivo para a sociedade. O mestrado me levou a rumos interdisciplinares que foram essenciais para construção da minha pesquisa. Estudando filosofia da tecnologia, raça e gênero,

sociologia etc., entendi que faltava incentivo à colaboração entre essas áreas e a computação. Inteligência Artificial trata-se de uma área altamente interdisciplinar. Tolice acreditar que seria possível resolver problemas sociais apenas com ferramentas tecnológicas.

Espero quero que essa dissertação sirva de referência para próximas gerações de cientistas para mostrar que a tecnologia não é neutra e, por isso, a importância de uma pesquisa implicada na computação. Meu objetivo nunca foi desenvolver otimizações de algoritmo nem defender aqui que a explicabilidade é a solução. Meu objetivo é reforçar a importância da interdisciplinaridade nessa discussão e tentar, de alguma forma, construir pontes com outras áreas.

Resumo

VIEIRA, Carla Piazzon Ramos. **Uso de agrupamento para alcançar explicabilidade global de modelos de aprendizado de máquina**. 2023. 68 f. Texto para o Exame de Qualificação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2023.

Atualmente, modelos de aprendizado de máquina são utilizados para tomar decisões nas mais diversas aplicações da sociedade. Erros humanos se tornam parte da inteligência artificial provando que nossas tendências de generalização são responsáveis pela seleção de uma regra que molda a criação desses modelos. Desta forma, é possível que, em vez de estar criando inovação, se está reforçando comportamentos discriminatórios existentes na sociedade. Nesse contexto, aumentar a interpretabilidade e identificar vieses discriminatórios nos modelos se tornam atividades extremamente relevantes. Acredita-se que a interpretabilidade pode ser uma ferramenta para garantir que não se violem requisitos éticos e legais. As explicações extraídas por técnicas de interpretabilidade facilitam que humanos identifiquem as possíveis deficiências dos modelos e analisem suas limitações. Desde 2016, o número de trabalhos envolvendo interpretabilidade de algoritmos e análises de "justiça" tem aumentado. No entanto, a área ainda carece de mais estudos que se dediquem a interpretabilidade global de modelos a fim de garantir o princípio de justiça. A presente dissertação apresenta os resultados de implementação de um método de explicabilidade global que tem como base o agrupamento de explicações locais. Os resultados mostraram que uma única explicação pode não ser suficiente para interpretar o modelo com um todo. De forma que o uso da técnica de agrupamento permitiu identificar comportamentos discriminatórios entre diferentes subgrupos. A análise dos resultados foi realizada por meio de comparações das explicações globais do modelo e explicações globais de cada grupo tendo como suporte análises exploratórias dos dados.

Palavras-chaves: Interpretabilidade. Explicabilidade. Aprendizado de máquina.

Abstract

VIEIRA, Carla Piazzon Ramos. **Using clustering methods to achieve global explainability for machine learning models**. 2023. 68 p. Dissertation project (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2023.

Nowadays, machine learning models are utilized to make decisions in the most diverse applications in society. Human errors become part of artificial intelligence, proving that our generalization tendencies are responsible for selecting the rules that shape the creation of these models. Thus, it is possible that, instead of creating innovation, existing discriminatory behaviors in society are being reinforced. In this context, increasing interpretability and identifying discriminatory biases in the models become extremely relevant activities. It is believed that interpretability can be a tool to ensure that ethical and legal requirements are not violated. The explanations extracted by interpretability techniques make it easier for humans to identify the possible deficiencies of the models and analyze their limitations. After 2016, the number of works involving the interpretability of algorithms and analyses of fairness increased. However, the area still needs more studies that are dedicated to the global interpretability of models to guarantee the principle of fairness. This dissertation aimed to generate global explanations by clustering local explanations. The results have shown that a single explanation may not be sufficient to interpret the model as a whole. Thus, the use of clustering techniques allowed us to compare the behavior of the models in different subgroups and identify discriminatory biases. As a way of evaluating the results, a comparative analysis was conducted between the model single global expression and the expression generated for each group with the support of exploratory data analysis.

Keywords: Interpretability. Explainability. Machine Learning.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

SIGLA	SIGNIFICADO
CERTIFAI	Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
DARPA	Defense Advanced Research Projects Agency
DeepLIFT	Deep Learning Important FeaTures
DiCE	Diverse Counterfactual Explanations for Machine Learning Classifiers
DNNs	Deep Neural Networks
FAccT	Fairness, Accountability, and Transparency
Fair-MAML	Fair-Model Agnostic Meta Learning
G-REX	Genetic Rule EXtraction
GAM	Global Attribution Method
GDPR	Regulação Geral de Proteção de Dados da União Europeia
IA	Inteligência Artificial
LGPD	Lei Geral de Proteção de Dados
LIME	Local Interpretable Model-agnostic Explanations
MAPLE	Model Agnostic SuPervised Local Explanations
ML	Machine Learning
MUSE	Model Usability Evaluation
NeurIPS	Neural Information Processing Systems
SHAP	SHapley Additive exPlanations
SLIM	Supersparse Linear Integer Model
XAI	Explainable Artificial Intelligence

Sumário

1 Introdução

Na sociedade contemporânea, a informação tornou-se elemento central para o desenvolvimento humano, configurando novas formas de organização social e cultural, acompanhada pela crescente evolução tecnológica. Essa sociedade é o resultado do trabalho de pesquisadores, da indústria e de governos que, ao transformar informação em conhecimento, desenvolveram, encorajaram e implementaram tecnologias e alternativas para automatizar inúmeras decisões. Decisões humanas passaram a ser automatizadas pelas máquinas por meio de algoritmos de aprendizado de máquina e, em muitas tarefas, têm performado melhor que seres humanos. Em 2016, o programa de computador *AlphaGO* (desenvolvido pela *Google DeepMind*) derrotou 18 vezes o profissional de Go sul-coreano Lee Sedol (SILVER *et al.*, 2016). O antigo jogo de tabuleiro chinês é há muito tempo considerado um teste da alta dificuldade para inteligência artificial dadas sua alta complexidade e necessidade de visão estratégica.

No entanto, ao mesmo tempo em que os algoritmos trazem possibilidades magníficas de otimização do trabalho e outras transformações positivas, também trazem riscos e desafios que precisam ser considerados antes de sua adoção. Sistemas utilizando aprendizado de máquina têm sido empregados em diferentes aplicações como: análise de crédito, reconhecimento facial, seleção de currículos e até decisões judiciais.

Um caso polêmico na área de segurança pública foi a condenação de Eric Loomis a seis anos de prisão auxiliada pelo resultado de um algoritmo chamado *COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)* (ANGWIN *et al.*, 2016). O *COMPAS* foi desenvolvido pela empresa privada Northpointe (agora Equivant) cujo slogan é "Software para Justiça". O algoritmo calcula a probabilidade de reincidência criminal do réu. No caso de Loomis, o resultado do algoritmo foi de que ele teria alta probabilidade de reincidir e, portanto, foi associado a uma sentença mais severa. Loomis discordou desta sentença, porém não teve como recorrer porque o algoritmo não era capaz de gerar explicação do resultado gerado. O direito à explicação foi negado.

Os exemplos apresentados (*AlphaGo* e *COMPAS*) ilustram as oportunidades e desafios postos dados os grandes avanços no desenvolvimento da Inteligência Artificial. Sistemas como o *COMPAS* dividem opiniões e vêm sendo alvo de críticas considerando pesquisas recentes que demonstraram os vieses e falhas existentes nessas tecnologias.

As implicações de se introduzir uma ferramenta de inteligência artificial que reproduza comportamentos discriminatórios ou ilegais vai além de questões financeiras, como perda de receita e multas por problemas de não conformidade. As falhas geradas por esses sistemas podem representar constrangimentos, discriminações e violações de privacidade e direitos humanos.

Leslie (2019) desenvolveu um guia pelo Instituto Alan Turing que apresenta uma lista de potenciais danos algorítmicos para a sociedade. O restante do presente capítulo é composto de duas seções a fim de apresentar um breve contexto sobre os seguintes danos: (i) Discriminação Algorítmica e (ii) Predições inexplicáveis ou injustificáveis.

1.1 Problema 1: Discriminação algorítmica

Discriminação algorítmica refere-se ao fenômeno em que modelos de aprendizado de máquina reproduzem e amplificam desigualdades existentes na sociedade. Por conta disso, a preocupação com justiça tem atraído cada vez mais atenção na área de aprendizado de máquina.

Uma tentativa para evitar discriminação nos resultados dos modelos é a proibição de uso de dados sensíveis prevista na Lei Geral de Proteção de Dados Pessoais (LGPD) (Brasil, 2018). Sendo que no Artigo 5, parágrafo II, defini-se dados sensíveis como: "São dados pessoais referentes a origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dados genético ou biométrico, quando vinculado a uma pessoa natural". Além disso, a LGPD busca garantir direitos de não-descriminação e privacidade.

Apenas não utilizar variáveis sensíveis pode não ser suficiente para garantir o princípio de não discriminação. Nos dados de treinamento do modelo, podem existir variáveis latentes que permitam identificar a variável sensível. Por exemplo, já existem diversos estudos demonstrando que o sobrenome de uma pessoa pode ser um indicativo da sua raça ou etnia (KOZLOWSKI *et al.*, 2022).

Avaliar se um algoritmo é justo não é uma tarefa fácil e diversas definições matemáticas foram propostas. (RUBACK; CARVALHO; AVILA, 2022) apresentam um artigo que realizou uma análise sociotécnica sobre vieses inseridos no aprendizado de máquina, impactos culturais gerados e possíveis caminhos de mitigação de vises. Verma e Rubin

(2018) fazem uma revisão da literatura e mostram que existem mais de vinte definições de justiça definidas nos últimos anos, sendo possível obter classificações contradizentes de justiça de acordo com a definição usada. Logo, ainda não há um consenso sobre as definições. No geral, os métodos escolhem uma definição de justiça para avaliar se as predições do modelo estão sendo justas para diferentes grupos representados no conjunto de dados. As definições de justiça podem ser divididas em dois principais tipos: justiça entre grupos e justiça entre indivíduos. Essas definições quantificam a relação entre grupos "privilegiado" e "desprivilegiado" para determinado atributo sensível.

1.2 Problema 2: Predições inexplicáveis ou injustificáveis

O problema com o estado da arte de vários modelos é a falta de transparência e interpretabilidade o que é denominado na literatura como modelos caixa-preta (PAPADOPOULOS; WALKINSHAW, 2015). De forma que mesmo quando as entradas e saídas são conhecidas, muitos algoritmos conseguem sugerir respostas, mas não dizer o porquê de suas decisões 1. Isto dificulta que uma empresa explique seu processo de tomada de decisão a reguladores, clientes, membros do conselho e outras partes interessadas, ou, ainda, que médicos tenham confiança sobre os resultados produzidos por um algoritmo. Com o uso de modelos menos complexos (modelos lineares, árvores de decisão etc.) pode-se alcançar um maior grau de interpretabilidade dos resultados gerados. No entanto, em determinados contextos e aplicações, que lidam com uma maior volume e complexidade de dados, modelos mais complexos apresentam melhor desempenho.

Figura 1 – Diagrama de um algoritmo caixa-preta

A ausência de explicação tem motivado um levante crescente de discussões nas mais diferentes esferas da sociedade. A fim de construir confiança em sistemas de aprendizado de máquina e avançar para a sua integração significativa na sociedade, em 2017, a DARPA *(Defense Advanced Research Projects Agency)* criou o programa XAI, *Explainable Artificial Intelligence* com o objetivo de desenvolver modelos de aprendizado de máquina explicáveis

(GUNNING, 2017). Inteligência Artificial Explicável refere-se a área de pesquisa que desenvolve métodos e técnicas aplicados a algoritmos de inteligência artificial de forma que os resultados e predições destes sejam compreensíveis por humanos. Do ponto de vista de desenvolvedores e pesquisadores de aprendizado de máquina, as explicações fornecidas podem ajudá-los a entender melhor o problema, os dados e por que um modelo pode falhar.

Na literatura, diversos livros sobre as implicações da Inteligência Artificial na sociedade e na democracia têm sido publicados (BENJAMIN, 2019; O'NEIL, 2016; PASQUALE, 2015; SILVA, 2020). Na indústria, a comoção pública diante dos casos de discriminação algorítmica tem feito com que muitas empresas se preocupem com os possíveis impactos dos seus produtos de IA na sociedade (SILVEIRA; SILVA, 2020). Na esfera acadêmica e científica, diversas conferências na área de aprendizado de máquina têm dado destaque ao tema. A conferência *NeurIPS* tornou requisito que os autores enviem uma declaração sobre o "impacto do seu trabalho na sociedade". Também surgiram conferências cujo principal tema é a ética, com destaque para a *FAccT (Fairness, Accountability, and Transparency)* que ocorre desde 2018.

Diversos métodos de interpretabilidade foram desenvolvidos com o objetivo de serem aplicados a diferentes modelos de aprendizado de máquina (JOHANSSON; KöNIG; NIKLASSON, 2010; LAKKARAJU *et al.*, 2019; EVANS; XUE; ZHANG, 2019; PLUMB; MOLITOR; TALWALKAR, 2018; MESSALAS; KANELLOPOULOS; MAKRIS, 2019; RIBEIRO; SINGH; GUESTRIN, 2016b; LUCIC; HANED; RIJKE, 2020; SHARMA; HENDERSON; GHOSH, 2020; MOTHILAL; SHARMA; TAN, 2020; IBRAHIM *et al.*, 2019; SLACK *et al.*, 2020a). Apesar de existirem diversas técnicas de interpretabilidade, ainda não existe um acordo sobre como avaliar as explicações geradas ou quais requisitos essas explicações devem atender. Dificultando a escolha de qual técnica de interpretabilidade seria mais adequada dentro de cada contexto. Além disso, um dos grandes desafios da área é oferecer explicações centradas no usuário de forma a permitir alguma ação a partir delas.

Como será argumentado de forma mais abrangente nas páginas seguintes, existem três objetivos pelos quais normalmente procura-se explicar as predições de algoritmos caixa-preta: auditar, validar e descobrir.

O primeiro objetivo surge com mais frequência em ambientes de alto risco, como justiça criminal e recrutamento, em que é importante garantir que os modelos de aprendizado de máquina não discriminem grupos historicamente marginalizados. Testar a confiança

de algoritmos como o *COMPAS* em atributos protegidos como raça requer técnicas sofisticadas de explicabilidade e *fairness*. Casos como esses tendem a interessar pesquisadores de diferentes áreas que enfatizam as implicações sociais, econômicas e políticas do uso de algoritmos para acelerar ou automatizar decisões sensíveis.

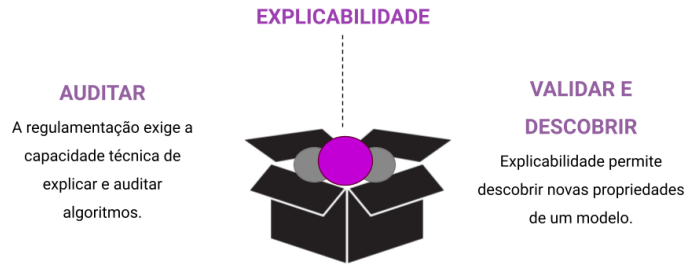


Figura 2 – Uso de explicabilidade para auditar, validar e descobrir

A validação, por outro lado, é mais um objetivo epistemológico do que ético. Mesmo algoritmos de alto desempenho são propensos a ter *overfitting* sobre os dados de treinamento, resultando em um comportamento inesperado em conjuntos de teste e validação. Por conta disso, que muitas ferramentas de interpretabilidade são projetadas para fornecer aos engenheiros maneiras de analisar a lógica interna de um modelo.

Por último, em menor destaque que os dois anteriores, está descobrir novas propriedades de um sistema. Imagine se o *AlphaGo* não só pudesse vencer os campeões mundiais em Go, mas ensinasse como melhorar seu próprio jogo explicando as decisões estratégicas enquanto ele joga? A perspectiva não é tão fantasiosa quanto parece. Quando os modelos estatísticos superam os especialistas humanos, há motivos para acreditar que eles aprenderam informações e correlações significativas desconhecidas.

O potencial científico dos avanços em interpretabilidade podem, portanto, ser bem significativos independente do objetivo.

1.3 Questões e proposta de pesquisa

Benjamin (2019) entende que "abrir a caixa-preta" também significa entender que qualquer ordem social é impactada pelo desenvolvimento tecnológico e que normas, ideologias e práticas sociais devem ser parte do desenvolvimento de novas tecnologias. Portanto, para enfrentar as discriminações algorítmicas socialmente e democraticamente inaceitáveis é necessário que esses sistemas sejam transparentes, explicáveis e auditáveis.

Epstein *et al.* (2018) apresentam um estudo que demonstrou que a quantidade de artigos com propostas de novos modelos supera em 10 vezes o total de artigos que apresentam análises críticas de modelos existentes. Os autores sugerem que a aceleração de estudos críticos aos modelos existentes requer incentivos tanto na academia quanto na indústria. De forma que, de um lado, existiria um primeiro grupo de pesquisadores que criam novos algoritmos; e, por outro lado, um segundo grupo, de pesquisadores que seriam como auditores de IA e visariam a desenvolver ferramentas para avaliar e interpretar os novos algoritmos propostos.

A presente dissertação encaixa-se no segundo grupo à medida que busca trazer rigor crítico e sociotécnico ao problema de explicabilidade para modelos caixa-preta. Por toda a atenção que a área de Interpretabilidade tem recebido recentemente, ambiguidades importantes em relação aos seus verdadeiros objetivos e critérios de sucesso permanecem sem resposta.

1.3.1 Hipótese

Conforme será detalhado no Capítulo 3, as técnicas de explicabilidade podem ser agrupadas de acordo com seu escopo: local ou global. As técnicas de escopo local buscam entender o comportamento do modelo em uma predição específica, já as técnicas globais visam a entender o comportamento do modelo com um todo. Nesta dissertação, a proposta implementada utilizou a técnica de explicabilidade local SHAP, uma técnica local de interpretabilidade, considerada referência.

No entanto, as técnicas de interpretabilidade global, em sua grande maioria, são limitadas porque atribuem a mesma explicação a múltiplas observações (MESSALAS; KANELLOPOULOS; MAKRIS, 2019). Além disso, podem existir contradições entre os resultados gerados pela técnica utilizada e a análise de *fairness* do modelo. A principal hipótese levantada no âmbito desta dissertação é de que o uso de algoritmos de agrupamento no desenvolvimento de técnicas de interpretabilidade, conforme o estudo de Ibrahim *et al.* (2019), pode permitir identificar subgrupos em que o modelo se comporte de maneiras diferentes ou até discriminatórias.

A proposta apresentada segue uma visão agnóstica de modelo que busca construir ferramentas para gerar explicações com pouca ou nenhuma suposição sobre o sistema

caixa-preta a ser explicado. O escopo da presente dissertação está limitado a algoritmos de aprendizado supervisionado utilizando dados tabulados. Algoritmos não supervisionados e de aprendizado por reforço apresentam desafios explanatórios únicos que estão além do escopo desta dissertação.

1.3.2 Objetivos

O objetivo geral desta pesquisa é propor o uso de técnicas de agrupamento para extrair explicações globais de modelos de aprendizado de máquina, buscando identificar potenciais comportamentos discriminatórios e aumentar o conhecimento do funcionamento do modelo. As principais contribuições dessa dissertação são:

- proposta de um método que permite identificar comportamentos discriminatórios em um modelo de aprendizado de máquina entre diferentes subgrupos de dados por meio do agrupamento de explicações locais
- proposta de avaliação de métodos de explicabilidade utilizando dados do contexto dos dados e análises demográficas

1.4 Organização deste documento

Este documento é composto por seis capítulos, incluindo o presente capítulo de introdução. O Capítulo 2 apresenta uma discussão de interpretabilidade pautada na importância da visão sociotécnica e da teoria crítica da tecnologia como lentes teóricas no desenvolvimento de ferramentas de interpretabilidade a algoritmos de IA. Após isto, o Capítulo 3 apresentamos conceitos, definições e desafios na construção de métodos de explicabilidade. No Capítulo 4, detalhamos a proposta da presente dissertação. Por fim, os Capítulos 5 e 6 apresentam os experimentos, dados e resultados realizados e considerações finais da dissertação, respectivamente.

1.5 Ética e reprodutibilidade

A presente dissertação não utiliza quaisquer dados pessoais, conforme definido na LGPD. Todos os dados analisados aqui são simulados ou retirados de repositórios de

dados públicos, como o repositório *Irvine Machine Learning* da Universidade da Califórnia (DUA; GRAFF, 2017). Todas as análises foram realizadas utilizando a linguagem Python. O código para reproduzir todos os resultados pode ser encontrado no repositório GitHub: ⟨https://github.com/carlaprv/dissertation⟩.

2 Em defesa de uma visão crítica da Inteligência Artificial

Há décadas, cientistas da computação vêm buscando fazer com que máquinas aprendam como humanos. Acredita-se que a história da inteligência artificial teve início após a Segunda Guerra Mundial. Em 1943, McCulloch e Pitts (1943) apresentaram um artigo inédito propondo um modelo matemático de estruturas de raciocínio artificiais que imitava o nosso sistema nervoso humano. Em 1950, Shannon (1988) apresentou um estudo sobre como programar uma máquina para jogar xadrez. Também em 1950, Turing (1950) desenvolveu o teste de Turing: uma forma de avaliar se uma máquina consegue se passar por um humano em uma conversa por escrito. Já em 1956, o cientista da computação John McCarthy cunhou o termo Inteligência Artificial como "a ciência e a engenharia de fazer máquinas inteligentes". Algo em comum na história e concepção da Inteligência Artificial é a sua suposta neutralidade e a visão instrumentalista de seus criadores.

Na última década, autores de diferentes áreas de formação (GEBRU, 2019; O'NEIL, 2016; FEENBERG, 1992) têm chamado atenção sobre os rumos para os quais a sociedade está se encaminhando se não refletir profundamente sobre o desenvolvimento científico e tecnológico. Suas ponderações e pesquisas alertam sobre a necessidade de, na área de Computação, serem investidos esforços no desenvolvimento de pesquisas interdisciplinares.

Considerando a emergência de discussões mais abrangentes e críticas sobre a tecnologia, uma alternativa que se apresenta é a Teoria Crítica da Tecnologia proposta pelo filósofo Andrew Feenberg (FEENBERG, 1992). Por defender que o entendimento da tecnologia está ligado a aspectos funcionais e sociais, os alicerces da teoria proporcionam embasamento para analisar sistemas tecnológicos com foco no papel dos indivíduos na construção desses sistemas. A seções a seguir exploram alguns fundamentos da teoria.

2.1 Teoria Crítica da Tecnologia

Assim como a filosofia política problematiza as formações culturais que fundamentaram as leis, a filosofia da tecnologia problematiza as formações que sucessivamente fundamentaram o desenvolvimento tecnológico. Feenberg (BORDIN, 2018) divide as teorias desenvolvidas na Filosofia da Tecnologia em três grupos: Instrumentalismo, Substantivismo

e Determinismo. As limitações dessas teorias tradicionais constituem os fundamentos da Teoria Crítica da Tecnologia de Andrew Feenberg.

2.1.1 Instrumentalismo, Determinismo e Substantivismo

O Instrumentalismo é a teoria tradicional da tecnologia na qual o controle humano e a neutralidade de valor se encontram. A concepção instrumentalista preconiza que, para além da neutralidade como meio instrumental, a tecnologia possui uma neutralidade política.

Segundo Feenberg (1992), "ao fim do século XIX, sob a influência de Marx e Darwin, o progressismo se tornou determinismo tecnológico". Esse determinismo afirma que o avanço tecnológico é uma força impulsionadora da história, e que não é controlada pela humanidade, mas, o contrário: "a tecnologia está enraizada, por um lado, no conhecimento da natureza e, por outro, nas características gerais da espécie humana. Não cabe a nós adaptar a tecnologia a nossos caprichos, mas, ao contrário, nós devemos nos adaptar à tecnologia como a expressão mais significativa de nossa humanidade" (FEENBERG, 1992). O grande sucesso da tecnologia moderna parece ter confirmado essa visão evolucionista e progressista do determinismo. Ao mesmo tempo, ela criou um sistema tecnocrático, isto é, "um sistema administrativo amplo, que é legitimado pela referência aos conhecimentos científicos" (FEENBERG, 1992).

O movimento substantivista surge contra essa tendência tecnocrática. Diferentemente da neutralidade do instrumentalismo e determinismo, ele "atribui valores substantivos à tecnologia" que envolvem o compromisso ético-político. De modo que a tecnologia não é meramente instrumental. Para a teoria substantivista, a teoria instrumentalista erra por ignorar as implicações culturais da tecnologia.

O termo "substantivismo" foi escolhido para descrever uma posição que atribui valores substantivos à tecnologia, em contraste com as visões como a do instrumentalismo e a do determinismo, nos quais a tecnologia é vista como neutra. A tese da neutralidade atribui um valor à tecnologia, mas é um valor meramente formal, a eficiência, que pode servir a diferentes concepções de "bem social". Um valor substantivo, pelo contrário, envolve um compromisso com uma concepção específica de "bem social". Se a tecnologia incorpora um valor substantivo, não é meramente instrumental e não pode ser usada

segundo diferentes propósitos de indivíduos ou sociedades com ideias diferentes de "bem social". Existe semelhança entre a teoria substantivista da tecnologia e o determinismo. Na realidade, a maioria dos teóricos substantivistas também são deterministas. Entretanto, a posição determinista é geralmente otimista. A teoria substantiva não faz tal suposição sobre as necessidades a que a tecnologia serve e não é otimista, mas crítica. Nesse contexto, a autonomia da tecnologia é ameaçadora.

A teoria crítica reconhece as consequências catastróficas do desenvolvimento tecnológico ressaltadas pelo substantivismo, mas ainda acredita em uma promessa de maior liberdade e democratização da tecnologia. Feenberg (1992) entende que a evolução tecnológica é guiada por interesses políticos e econômicos dominantes da sociedade. Com a participação desses interesses, as tecnologias não são totalmente autônomas na determinação do seu próprio desenvolvimento, pois, como já evidenciado ao longo da história humana, a evolução das tecnologias se encontra dependente dos interesses sociais de quem as guiam. Como elementos estruturais de sua Teoria, Andrew Feenberg destaca os conceitos de Código Técnico, Teoria da Instrumentalização e Racionalização Subversiva.

2.1.2 Código Técnico

Feenberg (1992) entende que qualquer tecnologia em uso na sociedade moderna se constrói obedecendo a um *design* que estabelece normas que determinam as possíveis aplicações da mesma. Nas filosofias tradicionais expostas anteriormente, esse *design* é ditado pelo paradigma da eficiência de forma que a tecnologia mais eficiente é a que prevalece. No entanto, os questionamentos e limitações levantados por Feenberg mostram que até mesmo o paradigma da eficiência encontra-se sujeito a uma relatividade sociocultural e política. Feenberg (1992) acredita que esta relatividade se encontra no conceito de **código técnico**. Portanto, código técnico é o conjunto que engloba as normas e interesses sociais que estão em jogo na construção e desenvolvimento de uma determinada tecnologia.

Em uma sociedade tecnológica como a capitalista, os códigos técnicos são enviesados a partir dos valores dos atores dominantes, e caso passem despercebidos, ou mesmo reforçados, pela própria sociedade, eles se tornam hegemônicos. Poderíamos tomar como exemplo, neste ponto, o quanto de reivindicações e pressões políticas estão presentes no uso de reconhecimento facial para segurança pública. As preocupações com o uso de

reconhecimento facial e os impactos negativos da tecnologia acabam sendo racionalizadas à medida em que o fator eficiência passa a ser determinante para novos investimentos.

Sendo assim, a tecnologia incorpora, para além do aspecto funcional, uma dimensão 'subjetiva' que se mostra por meio da participação que os interesses sociais desempenham na sua construção. No modelo de sociedade em que vivemos, há diferenças muito significativas nos interesses e nas influências exercidas pelas classes dominantes e grupos historicamente marginalizados. Feenberg acredita que o desenvolvimento tecnológico deveria ser dependente da participação dos interesses sociais e que a estrutura tecnológica existente está sujeita a uma transformação que pode garantir um caráter democrático à tecnologia (QUEIROZ, 2013). A base para essa transformação está na teoria da instrumentalização.

2.1.3 Teoria da Instrumentalização

A Teoria da Instrumentalização busca entender de que forma os interesses sociais são sistematizados na tecnologia e como a tecnologia pode ser democratizada a ponto de se libertar de um poder/controle sociopolítico. O autor acredita que, para compreender a tecnologia em toda a sua extensão, é preciso considerar duas dimensões: instrumentalização primária (correspondente à dimensão funcional da tecnologia) e secundária (correspondente à dimensão social da tecnologia).

A instrumentalização primária corresponde ao processo de funcionalização de sistemas tecnológicos. O processo pode ser divido em quatro momentos (Tabela 1), nos quais o sistema é analisado e construído a partir apenas dos seus aspectos funcionais.

Tabela 1 – Momentos da instrumentalização primária (MILHANO, 2010)

Momento	**Descrição**
Descontextualização	os objetos são descontextualizados do seu mundo, ou seja, são anuladas todas as relações que com ele se estabelecem
Reducionismo	os objetos já descontextualizados são simplificados e reduzidos a suas propriedades instrumentais (de utilidade)
Automatização	o objeto da ação tecnológica é abstraído dos seus possíveis impactos no mundo através da introdução da autonomia na sua estrutura
Posicionamento	o objeto é posicionado na esfera tecnológica com uma aplicação que está determinada nas leis funcionais que regem a sua utilização

A instrumentalização secundária, por sua vez, apresenta a tecnologia sob o aspecto social. É nesse ponto que surge a possibilidade de participação dos interesses sociais tanto na atribuição de funções à tecnologia quanto na orientação das escolhas que dizem respeito ao seu desenvolvimento e às suas implicações sociais. Da mesma forma que a instrumentalização primária, a secundária pode ser entendida a partir de quatro momentos como descritos na Tabela 2.

Tabela 2 – Momentos da instrumentalização secundária (MILHANO, 2010)

Momento	**Descrição**
Sistematização	estabelecimento das ligações necessárias para o funcionamento dos objetos tecnológicos, sendo esses recontextualizados no meio social do qual foram extraídos
Mediação	momento no qual são associados atributos sociais aos objetos da ação tecnológica
Vocação	os objetos da ação tecnológica não são autônomos; pelo contrário: estabelecem efeitos com os sujeitos que com eles se relacionam
Iniciativa	momento em que as aplicações atribuídas aos objetos são redefinidas a partir da sua implementação no meio social; Andrew Feenberg entende que os aspectos funcionais que regem a aplicação (ou o posicionamento) destes objetos se redefinem por meio das relações estabelecidas pelos sujeitos

Por meio da teoria da instrumentalização, Andrew Feenberg (FEENBERG, 1992) apresenta uma concepção reflexiva da tecnologia a partir da qual a sua transformação possa ser possível.

2.1.4 Racionalização subversiva

Andrew Feenberg entende que através da luta social conduzida por grupos sociais minoritários se institui, de forma democrática, uma racionalização subversiva na tecnologia, a qual coloca em debate o controle exercido pela tecnologia sobre esses grupos, assim como suas necessidades não contempladas.

No mundo moderno, assim como na sociedade brasileira, as classes sociais dominantes possuem maior poder sociopolítico que as subordinadas e isso pode vir a significar uma maior influência das primeiras sobre o processo de instrumentalização secundária. Feenberg (FEENBERG, 1992) argumenta que a tendência tecnocrática das sociedades

modernas suprime os potenciais benefícios da tecnologia que poderiam emergir de uma lógica diferente de desenvolvimento.

O que acontece é que se, por um lado, os atores estratégicos podem conceber seu projeto técnico de maneira descontextualizada e reducionista, por outro, os excluídos ou afetados negativamente por esses projetos podem perceber as consequências e, em razão disso, serem resistência.

A Teoria Crítica da Tecnologia nos mostra que o determinismo tecnológico tem sido tomado como dispositivo para sustentar a suposta neutralidade de uma tecnologia concebida a partir da colonial-modernidade. Nesse sentido, o "código técnico tecnocrata" pode e deve dar lugar, segundo Feenberg, a um "código técnico socialista". E, como orientação para uma política tecnológica, a teoria crítica da tecnologia tem como uma de suas funções identificar exatamente os limites dos códigos técnicos criados pela autonomia operacional, tentando abrir espaço para uma "democratização da tecnologia", na qual os valores dos atores subordinados também possam ter voz regulativa na dinâmica tecnológica.

Feenberg (FEENBERG, 1992), por meio da teoria crítica da tecnologia, demonstrou que os impactos sociais gerados por qualquer tecnologia precisam ser estudados em sua complexidade, pois envolvem aspectos funcionais e sociais. Isso implica questionar a própria realidade justificada ainda com base no determinismo tecnológico. Feenberg propõe uma racionalização subversiva para a tecnologia. Esta proposição contradiz o determinismo tecnológico que opera sob uma lógica linear segundo a qual a tecnologia necessariamente implica em progresso e que, os processos tecnológicos são independentemente de quaisquer fatores sociais. Portanto, é sob perspectivas dos riscos gerados pelo desenvolvimento não questionado e mercadológico que ameaça a dimensão crítica, que as tecnologias precisam ser desafiadas, reexaminadas e ressignificadas.

Talvez também seja necessário reformular e reforçar a relação entre trabalho acadêmico e ativismo social e político. Tudo isto se deve fazer tendo um horizonte amplo que inclua referência à necessidade de criar uma nova perspectiva em pesquisas de computação que enxergue a tecnologia de um ponto de vista crítico.

A presente dissertação buscou embasamento científico para que acadêmicos e profissionais da área de computação possam estar atualizados em sua prática. Assim como levantar uma crítica sobre os estudos atuais da área e a necessidade de uma renovação dos referenciais teóricos e críticas ao modelo de pensar existente.

3 O jogo da explicação

A explicabilidade não é um problema novo para os sistemas de IA, mas cresceu junto com o sucesso e a adoção de técnicas de Aprendizado Profundo (*Deep Learning*), que deram origem tanto a grandes avanços tecnológicos quanto à maior opacidade desses sistemas. A geração atual de sistemas baseados em aprendizado de máquina são o que chamamos de modelos caixa-preta. Conforme antecipado na introdução, mesmo quando as entradas e saídas são conhecidas, esses sistemas podem sugerir respostas, mas não dizer o "porquê" por trás de suas decisões.

"*Se os projetistas e usuários finais de um sistema de aprendizado quiserem ter confiança no desempenho do sistema, eles devem entender como ele chega a suas decisões. Os sistemas de aprendizado também podem desempenhar um papel importante no processo de descoberta científica.*" (CRAVEN; SHAVLIK, 1995)

É difícil imaginar uma pessoa que se sentiria confortável em concordar cegamente com a decisão de um sistema em situações de alto risco sem uma compreensão da lógica de tomada de decisão do sistema. Para tratar esses riscos, é necessário que uma IA forneça não apenas uma saída, mas também uma explicação compreensível por humanos que expresse a lógica da sua tomada de decisão.

"*A geração atual de sistemas de IA oferece enormes benefícios, mas sua eficácia será limitada pela incapacidade da máquina de explicar suas decisões e ações aos usuários*" (GUNNING, 2017).

Neste capítulo, temos como objetivo apresentar as diferentes definições propostas de explicabilidade, argumentar porque a explicabilidade é uma questão importante em IA e ML, assim como apresentar uma proposta de classificação geral das abordagens de explicabilidade que conduzirão os experimentos realizados posteriormente.

3.1 Terminologia e definições

Inteligência Artificial Explicável refere-se a área de pesquisa que desenvolve métodos e técnicas aplicados a algoritmos de inteligência artificial de forma que os resultados e predições destes sejam compreensíveis por humanos. Do ponto de vista de desenvolvedores e pesquisadores de aprendizado de máquina, as explicações fornecidas podem ajudá-los a

entender melhor o problema, os dados e por que um modelo pode falhar. No entanto, ainda não existe uma definição formal de explicabilidade. Portanto, antes de prosseguir com a apresentação dos métodos e técnicas de interpretabilidade, é conveniente estabelecer um ponto comum de entendimento sobre o que significa o termo explicabilidade no contexto de IA e, mais especificamente, ML.

Uma das questões que dificulta o estabelecimento de conceitos comuns é o uso intercambiável de interpretabilidade e explicabilidade na literatura. Não existe uma definição matemática de interpretabilidade. Biran e Cotton (2017) assumem que um sistema interpretável seria aquele cujas decisões são compreensíveis para nós humanos, seja por meio da inspeção do sistema ou por meio de alguma explicação produzida durante seu funcionamento. De forma similar, Miller (2017) define que "Interpretabilidade é o grau em que um ser humano pode entender a causa de uma decisão". Portanto, quanto maior a interpretabilidade de um modelo de aprendizado de máquina, mais fácil é alguém compreender por que certas decisões ou previsões foram feitas. Podemos assumir que interpretabilidade é uma propriedade do modelo de IA, podendo ele possuir alta ou baixa interpretabilidade. Já explicabilidade faz referência a técnicas externas ao modelo que buscam explicar modelos com baixo nível de interpretabilidade.

3.1.1 O que é uma explicação?

Segundo Miller (2017), "Uma explicação é a resposta a uma pergunta."

- Por que o tratamento não funcionou em um determinado paciente?
- Por que o empréstimo de um determinado cliente foi rejeitado?
- Por que um determinado cidadão foi classificado com alto risco de reincidência criminal?

3.1.2 O que é considerada uma boa explicação?

Os humanos geralmente não costumam questionar o porque de uma decisão por si só, mas sim buscam entender o porque uma decisão X foi tomada ao invés de uma decisão Y. Tendemos a pensar em casos contrafactuais e comparar diferentes cenários. Tomemos como exemplo um modelo de aprovação de empréstimos financeiros. Se um pedido de

empréstimo for rejeitado, o usuário, tipicamente, não se importa em conhecer todos os fatores que favoreceram a rejeição. O usuário estará interessado nas mudanças necessárias para que seu empréstimo seja aprovado no futuro. Explicações contrafactuais são mais fáceis de entender do que explicações completas. A melhor explicação é aquela que destaca a maior diferença entre o cenário atual e o cenário desejado.

3.2 Objetivos da explicabilidade

Como observado no capítulo 1, existem três objetivos principais que orientam o trabalho no campo de explicabilidade: auditar, validar e descobrir. Esses objetivos ajudam a motivar e focar a discussão, fornecendo uma tipologia para os tipos de explicações que provavelmente procuramos e valorizamos em cada contexto.

3.2.1 Auditar

Provavelmente a razão mais popular para explicar algoritmos seja o imenso impacto gerado pelo uso de Inteligência Artificial na sociedade. Centenas de aplicações surgem a cada dia, levantando um debate ético sobre discriminação algorítmica, falta de transparência e práticas de vigilância. O'Neil (2016) diz que vivemos em uma "sociedade algorítmica" que incorpora novas tecnologias em seu cotidiano sem ter uma visão crítica sobre elas, enxergando somente do ponto de vista da utilidade, sem pensar em questões de privacidade, segurança e como os conceitos de classificação, meritocracia e vigilância são automatizados e consolidados em caixas-pretas que ainda não se tem acesso.

Logo, vieses e preconceitos humanos presentes nos dados de treinamento de modelos algorítmicos passam despercebidos devido à falta de explicabilidade e à suposta "neutralidade e imparcialidade tecnológica". Estudos (RUBACK; AVILA; CANTERO, 2021; BIRHANE *et al.*, 2021) já discutiram como modelos de aprendizado de máquina são alimentados com as visões de quem os cria em todo o processo, desde o treinamento, parametrização até sua execução. Portanto, ao tentar parecer neutro, algoritmos potencializam ainda mais as desigualdades e os preconceitos.

> "É capciosa a noção de que tais tecnologias podem ser neutras quando, diariamente, são inseridos dados totalmente subjetivos nos algoritmos, depositando

> exatamente aquilo que nos diferencia das máquinas: a incapacidade de sermos moralmente neutros." (SCHNEIDER CAMILA B. E MIRANDA, 2020)

Estudos recentes já demonstraram que ferramentas de reconhecimento facial são frequentemente treinadas predominantemente com rostos brancos, tornando-as classificadores imprecisos para pessoas negras (BUOLAMWINI; GEBRU, 2018); o caso *COMPAS* (ANGWIN *et al.*, 2016) demonstrou como algoritmos de reincidência criminal têm reforçado preconceitos já existentes na sociedade, assim como a Amazon identificou que seu sistema de triagem de recrutamento privilegiava candidatos homens (IRIONDO, 2018). Esses são alguns de inúmeros casos que podemos listar para exemplificar problemas de vigilância e discriminação algorítmica.

A atual Regulação Geral de Proteção de Dados da União Europeia (GDPR) respalda o direito à explicação no contexto de decisões automatizadas (EUROPEAN COMMISSION, 2018). Da mesma forma, a Lei Geral de Proteção de Dados (LGPD) (Brasil, 2018): "prevê o direito à explicação no caso de decisões totalmente automatizadas que possam ter um impacto na vida do titular dos dados, principalmente no contexto de formação e uso de perfis comportamentais. A explicação deve incluir não somente informações sobre os dados pessoais que serviram de substrato para o algoritmo, mas também sobre a lógica por trás de tais decisões" (MONTEIRO, 2018).

Logo, não há dúvida de que os formuladores de políticas regulatórias devem considerar seriamente o impacto social da IA e discutir como regulamentar as indústrias que dependem de tal tecnologia (MONTEIRO, 2018). No entanto, qualquer tentativa de regulamentação exigirá a capacidade técnica de explicar e auditar algoritmos para testar rigorosamente se eles discriminam com base em atributos protegidos, como raça e gênero.

3.2.2 Validar e descobrir

Para além de preocupações éticas, existe um outro cenário em que pesquisadores de explicabilidade afirmam que suas ferramentas podem ajudar a debugar algoritmos que não estão funcionando conforme o esperado e descobrir novas propriedades de um sistema.

Um caso clássico que exemplifica esse cenário foi apresentado no artigo de Ribeiro, Singh e Guestrin (2016a) que propuseram uma das técnicas mais populares de explicabilidade atualmente: LIME. Os autores realizaram um experimento no qual criaram um

modelo de ML que diferenciava huskies de lobos e possuía 100% de acurácia. O modelo foi treinado propositalmente para a diferenciar um lobo de um husky não pelas características dos dois animais, como era esperado, mas pela presença ou ausência de neve nas imagens. Os resultados de predição do modelo foram apresentados para especialistas humanos a fim de verificar se eles conseguiriam identificar essa "falha". Antes de observar as explicações, mais de um terço dos especialistas confiava no classificador e um pouco menos da metade mencionou o padrão de neve como algo que a rede neural estava usando. No entanto, depois de examinar as explicações (Figura 3) quase todos os especialistas identificaram com muito mais certeza de que a neve era um fator determinante para o modelo. Além disso, a confiança no modelo caiu significativamente. Esse experimento demonstra a utilidade de explicar predições individuais para obter *insights* sobre classificadores, saber se podemos ou não confiar neles e corrigir problemas como o de *overfitting*.

Figura 3 – Imagem de treinamento e explicação gerada na tarefa "Husky vs Lobo" (RIBEIRO; SINGH; GUESTRIN, 2016a)

(a) Husky classified as wolf (b) Explanation

Os avanços na área de explicabilidade permite que analisemos o comportamento interno de um modelo em predições específicas. Esse é o objetivo, por exemplo, de técnicas de aproximação linear local, incluindo algoritmos populares como LIME (RIBEIRO; SINGH; GUESTRIN, 2016a) e SHAP (MESSALAS; KANELLOPOULOS; MAKRIS, 2019) que atribuem pesos às variáveis de entrada para que os usuários possam verificar se

o modelo não deu maior importância indevidamente a variáveis pouco informativas, como a neve informada anteriormente.

3.3 Categorização de métodos

Métodos de explicabilidade podem ser categorizados e organizados de acordo com diferentes critérios. Nesta dissertação, propomos os seguintes critérios:

1. **Estágio:** refere-se ao estágio em que um método gera explicações.
2. **Agnosticidade:** o método pode ser independente do modelo ou específico do modelo.
3. **Escopo:** trata-se do escopo da explicação gerada pelo método (global ou local).

A vantagem da divisão proposta é que ela destaca as características de diferentes abordagens e ajuda a encontrar o método de interpretabilidade mais adequado para uma determinada tarefa.

3.3.1 Modelos transparentes x caixa-preta

Embora os primeiros sistemas de IA fossem facilmente interpretáveis, os últimos anos testemunharam o surgimento de sistemas de decisão caixa-preta, como Redes Neurais Profundas (DNNs). Já modelos de regressão linear ou árvores de decisão oferecem maior interpretabilidade, mas têm desempenho limitado em dados de grandes dimensões, enquanto um modelo *Random forest* ou DNN poderá ter um desempenho muito melhor, mas será menos compreensível. De acordo com Mohseni, Zarei e Ragan (2020), a interpretabilidade de um modelo de aprendizado de máquina é inversamente proporcional ao seu tamanho e complexidade.

3.3.2 Estágio: Explicabilidade intrínseca x post-hoc

Os métodos de interpretabilidade podem ser agrupados em duas categorias: interpretabilidade intrínseca e interpretabilidade *post-hoc*. Essa classificação diferencia se a interpretabilidade é alcançada restringindo a complexidade do modelo de aprendizado de máquina (intrínseco) ou aplicando métodos que analisam o modelo após o treinamento (post-hoc).

Interpretabilidade intrínseca trata sobre o uso de modelos de aprendizado de máquina inerentemente explicáveis como: árvores de decisão, modelos lineares etc. Em contraste, interpretabilidade *post-hoc* engloba técnicas que criam um segundo modelo que deverá prover explicações sobre o comportamento e decisões do modelo caixa-preta original.

A principal diferença entre esses dois grupos reside no *trade-off* entre a acurácia do modelo e a fidelidade da explicação. Modelos intrinsecamente interpretáveis podem fornecer explicações, mas podem sacrificar a acurácia. Já o uso de técnicas *post-hoc* permite manter a acurácia do modelo original, mas as explicações geradas podem não ter alta fidelidade ou confiabilidade. Isto é, as explicações geradas podem não se aproximar da predição realizada pelo modelo.

3.3.3 Agnosticidade: abordagens agnósticas x específicas

Conforme Ribeiro, Singh e Guestrin (2016a), abordagens agnósticas ao modelo permitem explicar predições, independentemente da implementação utilizada no modelo. Essa abordagem trata os modelos como caixas pretas, de forma que explicações podem ser geradas mesmo sem acesso aos parâmetros internos do modelo. Em contraste, existem as explicações específicas ao modelo: abordagens projetadas exclusivamente para um determinado modelo ou classe de modelos de aprendizado de máquina (DU; LIU; HU, 2019).

3.3.4 Escopo: Explicações locais x globais

Com base nas categorizações anteriores, podemos ainda diferenciar as técnicas de explicabilidade em dois tipos: explicabilidade global e explicabilidade local. No primeiro caso, o objetivo é tornar todo o processo de decisão de um modelo explicável e compreensível. No último caso, o objetivo é explicar explicitamente um determinado resultado gerado pelo modelo (MOHSENI; ZAREI; RAGAN, 2020).

3.4 Técnicas de explicabilidade post-hoc e agnósticas

O escopo dessa dissertação está em técnicas de explicabilidade post-hoc e agnósticas. Essa seção dedica-se a explorar em maior profundidade as técnicas existentes dentro dessa categoria. A partir da revisão sistemática da literatura realizada no âmbito desta dissertação (), propõe-se agrupar as técnicas existentes em três grupos distintos: (i) explicações baseadas em regras, (ii) explicações baseadas na importância dos atributos e (iii) explicações baseadas em exemplos contrafactuais. As seções subsequentes foram dedicadas ao detalhamento de cada uma dessas estratégias.

3.4.1 Explicações baseadas em regras

Alguns métodos de interpretabilidade agnósticos ao modelo produzem explicações baseadas em regras ou conjuntos de decisões, explorando diferentes técnicas de extração de regras (JOHANSSON; KöNIG; NIKLASSON, 2010; LAKKARAJU *et al.*, 2019; PLUMB; MOLITOR; TALWALKAR, 2018; EVANS; XUE; ZHANG, 2019).

O método *Genetic Rule EXtraction* (G-REX) (JOHANSSON; KöNIG; NIKLASSON, 2010) utiliza algoritmos genéticos para gerar regras no formato *IF-THEN* com operadores *AND / OR*. O algoritmo apresentado parte de um conjunto vazio de regras e adiciona, a cada iteração, uma regra para cada predicado. Este método identifica as regras candidatas com a maior precisão estimada em um conjunto de dados, em que a precisão representa a proporção de predições corretas.

MUSE (LAKKARAJU *et al.*, 2019) é um método que também cria conjuntos de regras no formato *IF-THEN*. O MUSE é baseado em uma função objetivo que otimiza, simultaneamente, a acurácia e a interpretabilidade, aprendendo conjuntos de decisões curtos que capturam o comportamento de um determinado modelo caixa-preta e cobrem todo o espaço dos dados de entrada, considerando também as classes em menor proporção.

MAPLE *(Model Agnostic SuPervised Local Explanations)* (PLUMB; MOLITOR; TALWALKAR, 2018) combina a ideia de usar *random forests* como método de seleção dos vizinhos mais próximos para criação do modelo de interpretabilidade local, introduzido por Bloniarz *et al.* (2016) como SILO, com o método de seleção de características proposto por Kazemitabar *et al.* (2017) como DStump. Este método gera explicações locais fornecidas

pela abordagem de vizinhança supervisionada e explicação global através de extração de regras.

Apesar de Árvores de Decisão serem amplamente utilizadas em métodos de interpretabilidade, geralmente, algoritmos de indução de árvores de decisão utilizam uma abordagem gulosa, *top-down* e recursiva para a construção das árvores. No entanto, essa abordagem pode degradar de acordo com a qualidade dos dados de entrada; e estratégias gulosas geralmente produzem soluções ótimas locais, mas não globais.

Por fim, Evans, Xue e Zhang (2019) propõem um método de interpretabilidade global que utiliza programação genética na construção de árvores de decisão que serão representações das predições do modelo caixa-preta. Segundo os autores, esse método reduz a complexidade de toda a estrutura da árvore e das expressões resultantes, favorecendo a compreensibilidade.

3.4.2 Explicações baseadas na importância dos atributos

Em sua maioria, os métodos que produzem explicações numéricas utilizam a abordagem de estimar a importância de cada atributo para a predição final do modelo. Dessa forma, é possível descobrir quais atributos estão causando o maior impacto na tomada de decisão do modelo, seja o impacto negativo ou positivo. Contudo, uma desvantagem é que, se o conjunto de dados apresentar atributos correlacionados, a abordagem pode trazer resultados enviesados. Esta estratégia baseada na importância dos atributos foi aplicada por Messalas, Kanellopoulos e Makris (2019), Ribeiro, Singh e Guestrin (2016b), Ibrahim *et al.* (2019), Slack *et al.* (2020a).

SHAP, proposto por Messalas, Kanellopoulos e Makris (2019), é um método de explicabilidade que utiliza conceitos de teoria dos jogos e valor de Shapley (LUNDBERG; LEE, 2017) para calcular a importância dos atributos na predição do modelo. No contexto de teoria dos jogos, o resultado de uma predição é o jogo, os atributos do modelo são os jogadores e a importância de cada atributo é a recompensa distribuído de forma justa. SHAP possui diversas variações sendo uma delas agnóstica ao modelo (KernelShap) e outras específicas ao modelo (TreeShap, DeepShap e LinearSHAP) a fim de uma melhor aproximação local que explique o modelo caixa-preta. O Kernel SHAP é um método

agnóstico ao modelo; o Tree SHAP é específico para modelos baseados em árvore; o Deep SHAP para modelos de Deep Learning e o Linear SHAP para modelos lineares.

Já o estudo de Ribeiro, Singh e Guestrin (2016b) apresenta a técnica LIME: uma técnica de interpretação local, que utiliza a importância dos atributos do modelo e gera um modelo interpretável que garanta fidelidade local. Para gerar as instâncias de treinamento do modelo local, LIME seleciona aleatoriamente um subconjunto dos atributos de uma determinada instância. Em seguida, as instâncias são perturbadas e utilizadas como entrada no modelo caixa-preta para obter as predições/rótulos, que são então usadas para treinar um modelo linear interpretável. O objetivo é entender como o comportamento do modelo é afetado pelas perturbações nos dados. Por fim, o modelo caixa-preta pode ser explicado através dos pesos dos atributos do modelo interpretável criado, que não necessariamente precisa funcionar globalmente, mas deve se aproximar das predições do caixa-preta localmente para uma única instância. Existe também uma extensão da técnica LIME, utilizada para explicação global, chamada *Sub-modular Pick* (SP-LIME). SP-LIME seleciona, a partir do conjunto de dados de entrada, um conjunto representativo de explicações de diferentes instâncias que poderiam ser uma representação global de como o modelo toma decisões.

Ibrahim *et al.* (2019) propõem o Global Attribution Method (GAM) capaz de explicar as predições de redes neurais em subpopulações. Essa abordagem também endereça uma das limitações da técnica LIME ao lidar com a complexidade de modelos não lineares. O método GAM utiliza uma adaptação do algoritmo de *k-means* de clusterização para agrupar explicações locais semelhantes e escolher explicações representativas de cada subgrupo.

Outros métodos encontrados durante a revisão dedicam-se a garantir "justiça" (*fairness*) em modelos de aprendizado de máquina. Os autores Slack *et al.* (2020a) apresentam dois algoritmos: *Fairness Warnings* e *Fair-MAML*. Similar à técnica LIME, *Fairness Warnings* é um modelo agnóstico interpretável que utiliza perturbação de dados para identificar se um modelo está se comportando de maneira "injusta" para tarefas semelhantes em um determinado domínio. A abordagem utiliza *Supersparse Linear Integer Model* (SLIM), proposto por Ustun e Rudin (2015), como o modelo interpretável. Já *Fair-MAML* realiza o treinamento de modelos justos utilizando um pequeno conjunto de dados através de meta-aprendizado. O objetivo da técnica de meta-aprendizado é treinar modelos de forma

que eles possam ser treinados em novas tarefas usando um pequeno conjunto de dados e um treinamento reduzido.

3.4.3 Explicações baseadas em exemplos contrafactuais

Destaca-se que, nos estudos publicados durante o ano de 2020, passou-se a explorar técnicas de interpretabilidade baseadas em exemplos contrafactuais (SHARMA; HENDERSON; GHOSH, 2020; MOTHILAL; SHARMA; TAN, 2020). Esse conceito foi primeiramente introduzido por Wachter, Mittelstadt e Russell (2018). Os autores afirmam que explicações contrafactuais são uma forma de explicar os resultados do modelo aos usuários de forma que eles possam entender por que uma determinada decisão foi tomada e identificar o que precisaria ser alterado para receber um resultado desejado no futuro, com base no modelo de tomada de decisão atual.

Como exemplo, pode-se pensar em um usuário que tem sua solicitação de empréstimo rejeitada pelo modelo de crédito utilizado no banco. O usuário irá se perguntar por que seu pedido foi rejeitado e o que ele poderia fazer para melhorar suas chances de obter um empréstimo no futuro. A questão do "por que" pode ser formulada de forma contrafactual: qual é a menor alteração nos atributos (renda, pontuação de crédito, idade etc.) que alterariam o resultado da solicitação, i.e., a predição do modelo, de rejeitada para aprovada?

Sharma, Henderson e Ghosh (2020) introduzem o método *Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models* (CERTIFAI) que gera explicações contrafactuais utilizando um algoritmo evolutivo genético. O algoritmo inicia gerando um conjunto aleatório de dados de forma que tenham diferentes previsões do conjunto de entrada. Em sequência, um processo evolutivo é aplicado a esse conjunto e resulta em um novo conjunto de dados - mantendo os valores das predições iniciais.

Ressalta-se, ainda, que o estudo feito por Sharma, Henderson e Ghosh (2020), apresentou a possibilidade de o usuário do método criar restrições e, portanto, personalizar seu funcionamento de acordo com o domínio do problema:

1. **Muting features:** define quais atributos podem ter seus valores alterados;
2. **Feature range:** permite definir intervalos de valores para alteração dos atributos;

3. **Number of explanations:** o método pode gerar múltiplas explicações contrafactuais, então os usuários podem especificar a quantidade de explicações a serem geradas.

Usando um exemplo de decisão de empréstimo, uma explicação gerada pelo método pode sugerir "reduzir o aluguel da casa", porém pode ser que essa sugestão não seja factível para os usuários finais da empresa que utiliza aquele sistema. Idealmente, as explicações devem apresentar uma ampla gama de mudanças sugeridas (diversidade), assim como seguir o contexto da sociedade. Dessa forma, as restrições que o método CERTIFAI produz destacam-se por possibilitar explicações que serão transformadas em soluções viáveis no mundo real.

Mothilal, Sharma e Tan (2020) apresentam um método, DiCE, que gera conjuntos diversos de explicações contrafactuais para qualquer modelo de aprendizado de máquina. A ideia central é transformar a busca por explicações locais em um problema de otimização, semelhante a encontrar exemplos adversários. Similarmente aos métodos anteriores, DiCE utiliza perturbação nos valores dos atributos para verificar as fronteiras de decisão do modelo e fornece explicações no formato "what-if". Além disso, DiCE também suporta o uso de restrições para garantir viabilidades das explicações geradas.

Lucic, Haned e Rijke (2020) apresentam uma solução de interpretabilidade local para *tree ensembles*: *Monte Carlo Bounds for Reasonable Predictions* (MC-BRP). Apesar de árvores de decisão serem consideradas interpretáveis, a combinação de múltiplas árvores culmina na perda de transparência do modelo e necessidade de criação de técnicas de interpretabilidade. O método apresentado difere dos outros descritos nesta seção, pois não tem como objetivo identificar uma explicação contrafactual, mas sim um intervalo de valores de atributos para os quais a predição final de uma determina instância seria diferente.

Mediante o exposto, observa-se que cada método apresentado aborda diferentes aspectos e necessidades de interpretabilidade. Acredita-se que não existe uma única abordagem adequada para cada cenário, pois utilizar apenas um método dará entendimento parcial do funcionamento do modelo. Dessa forma, combinar diferentes abordagens pode fornecer maior confiabilidade no modelo.

3.5 Avaliação de métodos explicabilidade

A crescente quantidade de publicações e proposta de métodos de interpretabilidade levou os autores a também pensarem em formas de avaliação das técnicas. Diferentes métricas de avaliação foram propostas e encontradas na literatura, bem como diferentes tipos de avaliação foram realizados. Em sua maioria, as métricas devem avaliar o quão satisfatórias foram as explicações, o impacto das explicações no desempenho do modelo e também na confiança e segurança de seus usuários.

3.5.1 Avaliações quantitativas

Embora inicialmente considerados para métodos de extração de regras, Craven e Shavlik (1995) propuseram as seguintes dimensões para avaliar explicabilidade:

Fidelidade

Fidelidade (RIBEIRO; SINGH; GUESTRIN, 2016a; MESSALAS; KANELLOPOULOS; MAKRIS, 2019) se refere a quão bem as explicações geradas se aproximam da predição do modelo caixa-preta. É medida em termos de acurácia, F1-score e assim por diante, mas com relação ao resultado da caixa-preta. Dado um conjunto de dados D = X podemos aplicar a cada registro x ∈ X ambos os modelos: (i) para a caixa-preta b obtemos o conjunto de previsões Y, e (ii) para o preditor interpretável i obtemos o conjunto de predições Z. Assim, a pontuação de fidelidade pode ser calculada aplicando o mesmo cálculo da função de precisão onde os valores alvo são as predições Y da caixa-preta b contra os valores preditos Z. Acurácia e fidelidade estão relacionadas de forma que se o modelo de caixa-preta tiver alta acurácia e as explicações geradas tiverem alta fidelidade, essas explicações também terão alta acurácia.

No entanto, Messalas, Kanellopoulos e Makris (2019) argumentam que uma fidelidade alta não implica necessariamente que o processo de decisão dos dois modelos (do modelo caixa-preta e do modelo explicável gerado) seja o mesmo. Por processo de decisão, os autores referem-se à importância individual dos atributos para cada modelo. Diante disso, os autores introduziram uma nova métrica (*"Top Similarity"*) que calcula a

fidelidade interna entre o modelo caixa-preta e o modelo explicável. O cálculo da métrica é realizado da seguinte forma: os valores SHAP absolutos dos atributos de cada modelo (do caixa-preta e do explicável) são ordenados e, para cada instância, escolhe-se um conjunto de j atributos mais importantes. A partir disso, calcula-se a intersecção das decisões de ambos os modelos: $commom_i = ORIG_j(i) \cap SUR_j(i), \forall i \in N$, em que N é a quantidade de instâncias. Por fim, a fórmula da métrica proposta é:

$$Top_jSimilarity = \frac{avg(commom)}{j} \tag{1}$$

Dessa forma, uma $TopSimilarity = 80\%$ significa que os dois modelos concordam quanto aos atributos de maior importância para 80% das instâncias.

Compreensibilidade

Trata-se de avaliar o quanto as representações extraídas são humanamente compreensíveis. Na prática, a compreensibilidade é normalmente considerada uma propriedade binária; um determinado modelo é compreensível ou não, dada uma situação específica. De modo geral, porém, a escolha de avaliar a compreensibilidade usando o tamanho (complexidade) do modelo extraído é a mais aceita. Ribeiro, Singh e Guestrin (2016b) simularam a confiança do usuário nas representações extraídas pela técnica LIME definindo propositalmente explicações e modelos "não confiáveis". Eles testaram uma hipótese de como usuários reais prefeririam explicações mais confiáveis e escolheriam modelos melhores. Outro exemplo é o trabalho de Messalas, Kanellopoulos e Makris (2019), que realizou a comparação entre os métodos SHAP, LIME e DeepLIFT - com a suposição de que boas explicações do modelo devem ser consistentes com as explicações de humanos que conhecem o modelo.

Robustez

Garantir que entradas semelhantes tenham explicações semelhantes - e, portanto, confiabilidade. Sharma, Henderson e Ghosh (2020) utilizam o método contrafactual onde uma explicação contrafactual é um ponto gerado próximo a uma entrada que altera a previsão e pode, portanto, ser considerado um exemplo contraditório. Usando essa

noção de contrafactuais como exemplos adversários, os autores definem o *Counterfactual Explanation-based Robustness Score (CERScore)*, que é a distância esperada entre a instância de entrada e seu contrafactual correspondente. Modelos robustos resistem a exemplos adversários tentando alcançar robustez local em tantos pontos quanto possível.

Lakkaraju *et al.* (2019) argumentam que para descrever o comportamento de um determinado modelo, é importante construir uma explicação que seja não apenas fiel ao modelo original, mas também não-ambígua e interpretável. Os autores definem novas métricas da seguinte forma:

Divergência

Representa o número de instâncias para as quais a classe atribuída pela técnica de interpretabilidade não corresponde a classe original das predições do modelo caixa-preta. Nota-se que podemos considerá-la o complemento da métrica de fidelidade.

Não-ambiguidade

Uma explicação não-ambígua deve fornecer motivos específicos para descrever como o modelo se comporta em diferentes partes do conjunto de dados. A métrica pode ser calculada usando as sobreposições entre as regras de decisão extraídas.

3.5.2 Avaliações qualitativas e testes com usuário

Explicações somente são eficazes quando ajudam os usuários finais a construir uma representação mental e correta do processo de decisão de um determinado modelo. Alguns estudos exploraram maneiras pelas quais os usuários finais podem contribuir para minimizar erros de classificação.

Outro exemplo é o trabalho de Messalas, Kanellopoulos e Makris (2019), que realizou a comparação entre os métodos SHAP, LIME e DeepLIFT - com a suposição de que boas explicações do modelo devem ser consistentes com as explicações de humanos que conhecem o modelo.

Mothilal, Sharma e Tan (2020) partiram seus estudos da hipótese de que explicações contrafactuais eficazes devem satisfazer duas propriedades: viabilidade das ações contrafactuais, dado o contexto e as restrições do usuário, e diversidade de exemplos apresentados. Os autores desenvolveram um *framework* de interpretabilidade que utiliza interação com usuário para criar explicações.

De forma similar, Lakkaraju *et al.* (2019) apresentam uma forma interativa de geração de explicações contrafactuais.

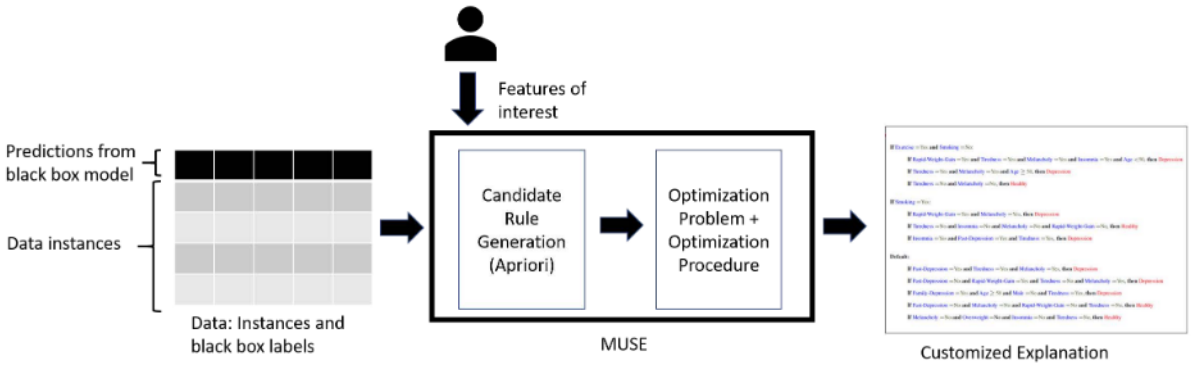


Figura 4 – Fluxo da abordagem MUSE (LAKKARAJU *et al.*, 2019)

A Figura 4 apresenta de forma visual a abordagem MUSE que: (a) desenha uma representação que permite incorporar sugestões do usuário; (b) quantifica as métricas de fidelidade, desambiguidade e interpretabilidade; (c) formula um problema de otimização do *trade-off* entre interpretabilidade e acurácia e o resolve de forma eficiente; e, por último, (d) customiza explicações extraídas de acordo com o contexto e preferências do usuário (atributos do conjunto de dados).

Em boa parte dos estudos revisados, foi mencionado que um dos problemas relacionados à interpretabilidade é calcular a compreensibilidade da técnica proposta. Diversos estudos apontaram na seção de trabalhos futuros a necessidade de comparar as explicações extraídas com outras técnicas e realizar testes com usuários reais para verificar a compreensibilidade da proposta.

3.6 Desafios explicabilidade

Apesar do progresso recente no aprendizado de máquina interpretável, ainda existem alguns desafios relacionados às técnicas e avaliação dos métodos. Nas seções anteriores,

cobrimos e antecipamos alguns desses desafios. Nessa seção, listamos uma série de desafios previamente identificados em um artigo nosso publicado em 2022 (VIEIRA; DIGIAMPIETRI, 2022).

3.6.1 Ausência de métodos de explicabilidade global

Os estudos da área, em sua imensa maioria, focam nas explicações locais que podem não ser suficientes para assegurar confiança no modelo antes de colocá-lo em produção. Ribeiro (2018) ressalta que apesar de interpretabilidade global ser difícil de ser alcançada na prática, existe uma oportunidade inexplorada em apresentar explicações de natureza global. Neste sentido, torna-se interessante o estudo e desenvolvimento de técnicas que expliquem o modelo de forma global a partir da combinação ou melhoria das técnicas existentes. Além disso, as técnicas de interpretabilidade global são limitadas porque atribuem a mesma explicação a múltiplas observações (MESSALAS; KANELLOPOULOS; MAKRIS, 2019) e utilizam modelos lineares que podem não ser suficientes para explicar o modelo como um todo (RIBEIRO; SINGH; GUESTRIN, 2016b).

3.6.2 Como evitar *ground-truth unjustification*?

As abordagens de interpretabilidade post-hoc são populares por uma série de fatores como: (i) escalabilidade: mudanças na arquitetura do modelo não implicam uma mudança na construção do método; (ii) flexibilidade de modelos: o método pode funcionar com diferentes modelos de aprendizado de máquina; e (iii) flexibilidade de representações: o método deve ser capaz de gerar uma representação diferente conforme o modelo caixa-preta que está sendo explicado. No entanto, existe alguns riscos associados: (i) o uso de modelos substitutos lineares pode reduzir a qualidade das explicações fornecidas; (ii) *ground-truth unjustification* (LAUGEL *et al.*, 2019): como métodos agnósticos *post-hoc* não têm acesso aos dados de treinamento, existe um risco de gerar explicações que são resultado de padrões aprendidos pelo modelo em vez do conhecimento real dos dados de entrada.

3.6.3 Como podemos melhor avaliar as explicações?

Avaliar explicações é, talvez, o aspecto mais imaturo da pesquisa sobre IA explicável (RIBEIRO, 2018). No que se trata de avaliar os métodos de interpretabilidade, nota-se que diversas métricas foram propostas e apresentadas, mas ainda não há consenso do que seria explicabilidade e como medi-la. Dessa forma, avaliar os métodos e garantir certos aspectos (estabilidade, compreensibilidade, consistência, similaridade, robustez etc.) ainda é um desafio e uma questão em aberto. Como podemos dizer que um método de explicabilidade é melhor do que outro se não sabemos por quê?

3.6.4 Explicável para quem? Podemos construir explicações melhores?

Com relação às representações geradas pelos métodos, foi observado que a maioria dos estudos utiliza representações numéricas e gráficas, sendo estas complexas para usuários não especialistas. Parte do desafio é que o campo de pesquisa de explicabilidade é dominado por pesquisadores de aprendizado de máquina. Mothilal, Sharma e Tan (2020) destacam a importância de envolver o usuário final na construção de um modelo poderoso, interativo e eficiente. Lakkaraju *et al.* (2019) apresentaram uma forma de geração interativa de explicações contrafactuais. Ambos os estudos (MOTHILAL; SHARMA; TAN, 2020; LAKKARAJU *et al.*, 2019) apresentaram técnicas que permitem que os modelos sejam controlados de acordo com as preferências de seus usuários. Os autores defendem que os usuários dos sistemas de IA devem fazer parte do processo de construção de interpretabilidade desde o início e diferentes usuários precisam de diferentes tipos de explicações.

3.6.5 Como garantir robustez?

Conforme apresentado, um grande grupo de técnicas de explicabilidade baseia-se na importância de atributos. Mas esse grupo de técnicas tem caído em desuso devido às limitações do uso de perturbação de dados como falta de robustez e consistência das explicações geradas, assim como sua maior sensibilidade a ataques adversários (SLACK *et al.*, 2020b).

As técnicas de explicações contrafactuais propõem maneiras interessantes de entender a importância dos atributos, no entanto, também sofrem com falta de robustez relacionada à perturbação de dados. O desempenho das técnicas depende bastante da função de otimização utilizada para localizar fronteiras de decisão. Apenas um grupo pequeno de estudos analisados apresenta métodos que buscam gerar diferentes explicações contrafactuais conforme alguma métrica de diversidade. Diante dessa tendência recente, acredita-se que há um caminho a seguir na construção de explicações diversas, robustas e factíveis para usuários finais.

Alvarez-Melis e Jaakkola (2018) ressaltam que a robustez das explicações é uma característica necessária nos métodos de interpretabilidade e, em seu artigo, introduziram métricas para quantificar a robustez e apresentar maneiras de incorporar o conceito às abordagens de interpretabilidade existentes.

3.6.6 Como *fairness* interage com a interpretabilidade?

O tema justiça (*"fairness"*) tem recebido atenção significativa na literatura de explicabilidade. No entanto, esta área de pesquisa vem com o desafio de identificar como as diferentes medidas de *"fairness"* se relacionam entre si, bem como em que medida são compatíveis ou mutuamente exclusivas.

3.6.7 Como combinar diferentes modelos de explicabilidade e colocá-los em produção?

Se examinarmos com atenção as abordagens existentes, descobriremos que, embora haja alguma sobreposição entre os vários tipos de explicação, na maioria das vezes, elas parecem ser segmentadas, cada uma abordando uma questão diferente. Neste ponto, gostaríamos de destacar que não existe uma forma estabelecida de combinar técnicas, portanto, há espaço para experimentá-las e ajustá-las, de acordo com a aplicação em questão. Essa direção pode não apenas ajudar a preencher a lacuna entre os modelos caixa-preta e transparentes, mas também pode ajudar no desenvolvimento de modelos explicáveis de maior desempenho e confiabilidade. Acredita-se que uma oportunidade de trabalho seria combinar diferentes estratégias de explicabilidade local da literatura para extrair explicações globais dos modelos.

4 Proposta

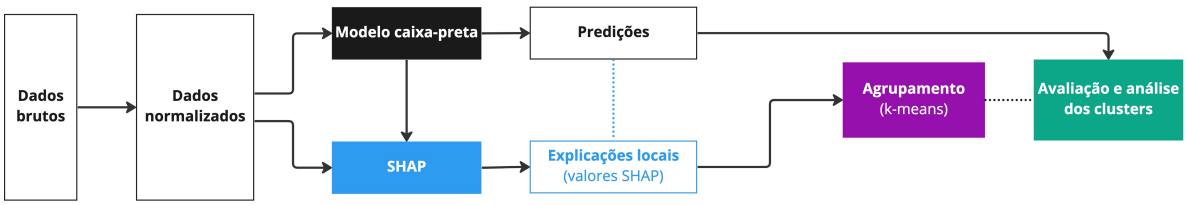
Este capítulo contém a descrição do arcabouço proposto para extrair explicações globais de um modelo caixa-preta por meio do agrupamento de explicações locais obtidas com o SHAP, assim como detalhes das métricas utilizadas e método de agrupamento. Uma das dificuldades de análise de viés em modelos caixa-preta é que métodos como SHAP geram explicações locais por amostra de dados e medidas de *fairness* buscam capturar o comportamento global do modelo. Portanto, explicações locais podem não ser suficientes para garantir *fairness* antes do modelo ser colocado em produção. O propósito do agrupamento das explicações é evitar gerar apenas uma explicação global que generalize o comportamento do modelo.

4.1 Arcabouço

Nessa dissertação, o foco foi alcançar explicabilidade global a partir de métodos locais de explicabilidade. As duas técnicas mais conhecidas de explicabilidade são LIME (*Locally Interpretable Model-Agnostic Explanations*) e SHAP (*Shapley Additive Explanations*). Foi escolhido utilizar o SHAP já que ele garante maior robustez.

A Figura 5 resume o arcabouço implementado. Inicialmente, as variáveis categóricas são transformadas em binárias; os dados são divididos em conjuntos de treinamento (80%) e teste (20%) e é realizada uma normalização dos atributos. Após a preparação dos dados, é feito o treinamento de um modelo (*Gradient Boosting* e Regressão Logística) com os dados normalizados e extração da técnica de explicabilidade SHAP. A etapa final consiste no uso do algoritmo *k-means* para realizar o agrupamento das explicações locais extraídas de acordo com três possíveis tratamentos.

Figura 5 – Diagrama da proposta sugerida



4.1.1 Agrupamento

Com relação à técnica de agrupamento, foi utilizado o método k-means tendo como objetivo analisar o comportamento do modelo em diferentes grupos. A quantidade de grupos é definida de acordo com a *silhouette score* e o uso do método do cotovelo. Para realizar o agrupamento, existem diferentes medidas de distância que podem ser utilizadas, sendo as mais comuns as distâncias euclidianas e Manhattan. Na presenta dissertação, foram analisadas três propostas para agrupamento de valores SHAP, descritas a seguir.

Proposta 1: agrupar valores SHAP usando distância euclidiana

A vantagem de usar valores SHAP para agrupamento é que os valores SHAP para todos os atributos estão na mesma escala, tendo maior potencial de gerar grupos significativos.

Proposta 2: agrupar rank de valores SHAP usando distância euclidiana

Ibrahim *et al.* (2019) propõem a distância (*Rank Distance*). Esta distância aplicada aos valores SHAP considera apenas a importância dos atributos (módulo do valor SHAP) independente se ele contribui positivamente ou negativamente para a predição. Com base nessa proposta, uma abordagem simplificada foi implementada: os valores SHAP do atributo são substituídos pela posição daquele atributo no ranking de características (*features*) da amostra.

Proposta 3: agrupar menor e maior valor SHAP usando similaridade cosseno

Uma proposta de maior simplificação foi agrupar considerando apenas os atributos que mais contribuem positivamente (*max shap value*) e negativamente (*min shap value*). Para cada amostra, os atributos foram ordenados de forma decrescente para obter o atributo de maior impacto positivo e ordem crescente para obter o atributo de maior impacto negativo. O uso da similaridade cosseno está relacionado a alta esparsidade da normalização dos dados gerados para essa proposta.

4.1.2 Avaliação dos resultados

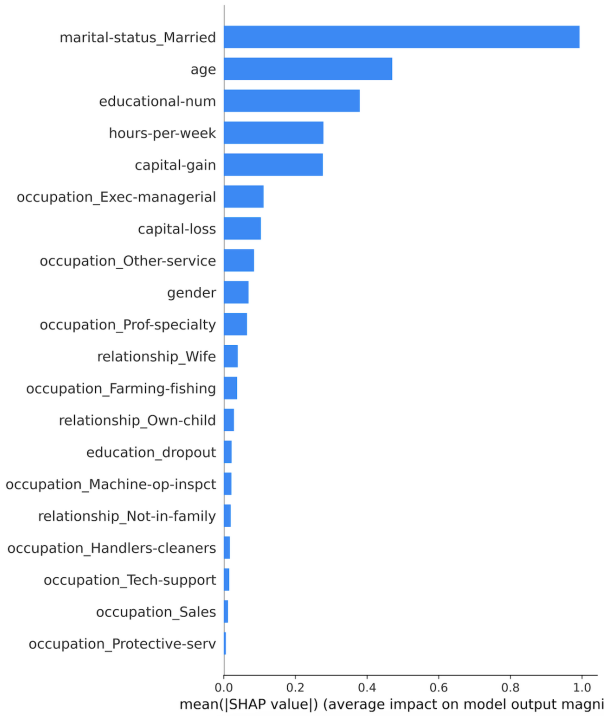
Embora tenha havido pesquisas significativas na validação de algoritmos de agrupamento, a interpretação de agrupamentos envolve análise qualitativa e conhecimento de domínio. Nesta dissertação, validamos a metodologia das seguintes maneiras:

Interpretabilidade global com SHAP

Além de propor diferentes abordagens para explicar localmente o resultado do modelo, o método SHAP também propões metodologias de agregação dos resultados a fim de extrair uma explicação global. Iremos utilizar a mesma equação para avaliar a importância global dos atributos, mas aplicada de uma maneira diferente. Primeiramente, a equação será aplicada para identificar a posição do atributo sensível no ranking. Para obter esse ranking, será calculada a importância global de cada atributo segundo a equação e ordenando os valores obtidos em ordem decrescente.

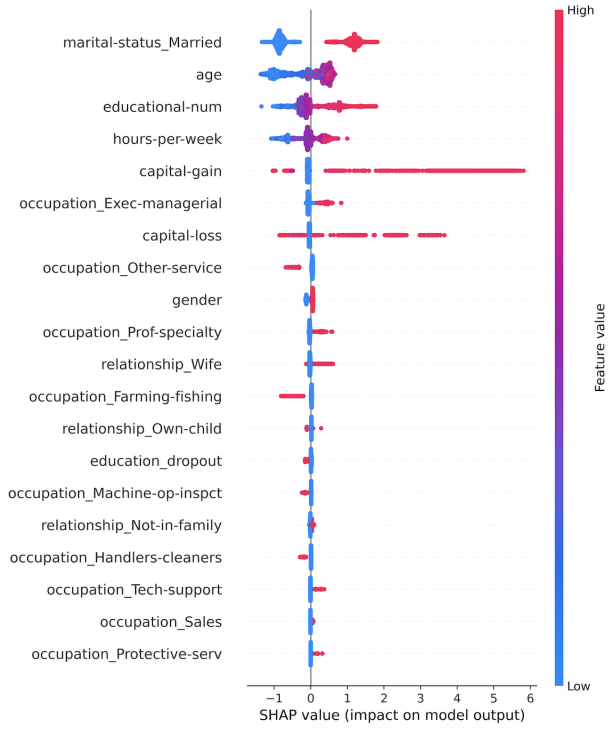
Uma forma de visualizar essa proposta é o gráfico de barras, mostrado na Figura 6 que apresenta um ranking da importância do atributo para a predição final utilizando-se o módulo do valor SHAP. A lista representa os atributos mais importantes independente se eles impactam positivamente ou negativamente na predição.

Figura 6 – Exemplo do SHAP Bar plot



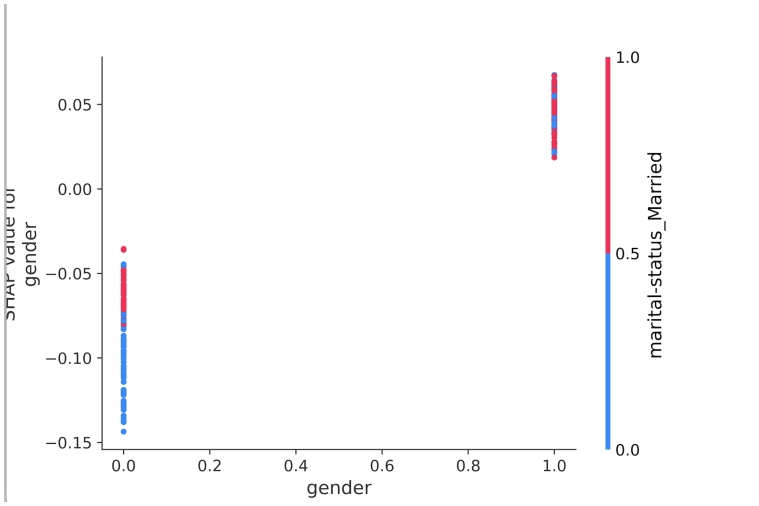
Uma outra forma de visualização é o *Summary plot* (Figura 7), em que os valores SHAP obtidos são representados por círculos e são ordenados e plotados horizontalmente. Cada valor SHAP é colorido de acordo com o valor do atributo, sendo associada a cor azul aos valores menores e vermelho aos maiores.

Figura 7 – Exemplo do SHAP Summary plot



Outro gráfico interessante do SHAP é o *Dependence Plot* (Figura 8), que permite entender com mais detalhes como um atributo específico afeta o resultado do modelo. Ele permite avaliar a relação entre os valores do atributo (eixo x) e os valores do SHAP (eixo y). Os pontos nos gráficos são coloridos de acordo com os valores do atributo correlacionado seguindo a escala de cores representada no lado esquerdo do gráfico. Avaliando esse gráfico de exemplo, podemos notar uma relação importante entre os atributos *gender* e *marital-status Married*. O gráfico nos diz que homens casados, *gender* = 1 e cor rosa, possuem valores SHAP maiores do que mulheres solteiras, *gender* = 0 e cor azul.

Figura 8 – Exemplo do SHAP Dependence plot



Na presente dissertação, os três gráficos foram utilizados para análises comparativas entre as explicações geradas para o modelo caixa-preta e as explicações globais de cada grupo. O objetivo é verificar se existem explicações locais importantes ou vieses que o cálculo global pode estar mascarando.

Análise dos grupos

Para análise dos grupos, foram utilizados os rankings de *features* de cada grupo, assim como informações demográficas. Nesse caso, buscamos avaliar os grupos para identificar possíveis semelhanças e diferenças do comportamento do modelo com relação ao atributo protegido.

5 Experimentos, Dados e Resultados

5.1 Experimentos e técnicas

Para testar o arcabouço descrito no capítulo 4, foram utilizados os modelos *Gradient Boosting* e Regressão Logística. A biblioteca scikit-learn[1] foi utilizada com dois conjuntos de dados enviesados. Para gerar os resultados do SHAP, foi utilizada a biblioteca python de código aberto disponível[2]. Todos os testes e os dados utilizados nos experimentos estão disponíveis em um repositório no GitHub[3]. Conforme detalhado anteriormente, existem diferentes variações do método SHAP. Para o modelo de *Gradient Boosting*, utilizamos TreeSHAP; já para a Regressão Logística, utilizamos LinearSHAP.

5.2 Dados

Os conjuntos de dados utilizados nesta dissertação são: *Adult-Income* e *COMPAS* os quais são comumente utilizados para avaliar viés em modelos de aprendizado de máquina. O conjunto de dados *COMPAS* foi obtido do ProPublica (ANGWIN *et al.*, 2016) e o conjunto *Adult-Income* foi obtido do repositório público da UCI (DUA; GRAFF, 2017).

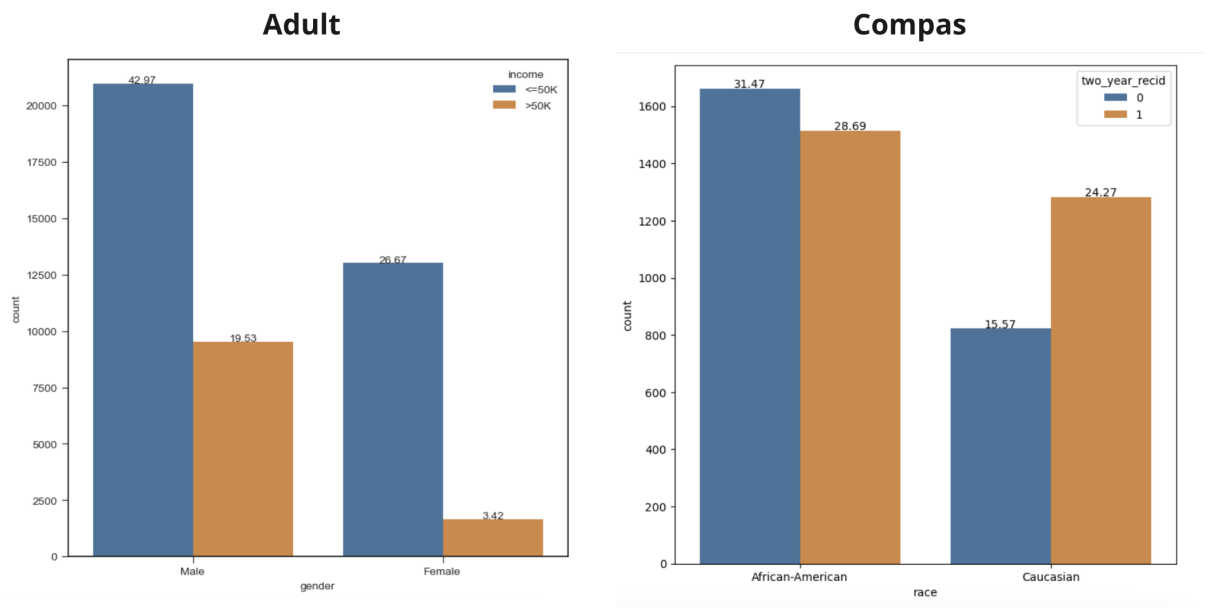
A Tabela 3 apresenta um resumo de cada conjunto de dados. A Figura 9 mostra, por grupo, o percentual de casos em que a variável alvo é favorável (valor 1) e desfavorável (valor 0). Nota-se que nos conjuntos apresentados o percentual de casos em que a variável alvo tem valor 1 é maior dentre o grupo privilegiado se comparado ao grupo desprivilegiado.

Tabela 3 – Informações dos conjuntos de dados

Dataset	**Atributo alvo**	**Atributo sensível**	**Grupo privilegiado**	**Grupo desprivilegiado**
Adult-Income	Salário anual acima de 50 mil dólares	Gênero	Masculino	Feminino
COMPAS	Não reincidir em 2 anos	Raça	Caucasianos	afro-americanos

[1] ⟨https://scikit-learn.org/stable/⟩
[2] ⟨https://github.com/slundberg/shap⟩
[3] ⟨https://github.com/carlaprv/dissertation⟩

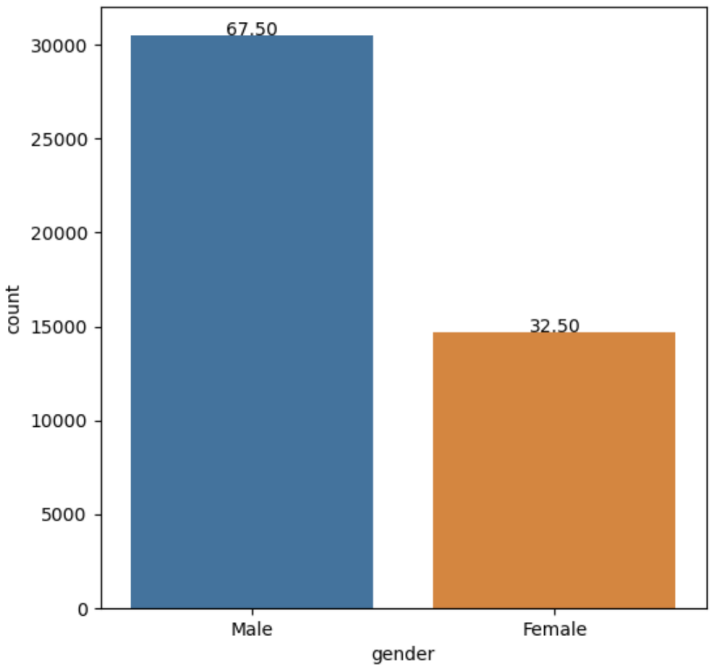
Figura 9 – Gráfico de comparação das amostras para grupos privilegiados e desprivilegiados



O conjunto de dados *Adult-Income* é resultado de um censo realizado em 1994 nos Estados Unidos. A partir desses dados, pode-se explorar a possibilidade de predizer a renda com base em informações pessoais. A tarefa de previsão é determinar se uma pessoa ganha mais ou menos de 50 mil ao ano. O gênero será usado como atributo protegido. A análise da distribuição do conjunto de acordo com a variável gênero, apresentada na Figura 10, explica que (i) gênero tem duas categorias únicas (masculino e feminino) e (ii) o conjunto de dados é enviesado para o sexo masculino com quase 67%.

Uma outra forma de visualização é o *Summary plot* (Figura 7), em que os valores SHAP obtidos são representados por círculos e são ordenados e plotados horizontalmente. Cada valor SHAP é colorido de acordo com o valor do atributo, sendo associada a cor azul aos valores menores e vermelho aos maiores.

Figura 10 – Gráfico de distribuição das amostras do conjunto Adult-Income por gênero



O conjunto de dados *COMPAS* (*Correctional Offender Management Profiling for Alternative Sanctions*) é utilizado por um algoritmo usado por juízes e oficiais de condicional para pontuar a probabilidade de reincidência do réu criminal nos EUA. A ProPublica realizou um estudo sobre o sistema *COMPAS* e os dados utilizados pelo algoritmo. Foram analisadas as notas de risco definidas pelo programa para mais de 7 mil pessoas presas em Broward County, na Flórida, de 2013 a 2014. Em seguida, os jornalistas verificaram quantos desses réus foram condenados por novos crimes nos dois anos seguintes. Os dados dos *COMPAS* utilizados são os analisados pelos jornalista da ProPublica (ANGWIN *et al.*, 2016). A raça será utilizada como atributo sensível, mas reclassificada em dois grupos: caucasianos e afro-americanos.

A análise da distribuição do conjunto de acordo com a variável raça, apresentada na Figura 11, explica que o conjunto de dados é enviesado para a raça *afro-americano*.

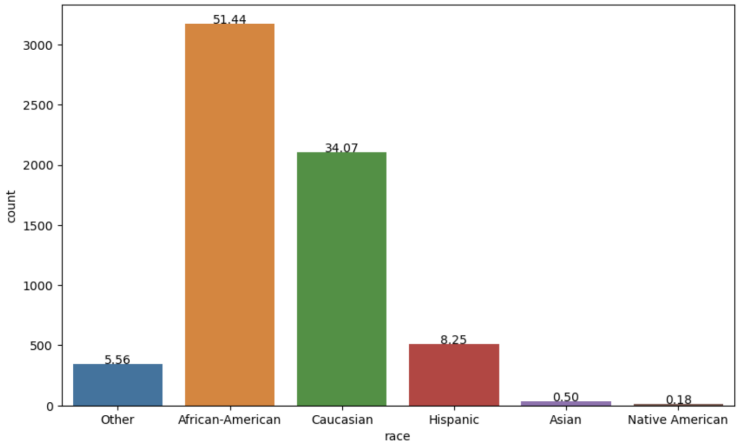


Figura 11 – Gráfico de distribuição das amostras do conjunto COMPAS por raça

5.3 Discussão dos resultados

Nesta seção, os resultados obtidos com o agrupamento das explicações são comparados e analisados a fim de compreender se houve algum ganho na proposta apresentada. Destaca-se que diversos experimentos foram realizados para garantir consistência da proposta dado o uso da técnica *k-means*, mas, como não houve grande variação nos resultados, não se faz necessário detalhamento de todos os experimentos realizados.

5.3.1 Métricas treinamento modelos caixa-preta

A Tabela 4 apresenta informações da acurácia dos modelos caixa-preta treinado.Dada a maior acurácia do modelo *Gradient Boosting* para ambos os casos, o detalhamento é realizado com base nos resultados deste modelo.

Tabela 4 – Tabela de acurácia dos modelos caixa-preta

Conjunto de dados	**Modelo**	**Acurácia**
Adult-Income	Gradient Boosting	0,86
Adult-Income	Regressão Logística	0,82
COMPAS	Gradient Boosting	0,71
COMPAS	Regressão Logística	0,68

Na seção 4.1.1, apresentamos três propostas de agrupamento de valores SHAP. As propostas 2 e 3 não geraram grupos significativos possivelmente devido à alta esparsidade dos dados e simplificação, respectivamente. A proposta número 1 de agrupamento dos

valores SHAP usando distância euclidiana resultou em melhores explicações e os resultados são apresentados a seguir.

5.3.2 Adult-Income

O modelo *Gradient Boosting* foi treinado com o conjunto de dados normalizados do dataset *Adult-Income Income*. Para o agrupamento dos valores SHAP calculados, foi determinado um número de clusters $k = 2$ devido a um maior valor de *silhouette score*. A Tabela 5 apresenta o tamanho dos grupos extraídos.

Tabela 5 – Tamanho dos grupos - conjunto de dados *Adult*

Dados	**Modelo**	**Grupo**	**Tamanho**
Adult-Income	Gradient Boosting	0	7795
Adult-Income	Gradient Boosting	1	7129

Uma primeira análise dos valores SHAP (Figura 12, 13 e 14) revelou que o atributo protegido (*gender*) ocupa a 8^a posição no ranking de importância de atributos e possui valor de $0,07$. Além disso, a média do valor SHAP do atributo protegido para a classe privilegiada ($renda >= \$50k$) é maior que a média da classe desprivilegiado - implicando que é uma variável importante para determinar uma predição favorável.

Figura 12 – Gráfico dos valores SHAP global do modelo

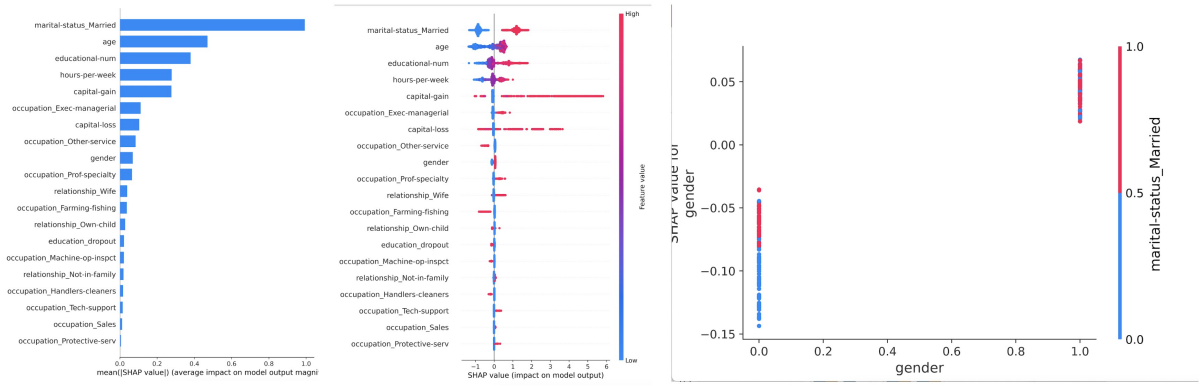


Figura 13 – Gráfico dos valores SHAP global do grupo 0

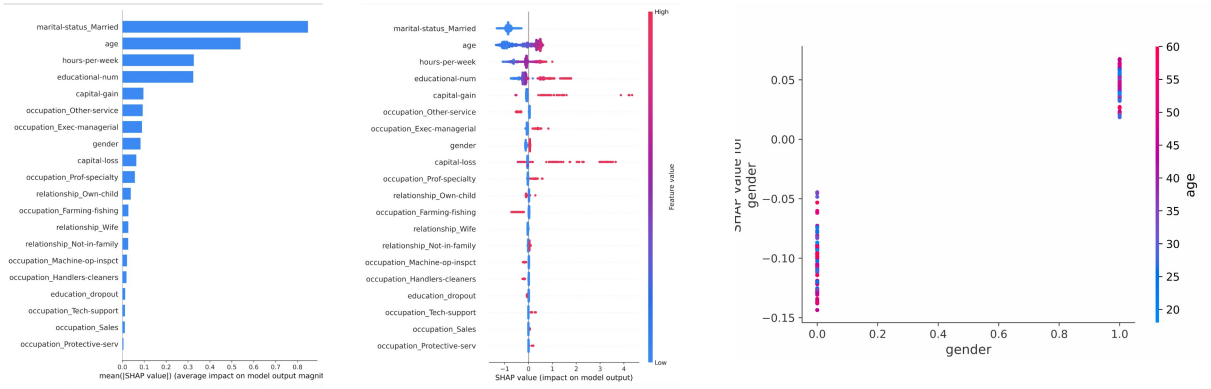


Figura 14 – Gráfico dos valores SHAP global do grupo 1

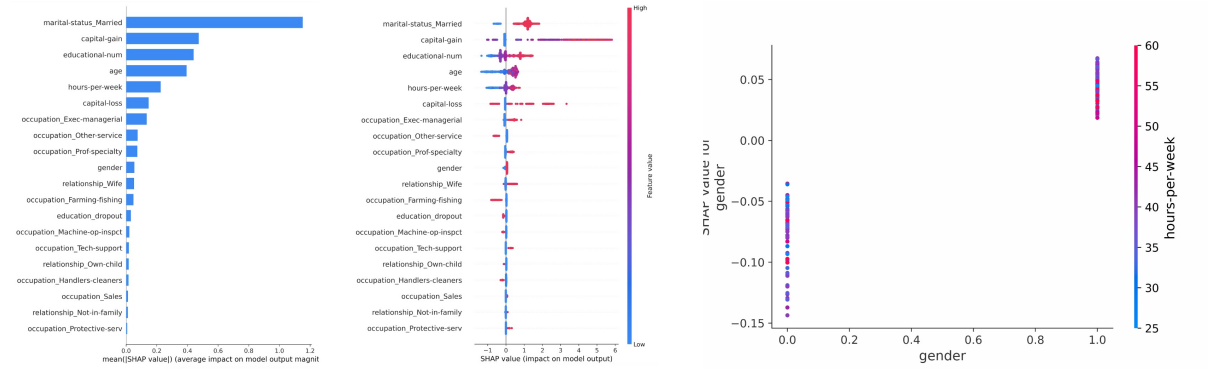


Tabela 6 – Informação dos grupos (Adult-Income)

Dataset	**Grupo**	**Idade Mulheres (média)**	**Idade Homens (média)**	**Mulheres Casadas**	**Homens casados**	**Mulheres (%)**	**Homens (%)**
Adult-Income	0	36	32	0	0	52%	47%
Adult-Income	1	43	43	705	6278	10%	89%

No grupo 0, as informações demográficas extraídas (Tabela 6) considerando os atributos de mais importância nos mostram que o grupo 0 é um grupo mais jovem e solteiro. Calculamos também a proporção de homens e mulheres com relação ao atributo alvo (*income*) e os dois grupos apresentam proporções muito próximas: homens com valor 0 para o atributo *income* representam 44% do grupo e mulheres na mesma classificação representam 50% do grupo. Nesse caso, podemos notar que os fatores de maior influência

para uma predição negativa foram o estado civil e a idade, conforme os gráficos SHAP gerados.

O *Summary plot* nos mostra o contraste entre os grupos. Apesar de o atributo estado civil casado ser o mais importante para os dois grupos, os valores SHAP são opostos. Para o grupo 0, o estado civil casado é uma variável que contribui negativamente para a predição favorável, enquanto que para o grupo 1, existe um cenário oposto. As características demográficas do grupo 1 explicam as divergências: trata-se de um grupo predominantemente composto por homens (89%) mais velhos (média de 43 anos) e casados.

Importante ressaltar que apenas observando os gráficos de barra do Gradient Boosting, a variável *gender* parece não ser tão relevante se comparada com os outros atributos. As análises dos grupos realizadas sob uma perspectiva interseccional permitiram identificar que não é somente a idade e o estado civil que influenciam no modelo, mas também o gênero.

5.3.3 COMPAS

O modelo *Gradient Boosting* foi treinado com o conjunto de dados normalizados do dataset COMPAS. Para o agrupamento dos valores SHAP calculados, foi determinando um número de clusters $k = 2$ devido ao maior valor de *silhouette score*. A Tabela 5.3.3 apresenta o tamanho dos grupos extraídos.

Tabela 7 – Tamanho dos grupos - conjunto de dados *Compass*

Dados	**Modelo**	**Grupo**	**Tamanho**
COMPAS	Gradient Boosting	0	680
COMPAS	Gradient Boosting	1	1357

Uma primeira análise dos valores SHAP revelou que o atributo protegido (*race*) ocupa a 7^a posição no ranking de importância de atributos e possui valor de $0,038$. Além disso, a média do valor SHAP do atributo protegido para a classe privilegiada (sem probabilidade de reincidência nos próximos dois anos) é maior que a média da classe desprivilegiada - implicando que é uma variável importante para determinar uma predição favorável.

O ranking de valores SHAP nos mostra que a quantidade de prisões anteriores (*priors count*), a idade e o score do *COMPAS* (*decile score*) são os atributos de maior

importância. De forma que, um número maior de prisões e menor idade indicam um maior risco de reincidência. Um maior score calculado pelo *COMPAS* também representa maior risco de reincidência, assim como ser do sexo masculino.

Figura 15 – Gráfico dos valores SHAP global do modelo

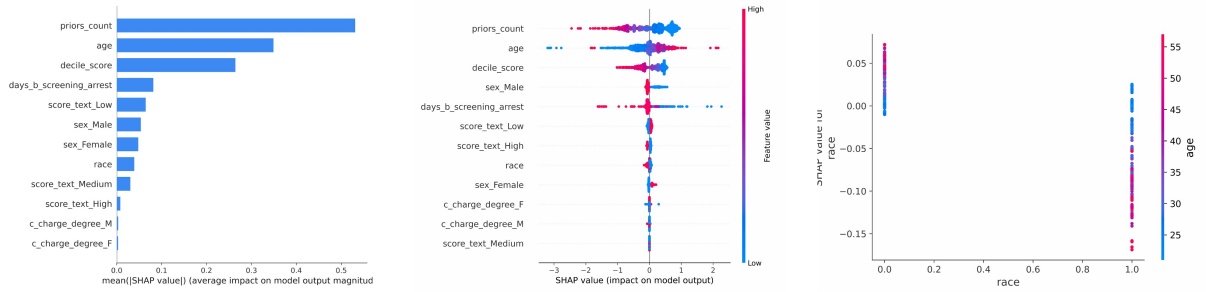


Figura 16 – Gráfico dos valores SHAP global do grupo 0

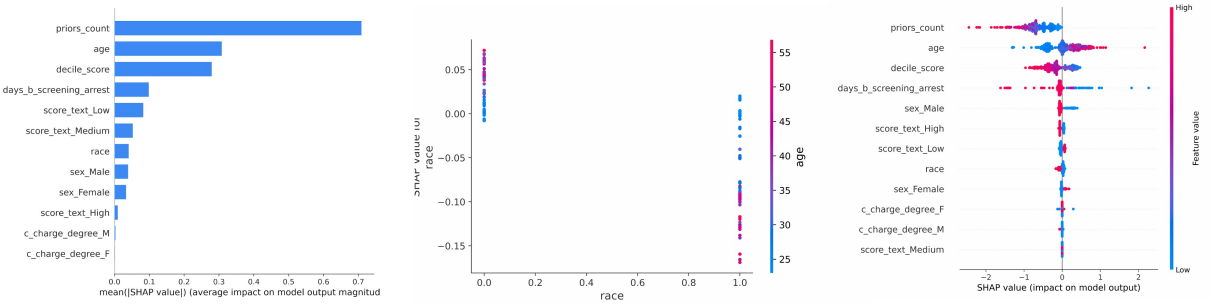
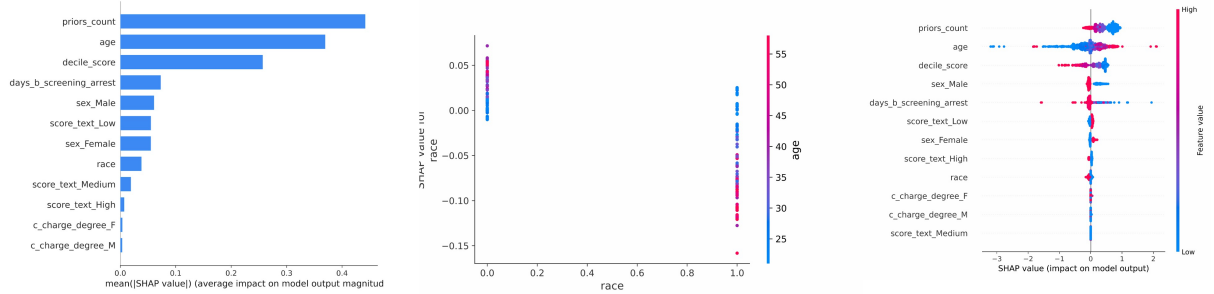


Figura 17 – Gráfico dos valores SHAP global do grupo 1



Os gráficos SHAP (Figuras 15, 16 e 17) de ambos os grupos são muito similares, assim como a importância dos atributos que é muito semelhante a explicação global. No entanto, uma exploração da demografia dos grupos explica um pouco melhor suas diferenças ilustradas também no *Dependence Plot*.

No grupo 0, há uma predominância de afro-americanos (72%) e afro-americanos com chance de reincidência em dois anos representam 52% dos integrantes do grupo. Já no grupo 1, existe uma divisão mais proporcional da quantidade de caucasianos e afro-americanos, assim como da proporção com chance maior de reincidência em cada um dos grupos 22% para afro-americanos e 14% para caucasianos. Um atributo de relevância no grupo 1 é a idade (mostrada na Tabela 8) de forma que caucasianos têm uma menor porcentagem de prováveis reincidentes já que a idade média dos membros caucasianos desse grupo é de 37 anos comparada com a média de 30 anos dos afro-americanos. Por outro lado, no grupo 0, a idade é semelhante entre as duas raças, portanto a raça se torna um fator mais determinante.

Tabela 8 – Informação dos grupos (COMPAS)

Dataset	**Grupo**	**Idade média (caucasian)**	**Idade média (afro-american)**	**Caucasian (%)**	**Afro-american (%)**
COMPAS	0	36	34	27%	72%
COMPAS	1	37	30	44%	55%

Apenas observado os rankings de atributos e gráficos pode parecer não haver diferenças na importância dos atributos para ambos os grupos. No entanto, conforme observamos, o atributo idade é determinante para o grupo 1 e o atributo raça para o grupo 0.

5.4 Discussão dos resultados

Na avaliação dos resultados, comparamos e analisamos as explicações agrupadas com as explicações globais do modelo. Foram gerados gráficos de ranking de atributos e dependência de atributos para ajudar na comparação. Também foram utilizados dados demográficos para auxiliar no entendimento do comportamento do modelo em cada um dos grupos gerados.

Nota-se também que as explicações globais de cada grupo estão alinhadas com os valores SHAP globais do modelo já que 100% das atribuições globais corresponderam aos principais atributos do *Gradient Boosting*, garantindo fidelidade e robustez.

As avaliações permitiram ver que o agrupamento de valores SHAP pode ser benéfico para análises de *fairness* já que permite descobrir novas informações sobre o processo de

tomada de decisão do modelo. No entanto, a descoberta de novas informações requer uma análise interseccional de como os atributos se relacionam e não apenas analisá-los de forma independente.

6 Conclusões

Em aplicações de Inteligência Artificial em que existem preocupações de garantir o princípio de não-discriminação e direito à explicação, os modelos costumam ser avaliados com medidas de justiça e com técnicas de interpretabilidade. São duas áreas complementares, pois, à medida que a justiça permite comparar o quanto um modelo é mais justo do que outro, a interpretabilidade permite garantir confiança no modelo já que garante maior entendimento de seu processo de tomada de decisão.

A principal meta desta dissertação foi desenvolver uma solução que combinasse diferentes técnicas de interpretabilidade a fim de alcançar explicações globais de modelos de aprendizado e máquina. Acreditava-se que, a partir da solução proposta, seria possível aprimorar e aumentar o desempenho observado no estado-da-arte de técnicas globais e apoiar estudos futuros nessa área. O princípio apresentado no trabalho de agrupamento confirmou que técnicas de interpretabilidade podem ser aliadas na identificação de comportamentos discriminatórios e análises de *fairness*. No entanto, a proposta requer mais do que apenas o conhecimento técnico, mas também conhecimento do contexto do problema ao qual o modelo se propõe resolver, requer maior investigação da *ground truth* existente nos dados e até possivelmente uma investigação da origem dos dados que podem estar causando discriminação.

Nesta dissertação, utilizamos o atributo sensível diretamente no modelo, pois com o método SHAP conseguimos interpretar apenas os resultados das variáveis usadas no modelo. No entanto, o uso não é recomendado em todos os casos já que pode gerar enviesamento. Por outro lado, a ausência do atributo protegido pode dificultar a detecção de comportamentos discriminatórios. O trabalho foi uma primeira versão de validação da proposta e trabalhos futuros são necessários para compreender como incorporar métodos de explicabilidade global em pipelines de produção de modelos de aprendizado de máquina.

A Inteligência Artificial Explicável é um campo promissor com inúmeras abordagens surgindo a cada ano. No entanto, apesar desses avanços, alguns desafios importantes permanecem sem solução e são necessárias soluções futuras para promover ainda mais o progresso desse campo. Nesta dissertação também apresentamos dados e reflexões sobre Inteligência Artificial sob uma perspectiva crítica baseada na teoria de Feenberg (1992). Feenberg (1992), por meio da teoria crítica da tecnologia, demonstrou que os impactos

sociais gerados por qualquer tecnologia precisam ser estudados em sua complexidade, pois envolvem aspectos funcionais e sociais. Isso implica questionar a própria realidade justificada ainda com base no determinismo tecnológico. Feenberg propõe uma racionalização subversiva para a tecnologia. Esta proposição contradiz o determinismo tecnológico que opera sob uma lógica linear segundo a qual a tecnologia necessariamente implica em progresso e que, os processos tecnológicos são independentemente de quaisquer fatores sociais. Portanto, é sob perspectivas dos riscos gerados pelo desenvolvimento não questionado e mercadológico que ameaça a dimensão crítica, que as tecnologias precisam ser desafiadas, reexaminadas e ressignificadas. Para tanto, é necessário avanço nos estudos que dialoguem com movimentos sociais. Talvez também seja necessário reformular e reforçar a relação entre trabalho acadêmico e ativismo social e político.

Ressalta-se, porém, que a discussão exposta nesta dissertação não representa uma tentativa de frear o desenvolvimento tecnológico. O problema não são as novas tecnologias de inteligência artificial em si, mas, antes, a estrutura da sociedade e os poderes que atuam em seu desenvolvimento. A presente dissertação buscou embasamento científico para que acadêmicos e profissionais da área de computação possam estar atualizados em sua prática e orientando sobre as possíveis complicações e impactos sociais existentes. Assim como, levantar uma crítica sobre os estudos atuais da área e a necessidade de uma renovação dos referenciais teóricos e críticas ao modelo de pensar existente.

Referências

ALVAREZ-MELIS, D.; JAAKKOLA, T. S. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018. Disponível em: ⟨http://arxiv.org/abs/1806.08049⟩. Citado na página 45.

ANGWIN, J.; LARSON, J.; MATTU, S.; KIRCHNER, L. *Machine Bias*. 2016. ⟨https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing⟩. Citado 4 vezes nas páginas 13, 30, 51 e 53.

BENJAMIN, R. *Race After Technology*: Abolitionist tools for the new jim code. [S.l.]: John Wiley & Sons, 2019. Citado 2 vezes nas páginas 16 e 17.

BIRAN, O.; COTTON, C. Explanation and justification in machine learning: a survey. *IJCAI-17 workshop on explainable AI (XAI)*, 2017. ⟨http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf⟩. Citado na página 28.

BIRHANE, A.; KALLURI, P.; CARD, D.; AGNEW, W.; DOTAN, R.; BAO, M. The values encoded in machine learning research. *CoRR*, abs/2106.15590, 2021. Disponível em: ⟨https://arxiv.org/abs/2106.15590⟩. Citado na página 29.

BLONIARZ, A.; TALWALKAR, A.; YU, B.; WU, C. Supervised neighborhoods for distributed nonparametric regression. In: *AISTATS*. [S.l.: s.n.], 2016. Citado na página 34.

BORDIN, W. A. B. L. Essa "tal" filosofia: sobre as concepções de tecnologia e seus reflexos no processo formativo em engenharia. *Revista Brasileira de Ensino de Ciência e Tecnologia*, v. 11, n. 1, p. 228–249, 2018. Disponível em: ⟨https://periodicos.utfpr.edu.br/rbect/article/view/5728⟩. Citado na página 21.

Brasil. *Lei Geral de Proteção de Dados Pessoais*. 2018. Disponível em: ⟨http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm⟩. Citado 2 vezes nas páginas 14 e 30.

BUOLAMWINI, J.; GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: FRIEDLER, S. A.; WILSON, C. (Ed.). *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. New York, NY, USA: PMLR, 2018. (Proceedings of Machine Learning Research, v. 81), p. 77–91. Disponível em: ⟨http://proceedings.mlr.press/v81/buolamwini18a.html⟩. Citado na página 30.

CRAVEN, M. W.; SHAVLIK, J. W. Extracting tree-structured representations of trained networks. In: *Proceedings of the 8th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 1995. (NIPS'95), p. 24–30. Available at ⟨http://dl.acm.org/citation.cfm?id=2998828.2998832⟩. Citado 2 vezes nas páginas 27 e 39.

DU, M.; LIU, N.; HU, X. Techniques for interpretable machine learning. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 63, n. 1, p. 68–77, dez. 2019. ISSN 0001-0782. Disponível em: ⟨https://doi.org/10.1145/3359786⟩. Citado na página 33.

DUA, D.; GRAFF, C. *UCI Machine Learning Repository*. 2017. Disponível em: ⟨http://archive.ics.uci.edu/ml⟩. Citado 2 vezes nas páginas 20 e 51.

EPSTEIN, Z.; PAYNE, B. H.; SHEN, J. H.; DUBEY, A.; FELBO, B.; GROH, M.; OBRADOVICH, N.; CEBRIÁN, M.; RAHWAN, I. Closing the AI knowledge gap. *CoRR*, abs/1803.07233, 2018. Disponível em: ⟨http://arxiv.org/abs/1803.07233⟩. Citado na página 18.

EUROPEAN COMMISSION. *2018 reform of EU data protection rules*. 2018. Disponível em: ⟨https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf⟩. Citado na página 30.

EVANS, B. P.; XUE, B.; ZHANG, M. What's inside the black-box? a genetic programming method for interpreting complex machine learning models. In: *Proceedings of the Genetic and Evolutionary Computation Conference*. New York, NY, USA: Association for Computing Machinery, 2019. (GECCO '19), p. 1012–1020. ISBN 9781450361118. Disponível em: ⟨https://doi.org/10.1145/3321707.3321726⟩. Citado 3 vezes nas páginas 16, 34 e 35.

FEENBERG, A. Critical theory of technology. *Social Science Computer Review*, v. 10, n. 3, p. 447–448, 1992. Disponível em: ⟨https://doi.org/10.1177/089443939201000339⟩. Citado 6 vezes nas páginas 21, 22, 23, 25, 26 e 61.

GEBRU, T. *Oxford Handbook on AI Ethics Book Chapter on Race and Gender*. 2019. Citado na página 21.

GUNNING, D. *Explainable Artificial Intelligence (XAI)*. [S.l.], 2017. Available at ⟨https://www.darpa.mil/attachments/XAIProgramUpdate.pdf⟩. Citado 2 vezes nas páginas 16 e 27.

IBRAHIM, M.; LOUIE, M.; MODARRES, C.; PAISLEY, J. Global explanations of neural networks: Mapping the landscape of predictions. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery, 2019. (AIES '19), p. 279–287. ISBN 9781450363242. Disponível em: ⟨https://doi.org/10.1145/3306618.3314230⟩. Citado 5 vezes nas páginas 16, 18, 35, 36 e 47.

IRIONDO, R. *Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women*. 2018. [Online; posted 11-October-2018]. Disponível em: ⟨https://medium.datadriveninvestor.com/amazon-scraps-secret-ai-recruiting-engine-that-showed-biases-against-women-995c505f5c6f⟩. Citado na página 30.

JOHANSSON, U.; KöNIG, R.; NIKLASSON, L. Genetic rule extraction optimizing brier score. In: *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation*. New York, NY, USA: Association for Computing Machinery, 2010. (GECCO '10), p. 1007–1014. ISBN 9781450300728. Disponível em: ⟨https://doi.org/10.1145/1830483.1830668⟩. Citado 2 vezes nas páginas 16 e 34.

KAZEMITABAR, J.; AMINI, A.; BLONIARZ, A.; TALWALKAR, A. S. Variable importance using decision trees. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). *Advances in Neural*

Information Processing Systems 30. Curran Associates, Inc., 2017. p. 426–435. Disponível em: ⟨http://papers.nips.cc/paper/6646-variable-importance-using-decision-trees.pdf⟩. Citado na página 34.

KOZLOWSKI, D.; MURRAY, D. S.; BELL, A.; HULSEY, W.; LARIVIÈRE, V.; MONROE-WHITE, T.; SUGIMOTO, C. R. Avoiding bias when inferring race using name-based approaches. *PLoS One*, United States, v. 17, n. 3, p. e0264270, mar. 2022. Citado na página 14.

LAKKARAJU, H.; KAMAR, E.; CARUANA, R.; LESKOVEC, J. Faithful and customizable explanations of black box models. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery, 2019. (AIES '19), p. 131–138. ISBN 9781450363242. Disponível em: ⟨https://doi.org/10.1145/3306618.3314229⟩. Citado 6 vezes nas páginas 8, 16, 34, 41, 42 e 44.

LAUGEL, T.; LESOT, M.-J.; MARSALA, C.; RENARD, X.; DETYNIECKI, M. *The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations*. 2019. Citado na página 43.

LESLIE, D. *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. [S.l.], 2019. Citado na página 14.

LUCIC, A.; HANED, H.; RIJKE, M. de. Why does my model fail? contrastive local explanations for retail forecasting. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery, 2020. (FAT* '20), p. 90–98. ISBN 9781450369367. Disponível em: ⟨https://doi.org/10.1145/3351095.3372824⟩. Citado 2 vezes nas páginas 16 e 38.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017. p. 4765–4774. Disponível em: ⟨http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf⟩. Citado na página 35.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, v. 5, n. 4, p. 115–133, Dec 1943. ISSN 1522-9602. Disponível em: ⟨https://doi.org/10.1007/BF02478259⟩. Citado na página 21.

MESSALAS, A.; KANELLOPOULOS, Y.; MAKRIS, C. Model-agnostic interpretability with shapley values. In: *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. [S.l.: s.n.], 2019. p. 1–7. Citado 8 vezes nas páginas 16, 18, 31, 35, 39, 40, 41 e 43.

MILHANO Ângelo S. N. *A Emergência da Teoria Crítica da Tecnologia de Andrew Feenberg*. Dissertação (Mestrado) — Faculdade de Letras da Universidade do Porto, Portugal, 2010. Citado 3 vezes nas páginas 9, 24 e 25.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. *Cornell University*, Jun 2017. Available at ⟨https://arxiv.org/abs/1706.07269⟩. Citado na página 28.

MOHSENI, S.; ZAREI, N.; RAGAN, E. D. *A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems*. 2020. Citado 2 vezes nas páginas 32 e 33.

MONTEIRO, R. L. *Existe direito à explicação na Lei Geral de proteção de Dados no Brasil?* 2018. Disponível em: ⟨https://igarape.org.br/existe-um-direito-a-explicacao-na-lei-geral-de-protecao-de-dados-no-brasil/⟩. Citado na página 30.

MOTHILAL, R. K.; SHARMA, A.; TAN, C. Explaining machine learning classifiers through diverse counterfactual explanations. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery, 2020. (FAT* '20), p. 607–617. ISBN 9781450369367. Disponível em: ⟨https://doi.org/10.1145/3351095.3372850⟩. Citado 5 vezes nas páginas 16, 37, 38, 42 e 44.

O'NEIL, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. USA: Crown Publishing Group, 2016. ISBN 0553418815. Citado 3 vezes nas páginas 16, 21 e 29.

PAPADOPOULOS, P.; WALKINSHAW, N. Black-box test generation from inferred models. In: *Proceedings of the Fourth International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering*. [S.l.]: IEEE Press, 2015. (RAISE '15), p. 19–24. Citado na página 15.

PASQUALE, F. Front matter. In: _____. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015. ISBN 9780674368279. Disponível em: ⟨http://www.jstor.org/stable/j.ctt13x0hch.1⟩. Citado na página 16.

PLUMB, G.; MOLITOR, D.; TALWALKAR, A. Model agnostic supervised local explanations. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2018. (NIPS'18), p. 2520–2529. Citado 2 vezes nas páginas 16 e 34.

QUEIROZ, I. P. de. *Fanon, o reconhecimento do negro e o novo humanismo: horizontes descoloniais da tecnologia*. Tese (Doutorado) — Universidade Tecnológica Federal do Paraná, 2013. Disponível em: ⟨http://repositorio.utfpr.edu.br/jspui/handle/1/492⟩. Citado na página 24.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Model-agnostic interpretability of machine learning. *Cornell University*, Jun 2016. Available at ⟨https://arxiv.org/abs/1606.05386⟩. Citado 5 vezes nas páginas 8, 30, 31, 33 e 39.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 1135–1144. ISBN 9781450342322. Disponível em: ⟨https://doi.org/10.1145/2939672.2939778⟩. Citado 5 vezes nas páginas 16, 35, 36, 40 e 43.

RIBEIRO, M. T. C. *Model-Agnostic Explanations and Evaluation of Machine Learning*. Tese (PhD dissertation) — University of Washington, 2018. Citado 2 vezes nas páginas 43 e 44.

RUBACK, L.; AVILA, S.; CANTERO, L. Vieses no aprendizado de máquina e suas implicações sociais: Um estudo de caso no reconhecimento facial. In: *Anais do II Workshop sobre as Implicações da Computação na Sociedade*. Porto Alegre, RS, Brasil: SBC, 2021. p. 90–101. ISSN 2763-8707. Disponível em: ⟨https://sol.sbc.org.br/index.php/wics/article/view/15967⟩. Citado na página 29.

RUBACK, L.; CARVALHO, D.; AVILA, S. Mitigating bias in machine learning: A socio-technical analysis. *iSys - Brazilian Journal of Information Systems*, Aug. 2022. Disponível em: ⟨https://sol.sbc.org.br/journals/index.php/isys/article/view/2396⟩. Citado na página 14.

SCHNEIDER CAMILA B. E MIRANDA, P. F. M. Vigilância e segurança pública: preconceitos e segregação social ampliados pela suposta neutralidade digital (surveillance and public security: prejudices and social segregation widened by the alleged digital neutrality). *Emancipação*, v. 20, p. 1–22, 2020. Disponível em: ⟨https://revistas2.uepg.br/index.php/emancipacao/article/view/14258⟩. Citado na página 30.

SHANNON, C. E. Programming a computer for playing chess. In: _____. *Computer Chess Compendium*. New York, NY: Springer New York, 1988. p. 2–13. ISBN 978-1-4757-1968-0. Disponível em: ⟨https://doi.org/10.1007/978-1-4757-1968-0_1⟩. Citado na página 21.

SHARMA, S.; HENDERSON, J.; GHOSH, J. Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery, 2020. (AIES '20), p. 166–172. ISBN 9781450371100. Disponível em: ⟨https://doi.org/10.1145/3375627.3375812⟩. Citado 3 vezes nas páginas 16, 37 e 40.

SILVA, T. Visão computacional e racismo algorítmico: Branquitude e opacidade no aprendizado de máquina. v. 12, p. 428–448, 02 2020. Citado na página 16.

SILVEIRA, S. A. d.; SILVA, T. R. d. CONTROVÉRSIAS SOBRE DANOS ALGORÍTMICOS: discursos corporativos sobre discriminação codificada. *Revista Observatório*, v. 6, n. 4, p. a1pt, jul. 2020. Disponível em: ⟨https://sistemas.uft.edu.br/periodicos/index.php/observatorio/article/view/11069⟩. Citado na página 16.

SILVER, D.; HUANG, A.; MADDISON, C. J.; GUEZ, A.; SIFRE, L.; DRIESSCHE, G. van den; SCHRITTWIESER, J.; ANTONOGLOU, I.; PANNEERSHELVAM, V.; LANCTOT, M.; DIELEMAN, S.; GREWE, D.; NHAM, J.; KALCHBRENNER, N.; SUTSKEVER, I.; LILLICRAP, T.; LEACH, M.; KAVUKCUOGLU, K.; GRAEPEL, T.; HASSABIS, D. Mastering the game of go with deep neural networks and tree search. *Nature*, v. 529, n. 7587, p. 484–489, Jan 2016. ISSN 1476-4687. Disponível em: ⟨https://doi.org/10.1038/nature16961⟩. Citado na página 13.

SLACK, D.; FRIEDLER; A., S.; GIVENTAL, E. Fairness warnings and fair-maml: Learning fairly with minimal data. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing

Machinery, 2020. (FAT* '20), p. 200–209. ISBN 9781450369367. Disponível em: ⟨https://doi.org/10.1145/3351095.3372839⟩. Citado 3 vezes nas páginas 16, 35 e 36.

SLACK, D.; HILGARD, S.; JIA, E.; SINGH, S.; LAKKARAJU, H. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery, 2020. (AIES '20), p. 180–186. ISBN 9781450371100. Disponível em: ⟨https://doi.org/10.1145/3375627.3375830⟩. Citado na página 44.

TURING, A. M. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX, n. 236, p. 433–460, 10 1950. ISSN 0026-4423. Disponível em: ⟨https://doi.org/10.1093/mind/LIX.236.433⟩. Citado na página 21.

USTUN, B.; RUDIN, C. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, Springer Science and Business Media LLC, v. 102, n. 3, p. 349–391, Nov 2015. ISSN 1573-0565. Disponível em: ⟨https://doi.org/10.1007/s10994-015-5528-6⟩. Citado na página 36.

VERMA, S.; RUBIN, J. Fairness definitions explained. In: *Proceedings of the International Workshop on Software Fairness*. New York, NY, USA: Association for Computing Machinery, 2018. (FairWare '18), p. 1–7. ISBN 9781450357463. Disponível em: ⟨https://doi.org/10.1145/3194770.3194776⟩. Citado na página 15.

VIEIRA, C. P.; DIGIAMPIETRI, L. A. Machine learning post-hoc interpretability: A systematic mapping study. In: *XVIII Brazilian Symposium on Information Systems*. New York, NY, USA: Association for Computing Machinery, 2022. (SBSI). ISBN 9781450396981. Disponível em: ⟨https://doi.org/10.1145/3535511.3535512⟩. Citado na página 43.

WACHTER, S.; MITTELSTADT, B.; RUSSELL, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard journal of law & technology*, v. 31, p. 841–887, 04 2018. Citado na página 37.