UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I - CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIAS
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE GRADUAÇÃO EM BACHARELADO EM COMPUTAÇÃO

MARCELO VICTOR ARAÚJO PEDROSA

DESVENDANDO VIÉS: MAPEAMENTO DE PROCESSOS E FERRAMENTAS NA DETECÇÃO DE VIÉS E MITIGAÇÃO DE VIÉS EM MODELOS DE APRENDIZADO DE MÁQUINA

CAMPINA GRANDE
2024

MARCELO VICTOR ARAÚJO PEDROSA

DESVENDANDO VIÉS: MAPEAMENTO DE PROCESSOS E FERRAMENTAS NA DETECÇÃO DE VIÉS E MITIGAÇÃO DE VIÉS EM MODELOS DE APRENDIZADO DE MÁQUINA

Trabalho apresentado ao Curso de Bacharelado em ciências da computação da Universidade Estadual da Paraíba, em cumprimento à exigência para obtenção do Bacharel em Computação.

Área de concentração: Inteligência Artificial.

Orientador : Profª. Ana Isabella Muniz Leite

CAMPINA GRANDE
2024

P372d Pedrosa, Marcelo Victor Araujo.

Desvendando viés [manuscrito] : mapeamento de processos e ferramentas na detecção de viés e mitigação de viés em modelos de aprendizado de máquina / Marcelo Victor Araujo Pedrosa. - 2024.

50 f. : il. color.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Ciência da computação) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2024.

"Orientação : Prof. Ma. Ana Isabella Muniz Leite, Departamento de Computação - CCT".

1. Aprendizado de máquina. 2. Métricas de equidade. 3. Mitigação de viés. I. Título

21. ed. CDD 004.015 1

Elaborada por Lêda Cristina Diniz Andrade - CRB - 15/1032 BC

MARCELO VICTOR ARAUJO PEDROSA

DESVENDANDO VIÉS : MAPEAMENTO DE PROCESSOS E FERRAMENTAS NA DETECÇÃO DE VIÉS E MITIGAÇÃO DE VIÉS EM MODELOS DE APRENDIZADO DE MÁQUINA

Monografia apresentado à Coordenação do Curso de Ciência da Computação da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Computação

Aprovada em: 19/11/2024.

Documento assinado eletronicamente por:

- **Kézia de Vasconcelos Oliveira Dantas** (***.714.244-**), em **05/12/2024 14:54:57** com chave **0a170b0cb33211ef9ece2618257239a1**.
- **Ana Isabella Muniz Leite** (***.834.864-**), em **05/12/2024 14:52:35** com chave **b5b2e02cb33111ef9fb11a7cc27eb1f9**.
- **Wellington Candeia de Araujo** (***.655.074-**), em **06/12/2024 06:53:08** com chave **e5b51d22b3b711ef8fc22618257239a1**.

Documento emitido pelo SUAP. Para comprovar sua autenticidade, faça a leitura do QrCode ao lado ou acesse https://suap.uepb.edu.br/comum/autenticar_documento/ e informe os dados a seguir.
Tipo de Documento: Termo de Aprovação de Projeto Final
Data da Emissão: 06/12/2024
Código de Autenticação: 7925e3

AGRADECIMENTOS

Primeiramente, à Deus, por estar presente em todos os momentos da minha vida me dando forças.

Aos meus pais Andreia e Marcos, por terem sempre acreditado no meu potencial, até mesmo quando eu duvidava. Sou grato por tudo.

À toda minha família, em especial aos meus avós e ao meu tio, pelo apoio incondicional.

Aos meus amigos e colegas e colegas de turma, que juntos aprendemos, enfrentamos e superamos desafios ao longo da graduação.

À minha professora orientadora Ana Isabella, cujos conselhos, dedicação e paciência foram essenciais para a realização deste trabalho.

RESUMO

O crescimento e inovação tecnológica em campos da inteligência artificial como o aprendizado de máquina, culminou na implementação desses sistemas em diversas áreas, como : a financeira, de saúde, trabalho, criminal, entre outros. Todavia, após a presença dessas ferramentas em áreas sensíveis da sociedade, questões como equidade e viés ganharam cada vez mais atenção dos desenvolvedores e da mídia como um todo, devido aos possíveis danos que esses sistemas podem acarretar a grupos minoritários, indivíduos e a sociedade em geral. Diante disso, pesquisadores se empenharam para o desenvolvimento de técnicas e ferramentas voltadas para a detecção, medição e mitigação de possíveis vieses em modelos de aprendizado de máquina, a fim de garantir a equidade nesses sistemas. Este trabalho, portanto, pretende revisar os principais meios e ferramentas utilizados para identificar, medir e mitigar viés, através de uma revisão bibliográfica não sistemática. Inicialmente, são apresentados conceitos fundamentais de aprendizado de máquina, bem como a problemática do viés, seguido pela análise e discussão de diferentes métricas de equidade que atendem a necessidades únicas e oferecendo diferentes percepções, sendo elas : *Equalized Odds, Equality of Opportunity, Demographic Parity* e *Disparate Impact.* Além disso, são exploradas as seguintes ferramentas: *AI Fairness 360 (IBM), Fairlearn (Microsoft)* e *Aequitas (UChicago)*, analisando suas funcionalidades e características. Embora as ferramentas possuam pontos semelhantes, a *AI Fairness 360* se destaca pela ampla variedade de métricas e algoritmos de mitigação disponíveis, além do suporte a múltiplas linguagens. O *Fairlearn* disponibiliza interfaces interativas e é capaz de se integrar a diversas bibliotecas do ecossistema *Python.* Já o *Aequitas* se destaca por sua facilidade de uso para usuários não técnicos, sendo, entretanto, a única ferramenta dentre as estudadas que não fornece métodos para mitigação de viés.

Palavras-Chave: aprendizado de máquina; métricas de equidade; detecção de viés.

ABSTRACT

The growth and technological innovation in fields of artificial intelligence such as machine learning has culminated in the implementation of these systems in several areas, such as finance, health, labor, criminal law, among others. However, after the presence of these tools in sensitive areas of society, issues such as equity and bias have gained increasing attention from developers and the media as a whole, due to the possible harm that these systems can cause to minority groups, individuals and society in general. In view of this, researchers have strived to develop techniques and tools aimed at detecting, measuring and mitigating possible biases in machine learning models, in order to ensure equity in these systems. This work, therefore, aims to review the main means and tools used to identify, measure and mitigate bias, through a non-systematic bibliographic review. Initially, fundamental concepts of machine learning are presented, as well as the issue of bias, followed by the analysis and discussion of different fairness metrics that meet unique needs and offer different insights, namely: Equalized Odds, Equality of Opportunity, Demographic Parity and Disparate Impact. In addition, the following tools are explored: AI Fairness 360 (IBM), Fairlearn (Microsoft) and Aequitas (UChicago), analyzing their functionalities and characteristics. Although the tools have similar points, AI Fairness 360 stands out for the wide variety of metrics and mitigation algorithms available, in addition to support for multiple languages. Fairlearn provides interactive interfaces and is capable of integrating with several libraries in the Python ecosystem. Aequitas stands out for its ease of use for non-technical users, however, it is the only tool among those studied that does not provide methods for bias mitigation.

Keywords: machine learning; fairness metrics; bias detection.

LISTA DE ILUSTRAÇÕES

TABELAS

SUMÁRIO

1 INTRODUÇÃO

Desde o início do século XXI, o crescimento acelerado da inteligência artificial tem transformado profundamente a vida social e econômica da sociedade contemporânea (Cozman; Plonski; Neri, 2021). A adoção de novas tecnologias ligadas a plataformas digitais, bem como a utilização de algoritmos, *big data* e inteligência artificial se tornaram cada vez mais comuns no mercado de trabalho (Reis; Graminho, 2019). Além disso, a utilização de algoritmos de previsão para tomada de decisões tem se espalhado nos setores industrial e público (Mitchell *et al.*, 2020).

Esse cenário se desenvolveu em busca da otimização de tarefas e serviços, substituindo funções que eram primordialmente desempenhadas por pessoas e que atualmente são encarregadas às máquinas (Júnior, 2021). Embora fosse esperado que a implementação dessas tecnologias proporcionasse um melhor desempenho e imparcialidade, a inteligência artificial pode refletir e ampliar preconceitos existentes na sociedade (Gonçalves, 2024). Diante dessa situação, a privacidade dos dados, a transparência de algoritmos e vieses da inteligência artificial em áreas relacionadas à raça, gênero e cor se tornaram questões de atenção (Vieira, 2019).

Embora o fenômeno de viés algorítmico não possua uma definição objetiva (Simões-Gomes; Roberto; Mendonça, 2020). Este trabalho adotou o conceito de viés algoritmo como a reprodução de preconceitos sociais pelos algoritmos em processos decisórios, gerando resultados que podem beneficiar ou prejudicar determinados grupos de forma desproporcional (Brissant, 2023; Gonçalves, 2023).

O fenômeno de viés algorítmico está principalmente relacionado ao uso de algoritmos de aprendizado de máquina, apesar da existência de diversas categorias de algoritmos (Simões-gomes; Roberto; Mendonça, 2020). Algoritmos de aprendizado de máquina são capazes de identificar padrões a partir de grandes conjuntos de dados, se baseando em modelos estatísticos (Ruback; Avila; Cantero, 2021).

A análise das variáveis iniciais e dos hiperparâmetros contribui para a identificação da origem de possíveis enviesamentos (Cozman; Kaufman, 2022). A detecção de possíveis vieses presentes em modelos aprendizado de máquina, está relacionado a presença de atributos sensíveis, nos quais, são identificados através de métricas de equidade (Pagano *et al.*, 2023).

Dessa forma, este trabalho tem como objetivo fazer um mapeamento dos métodos e tecnologias utilizados na detecção e mitigação de vieses em modelos de aprendizado máquina, a fim de servir como um guia acessível para indivíduos, apresentando e analisando as ferramentas utilizadas e os métodos implementados. De maneira mais específica, pretende-se:

- Analisar quais são os métodos utilizados para a detecção de viés em modelos de aprendizado de máquina.
- Analisar quais são as ferramentas utilizadas para a detecção e mitigação de viés em modelos de aprendizado de máquina.
- Analisar estudos de casos específicos, que ilustram a detecção de vieses em modelos de aprendizado de máquina, como o viés de etarismo em sistemas de recrutamento automatizado (Harris, 2023), o viés de gênero em sistemas de aprovação de crédito bancário (Dudík *et al.*, 2020) e discriminação racial em modelos utilizados em julgamentos (Rodolfa *et al.*, 2020).

Para isso, foi conduzida uma pesquisa bibliográfica estruturada, composta por estratégias de pesquisas distintas, como buscas em bases de dados acadêmicos, seleção de artigos relevantes por meio de palavras-chave específicas e análise de citações de estudos sobre o tema. Deste modo, este trabalho pretende contribuir para a literatura existente na área de inteligência artificial. Em princípio, ao fornecer uma análise sobre as ferramentas e métodos utilizados para a detecção e mitigação de viés em modelos de aprendizado de máquina empregadas em áreas sensíveis de tomadas de decisão, espera-se servir como um guia acessível elucidando como essas ferramentas e processos são implementadas no processo de detecção de viés.

Este trabalho está dividido em 6 capítulos. O primeiro capítulo apresenta a introdução, objetivos, metodologia, contribuição prevista e estrutura do trabalho.

O segundo capítulo aborda a fundamentação teórica. Onde serão discutidos o treinamento de modelos em IA, a detecção de viés e os trabalhados relacionados ao tema.

O terceiro capítulo aborda a metodologia adotada para o estudo. Onde descreve como o trabalho foi desenvolvido e estruturado em cada etapa .

O quarto capítulo vai abordar os resultados do estudo. Onde será discutido as resoluções da análise dos objetivos específicos do trabalho.

O quinto capítulo aborda a discussão relacionada ao estudo. Onde serão discutidas as lições aprendidas, os desafios e uma avaliação sobre o que foi observado no trabalho nas resoluções dos objetivos.

O sexto capítulo aborda a conclusão do trabalho. Onde será retomado o objetivo do trabalho e os benefícios que o mesmo trás para a discussão e literatura envolvendo inteligências artificiais.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Aprendizado de máquina

Os seres humanos possuem a capacidade natural de aprender através de experiências. De maneira parecida, as máquinas também possuem capacidade de aprender a partir de dados (Manakitsa *et al.*, 2024). Esse processo é conhecido como aprendizado de máquina, um subcampo da inteligência artificial que surgiu em 1959 (Neoway, 2020).

O aprendizado de máquina pode ser descrito como uma área de estudo que permite que os computadores tenham a capacidade de aprender de forma autônoma, sem a necessidade de uma programação específica (Mahesh, 2019). A implementação desses sistemas já é presente em áreas como : reconhecimento de voz, propagandas personalizadas, carros autônomos e reconhecimento facial (Foster *et al.*, 2021). Os modelos de aprendizado de máquina podem ser classificados em quatro categorias, sendo elas : aprendizado supervisionado, não supervisionado, semi-supervisionado e por reforço (IBM, s.d).

Na categoria de aprendizado de máquina supervisionado, o conjunto de dados de treinamento possui as soluções almejadas, que são chamadas de rótulos. O objetivo do algoritmo é aprender a mapear os *inputs* para os devidos *outputs*. Algoritmos como regressão linear, árvores de decisão e redes neurais são exemplos de aprendizado supervisionado (Géron, 2019; Mahesh, 2019).

Já a categoria de aprendizado não supervisionado, o conjunto de dados de treinamento não são rotulados, ou seja, sem as respostas corretas conhecidas. Sem o auxílio de um “professor”, o objetivo desse algoritmo é descobrir sozinho padrões no conjunto de dados. Os algoritmos de clustering e de redução de dimensionalidade são exemplos de aprendizado não supervisionado (Mahesh, 2019).

A categoria de aprendizado semi-supervisionado, possui características do aprendizado supervisionado e não supervisionado. No qual, o conjunto de dados de treinamento é predominantemente composto por dados não rotulados, sendo complementado por uma quantidade limitada de dados rotulados. Em sua maioria os algoritmos semi-supervisionados também são uma combinação entre algoritmos supervisionados e não supervisionados (Géron, 2019).

Por fim, a categoria de aprendizado por reforço, o sistema aprende através de interações com o ambiente, ao selecionar e realizar suas ações, o agente de aprendizado obtém recompensas ou penalidades de acordo com as ações executadas. O algoritmo tem como objetivo selecionar a melhor decisão a fim de conquistar o máximo de recompensas possíveis (Géron, 2019).

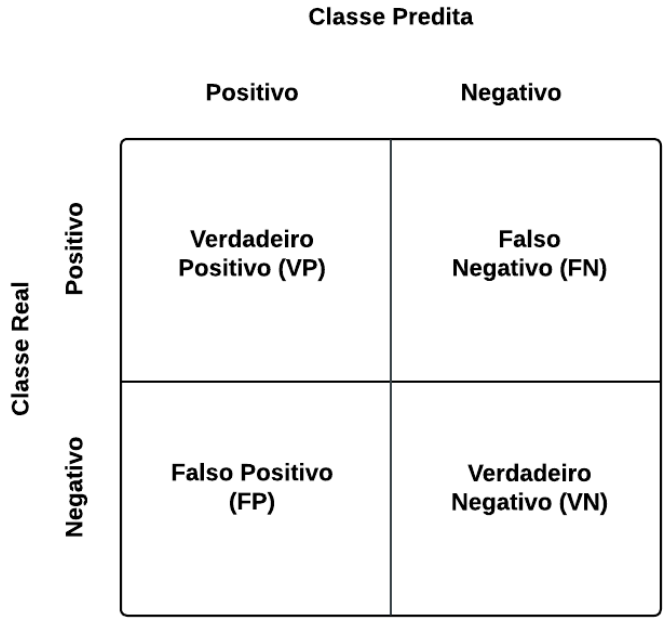
O aprendizado de máquina possui duas técnicas importantes, a classificação e a regressão. Em problemas de classificação o algoritmo deve aprender a atribuir um rótulo a um exemplo ainda não rotulado, ou seja, classificar os dados em categorias específicas. Um exemplo de classificação é a detecção de spam em e-mails, classificando eles como spam ou não spam. Já problemas relacionados à regressão estão voltados para a previsão de valores numéricos contínuos, ou seja, estimar um número real que é chamado de alvo (ou *target*) com base em dados passados. Um exemplo de regressão seria prever o preço de um imóvel com base nas suas características , como localização e número de quartos (Burkov, 2019).

Para avaliar a performance de modelos de aprendizado de máquinas, diversas ferramentas e métricas podem ser utilizadas (Burkov, 2019). Para modelos de regressão, o ideal é que os valores previstos estejam próximos dos valores reais, por exemplo, a métrica do Erro Quadrático Médio (*Mean Squared Error),* calcula a média da diferença quadrática entre o valor predito com o valor real, ou seja, quanto menor for o valor resultante da métrica, melhor é o desempenho do modelo (Burkov, 2019; Jordan, 2017; Júnior, 2021).

Para a avaliação de modelos de classificação, utilizamos a matriz de confusão (Ruback; Avila; Cantero, 2021). A matriz de confusão é uma tabela que auxilia na visualização dos resultados de um algoritmo de classificação, no qual o eixo vertical representa classe real, enquanto o eixo horizontal representa a classe predita pelo modelo. Embora a matriz de confusão não seja uma métrica de desempenho, ela é a base para o cálculo de diversas métricas (Murel; Kavlakoglu, 2024). Em uma classificação binária, a matriz é dividida em quatro partes, com os seguintes valores em cada célula : verdadeiro positivo (VP), são casos em que o modelo previu corretamente rótulos positivos. O verdadeiro negativo (VN), são casos que o modelo previu corretamente rótulos negativos. Falso positivo (FP), casos que o modelo previu um rótulo positivo, porém o resultado real é negativo. E por fim, o falso negativo (FN), são casos em que o modelo previu um rótulo negativo, porém o

resultado real é positivo (Bajaj, 2023; Foster *et al.*, 2021). A matriz de confusão é definida da seguinte forma na Figura 1:

Figura 1 - Matriz de confusão.



Fonte: Baseado em Murel e Kavlakoglu (2024)

Embora exista uma diversidade de métricas para calcular o desempenho de modelos de classificação, as duas mais frequentemente utilizadas são as de precisão e revocação (Burkov, 2019). A métrica de precisão é a proporção de casos positivos corretamente identificados, ou seja, é a razão entre os casos positivos previstos corretamente em relação ao total de casos positivos (Pagano *et al.*, 2023). É definida pela seguinte fórmula na Figura 2:

Figura 2 - Fórmula da Precisão.

$$Precisão = \frac{VP}{VP + FP}$$

Fonte: Baseado em Burkov (2019).

Já a métrica de revocação é a razão entre os casos positivos previstos corretamente em relação a todos os casos positivos que são verdadeiros (Bajaj, 2023). É definida pela seguinte fórmula na Figura 3:

Figura 3 - Fórmula da Revocação.

$$Revocação = \frac{VP}{VP + FN}$$

Fonte: Baseado em Burkov (2019).

2.2 Viés

Ao passar dos anos, a utilização de modelos de aprendizado de máquina em tomadas de decisão vem aumentando, impactando tanto indivíduos quanto a sociedade em geral (Amini *et al.*, 2019). Em decorrência desse crescimento, pesquisadores e desenvolvedores se tornaram mais atentos a possíveis vieses que podem estar presentes nesses sistemas (Mehrabi *et al.*, 2021). Vieses em sistemas de inteligência artificial, se manifestam de diversas maneiras, afetando diferentes grupos de maneiras distintas (Lee; Resnick; Barton, 2019). Diversos estudos já tratam de vieses em campos variados, como na aprovação de créditos (Kumar; Sharma; Mahdavi, 2021), na avaliação de detentos propensos à reincidência (Angwin *et al.*, 2016) e em propagandas direcionadas (Akter *et al.*, 2022). No contexto de aprendizado de máquina a definição de viés possui diversos significados (Helström; Dignum; Bensch, 2020).

Em seu estudo, Silber e Manyika (2019) definem viés como uma discriminação sistemática a grupos de indivíduos baseado nas suas características, como raça, gênero, orientação sexual, entre outras. Embora os sistemas de inteligência artificial não possuam consciência, eles podem herdar possíveis vieses dos engenheiros que modelam o *software* (Harari, 2018).

Em modelos de aprendizado de máquina, as previsões tendem a ser tão boas quanto os dados utilizados em seu treinamento (PROGRESS, 2023). A empresa norte-americana *Amazon*, por exemplo, implementou um sistema de inteligência artificial na triagem de candidatos a vagas de empregos. O sistema avaliava os currículos e os classificava através de uma nota. No entanto, foi identificado que candidatas do sexo feminino sofriam discriminação para vagas em funções técnicas. Esse enviesamento ocorreu pois o modelo foi treinado em cima dos padrões de currículos de candidatos dos últimos dez anos, que, em sua maioria, pertenciam a homens. Consequentemente, o sistema considerava candidatos homens mais aptos para ocuparem esses cargos (Reis; Graminho, 2019).

Os principais vieses que podem ser inseridos no ciclo de vida de um modelo de aprendizado de máquina são : viés histórico, que são preconceitos do passado que estão presentes nos dados de treinamento. O viés de representação, que ocorre quando a amostra coletada não representa a população adequadamente, ou seja, existem grupos sub representados. Viés de aprendizagem, que ocorre na fase de treinamento, quando há disparidade de desempenho entre grupos. O viés de avaliação ocorre quando dados utilizados para avaliar o modelo não representam grupos da população de maneira balanceada. Por fim, o viés de interpretação humana, ocorre quando há uma diferença entre o problema que um modelo foi projetado para resolver e a maneira como ele é utilizado na prática (González-Sendino *et al.*, 2023; Ruback; Avila; Cantero, 2021). O conceito de viés adotado por este trabalho é definido como a reprodução de preconceitos sociais pelos algoritmos em processos decisórios, gerando resultados que podem beneficiar ou prejudicar determinados grupos de forma desproporcional (Brissant, 2023; Gonçalves, 2023).

2.3 Trabalhos relacionados

O enviesamento e a sua detecção em modelos de inteligência artificial têm sido amplamente investigado nos últimos anos. Diversos estudos se concentram em investigar a causa, os impactos sociais e formas de mitigação desse enviesamento. Por exemplo, Simões-Gomes, Roberto e Mendonça (2020) realizam um balanço bibliográfico de artigos sobre viés algoritmo nas línguas portuguesa, inglesa e espanhola, onde detectam que este termo não possui uma definição clara. Além disso, identificaram que no aprendizado de máquina as maiores causas para esse fenômeno se encontram na construção do modelo e na utilização de dados enviesados.

Outro trabalho relevante que debruça sobre vieses é o de Avila, Ruback, e Cantero (2021), que realizam um estudo de caso no reconhecimento facial, onde detalham quais tipos de vieses podem ser inseridos nas etapas de coleta de dados, pré-processamento, criação do modelo, na sua avaliação e nos pós-processamento. Além disso, abordam o impacto social do enviesamento e como modelos de reconhecimento facial vem reforçando preconceitos existentes na sociedade. Por

outro lado, o estudo realizado por Cozman e Kaufman (2022) aborda o aprendizado de máquina baseado em redes neurais de aprendizado profundo, discutindo os tipos de vieses, incluindo o viés na escolha dos desenvolvedores, que são responsáveis por identificar o problema a ser resolvido e traduzi-los em variáveis observáveis e manipuláveis. Além disso, o trabalho apresenta meios de mitigação de vieses e as analisa.

Por fim, o trabalho realizado por Vieira (2019), que aborda o caso Compas, um programa utilizado no sistema judiciário americano para avaliar o risco de reincidência de réus em processos criminais. Onde o estudo destaca os vieses encontrados no sistema, as possíveis reparações dos enviesamentos detectados e os impactos sociais na utilização de algoritmos preditivos. Também há o estudo de Reis e Graminho (2019), que aborda a utilização da inteligência artificial em processos seletivos no caso da *Amazon*. Onde o sistema promovia uma discrimação de gênero contra mulheres que se candidataram para determinados cargos. Além disso, o estudo discute como a utilização de dados de trabalhadores infringe direitos básicos dos mesmos, como à privacidade e a igualdade, assim como aborda uma análise da legislação vigente e possíveis melhorias.

3 METODOLOGIA

O presente trabalho tem como objetivo fornecer um guia acessível que auxilie no entendimento de processos e ferramentas utilizadas na detecção de viés em modelos de aprendizado de máquina, com foco em áreas sensíveis de tomada de decisão. Para isso, o desenvolvimento do estudo ocorreu por meio de uma revisão bibliográfica abrangente, como buscas em bases de dados acadêmicos, seleção de artigos relevantes através de palavras-chave, análise de citações de estudos sobre o tema e trabalhos relacionados ao tema.

A princípio, foi realizada uma revisão bibliográfica exploratória não sistemática devido à gama e a diversidade de estudos sobre o tema. Foram realizadas buscas nas bases de dados das plataformas do *Google Scholar* e *Scielo*, utilizando palavras específicas como: aprendizado de máquina, viés, métricas de equidade, equidade, ferramentas e detecção de viés. Os materiais incluíram trabalhos científicos em inglês e português com foco nos publicados nos últimos seis anos. Além de notícias e artigos em portais, livros no formato de *ebooks* e outros documentos relevantes sobre o tema.

Em seguida, foi determinado as ferramentas a serem analisadas com base na frequência das mesmas em publicações em artigos científicos, sendo elas: *AI Fairness 360* (*IBM*), *Fairlearn* (*Microsoft)*, *Aequitas* (*UChicago*). Para cada ferramenta, foi realizado um levantamento de estudos de caso, com foco em um estudo representativo para cada ferramenta selecionada. O critério de seleção dos estudos foi a aplicação dessas ferramentas em modelos empregados em áreas sensíveis, como justiça, mercado de trabalho e crédito bancário. Essas áreas foram escolhidas devido aos potenciais danos que decisões enviesadas podem ocasionar a grupos vulneráveis e indivíduos.

Por fim, ao término do trabalho, será realizada uma análise comparativa das ferramentas e métricas utilizadas, com destaque para as principais características de cada ferramenta, incluindo as métricas de avaliação suportadas e a facilidade de uso.

4 ABORDAGENS EMPREGADAS NA DETECÇÃO DE VIÉS

Ao abordarmos métodos na detecção de viés em modelo de aprendizado de máquina, encontramos uma dificuldade decorrente da falta de transparência dos algoritmos, sua crescente complexidade e limitações técnicas relacionadas à interpretabilidade. Desse modo, os pesquisadores frequentemente precisam analisar os resultados dos algoritmos para detectar enviesamentos (Fu; Huang; Singh, 2020). Em auditorias realizadas nesses modelos, a abordagem utilizada para a detecção de possíveis vieses em dados, algoritmos e sistemas são através de medidas específicas de equidade (Orphanou *et al.*, 2022).

Diante disso, muitos pesquisadores desenvolveram abordagens relacionadas a métricas de equidade, que tem como objetivo mensurar de maneira quantitativa a equidade dos modelos através do seu desempenho entre diferentes grupos (Mikołajczyk-Bareła; Grochowski, 2023). Essas métricas podem ser divididas em duas categorias, equidade de grupo e equidade individual (Chakraborty; Majumder; Menzies, 2021; Chouldechova; Roth, 2018).

A equidade de grupo foca em garantir que as previsões ou decisões de um modelo de aprendizado de máquina sejam iguais ou similares entre grupos sensíveis. Ou seja, grupos protegidos e não protegidos devem ser tratados de maneira justa. Enquanto a equidade individual foca em garantir que indivíduos que possuam características semelhantes recebam resultados semelhantes, mesmo que pertençam a grupos distintos (Stevens, 2020; Xu; Strohmer, 2024).

Em métricas de equidade individual, a métrica *counterfactual*, por exemplo, garante que, mesmo que um atributo sensível de um indivíduo seja alterado, e preservando os demais atributos que não dependem deste atributo sensível, a saída do modelo deve permanecer a mesma, ou seja, garante que o modelo não depende de atributos sensíveis nas previsões (Caton; Haas, 2024).

A fim de abordar as métricas de equidade de grupo, as seguintes definições métricas estatísticas de equidade devem ser abordadas (Foster *et al.*, 2021; Prates, 2021; Verma; Rubin, 2018).

A métrica de *False Discovery Rate (FDR)* é a razão de indivíduos que o modelo prevê como positivos, mas na verdade são negativos. Ou seja, é a proporção de predições positivas feitas por um modelo que na verdade são negativas (Verma; Rubin, 2018). É definida pela seguinte fórmula na Figura 4:

Figura 4 - Fórmula de FDR.

$$FDR = \frac{FP}{VP + FP}$$

Fonte: Baseado em Foster et al. (2021).

A métrica de *Positive Predictive Value (PPV)* é a razão dos indivíduos modelo prevê como verdadeiro positivo, em relação a todos os casos positivos. Também chamado de precisão (Verma; Rubin, 2018; Prates, 2021). É definida pela seguinte fórmula na Figura 5:

Figura 5 - Fórmula da PPV.

$$PPV = \frac{VP}{VP + FP}$$

Fonte: Baseado em Burkov (2019) e Verma e Rubin (2018).

A métrica de *False Omission Rate (FOR)* é definida pela razão dos indivíduos que o modelo prevê como falsos negativos, em relação a todos os casos negativos. Ou seja, a probabilidade de um caso previsto como negativo seja na verdade positivo (Foster *et al.*, 2021). É definida pela seguinte fórmula na Figura 6:

Figura 6 - Fórmula de FOR.

$$FOR = \frac{FN}{VN + FN}$$

Fonte: Baseado em Foster et al. (2021).

A métrica de *False Positive Rate (FPR)* é definida pela razão de indivíduos que o modelo prevê como falsos positivos, em relação a todos os casos negativos (Foster *et al.*, 2021; Prates, 2021). É definida pela seguinte fórmula na Figura 7:

Figura 7 - Fórmula de FPR.

$$FPR = \frac{FP}{FP + VN}$$

Fonte: Baseado em Foster et al. (2018).

A métrica de *Negative Predictive Values (NPV)* é a razão de indivíduos que o modelo prevê como verdadeiros negativos, em relação a todos os casos negativos. (Verma; Rubin, 2018). É definida pela seguinte fórmula na Figura 8:

Figura 8 - Fórmula de NPV.

$$NPV = \frac{VN}{VN + FN}$$

Fonte: Baseado em Verma e Rubin (2018).

A métrica de *True Positive Rate (TPR)* é a razão de indivíduos que o modelo prevê como verdadeiros positivos, em relação a todos os casos positivos. É a probabilidade do modelo identificar um caso verdadeiramente positivo. Também chamada de revocação (Foster *et al.*, 2021; Verma; Rubin, 2018; Prates, 2021). É definida pela seguinte fórmula na Figura 9:

Figura 9 - Fórmula de TPR.

$$TPR = \frac{VP}{VP + FN}$$

Fonte: Baseado em Verma e Rubin (2018).

A métrica de *False Negative Rate (FNR)* é a razão de indivíduos que o modelo prevê como falso negativos, em relação a todos os casos que são positivos. É a probabilidade do modelo identificar um caso verdadeiramente positivo seja incorretamente previsto como negativo (Foster *et al.*, 2021). É definida pela seguinte fórmula na Figura 10:

Figura 10 - Fórmula do FNR.

$$FNR = \frac{FN}{FN + VP}$$

Fonte: Baseado em Foster et al. (2021).

A métrica de *True Negative Rate (TNR)* é a razão de indivíduos que o modelo prevê como verdadeiros negativos, em relação a todos os casos que são negativos.

É a probabilidade de um caso verdadeiramente negativo seja identificado como negativo pelo modelo (Verma; Rubin, 2021; Prates, 2021). É definida pela seguinte fórmula na Figura 11:

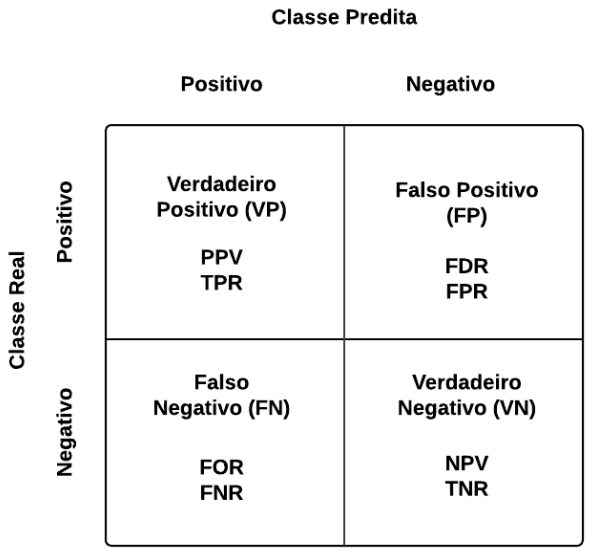
Figura 11 - Fórmula do TNR.

$$TNR = \frac{VN}{VN + FP}$$

Fonte: Baseado em Verma e Rubin (2018).

As métricas estatísticas abordadas se organizam da seguinte forma na matriz de confusão na Figura 12:

Figura 12 - Matriz de confusão com as métricas.



Fonte : Baseado em Verma e Rubin (2018).

Para a detecção de vieses em modelos de aprendizado de máquina diversas métricas de equidade foram desenvolvidas e aplicadas em prática. Entre as mais populares estão *Equalized Odds, Equality Of Opportunity, Demographic Parity, Disparate Impact* (González-Sendino, 2023; Pagano *et al.*, 2023).

A métrica de *Equalized Odds* refere-se a exigência de que as taxas de FPR e TPR sejam iguais para grupos protegidos e não protegidos. Ou seja, a probabilidade de uma pessoa na classe positiva receber corretamente um resultado positivo deve ser a mesma para todos os grupos, e a probabilidade de uma pessoa na classe negativa receber incorretamente um resultado positivo também deve ser igual para

todos os grupos (Hardt; Price; Srebro, 2016; Mehrabi *et al.*, 2021). É definida pela seguinte fórmula na Figura 13:

Figura 13 - Fórmula *Equalized Odds.*

$$EO = (FPR_{GP} - FPR_{GNP}) + (TPR_{GP} - TPR_{GNP})$$

Fonte: Baseado em González-Sendino (2023).

A métrica de *Equality of Opportunity* refere-se a exigência de que a taxa de TPR seja igual para grupos protegidos e não protegidos. Ou seja, a probabilidade de uma pessoa na classe positiva receber um resultado positivo é a mesma para grupos protegidos e não protegidos (Hardt; Price; Srebro, 2016; Pagano *et al.*, 2023). É definida pela seguinte fórmula na Figura 14:

Figura 14 - Fórmula *Equality of Opportunity*.

$$EOO = TPR_{GP} - TPR_{GNP}$$

Fonte: Baseado em Pagano et al. (2023) e González-Sendino (2023).

A métrica de *Demographic Parity* garante que a probabilidade de um resultado positivo seja a mesma para diferentes grupos, independentemente das características individuais desses grupos. Ou seja, assegura que a proporção de indivíduos que recebem um resultado positivo seja igual entre grupos protegidos e não protegidos (Mehrabi *et al.*, 2021). É definida pela seguinte fórmula na Figura 15:

Figura 15 - Fórmula *Demographic Parity.*

$$DP = \frac{TP + FP}{N}$$

Fonte: Baseado em Pagano et al. (2023) e González-Sendino (2023).

A métrica de *Disparate Impact* é uma métrica usada para avaliar se um modelo trata grupos de maneira justa, comparando a razão entre a taxa de resultado de favoráveis entre dois grupos, um grupo protegido e um grupo não protegido, para ser justo o resultado tem que ser igual ou próximo a 1 (González-Sendino, 2023). Pode ser definida pela seguinte fórmula na Figura 16:

Figura 16 - Fórmula *Disparate Impact.*

$$DI = \frac{\frac{VP_{GP} + FP_{GP}}{N_{GP}}}{\frac{VP_{GNP} + FP_{GNP}}{N_{GNP}}}$$

Fonte: Baseado em Pagano et al. (2023).

4.1 Ferramentas utilizadas na detecção de viés

Esta seção aborda ferramentas utilizadas para a detecção de vieses em modelos de aprendizado de máquina. Para este trabalho foram selecionadas três ferramentas amplamente reconhecidas na literatura devido à sua aplicação em diferentes cenários e ao suporte oferecido para diversas métricas e algoritmos de mitigação, sendo elas: *AI Fairness 360* (*Microsoft*), *Fairlearn* (*Microsoft*) e *Aequitas* (*UChicago*). Essa escolha foi baseada em estudos que destacam sua relevância, popularidade e características (Balayn *et al.*, 2023; Deng *et al.*, 2022; González-Sendino *et al.*, 2023; Lee; Singh, 2021; Pagano *et al.*, 2023; Richardson;Gilbert, 2021), além de critérios voltados para documentação robusta, abrangência e acessibilidade. A Tabela 1 demonstra algumas características de cada ferramenta levantadas nos estudos analisados.

Tabela 1 - Características das ferramentas.

Ferramenta	**Número de citações em estudos**	**Organização**	**Código Aberto**	**Métricas Equidade de Grupo**	**Métricas de Equidade Individual**	**Algoritmos de Mitigação**
AI Fairness 360	**6**	**IBM**	**X**	**X**	**X**	**X**
Fairlearn	**6**	**Microsoft**	**X**	**X**		**X**
Aequitas	**6**	**Universidade de Chicago**	**X**	**X**		

Fonte: Elaborado pelo autor, 2024.

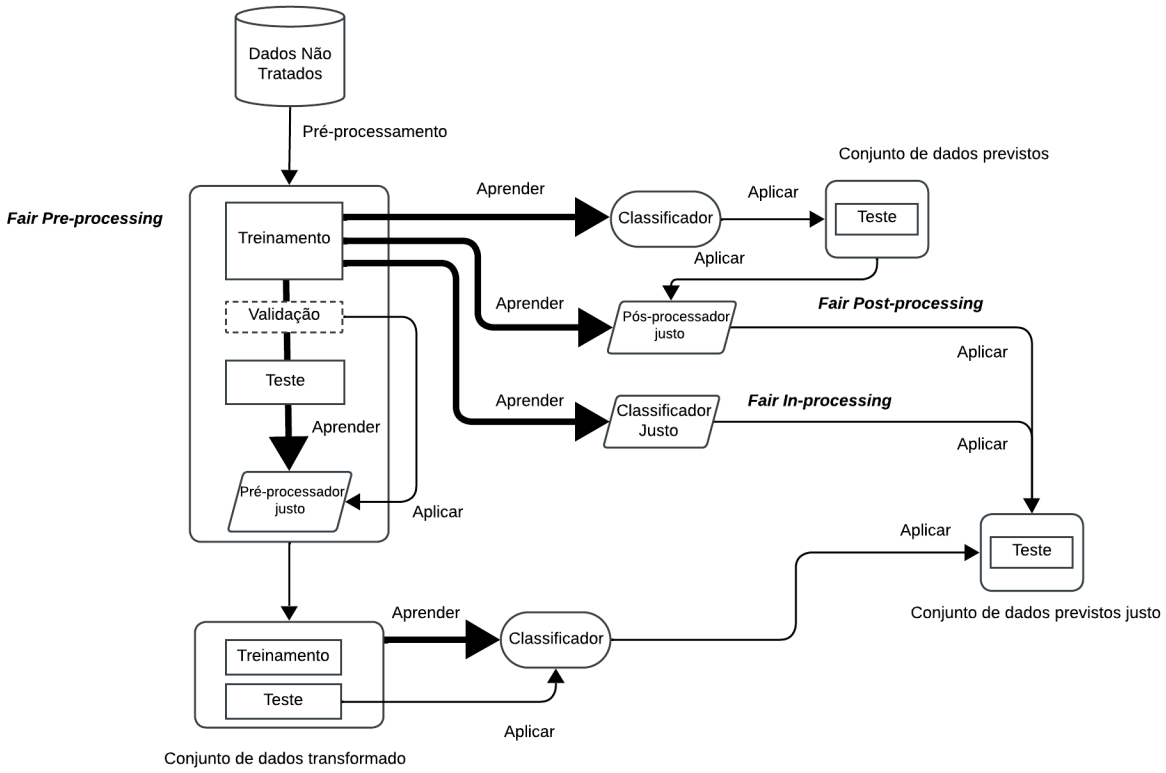
4.1.1 AI Fairness 360

A *AI Fairness 360* é um kit de ferramentas desenvolvido pela IBM de código aberto para a detecção e mitigação de viés. Sendo o primeiro sistema de código aberto a fornecer ferramentas para medir, entender e mitigar vieses em modelos de aprendizado de máquina. (Bellamy *et al.*, 2018; Hufthammer *et al.*, 2020).

Este kit de ferramentas tem como objetivo auxiliar no entendimento de métricas de equidade, conectar a comunidade de pesquisadores e facilitar a adoção das soluções em situações reais (Bellamy *et al.*, 2018). Em seu pacote inicial disponibiliza mais de 70 métricas de detecção de viés, 9 algoritmos de mitigação e recursos de explicação para os usuários, auxiliando no entendimento dos resultados obtidos (Bellamy *et al.*, 2018; Deng *et al.*, 2022). As métricas disponibilizadas pela ferramenta, podem ser divididas em métricas de equidade de grupo, equidade individual e de performance (Richardson; Gilbert, 2021).

A ferramenta *AI Fairness 360* foi projetada para ter um fluxo de trabalho ponta a ponta com os objetivos de extensibilidade e facilidade de uso. Assim, promovendo uma facilidade no processo de desenvolvimento, desde a manipulação de dados até criação de um modelo justo. (Bellamy *et al.*, 2018).

Figura 17 - Pipeline de mitigação.



Fonte: Baseado em Harris (2020) e Bellamy et al. (2018).

As estratégias de mitigação que o kit disponibiliza, podem ser classificadas em três categorias e utilizadas em diferentes etapas do *pipeline*, sendo elas : pré-processamento, processamento e pós-processamento (Richardson; Gilbert, 2021). A Figura 17 é um exemplo simplificado do *pipeline* de mitigação de viés, no qual cada retângulo representa um conjunto de dados. O fluxo dele pode ser exemplificado da seguinte forma : o primeiro passo é o carregamento dos dados, a fim de organizá-los. Em seguida, aplicar um algoritmo de pré-processamento, transformando os dados a fim de reduzir possíveis enviesamentos. Após isso, os dados transformados são utilizados para treinar um modelo de aprendizado de máquinas e obter as previsões deste modelo (Bellamy *et al.*, 2018; Harris, 2020).

O pipeline da *AI Fairness 360* é flexível, permitindo que cada etapa do pipeline possa ser modificada para as necessidades específicas do projeto ou do modelo, garantindo que as medidas de equidade sejam utilizadas de acordo com o contexto (Bellamy *et al.*, 2018).

A *AI Fairness 360* dispõe de uma grande variedade de métricas de fairness para conseguir auxiliar em diversos tipos de contextos, por isso o projeto foi desenvolvido para ser de código aberto, com a finalidade de que desenvolvedores e

pesquisadores possam contribuir. Desse modo, este trabalho irá abordar algumas métricas importantes.

Segundo Arcolezi, Makhlouf e Palamidessi (2023), a métrica de *Statistical Parity Difference* é a diferença na proporção de previsões positivas para ambos os grupos, protegidos e não protegidos, onde um resultado igual a 0 indica que ambos os grupos têm a mesma proporção de previsões positivas. Pode ser definida pela seguinte fórmula na Figura 18:

Figura 18 - Fórmula de *Statistical Parity Difference*.

$$SPD = \frac{VP_{GP} + FP_{GP}}{N_{GP}} - \frac{VP_{GNP} + FP_{GNP}}{N_{GNP}}$$

Fonte: Baseado em Pagano et al. (2023) e González-Sendino (2023).

A métrica de *Average Odds Difference* de acordo com Hufthammer et al. (2020), é a média da diferença absoluta entre as $TPRs$ e $FPRs$ dos grupos protegidos e não protegidos, o valor igual a 0 significa uma equidade ideal. Caso o resultado seja negativo é indício que o grupo não protegido sofre preconceito. Pode ser definida pela seguinte fórmula na Figura 19:

Figura 19 - Fórmula de *Average Odds Difference.*

$$AOD = \frac{(FPR_{GP} - FPR_{GNP}) + (TPR_{GP} - TPR_{GNP})}{2}$$

Fonte: Baseado em Pagano et al. (2023).

Por fim, a métrica de *Equal Opportunity Difference* que segundo González-Sendino et al. (2023), é utilizada para medir o desvio da igualdade de oportunidades, medindo a diferença na TPR (*True Positive Rate)* entre um grupo protegido e não protegido, ou seja, compara a probabilidade do modelo prever corretamente um resultado positivo para ambos os grupos. No qual, para o modelo ser considerado justo o resultado deve ser igual a 0. Pode ser definida pela seguinte fórmula na Figura 20:

Figura 20 - Fórmula *Equal Opportunity Difference.*

$$EOD = TPR_{GP} - TPR_{GNP}$$

Fonte: Baseado em González-Sendino (2023).

A *AI Fairness 360* possui diversos algoritmos de mitigação que são separados em categorias de acordo com onde são implementados no processo. Sendo eles, algoritmos de pré-processamento, processamento e pós-processamento (Harris, 2020). Os algoritmos de pré-processamento atuam na modificação dos dados utilizados no treinamento do modelo. Enquanto os algoritmos de processamento atuam no treinamento do modelo remodelando o seu processo de aprendizagem. Por fim, os algoritmos de pós-processamento atuam como uma caixa preta no modelo já treinado (Bellamy *et al.*, 2018). Devido ao número de algoritmos presentes na ferramenta, este trabalho optou por selecionar um de cada categoria.

Uma das abordagens presentes na categoria de pré-processamento é o método de *reweighing* (Bellamy *et al.*, 2018). De acordo com Kamiran e Calders (2012), este método ajusta pesos diferentes para cada instância com base em suas categorias de atributos protegidos e nos resultados, com o objetivo de reduzir vieses no conjunto de dados utilizados no treinamento do modelo. Consequentemente, isso resulta em métricas como *Disparate Impact* próxima de 1 e *Statistical Parity Difference* próxima de 0, apontando uma maior equidade entre os grupos.

Na categoria de algoritmos de processamento, o algoritmo de *Adversarial Debiasing* é um método que envolve dois modelos distintos, um preditivo e um adversário (Bellamy *et al.*, 2018) . O modelo preditivo, tem a tarefa de prever com precisão a variável de interesse. Em contrapartida, o segundo modelo, chamado de adversário, é treinado para prever um atributo sensível baseado nas previsões do primeiro modelo. Desse modo, o primeiro modelo tende a fazer suas previsões não tão influenciadas pelos atributos sensíveis, a fim de que esses atributos não sejam detectados pelo modelo adversário. Consequentemente, diminuindo o viés do modelo (Zhang *et al.*, 2018).

O algoritmo de *Reject Option Classification* é uma das abordagens presentes na categoria de algoritmos de pós-processamento (Bellamy *et al.*, 2018). Segundo Kamiran et al. (2012), esse método analisa as previsões e identifica aquelas que têm baixa confiabilidade, ou seja, aquelas que estão próximas do limiar de classificação,

ao identificar essas previsões o algoritmo opta por “rejeitar” a previsão original e reclassifica essas previsões atribuindo resultados a grupos protegidos e não protegidos com o objetivo de mitigar enviesamentos.

Em seu estudo, Harris (2020) conduz um experimento para examinar discriminações relacionadas a atributos como gênero, idade e raça, bem como os seus impactos em modelos de inteligência artificial utilizados para seleção de candidatos a vagas de emprego. Foi constatado que os candidatos mais velhos possuíam uma taxa de seleção menor, ou seja, a idade exerceu uma influência maior em comparação com os outros atributos no modelo criado. Uma das abordagens escolhidas para análise e mitigação foi o kit de ferramentas *AI Fairness 360*, no qual foi implementado um método de cada categoria de algoritmos de pré-processamento, processamento e pós-processamento, resultando na diminuição do viés em relação à idade. Todavia, em um cenário real com uma maior variação nos candidatos, a utilização desta ferramenta não seria tão prática e aplicável para a maioria das empresas.

4.1.2 Fairlearn

O *Fairlearn* é um kit de ferramentas de código aberto inicialmente desenvolvido pela Microsoft com o objetivo de auxiliar pesquisadores, desenvolvedores e empresas a melhorar e avaliar a equidade de modelos de inteligência artificial. Ele visa mitigar os impactos negativos relacionados à equidade que modelos podem ter sobre determinados grupos da sociedade, semelhantes aos que figuram em questões de raça, gênero ou situação de deficiência (Bird *et al.*, 2020; Deng *et al.*, 2022).

A ferramenta consiste em uma biblioteca em Python, que fornece uma API de fácil utilização que é capaz de se integrar com outras bibliotecas do ecossistema da linguagem, sendo elas: *pandas, matplotlib, scikit-learn, TensorFlow* e *Pytorch* (Weerts *et al.*, 2023). O desenvolvimento do *Fairlearn* foi fundamentado na perspectiva de que o viés é um desafio sociotécnico, ou seja, envolve tanto aspectos tecnológicos quanto sociais (Dudík *et al.*, 2020). Nesse sentido, no contexto do Fairlean, o entendimento de equidade se dá em termos de possíveis “danos” causados a determinados grupos, ao invés de atribuir a conceitos específicos de viés (Gayhardt *et al.*, 2024). Diante disso, o *Fairlearn* opta por definir a injustiça de

um modelo com base nos impactos negativos em determinados grupos de indivíduos, definindo dois tipos comuns de danos : danos de alocação e danos de qualidade de serviço (Deng *et al.*, 2022; Gayhardt *et al.*, 2024).

Danos de alocação são situações em que o modelo de inteligência artificial reduz o acesso de benefícios, recursos ou oportunidades entre diferentes grupos. Por exemplo, um sistema de promoção de cargos que recomenda mais candidatos homens do que mulheres. (Bird *et al.*, 2020)

Danos de qualidade de serviço são situações em que o modelo de inteligência artificial possui uma menor qualidade de previsões referente a diferentes grupos, ainda que o modelo esteja alocando as oportunidades de maneira balanceada, o modelo pode ser menos preciso ou cometer mais erros para alguns grupos em comparação com outros. Por exemplo, um sistema de reconhecimento facial que possui uma menor precisão para asiáticos do que para pessoas brancas (Bird *et al.*, 2020). Embora haja semelhança entre os dois tipos, danos de alocação se refere a quem recebe uma oportunidade ou recurso, enquanto os danos de qualidade de serviço refere-se a quão bem o modelo funciona para diferentes grupos. (Emanuilov; Yordanova, 2020; Weerts *et al.*, 2023).

Segundo Barocas et al. (2021) Métricas de desempenho agregadas, embora úteis para a avaliação do desempenho geral de um modelo, podem esconder problemas que afetam grupos menos representados nos dados de treinamento. Um modelo avaliado de forma global pode apresentar uma boa acurácia. Porém, se grupos específicos relacionados à raça, gênero ou idade estiverem sub-representados nos dados, o modelo pode apresentar um pior desempenho ao se tratar deles.

Nesse sentido, para a avaliação de equidade dos modelos, o *Fairlearn* adota métricas de desagregação. A ideia por trás dessa abordagem é avaliar a performance do modelo de aprendizado de máquina separadamente para os diferentes grupos presentes nos dados. Assim, permitindo identificar se o modelo está performando de maneira injusta ou prejudicando algum grupo em específico (MICROSOFT DEVELOPER, 2020a; MICROSOFT DEVELOPER, 2020b; Weerts *et al.*, 2023).

As principais métricas de equidade que estão inclusas no kit de ferramentas do *Fairlearn* são *equalized odds, demographic parity e equal opportunity* (Lee e Singh, 2021). Todas já explicadas em uma seção anterior deste artigo.

A biblioteca do *Fairlearn* inclui diversos algoritmos para mitigação de viés. Muitos desses métodos são definidos como meta-algoritmos, pois atuam como uma camada externa que visa corrigir ou mitigar possíveis vieses, sem modificar o algoritmo do modelo (MICROSOFT DEVELOPER, 2020a). Podemos dividir os algoritmos presentes no kit de ferramentas em três categorias, sendo elas: pré-processamento, pós-processamento e de processamento (Pandey, 2023).

O algoritmo presente na categoria de pré-processamento opera na transformação dos dados utilizados no treinamento a fim de atenuar possíveis enviesamentos (D'alessandro, 2019). O *Correlation Remover* é um algoritmo utilizado com o objetivo de reduzir ou eliminar a correlação entre atributos sensíveis e outros atributos do conjunto de dados, assim garantindo que estes atributos sensíveis não influenciam nas previsões do modelo (Weerts *et al.*, 2023).

No processamento os algoritmos desta categoria atuam no processo de treinamento do modelo (D'alessandro, 2019). O *Fairlearn* adota os algoritmos de *AdversarialFairnessClassifier* e *AdversarialFairnessRegressor.* Ambos os algoritmos treinam dois modelos. O modelo preditivo tem como objetivo fazer previsões corretas, evitando que o modelo adversário consiga identificar atributos protegidos (como raça, gênero, idade, etc). Enquanto o modelo adversário tem como objetivo detectar atributos sensíveis a partir das previsões do modelo preditivo (FAIRLEARN, 2023).

A categoria de pós-processamento atua nas previsões ou saídas do modelo já treinado (D'alessandro, 2019). No *Fairlearn*, esta categoria disponibiliza apenas o algoritmo de *Threshold Optimizer,* que tem como objetivo modificar as previsões do modelo ajustando o limiar de classificação para cada grupo sensível. Esse ajuste visa satisfazer os critérios de equidade estabelecidos previamente, como a métrica de *true positive rate parity (TPR)*. Contudo, esse processo pode acarretar em uma queda no desempenho do modelo, como uma diminuição na acurácia (FAIRLEARN, 2023).

O kit de ferramentas do *Fairlearn* também possibilita a comparação de diversos modelos através de um *dashboard,* onde cada modelo estará representado como um ponto em um gráfico de dispersão, com base em uma métrica de desempenho geral e uma métrica de equidade definida previamente. Esse recurso permite que o desenvolvedor examine cada modelo e identifique qual supre melhor suas necessidades em relação ao desempenho e equidade (Weerts *et al.*, 2023)

Em seu trabalho, Dudík et al. (2020) aborda como um modelo de aprendizado de máquina treinado por algoritmos padrões pode conter vieses em um cenário de análise de empréstimo bancário. O modelo tem como objetivo prever a probabilidade de inadimplência de um cliente que solicita um empréstimo ao banco, os clientes que possuem uma probabilidade de inadimplência acima de um limite determinado são rejeitados, enquanto clientes que possuem uma probabilidade de inadimplência abaixo do limite são considerados aptos à aprovação do empréstimo. Foi detectado um enviesamento relacionado ao sexo dos candidatos, causando danos que afetam grupos de homens e mulheres. Onde, o modelo favorece mulheres inadimplentes e desfavorece homens que poderiam receber o empréstimo.

Para a avaliação de equidade uma das métricas escolhidas foi a de *equalized odds difference.* Enquanto uma das abordagens escolhidas para mitigação foi o algoritmo de pós-processamento *Threshold Optimizer,* onde após sua implementação resultou em uma diminuição do viés, com pouco impacto do desempenho do modelo.

4.1.3 Aequitas

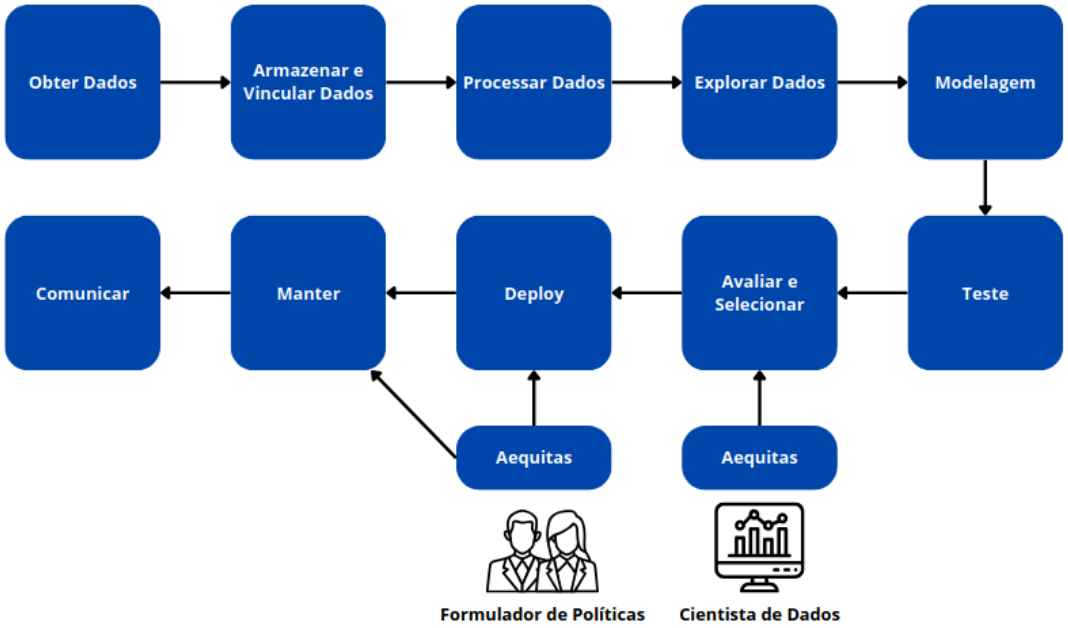
Lançado em 2018, o *Aequitas* é um kit de ferramentas de código aberto desenvolvido pelo Centro de Ciência de Dados e Políticas Públicas da Universidade de Chicago, criado com o objetivo de auxiliar na auditoria de viés, equidade e transparência em modelos de aprendizado de máquina (Ansari, 2024; Bilro, 2021; Venugopal *et al.*, 2023). A ferramenta possibilita que os usuários analisem seus modelos sob diversas perspectivas de vieses e métricas de equidade, gerando um relatório ao final, auxiliando nas tomadas de decisão sobre o desenvolvimento e implementação desses modelos (Bilro, 2021; Saleiro *et al.*, 2018).

Visando a acessibilidade a uma maior gama de usuários, não apenas a cientistas de dados, o *Aequitas* é uma ferramenta que pode ser utilizada de três maneiras : através de uma aplicação web para usuários considerados não técnicos, por meio da biblioteca em Python ou por uma interface de linha de comando (Foster *et al.*, 2021).

O *Aequitas* foi desenvolvido principalmente para dois tipos principais de usuários, sendo eles : cientistas de dados, que, durante o processo de construção e

escolha de modelos, vão utilizar a ferramenta para comparar medidas de equidade e verificar disparidades entre os modelos que estão criando. E o formulador de políticas (*policymaker*), que, antes de validar o uso do modelo, utilizam o Aequitas para entender que tipos de vieses podem existir no modelo e quais ações tomar para mitigá-los. Além disso, é necessário um monitoramento periódico para garantir a equidade do modelo ao longo do tempo (Foster *et al.*, 2021).

Figura 21 - Aequitas acoplado no contexto do ciclo de vida do aprendizado de máquina.



Fonte: Baseado em Saleiro et al. (2018).

Nesse sentido, a Figura 21 ilustra em que etapa o *Aequitas* e seus tipos de usuários atuam no ciclo de vida do aprendizado de máquina. A ferramenta visa estabelecer um padrão de auditagem, promovendo que seus usuários sempre tomem decisões relacionadas a escolha do modelo, o uso do modelo e a possível necessidade de um novo treinamento levando em consideração o viés e a equidade (DATA SCIENCE FOR SOCIAL GOOD, 2018; Saleiro *et al.*, 2018)

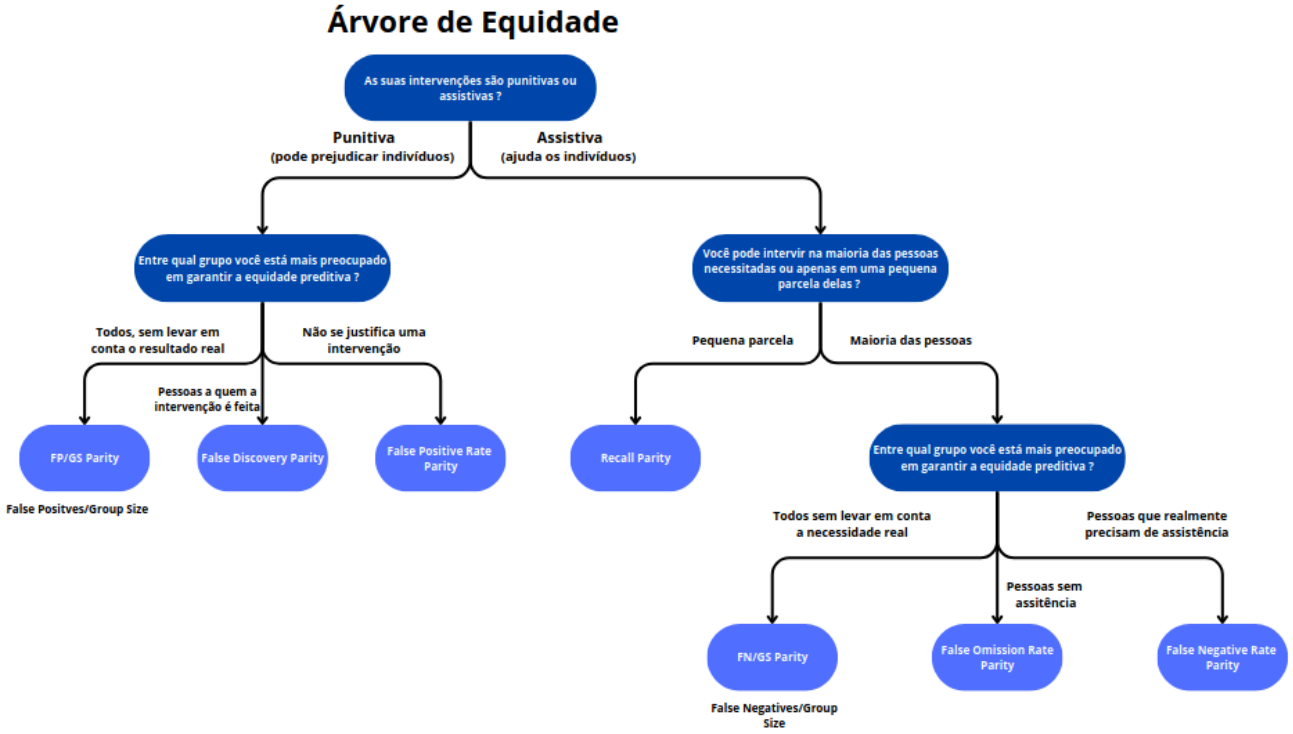
Os modelos de inteligência artificial que auxiliam seus usuários a tomarem decisões relacionadas a problemas de políticas públicas possuem características comuns entre si, uma delas é a implementação da métrica de precisão para os top “k”, ou seja, a capacidade do modelo de acertar os principais “k” casos mais importantes. Essa abordagem é essencial para priorizar os casos que necessitam de intervenção, uma vez que os recursos para intervir são limitados. O funcionamento

do modelo se dá da seguinte forma, o modelo prevê uma pontuação de risco para cada entidade, as que possuírem o maior risco são classificadas no topo. Normalmente, é alocado um humano para participar do processo revisando as classificações, e, por fim, as entidades que estão no top “k” são as escolhidas para a intervenção (Saleiro *et al.*, 2018; Rahul; Yao, 2019).

Existem dois tipos de intervenção, a assistiva e punitiva (PYDATA, 2020). De acordo com Foster et al. (2018), a intervenção assistiva visa ajudar os indivíduos que possam ter sido erroneamente excluídos do top “k”, ou seja, classificados de baixo risco quando deveriam receber a intervenção, assim focando principalmente na taxa de falsos negativos (FN). Enquanto, na intervenção punitiva o foco é evitar que os indivíduos possam ser erroneamente incluídos no top “k”, assim focando principalmente com a taxa de falsos positivos (FP) (PYDATA, 2020).

Diante da ampla gama de métricas disponibilizadas pelo *Aequitas* que podem ser utilizadas para a auditoria de viés e equidade dos modelos, a escolha de qual abordagem adotar em relação a um determinado contexto, pode se mostrar um processo confuso com uma navegação difícil (Lee e Singh, 2021). Em decorrência disso, o *Aequitas* disponibiliza uma árvore de equidade que pode ser entendida como um *framework* que tem como objetivo auxiliar os usuários a navegar através de todas as métricas disponíveis na ferramenta. Seu fluxo é guiado através de perguntas específicas direcionadas ao usuário, que ao responder o encaminha para as métricas que se encaixam no contexto do seu problema (DATA SCIENCE FOR SOCIAL GOOD, 2020; Foster *et al.*, 2018; Lee; Singh, 2021). Conforme pode ser observado na Figura 22:

Figura 22 - Árvore de Equidade.



Fonte: Baseado em Rodolfa et al. (2020).

As métricas presentes no *Aequitas* podem ser separadas em *Equal Parity* também conhecido como *Demographic Parity, Proportional Parity, False Positive Parity* e *False Negative Parity.*

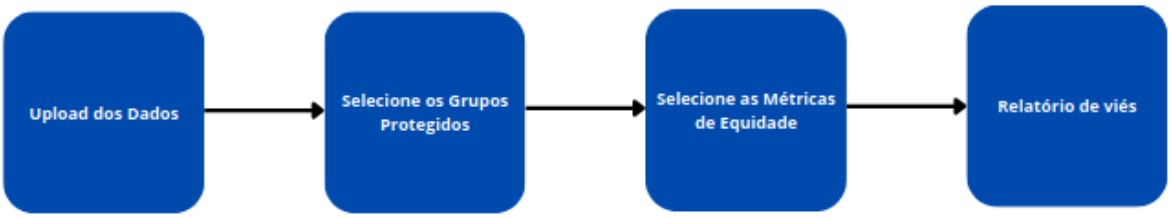
A implementação do *Aequitas* para auditar modelos de aprendizado de máquina pode ser dividido em 4 etapas (DATA SCIENCE FOR SOCIAL GOOD, 2018), sendo elas :

- *Upload* dos dados : O primeiro passo é fazer o *upload* de um arquivo dos dados (exemplo um arquivo CSV) contendo as colunas de previsões e os rótulos.
- Seleção de grupos protegidos : A segunda etapa é selecionar os grupos protegidos que serão analisados para identificar possíveis desigualdades. Podem ser definidos por raça, idade ou gênero.
- Seleção de métrica de equidade : A próxima etapa é selecionar quais métricas serão utilizadas para auditar o modelo, a árvore de equidade pode ser utilizada para auxiliar na escolha das métricas baseado na necessidade do usuário.

- Relatório do viés : Na última etapa a ferramenta fornece um relatório onde podemos explorar e detalhar o modelo por diferentes métricas de equidade e por subgrupos.

Conforme pode ser observado na Figura 23 :

Figura 23 - Pipeline de avaliação do Aequitas.



Fonte: Baseado em Aequitas (2018).

No estudo de Rodolfa et al. (2020), é discutida a colaboração entre a Procuradoria da cidade de Los Angeles e o Centro de Ciência de Dados e Políticas Públicas da Universidade de Chicago no desenvolvimento de um modelo de aprendizado de máquina que prevê a probabilidade de reincidência de um indivíduo. Esse projeto tem como ideia romper o ciclo de reincidência de pessoas que já foram detidas anteriormente, promovendo medidas de intervenção personalizadas ao invés de métodos de sentença tradicionais. O foco é abordar métodos relacionados a serviços sociais, como : serviços prestados à comunidade, plano de tratamento de reabilitação de dependentes químicos e outras formas de serviços sociais apropriadas ao contexto de cada indivíduo. Entretanto, adaptar essas intervenções a casos específicos exige muitos recursos.

Diante disso, foi desenvolvido um modelo preditivo para identificar 150 indivíduos com maior risco de reincidência, no qual o modelo foi otimizado para o desempenho em detrimento da equidade. A auditoria do modelo foi realizada através do kit de ferramentas do *Aequitas*. Considerando o contexto, a métrica selecionada para avaliação foi a *Recall Parity* por etnia, que identificou que os grupos étnicos hispânicos e indivíduos com etnia não declarada foram sub representados entre os 150 indivíduos de maior risco.

A medida adotada para a mitigação desse viés é uma abordagem de pós-processamento, ajustando ligeiramente o limiar de classificação utilizado pelo

modelo para selecionar os indivíduos de cada grupo étnico. O objetivo desse ajuste é melhorar a revocação de forma mais equilibrada entre os grupos. A precisão do modelo após o ajuste foi de 70.7%, enquanto a precisão do modelo sem o processo de mitigação foi de 72,7%.

5 DISCUSSÃO

No decorrer do desenvolvimento deste trabalho, foi possível realizar uma revisão bibliográfica não sistemática em artigos e pesquisas, que culminou em diversos campos de conhecimento, desde as ciências sociais até a inteligência artificial. A problemática do enviesamento em sistemas de inteligência artificial, abrange tanto questões sociais quanto técnicas. O crescimento da adesão destes sistemas em áreas sensíveis da nossa sociedade pode acarretar em diversos impactos negativos a grupos minoritários, indivíduos e a sociedade como um todo. Diante disso, essa problemática e suas repercussões na sociedade ganharam espaço de debate na mídia e se tornou objeto de estudo de pesquisadores de diversas áreas.

O viés é algo construído a partir das experiências do indivíduo, que acabam por determinar as suas tomadas de decisão. Embora o viés possa ter diferentes origens e formas de se manifestar no contexto de inteligência artificial, no caso dos modelos de aprendizado de máquina, o aprendizado é baseado em nossas experiências transmitidas por meio de bases de dados. Ou seja, o nosso viés é passado para a máquina. Assim, levantando questões como a falta de neutralidade das bases utilizadas, a possibilidade de vieses estarem ocultos nos dados e a validade dos conhecimentos presentes nessas bases.

Devido à falta da capacidade de análise crítica dos modelos de inteligência artificial, estratégias como métricas de equidade foram desenvolvidas para a detecção e mensuração de viés em modelos de aprendizado de máquina. Entretanto, para a implementar essas métricas, é necessário que o desenvolvedor ou auditor do modelo possua um conhecimento sólido sobre essas métricas e seja capaz de compreender o contexto do problema, pois cada métrica atende a um propósito específico. Ou seja, um modelo pode ser considerado justo em uma métrica e em outra não. Por exemplo, a métrica de *Equality of Opportunity* verifica se as taxas de *true positive rate* são iguais entre grupos protegidos e não protegidos, garantindo chances iguais para aqueles que realmente merecem. Já a métrica de *Equalized Odds*, além de verificar as taxas de *true positive rate* entre os dois grupos, também verifica se a taxa de *false positive rate* é igual entre grupos protegidos e não protegidos, assegurando equidade tanto em verdadeiros quanto em falsos positivos, caracterizando-se como uma métrica mais restritiva.

Dentre as ferramentas analisadas na detecção e mitigação de viés, a *AI Fairness 360* (*IBM*) e *Fairlearn* (*Microsoft*), além de auxiliarem na detecção de viés através de uma gama de métricas disponíveis, possuem também algoritmos para serem implementados na mitigação de possíveis vieses identificados, ambas as ferramentas classificam seus algoritmos em três categorias, sendo elas: pré-processamento, processamento e pós-processamento. A *AI Fairness 360* é uma ferramenta robusta, destacando-se por sua ampla gama de métricas de equidade e algoritmos de mitigação, além de oferecer suporte para múltiplas linguagens, como *Python* e R. Contudo, sua curva de aprendizado tende a ser mais acentuada, especialmente para usuários iniciantes ou não técnicos, com pouco conhecimento sobre métricas de equidade. O *Fairlearn*, por sua vez, possui um conjunto mais limitado de métricas e algoritmos de mitigação em comparação com o *AI Fairness 360*, mas se destaca por suas ferramentas interativas, como um *dashboard*, que auxilia na visualização e análise de resultados de diferentes modelos. Por fim, o *Aequitas* é o que mais se diferencia, pois é a única ferramenta dentre as analisadas projetada únicamente para auditoria e detecção de viés. Ou seja, não possui algoritmos de mitigação, mas se destaca pela facilidade de uso para usuários não técnicos, oferecendo um guia chamado de árvore de equidade que ajuda o usuário a selecionar a métrica mais adequada ao contexto do seu problema. Além disso, conta com uma aplicação *web* e a geração de relatórios após cada auditoria.

Ao analisarmos casos de uso voltados à implementação dessas ferramentas em cenários reais, observamos que o *AI Fairness 360*, *Fairlearn* e o *Aequitas* foram eficazes em identificar vieses nos modelos a qual foram aplicados. Além disso, o *AI Fairness 360* e o *Fairlearn* demonstraram sucesso na mitigação dos vieses detectados, utilizando seus algoritmos de mitigação disponibilizados, sem causar uma grande diminuição no desempenho do modelo. Por fim, todos os estudos pontuaram a importância do equilíbrio entre o desempenho do modelo e sua equidade.

6 CONCLUSÃO

Com o avanço tecnológico das inteligências artificiais e sua crescente presença no cotidiano das pessoas, elas participam de tarefas simples, como traduções de textos e indicações de músicas, até áreas mais sensíveis, como aprovações de crédito bancário e seleção de candidatos a vagas de emprego. Em decorrência disso, este trabalho tem como objetivo apresentar uma análise de métricas de equidade e ferramentas utilizadas na detecção e mitigação de viés em modelos de aprendizado de máquina.

Foi abordado os métodos que são utilizados para a identificação de viés, através de uma revisão bibliográfica observamos que, mesmo que uma métrica não identifique vieses, outras métricas podem detectá-lo. Demonstrando a necessidade de desenvolvimento de métricas com a capacidade de cobrir um escopo mais amplo. Em seguida, analisamos as ferramentas utilizadas na detecção e mitigação de enviesamentos, destacando suas principais características e particularidades.

Por fim, ao analisarmos os estudos relacionados a casos de uso dessas ferramentas, observamos que, na abordagem da mitigação, é necessário equilibrar equidade e desempenho do modelo. Além disso, as ferramentas abordadas obtiveram resultados positivos e se mostraram úteis à proposta de seu uso.

Através do desenvolvimento deste trabalho, foi possível compreender, por meio da revisão bibliográfica, os possíveis impactos que um modelo de inteligência artificial enviesado pode causar, além de concluir que os modelos repercutem vieses já existentes na sociedade, pois o fator humano está presente em todas as etapas. No entanto, um desafio recorrente foi a carência de pesquisas em português com um enfoque mais técnico sobre ferramentas, métricas e estudos de caso.

Por fim, este trabalho tem como objetivo contribuir para o campo da inteligência artificial, servindo como um guia acessível para leitores leigos, estudantes e profissionais da área, oferecendo uma base teórica para compreensão de métricas comumente utilizadas na detecção de viés, além de apresentar pontos relevantes das ferramentas discutidas ao longo do estudo.

Nesse sentido, para trabalhos futuros sugiro a realização das seguintes pesquisas:

- Análises em contextos nacionais, abordando particularidades culturais, sociais e econômicas, a fim de diminuir a carência de estudos técnicos sobre ferramentas de mitigação e métricas de equidade na língua portuguesa.
- Investigar como diferentes ferramentas podem ser combinadas para a detecção e mitigação de viés em modelos de aprendizado de máquina.
- Revisão e atualização dos resultados apresentados neste estudo, analisando novas funcionalidades, métricas e abordagens implementadas nas ferramentas estudadas.

REFERÊNCIAS

AKTER, S. et al. Algorithmic bias in machine learning-based marketing models. **Journal of Business Research**, v. 144, n. 144, p. 201–216, maio 2022. Disponível em: https://www.sciencedirect.com/science/article/pii/S0148296322000959. Acesso em: 08 jun. 2024.

ANSARI, T. **7 Toolkits To Build Your AI Ethically**. Disponível em:https://analyticsindiamag.com/top-ai-tools/7-toolkits-to-build-your-ai-ethically/. Acesso em: 03 set. 2024.

AMINI, A. et al. Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. **Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society**, 27 jan. 2019. Disponível em: https://dl.acm.org/doi/pdf/10.1145/3306618.3314243. Acesso em: 09 jun. 2024.

BRISSANT, Otavio. Viés algorítmico:: o uso de algoritmos em processos decisórios tem gerado preocupações acerca da possibilidade de discriminação. **Legal Bytes,** 2023. Disponível em:https://legalbytes.hurb.com/vies-algoritmico-o-uso-de-algoritmos-em-processos-decisorios-tem-gerado-preocupacoes-acerca-da-possibilidade-de-discriminacao/. Acesso em: 30 jun. 2024.

BURKOV, A. **The hundred-page machine learning book**. Quebec City, QC, Canada: Andriy Burkov, 2019.

BALAYN, A. *et al*. “Fairness Toolkits, A Checkbox Culture?” On the Factors that Fragment Developer Practices in Handling Algorithmic Harms. **Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society**. 2023. p. 482-495. Disponível em : https://dl.acm.org/doi/pdf/10.1145/3600211.3604674. Acesso em : 03 jul. 2024.

BILRO, A. R. O. **Creating a Product to Segment Donors and Predict Donor Churn-AI Ethics in NGO-s: Implications of Biased or Unfair Machine Learning in the Non-Profit Sector**. 2021. Dissertação de Mestrado. Universidade NOVA de Lisboa (Portugal). Disponível em : https://core.ac.uk/outputs/532997238/?source=2. Acesso em : 01 set. 2024.

BAJAJ, A. **Performance Metrics in Machine Learning [Complete Guide]**. Disponível em: https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide. Acesso em: 10 jun. 2024.

BIRD, Sarah *et al*. Fairlearn: A toolkit for assessing and improving fairness in AI. **Microsoft,** 2020. Disponível em: .https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf. Acesso em: 12 jun. 2024.

BARROCAS, S. et al. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. **Proceedings of the 2021 AAAI/ACM Conference**

on AI, Ethics, and Society, 21 jul. 2021. Disponível em : https://arxiv.org/pdf/2103.06076. Acesso em : 26 jul. 2024.

COZMAN, F. G.; PLONSKI, G. A.; NERI, H. (EDS.). **Inteligência artificial: avanços e tendências**. [s.l.] Universidade de São Paulo. Instituto de Estudos Avançados, 2021.

COZMAN, Fabio Gagliardi; KAUFMAN, Dora. Viés no aprendizado de máquina em sistemas de inteligência artificial: a diversidade de origens e os caminhos de mitigação. **Revista USP**, n. 135, p. 195-210, 2022. Disponível em : https://sol.sbc.org.br/index.php/wics/article/view/15967. Acesso em: 03 jun. 2024.

CHAKRABORTY, J.; MAJUMDER, S.; MENZIES, T. Bias in machine learning software: why? how? what to do? **Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering**, 18 ago. 2021. Disponível em : https://dl.acm.org/doi/pdf/10.1145/3468264.3468537. Acesso em : 25 jun. 2024.

CHOULDECHOVA, A.; ROTH, A. The Frontiers of Fairness in Machine Learning. **Cornell University**, 1 jan. 2018. Disponível em : https://arxiv.org/pdf/1810.08810. Acesso em : 25 jun. 2024.

CATON, S.; HAAS, C. Fairness in Machine Learning: A Survey. **ACM Computing Surveys**, 23 ago. 2023. Disponível em : https://dl.acm.org/doi/pdf/10.1145/3616865. Acesso em : 26 jun. 2024.

DENG, Wesley Hanwen et al. Exploring how machine learning practitioners (try to) use fairness toolkits. **Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency**. 2022. p. 473-484. Disponível em : https://dl.acm.org/doi/pdf/10.1145/3531146.3533113. Acesso em : 03 jul. 2024.

DATA SCIENCE FOR SOCIAL GOOD. **Data Science for Social Good Data Fest 2018 - Aequitas.** YouTube, 2018. 1 vídeo (6 min). Disponível em : https://www.youtube.com/watch?v=DjSYRb8lWd0&t=239s. Acesso : 10 set. 2024.

DATA SCIENCE FOR SOCIAL GOOD. **SSA | Aequitas Tutorial.** YouTube, 2020. 1 vídeo (36 min). Disponível em : https://www.youtube.com/watch?v=6-ceLhDBwxg&t=1116s. Acesso : 10 set. 2024.

D'ALESSANDRO, B.; O'NEIL, C.; LAGATTA, T. Conscientious Classification: A Data Scientist 's Guide to Discrimination-Aware Classification. **Big Data**, v. 5, n. 2, p. 120–134, jun. 2017. Disponível em : https://arxiv.org/pdf/1907.09013. Acesso : 05 ago. 2024.

DUDÍK, Miroslav et al. **Assessing and mitigating unfairness in credit models with the fairlearn toolkit**. Tech Rep. MSR-TR-2020-34, Microsoft, 2020. Disponível em : https://www.microsoft.com/en-us/research/uploads/prod/2020/09/Fairlearn-EY_WhitePaper-2020-09-22.pdf. Acesso : 05 jun. 2024.

EMANUILOV, Ivo; YORDANOVA, Katerina. Do You Believe in FAIR-y-tales? An Overview of Microsoft's New Toolkit for Assessing and Improving Fairness of Algorithms. **Ku Leuven,** 2020. Disponível em: https://www.law.kuleuven.be/citip/blog/do-you-believe-in-fair-y-tales-an-overview-of-microsofts-new-toolkit-for-assessing-and-improving-fairness-of-algorithms/. Acesso em: 20 jul. 2024.

FOSTER, I. *et al*. **Big data and social science: Data science methods and tools for research and practice**. 2. ed. Filadélfia, PA, USA: Chapman & Hall/CRC, 2020.
AMINI, A. et al. Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. **Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society**, 27 jan. 2019. Disponível em: https://dl.acm.org/doi/pdf/10.1145/3306618.3314243. Acesso em: 09 jun. 2024.

FU, R.; HUANG, Y.; SINGH, P. V. AI and Algorithmic Bias: Source, Detection, Mitigation and Implications. **SSRN Electronic Journal**, 2020. Disponível em : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3681517. Acesso em : 20 jun. 2024.

FAIRLEARN. User guide - Adversarial Mitigation. Disponível em : https://fairlearn.org/main/user_guide/mitigation/adversarial.html. Acesso em : 20 jul. 2024.

FAIRLEARN. User guide - Postprocessing. Disponível em : https://fairlearn.org/main/user_guide/mitigation/postprocessing.html. Acesso em : 20 jul. 2024.

GÉRON, A. **Mãos à obra: aprendizado de máquina com Scikit-Learn & TensorFlow:** Conceitos, Ferramentas e Técnicas Para a Construção de Sistemas Inteligentes. Tradução: Rafael Contatori. Rio de Janeiro: Alta Books, 2019.

GONÇALVES, Mariana Sbaite. Viés algorítmico e descriminação: Como os algoritmos de IA podem perpetuar e amplificar vieses sociais. **Migalhas,** 2024. Disponível em:https://www.migalhas.com.br/depeso/415125/vies-algoritmico-e-discriminacao-ia-pode-amplificar-vieses-sociais . Acesso em: 05 out. 2024.

GAYHARDT, Lauryn *et al*. Desempenho e imparcialidade do modelo. **Microsoft,** 2024. Disponível em:https://learn.microsoft.com/pt-br/azure/machine-learning/concept-fairness-ml?view=azureml-api-2. Acesso em: 07 ago. 2024.

GONZÁLEZ-SENDINO, Rúben et al. A Review of Bias and Fairness in Artificial Intelligence. **International Journal of Interactive Multimedia and Artificial Intelligence**, v. In press, n. In press, p. 1–1, 1 jan. 2023. Disponível em : https://reunir.unir.net/bitstream/handle/123456789/15693/ip2023_11_001.pdf?sequence=1&isAllowed=y. Acesso em : 14 jul. 2024.

HARRIS, Christopher. Mitigating age biases in resume screening AI models. **The International FLAIRS Conference Proceedings**. 2023. Disponível em : https://journals.flvc.org/FLAIRS/article/view/133236. Acesso em : 14 jul. 2024.

HARDT, M.; PRICE, E.; SREBRO, N. Equality of opportunity in supervised learning. **Advances in neural information processing systems**, v. 29, 2016. Disponível em : https://papers.nips.cc/paper_files/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html. Acesso em : 25 jun. 2024.

HARARI, Y. N. **21 lições para o século 21.** 1. ed. São Paulo: Companhia das Letras, 2018.

HUFTHAMMER, K. T. *et al*. Bias mitigation with AIF360: A comparative study. **NIKT: Norsk IKT-konferanse for forskning og utdanning 2020**. Norsk IKT-konferanse for forskning og utdanning, 2020. Disponível em : https://bora.uib.no/bora-xmlui/bitstream/handle/11250/2764230/833-Article%2bText-1888-1-10-20201116.pdf?sequence=1&isAllowed=y. Acesso em : 10 jul. 2024.

IBM. **O que é aprendizado de máquina (ML) ?.** Disponível em: https://www.ibm.com/br-pt/topics/machine-learning. Acesso em: 09 jun. 2024.

JÚNIOR, Ricardo Rodrigues Dos Santos. A explicabilidade como diretriz para as decisões automatizadas e o art. 20 da lei 13.079/18 (LGPD). **Migalhas,** 2021. Disponível em: https://www.migalhas.com.br/depeso/348841/a-explicabilidade-como-diretriz-para-as-decisoes-automatizadas . Acesso em: 30 jun. 2024.

JORDAN, J. **Evaluating a machine learning model.** Disponível em: https://www.jeremyjordan.me/evaluating-a-machine-learning-model/. Acesso em: 08 jun. 2024.

JÚNIOR, C. DE O. **Prevendo Números: Entendendo as métricas R^2, MAE, MAPE, MSE e RMSE**. Disponível em: .https://medium.com/data-hackers/prevendo-n%C3%BAmeros-entendendo-m%C3%A9tricas-de-regress%C3%A3o-35545e011e70. Acesso em: 12 jun. 2024.

KUMAR, A.; SHARMA, S.; MAHDAVI, M. Machine Learning (ML) Technologies for Digital Credit Scoring in Rural Finance: A Literature Review. **Risks**, v. 9, n. 11, p. 192, 31 out. 2021. Disponível em: https://www.mdpi.com/2227-9091/9/11/192. Acesso em: 09 jun. 2024.

KAMIRAN, F.; CALDERS, T. Data preprocessing techniques for classification without discrimination. **Knowledge and Information Systems**, v. 33, n. 1, p. 1–33, 3 dez. 2011. Disponível em : https://core.ac.uk/download/pdf/81728147.pdf. Acesso em : 10 jul. 2024.

KAMIRAN, F.; KARIM, A.; ZHANG, X. Decision theory for discrimination-aware classification. **2012 IEEE 12th international conference on data mining**. IEEE, 2012. p. 924-929. Disponível em : https://web.lums.edu.pk/~akarim/pub/decision_theory_icdm2012.pdf. Acesso em : 11 jul. 2024.

LEE, Nicol Turner; RESNICK, Paul; BARTON, Genie. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. 2019. Disponível em: https://policycommons.net/artifacts/4141276/algorithmic-bias-detection-and-mitigation/4949849/. Acesso em: 08 jun. 2024.

LEE, M. S. A.; SINGH, J. The Landscape and Gaps in Open Source Fairness Toolkits. **Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems**, 6 maio 2021. Disponível em : https://dl.acm.org/doi/pdf/10.1145/3411764.3445261. Acesso em : 03 jul. 2024.

MITCHELL, S. et al. Algorithmic Fairness: Choices, Assumptions, and Definitions. **Annual Review of Statistics and Its Application**, v. 8, n. 1, p. 141–163, 7 mar. 2021. Disponível em :https://www.annualreviews.org/docserver/fulltext/statistics/8/1/annurev-statistics-042720-125902.pdf?expires=1730806330&id=id&accname=guest&checksum=21C93C8642E548EC3F8E4D650754B665. Acesso em 30 jun. 2024.

MANAKITSA, N. et al. A Review of Machine Learning and Deep Learning for Object Detection, Semantic Segmentation, and Human Action Recognition in Machine and Robotic Vision. **Technologies**, v. 12, n. 2, p. 15, 1 fev. 2024. Disponível em: https://www.mdpi.com/2227-7080/12/2/15. Acesso em: 08 jun. 2024.

MAHESH, B. Machine Learning Algorithms -A Review. **International Journal of Science and Research (IJSR) ResearchGate Impact Factor**, v. 9, n. 1, 2018. Disponível em : https://www.ijsr.net/archive/v9i1/ART20203995.pdf. Acesso em: 08 jun. 2024.

MUREL, J.; KAVLAKOGLU, E. **O que é uma matriz de confusão?** Disponível em: https://www.ibm.com/br-pt/topics/confusion-matrix. Acesso em: 10 jun. 2024.

MEHRABI, N. et al. A Survey on Bias and Fairness in Machine Learning. **Cornell University**, 22 ago. 2019. Disponível em: https://arxiv.org/pdf/1908.09635. Acesso em: 11 jun. 2024.

MIKOŁAJCZYK-BAREŁA, A.; GROCHOWSKI, M. **A survey on bias in machine learning research**. Disponível em: https://arxiv.org/pdf/2308.11254. Acesso em: 25 jun. 2024.

MICROSOFT DEVELOPER. **Building fairer AI Systems with Fairlearn**. YouTube, 18 de maio de 2020a. 1 vídeo (26 min). Disponível em : https://www.youtube.com/watch?v=ZpvEY5BaZ8w&t=249s. Acesso em : 22 jul. 2024.

MICROSOFT DEVELOPER. **How to Test Models for Fairness with Deep-Dive**. YouTube, 16 de maio de 2020b. 1 vídeo (12 min). Disponível em : https://www.youtube.com/watch?v=Ts6tB2p97ek&t=2s. Acesso em 24 jul. 2024.

NEOWAY. **Machine learning: O que é, conceito e para que serve**. Disponível em: https://blog.neoway.com.br/machine-learning/. Acesso em: 09 jun. 2024.

ORPHANOU, K. et al. Mitigating Bias in Algorithmic Systems -- A Fish-Eye View. **Cornell University**, 1 jan. 2021. Disponível em : https://arxiv.org/pdf/2103.16953. Acesso em : 26 jun. 2024.

PAGANO, T. P. et al. **Bias and unfairness in machine learning models: a systematic literature review**. Disponível em: https://arxiv.org/pdf/2202.08176. Acesso em: 20 maio. 2024.

PYDATA. **Tutorial: Fairness in decision-making with AI: a practical guide & hands-on tutorial using Aequitas**. YouTube, 2020. 1 vídeo (1h). Disponível em : https://youtu.be/yOR71zBm3Uc?si=r3mCd6OsnThi13E0. Acesso em : 12 set. 2024.

PROGRESS. Data Bias: the hidden risk of AI - What it is and what you need to be doing about it. Burlington : Progress Software Corporation, 2023. Disponível em : https://d117h1jjiq768j.cloudfront.net/docs/default-source/papers/v6-data-bias-research-study.pdf?sfvrsn=29c8a0_1. Acesso em: 07 jun. 2024.

PANDEY, Harshitaa. Comparison of the usage of Fairness Toolkits amongst practitioners : AIF360 and Fairlearn. 2022. Dissertação (Bacharelado em Ciência da Computação e Engenharia) - Delft University of Technology, EEMCS, Holanda, 2022. Disponível em : https://repository.tudelft.nl/file/File_b3d267fe-5f77-4515-8c84-ab90d86513d9?preview=1. Acesso : 05 ago. 2024.

REIS, B. DE F.; GRAMINHO, V. M. C. A Inteligência Artificial no recrutamento de trabalhadores : O caso amazon analisado sob a ótica dos direitos fundamentais. **Seminário Internacional Demandas Sociais e Políticas Públicas na Sociedade Contemporânea**, n. 0, 8 maio 2019. Disponível em : https://online.unisc.br/acadnet/anais/index.php/sidspp/article/view/19599. Acesso em : 30 jun. 2024.

RAHUL, Saladi; TAO, Yufei. A guide to designing top-k indexes. **ACM SIGMOD Record**, v. 48, n. 2, p. 6-17, 2019. Disponível em : https://sigmodrecord.org/publications/sigmodRecord/1906/pdfs/03_Principles_Rahul.pdf. Acesso em : 07 set. 2024.

RODOLFA, Kit T. *et al*. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In: **Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency**. 2020. p. 142-153. Disponível em : https://arxiv.org/pdf/2001.09233. Acesso em : 10 set. 2024.

RUBACK, Lívia; AVILA, Sandra; CANTERO, Lucia. Vieses no aprendizado de máquina e suas implicações sociais: Um estudo de caso no reconhecimento facial. **Workshop sobre as Implicações da Computação na Sociedade (WICS)**. SBC, 2021. p. 90-101. Disponível em :https://sol.sbc.org.br/index.php/wics/article/view/15967/15808. Acesso em: 03 jun. 2024.

RICHARDSON, B.; GILBERT, J. E. A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions. **Cornell University**, 10 dez. 2021. Disponível em : https://arxiv.org/pdf/2112.05700. Acesso em : 04 jul. 2024.

SALEIRO. P. *et al.* Aequitas: A Bias and Fairness Audit Toolkit. **Cornell University**, 13 nov. 2018. Disponível em : https://arxiv.org/pdf/1811.05577. Acesso em : 10 ago. 2024.

SIMÕES-GOMES, L.; ROBERTO, E.; MENDONÇA, J. Viés algorítmico – um balanço provisório. **Estudos de Sociologia**, v. 25, n. 48, 24 jul. 2020. Disponível em: https://periodicos.fclar.unesp.br/estudos/article/view/13402/9352. Acesso : 30 jun. 2024.

SILBERG, Jake; MANYIKA, James. Notes from the AI frontier: Tackling bias in AI (and in humans). **McKinsey Global Institute**, v. 1, n. 6, p. 1-31, 2019. Disponível em : https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/mgi-tackling-bias-in-ai-june-2019.pdf. Acesso em : 07 jun. 2024.

STEVENS, A. et al. Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva. **2020 IEEE Symposium Series on Computational Intelligence (SSCI)**. IEEE, 2020. p. 1241-1248. Disponível em : https://ieeexplore.ieee.org/document/9308371. Acesso em : 25 jun. 2024.

SURESH, H.; GUTTAG, J. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. **Equity and Access in Algorithms, Mechanisms, and Optimization**, 5 out. 2021. Disponível em : https://dl.acm.org/doi/abs/10.1145/3465416.3483305. Acesso em : 24 jul. 2024.

VIEIRA, Leonardo Marques. A problemática da inteligência artificial e dos vieses algorítmicos: caso COMPAS. **Brazilian Technology Symposium**. 2019. Disponível em : https://lcv.fee.unicamp.br/images/BTSym-19/Papers/090.pdf. Acesso em: 20 jun. 2024.

VENUGOPAL, V. K. *et al*. **Navigating Fairness in Radiology AI: Concepts, Consequences,and Crucial Considerations**. Disponível em: https://arxiv.org/abs/2306.01333. Acesso em: 17 jul. 2024.

WEERTS, Hilde et al. Fairlearn: Assessing and improving fairness of ai systems. **Journal of Machine Learning Research**, v. 24, n. 257, p. 1-8, 2023. Disponível em: https://arxiv.org/pdf/2303.16626. Acesso em: 20 jul. 2024.

XU, S.; STROHMER, T. On the (In)Compatibility between Group Fairness and Individual Fairness. **Cornell University**, 1 jan. 2024. Disponível em : https://arxiv.org/pdf/2401.07174. Acesso em : 25 jun. 2024.

ZHANG, B. H.; LEMOINE, B.; MITCHELL, M.. Mitigating unwanted biases with adversarial learning. **Proceedings of the 2018 AAAI/ACM Conference on AI,**

Ethics, and Society. 2018. p. 335-340. Disponível em : https://arxiv.org/pdf/1801.07593. Acesso em : 11 jul. 2024.