On the evaluation of algorithm fairness strategies: a use case on gender bias

Samuel de Morais Lima[1], **Alex Sandro da Cunha Rêgo**[1]
Damires Yluska de Souza Fernandes[1]

[1] Instituto Federal da Paraíba (IFPB)

samuel.morais@academico.ifpb.edu.br

{alex, damires}@ifpb.edu.br

Abstract. Research Context: *The growing use of Machine Learning (ML) techniques in the development of predictive models to support decision-making and the development of information systems, while fostering advancements, has introduced significant ethical challenges, notably the emergence of biases that may lead to unfair decisions.* ***Scientific and/or Practical Problem:*** *The main scientific and practical challenge addressed is the identification and mitigation of algorithmic biases that may reproduce or exacerbate discrimination against minority groups, with a specific focus on the sensitive attribute gender.* ***Proposed Solution and/or Analysis:*** *This study presents an experimental evaluation of different bias mitigation strategies, including the use of the EqOddsPostprocessing and Reweighing methods from AI Fairness 360 toolkit, the application of weights, and the randomization of sensitive attribute values. Fairness performance was assessed using the Equal Opportunity and Demographic Parity metrics.* ***Related IS Theory:*** *This research is grounded in the theory of Algorithmic Fairness, aimed at ensuring impartiality, and in the concept of Socio-Technical Bias, which acknowledges that socially embedded prejudices are reflected in the outcomes produced by ML algorithms.* ***Research Method:*** *An experimental evaluation was conducted on a binary classification problem using the Portuguese SATDAP dataset. The baseline model was built with the Decision Tree algorithm, chosen for interpretability. Methodology comprised five experimental scenarios designed to test mitigation strategies and assess fairness through cross-validation.* ***Summary of Results:*** *Findings showed that Reweighing method fostered fairer predictions according to fairness metrics, in addition to yielding a slight but notable improvement in performance metrics such as accuracy, precision, recall, and f1-score.* ***Contributions and Impact to IS area:*** *This research reinforces the importance of integrating ethical guidelines and bias mitigation methodologies in developing ML systems, contributing to the construction of solutions fostering predictive fairness and countering discrimination without negatively impacting predictive performance.*

1. Introdução

A popularização do campo multidisciplinar da Ciência de Dados está muito associada ao uso de técnicas de Aprendizado de Máquina (AM) com foco na construção de modelos preditivos, amplamente usados no apoio à tomada de decisões e no desenvolvimento de sistemas de informação inteligentes. Um exemplo de aplicação está no setor da saúde,

em que o uso de modelos preditivos, treinados a partir de dados históricos de anamneses ou de exames como eletrocardiogramas, tem contribuído para a identificação de quadros clínicos adversos e apoiado a decisão de profissionais de saúde [Dutenhefner et al. 2024].

Ao mesmo tempo em que a construção de modelos preditivos vem ganhando notoriedade, também têm emergido desafios adicionais, especialmente no que se refere a questões éticas. Um desses desafios diz respeito aos vieses que podem ser induzidos durante o processo de aprendizado, os quais podem levar o modelo a realizar classificações incorretas e, em algumas situações, desencadear decisões injustas [Ruback et al. 2021]. Esse fenômeno pode favorecer inclusive a reprodução ou acentuação de discriminações contra grupos socialmente minoritários, contribuindo para a perpetuação de desigualdades históricas por meio das decisões automatizadas do modelo [Fernandes and Rêgo 2023].

Lidar com classificações equivocadas em modelos preditivos evidencia normalmente dois problemas cruciais: (i) vieses decorrentes de super-representação de uma classe, e (ii) necessidade de atenção à justiça algorítmica, particularmente no que se refere à promoção de igualdade e equidade nas decisões. Nesse contexto, consolidou-se uma nova área de estudo denominada *justiça algorítmica*, cujo objetivo é assegurar que algoritmos e modelos de AM tratem todos os indivíduos ou grupos de forma justa e imparcial [Martini and Berton 2024]. Para tanto, a literatura especializada dispõe de estratégias voltadas à identificação de vieses, bem como métricas destinadas a avaliar o quanto os modelos atuam de forma equitativa [Pagano 2023].

Este trabalho parte da análise de um problema particular de classificação binária no qual um modelo é treinado para prever a possibilidade de um estudante concluir ou abandonar um curso de graduação. Com base nesse cenário, busca-se examinar diferentes métricas e/ou técnicas de justiça existentes na literatura, de modo a identificar possíveis situações de desfavorecimento de grupos minoritários no processo decisório do modelo. Particularmente, a investigação concentra-se na avaliação de potenciais vieses relacionados ao atributo "gênero" visando responder à seguinte questão de pesquisa: *Como identificar e mitigar vieses associados ao atributo "gênero" no treinamento de um modelo preditivo, com base na aplicação de métricas de justiça?*

Para responder à questão de pesquisa, experimentos foram realizados com a intenção de explorar os seguintes pontos: (a) mensurar, de maneira aplicada em um problema preditivo binário, a presença de desfavorecimento de um grupo minoritário na tomada de decisão do modelo; e (b) avaliar diferentes estratégias de mitigação de vieses que possibilitem ao modelo realizar predições mais justas e equitativas, considerando o atributo sensível em estudo (gênero).

O artigo está organizado como segue: a Seção 2 provê um embasamento teórico; a Seção 3 aborda alguns trabalhos relacionados; a Seção 4 descreve a metodologia experimental; a Seção 5 mostra os resultados obtidos, e a Seção 6 promove sua discussão. Por fim, a Seção 7 tece as considerações finais e propostas para trabalhos futuros.

2. Fundamentação Teórica

Esta seção introduz conceitos associados a vieses, atributos sensíveis e métricas de justiça.

2.1. Viés em Aprendizado de Máquina

Um viés em Aprendizado de Máquina (AM) supervisionado consiste em uma tendência sistemática e previsível presente no processo de construção de um modelo preditivo. Esse viés pode comprometer a qualidade das predições, seja em virtude de uma má preparação do conjunto de dados, seja por um desequilíbrio na distribuição das classes ou pela ausência de representatividade populacional nos dados de treinamento.

Segundo [Ferrara 2024], os vieses em modelos preditivos são geralmente classificados em duas categorias principais: *vieses de dados* e *vieses algorítmicos*. Os vieses de dados ocorrem quando há ruído no processo de coleta das informações ou inconsistências nos dados adquiridos, sendo, em grande parte, consequências de falhas na etapa de curadoria e pré-processamento. Essas etapas incluem tratamentos relacionados a valores ausentes, detecção e correção de *outliers*, verificação de incompletude e/ou padronização nos dados. O viés algorítmico pode ser compreendido como um fenômeno de natureza sócio-técnica [Favaretto et al. 2019]. Sua dimensão social refere-se aos preconceitos historicamente enraizados na sociedade, os quais afetam de forma desproporcional, sobretudo, grupos minoritários e comunidades desfavorecidas. Do ponto de vista técnico, está relacionado à manifestação desses vieses nos resultados produzidos pelos algoritmos, refletindo assimetrias presentes nos dados ou introduzidas no processo de modelagem [Kordzadeh and Ghasemaghaei 2021]. Com base nessa perspectiva, o viés algorítmico ocorre quando um modelo distribui benefícios ou encargos de forma desigual entre diferentes indivíduos ou grupos populacionais [Wong 2019]. Nesse contexto, de acordo com as doutrinas da justiça e imparcialidade existentes na literatura, as pessoas devem ser tratadas de forma semelhante, especialmente em relação a questões sociais, políticas ou econômicas, de maneira que um sistema automatizado seja considerado justo e isento de discriminação [Binns 2018].

2.2. Atributos Sensíveis e Justiça Algorítmica

A Lei Geral de Proteção de Dados Pessoais (LGPD) [Brasil 2018] define dados pessoais sensíveis como uma categoria especial de dados pessoais (e.g., gênero, raça, convicção religiosa, opinião política), cuja exposição pode gerar algum tipo de discriminação, ferindo, assim, o direito fundamental à privacidade. A existência de atributos sensíveis em problemas preditivos requer uma atenção especial, uma vez que os algoritmos de AM tendem a reproduzir desigualdades sociais preexistentes. Um exemplo clássico documentado na literatura é o do viés racial identificado no sistema COMPAS, aplicado para prever eventos de reincidência criminal. Nesse caso, réus negros apresentaram o dobro da taxa de falsos positivos em comparação aos réus brancos, sendo classificados de forma desproporcional como mais propensos a reincidir no crime, mesmo em situações nas quais não cometeriam crime [Barocas et al. 2023].

No contexto de treinamento de modelos preditivos que incorporam atributos sensíveis em seus dados, é imprescindível empregar mecanismos de identificação, mensuração e mitigação de vieses algorítmicos, a fim de promover propriedades de justiça e equidade nas decisões automatizadas. Logo, tais características de cunho sensível não devem influenciar diretamente a predição de um modelo [Pagano 2023].

Do ponto de vista legal, legislações norte-americanas como o *Fair Housing Act* (FHA) e o *Equal Credit Opportunity Act* (ECOA) estabelecem que tais atributos não po-

dem favorecer, prejudicar ou modificar o resultado para indivíduos ou grupos em processos decisórios, como por exemplo, a seleção de candidatos ou julgamentos judiciais. Assim, a detecção de viés requer a definição explícita de um atributo sensível, sobre o qual se fundamenta a aplicação de métricas de justiça, tais como *Equal Opportunity* (EO) e *Demographic Parity* (DP), apresentadas e discutidas mais adiante.

Neste trabalho, o atributo sensível "gênero" foi escolhido como objeto de discussão para identificar e mitigar o viés algorítmico. Esse atributo historicamente está associado a diferentes formas de preconceitos e/ou discriminações em diversos contextos sociais (e.g., desigualdade salarial, sub-representação em cargos de liderança e em carreiras na área de ciência e tecnologia). O tratamento desigual observado no mundo real justifica a escolha do "gênero" como um atributo sensível relevante para investigação no âmbito da justiça algorítmica.

2.3. Estratégias de Mitigação de Injustiça em Modelos Preditivos

O viés discriminativo em modelos preditivos merece uma atenção cuidadosa diante de suas implicações éticas, sociais e regulatórias. Nessa perspectiva, a preocupação requer um olhar crítico que vai além do aspecto técnico relacionado ao desempenho geral do modelo.

De acordo com [Mehrabi 2021], as estratégias de tentativa de mitigação de vieses algorítmicos são enquadradas em três categorias: (a) pré-processamento: modificar os dados de treinamento visando remover eventuais disparidades; (b) em-processamento: interferir no objetivo do algoritmo durante o treino; e (c) pós-processamento: realizar ajustes na predição após o treinamento do modelo.

Dentre as técnicas de pré-processamento, ressaltam-se as que realizam o balanceamento dos dados entre diferentes grupos à neutralização do atributo sensível. A simples remoção do atributo, embora intuitiva, pode ser ineficaz, uma vez que variáveis correlacionadas podem atuar como *proxies* e manter o viés de forma implícita no modelo [Barocas and Selbst 2016]. Em contrapartida, a manutenção dos atributos sensíveis possibilita a aplicação de técnicas mais sofisticadas de mitigação, capazes de reduzir seus efeitos diretos e indiretos sobre as predições. Além disso, a preservação desses dados no treinamento do modelo favorece a análise de vieses associados a múltiplos grupos protegidos (interseção de diferentes identidades que se sobrepõem), fenômeno este reconhecido nas ciências sociais como *interseccionalidade* [Yang et al. 2020].

O AI Fairness 360 (AIF360) é um *toolkit* de código aberto em Python desenvolvido para detectar, compreender e mitigar vieses algorítmicos em modelos de AM. A ferramenta disponibiliza, em uma plataforma unificada, diversas métricas de justiça e algoritmos de mitigação. As estratégias de intervenção disponibilizadas são organizadas nas categorias de pré-processamento, durante o processamento e pós-processamento [Bellamy et al. 2018].

Entre os métodos disponíveis no AIF360 e abordados neste trabalho, o *Reweighing* (Pré-processamento) atua ajustando os pesos das instâncias do conjunto de treinamento sem modificar os valores originais dos atributos. O método estima fatores de pesos distintos para cada combinação de grupo sensível (privilegiado/não privilegiado) e rótulo, de forma a reduzir disparidades entre grupos antes da classificação. Diferentemente da abordagem tradicional de balanceamento de classes, que considera exclusivamente a dis-

tribuição da classe a ser balanceada, o *Reweighing* pondera as instâncias com base na intersecção entre o atributo sensível e a classe. A ideia central consiste em atribuir pesos maiores aos exemplos pertencentes a grupos desfavorecidos associados à classe positiva (e menor peso para grupos privilegiados) [Bellamy et al. 2018].

O método *Equalized Odds Post-processing* (Pós-processamento) atua apenas após o modelo ter realizado suas previsões. A técnica formula um problema de otimização linear com o objetivo de determinar probabilidades de ajustes dos rótulos preditos, de modo a equalizar as taxas de verdadeiros positivos e falsos positivos entre os diferentes grupos. Na prática, o algoritmo modifica as saídas do classificador e/ou ajusta seus limiares de decisão (*thresholds*) para satisfazer a restrição de igualdade de oportunidades, podendo, em alguns casos, reverter determinadas decisões individuais do modelo a fim de assegurar a propriedade de justiça [Hastings et al. 2024].

2.4. Métricas de Justiça

As métricas de justiça permitem identificar e quantificar vieses algorítmicos e desigualdades propagadas por modelos preditivos, possibilitando a adoção de estratégias de mitigação com foco na promoção de um tratamento equitativo entre diferentes grupos demográficos. Neste trabalho, são exploradas as métricas **Paridade Demográfica** (*Demographic Parity* - DP) e **Igualdade de Oportunidades** (*Equal Opportunity* - EO) para avaliar a equidade preditiva de um modelo. Essas métricas são comumente empregadas na literatura e baseiam-se na comparação das taxas de classificação positiva entre diferentes grupos definidos por atributos sensíveis [Pagano 2023]. Sua aplicação é relevante em contextos de impacto social como, por exemplo, no recrutamento de recursos humanos.

A Paridade Demográfica avalia se a proporção de instâncias classificadas como positivas é a mesma entre os diferentes grupos, independentemente se as predições foram realizadas corretamente ou não. De acordo com [Menezes 2021, Pagano 2023], a DP é calculada pela equação:

$$P(\hat{Y} = 1 \mid G = m) = P(\hat{Y} = 1 \mid G = f) \tag{1}$$

Onde:

- G representa o atributo sensível (neste trabalho, o gênero).
- $g \in \{m, f\}$ representa os grupos masculino (m) e feminino (f).
- $P(\hat{Y} = 1 \mid G = g)$ é a probabilidade do modelo prever um resultado positivo $(\hat{Y} = 1)$ para o exemplo pertencente ao grupo g.

A predição positiva é definida pela fração da soma dos Verdadeiros Positivos (VP) e Falsos Positivos (FP) para $G = g$, sobre o total de exemplos do grupo, independentemente da predição ter sido correta ou incorreta. Para fins de entendimento, se $g = masculino$ apresenta 800 predições positivas (VP=500 e FP=300) em um total de 1000 exemplos do grupo (positivos e negativos) e, para $g = feminino$ houve 200 predições positivas (VP = 110 e FP = 90) para um total de 800 exemplos do grupo (positivos e

negativos), tem-se:

$$P(\hat{Y} = 1 \mid G = m) = \frac{500 + 300}{1000} = 0{,}80 \quad (80\%)$$
$$P(\hat{Y} = 1 \mid G = f) = \frac{110 + 90}{800} = 0{,}25 \quad (25\%)$$

O resultado indica que o modelo está favorecendo o grupo masculino, com 80% de chance de obter um resultado positivo, mesmo que nem todas as predições sejam corretas. Como a DP não assegura a igualdade nas taxas de erro entre os grupos, é comum que sua análise seja complementada com outra métrica de justiça, como a Igualdade de Oportunidades (EO).

A métrica de Igualdade de Oportunidades avalia se a Taxa de Verdadeiros Positivos (TVP) é equivalente entre os grupos sensíveis, assegurando que todos tenham igual oportunidade de serem corretamente identificados como positivos. Sua fórmula é definida por [Menezes 2021, Pagano 2023]:

$$EO = TVP_{G=g} = \frac{VP_g}{VP_g + FN_g} \tag{2}$$

Onde:

- VP_g: Verdadeiros Positivos do grupo g.
- FN_g: Falsos Negativos do grupo g.

Considerando $VP_m = 100$, $FN_m = 30$, $VP_f = 80$ e $FN_f = 60$, tem-se:

$$TVP_m = \frac{100}{100 + 30} = 0{,}77 \quad (77\%)$$
$$TVP_f = \frac{80}{80 + 60} = 0{,}57 \quad (57\%)$$

Observa-se, neste caso, uma violação da métrica EO ($TVP_m \neq TVP_f$), indicando que as mulheres têm menos chances de serem corretamente classificadas como casos positivos com base no mérito.

A EO é particularmente mais indicada em contextos nos quais as decisões incorretas podem acarretar grandes impactos, como em diagnósticos médicos ou em apoio à decisão de julgamentos, por exemplo, do judiciário. Já a DP considera a distribuição global das predições positivas, independentemente da classe real. Cabe ressaltar que a otimização de uma métrica pode interferir no resultado da outra [Caton and Haas 2024]. Por exemplo, ao tentar forçar um melhor valor de DP, pode-se aumentar o número de falsos positivos para um grupo. Já a EO busca mitigar esse tipo de distorção ao se concentrar na igualdade de tratamento entre os indivíduos do grupo que pertencem à classe positiva real.

3. Trabalhos Relacionados

A imparcialidade na tomada de decisão em AM é retratada por [Uddin et al. 2024]. Nesse estudo, os autores conduzem um experimento de avaliação empírica da justiça algorítimica de um modelo com o objetivo de examinar a imparcialidade inerente a seis algoritmos de AM, dentre os quais Regressão Logística e a Árvore de Decisão. Para a análise, foram consideradas quatro métricas de justiça: *Equalised Odds*, *Equal Opportunity*, *Treatment Equality* e *Comprehensive* aplicadas em 4 (quatro) bases de dados distintas. O cenário desses dados abrange a concessão de crédito bancário, a previsão de reincidência criminal, o perfil de renda familiar e a adesão a serviços de marketing. Nessas bases, o atributo sensível é o gênero em três delas, e a raça na quarta base. Os resultados dos experimentos evidenciaram que, ainda usando o mesmo conjunto de dados, determinados algoritmos podem produzir predições relativamente mais justas, enquanto outros resultam em decisões enviesadas. Essa constatação suscita uma discussão relevante a respeito do comportamento diferenciado dos algoritmos de AM frente a atributos sensíveis, sugerindo que a equidade algorítmica pode estar relacionada não só aos dados, mas também à natureza intrínseca do algoritmo empregado.

A mitigação de vieses a partir de estratégias de pré-processamento é abordada em [Freitas 2024]. Utilizando o conjunto de dados *Adult Census Income* com o propósito de classificar se a renda anual de uma pessoa vai ultrapassar 50.000 dólares, o trabalho empregou o método *EqOddsPostprocessing* da biblioteca *AI Fairness 360*[1] para avaliar a redução de vieses quanto aos atributos "gênero" e "raça" (interseccional), sob a perspectiva das métricas de justiça *Disparate Impact Ratio*, *Statistical Parity Difference* e *Average Odds Difference*. Os resultados obtidos indicaram que a utilização do método *EqOddsPostprocessing* viabiliza a concepção de modelos mais justos e equitativos, contribuindo para o desenvolvimento de soluções éticas em AM com base na sua estrátegia de pós processamento.

Utilizando também o conjunto de dados *Adult Census Income*, [Castaneda et al. 2022] propuseram uma abordagem sócio-estatística voltada ao combate de injustiças algorítmicas, com foco nos atributos sensíveis "gênero" e "raça". A técnica consistiu na aleatorização dos dados das variáveis sensíveis, com o intuito de reduzir uma indução direta entre esses atributos e a variável-alvo. Dessa forma, esperava-se que o modelo direcionasse sua aprendizagem para características mais relevantes à tarefa de classificação. Apesar da simplicidade do método, os resultados indicaram que o modelo conseguiu atribuir decisões mais equilibradas, de maneira que homens brancos e mulheres não brancas recebessem as mesmas oportunidades.

[Ribeiro et al. 2024] destacam que um dos desafios em AM é o desbalanceamento de classes. O trabalho discute causas e implicações desse problema e como ele pode comprometer o desempenho de modelos preditivos, especialmente no tocante à imparcialidade e generalização. Técnicas como a reamostragem e o ajuste de pesos são abordadas para mitigar o viés na previsão das classes. Ao ajustar o peso das classes (aplicar uma penalidade para os erros cometidos na previsão das classes minoritárias), o modelo é orientado a atentar mais aos exemplos da classe minoritária para compensar o desequilíbrio na distribuição das classes, estimulando um tratamento mais igualitário entre elas. O ajuste de

[1] https://research.ibm.com/publications/ai-fairness-360-an-extensible-toolkit-for-detecting-and-mitigating-algorithmic-bias

pesos pode ser útil na mitigação de injustiças relacionadas a atributos sensíveis.

Os trabalhos descritos serviram de inspiração para elaborar um conjunto de experimentos que pudesse avaliar a aplicação de algumas métricas de justiça em combinação com diferentes abordagens de mitigação de vieses, com foco na equidade em um atributo sensível. As métricas são aplicadas no pré e pós-processamento do modelo, com o intuito de inspecionar como o modelo se comporta em termos de equidade para o atributo "gênero". São considerados, nesse trabalho, os cenários em que o modelo reflete o aprendizado obtido naturalmente a partir dos dados, sem qualquer intervenção na construção do modelo, e outros cenários que incluem a avaliação das métricas de justiça aliadas a estratégias de mitigação de vieses.

4. Metodologia

Esta seção descreve a metodologia planejada para a execução dos experimentos, assim como o conjunto de dados utilizado para a construção do modelo.

4.1. Conjunto de Dados e Problema Preditivo

O conjunto de dados português SATDAP [Martins 2021] reúne informações de estudantes de diferentes cursos de graduação. Seus atributos descrevem dados coletados no ato da matrícula e alusivos ao desempenho acadêmico ao longo do curso. Constituído por 4.424 instâncias, seus 32 atributos abrangem informações relacionadas a aspectos demográficos (e.g., nacionalidade), socioeconômicos (e.g., mensalidades em dia, existência de dívida) e acadêmicos (e.g., nota de admissão). O atributo *status do estudante* é um rótulo que indica a situação do estudante no curso: "concluído", "matriculado" ou "abandono". A distribuição natural das classes representa, respectivamente, 32,1%, 18,0% e 49,9% dos exemplos do conjunto de dados. A Tabela 1 apresenta uma amostra do dicionário de dados, com destaque ao atributo sensível "gênero" e à variável-alvo *status do estudante*.

Tabela 1. Dicionário de Dados

Nome da variável	Tipo	Descrição
Estado civil	Inteiro	Estado civil (1 – solteiro, 2 – casado, etc.)
Modo de aplicação	Inteiro	(ex: 1 - 1ª fase - contingente geral, etc.)
Ordem de aplicação	Inteiro	(entre 0 - primeira escolha e 9 - última escolha)
Curso	Inteiro	Código do curso
Atendimento diurno/noturno	Inteiro	Turno de atendimento
Escolaridade anterior	Inteiro	Nível de escolaridade anterior
Nota esc anterior	Contínuo	Nota da escolaridade anterior (entre 0 e 200)
Nacionalidade	Inteiro	Código da nacionalidade
Qualificação da mãe	Inteiro	Nível de escolaridade da mãe
Qualificação do pai	Inteiro	Nível de escolaridade do pai
Ocupação da mãe	Inteiro	Código da ocupação da mãe
Ocupação do pai	Inteiro	Código da ocupação do pai
Nota de admissão	Contínuo	Nota de admissão do aluno (escala de 0 a 200)
Necessidades educacionais especiais	Inteiro	Indica se possui necessidades educacionais especiais
Devedor	Inteiro	Indica se o estudante possui dívidas com a instituição
Mensalidades em dia	Inteiro	Indica se o estudante está com as mensalidades em dia
Gênero	**Inteiro**	**Gênero do(a) estudante: 0- feminino, 1 - masculino**
Bolsista	Inteiro	Indica se o estudante é bolsista
Idade	Inteiro	Idade do estudante no momento da matrícula
Internacional	Inteiro	Indica se o estudante é estrangeiro
Unid curric 1S	Inteiro	Nº de unidades curriculares creditadas no 1º semestre
Unidades curriculares 1º semestre (matriculado)	Inteiro	Número de unidades curriculares matriculadas no 1º semestre
Unidades curriculares 1º semestre (avaliações)	Inteiro	Número de avaliações às unidades curriculares do 1º semestre
Unidades curriculares 1º semestre (aprovado)	Inteiro	Número de unidades curriculares aprovadas no 1º semestre
Med notas 1S	Inteiro	Média das notas no 1º semestre
Unid curric 2S	Inteiro	Nº de unidades curriculares creditadas no 2º semestre
Unidades curriculares 2º semestre (matriculado)	Inteiro	Número de unidades curriculares matriculadas no 2º semestre
Unidades curriculares 2º semestre (avaliações)	Inteiro	Número de avaliações às unidades curriculares do 2º semestre
Unidades curriculares 2º semestre (aprovado)	Inteiro	Número de unidades curriculares aprovadas no 2º semestre
Med notas 2S	Inteiro	Média das notas no 2º semestre
Status do estudante	**Categórica**	**Indicação da situação atual do estudante no curso: concluído, abandono ou matriculado**

É importante destacar que o conjunto de dados SATDAP foi utilizado em sua versão pré-processada, conforme disponibilizado originalmente [Martins 2021]. Dessa forma, as etapas de limpeza de dados, normalização e engenharia de atributos não foram objeto de intervenção neste estudo. Diante disso, visando atender a uma decisão experimental de treinar um classificador para prever a situação de abandono ou conclusão de curso por parte dos estudante, foram eliminadas todas as instâncias pertencentes à classe "matriculado" no conjunto de dados original, haja vista que não é possível determinar se um ou outro estudante concluiu ou abandonou o curso. O conjunto de dados resultante passou a conter 3.640 instâncias, das quais 60,8% correspondem à classe "concluído"

e 39,2% à classe “abandono”. Nessa perspectiva, o objetivo é construir um modelo de classificação binária, capaz de prever se um estudante irá graduar-se (classe positiva) ou abandonar o curso (classe negativa).

Explorar um problema de classificação binária para analisar viés de gênero é particularmente apropriado quando o objetivo é avaliar a justiça algorítmica de maneira clara e mensurável, considerando os seguintes aspectos: (a) analítico: direciona o foco do estudo para decisões mais diretas, de modo a facilitar a identificação de previsões injustas entre grupos do atributo “gênero"; e (b) metodológico: reduz a complexidade na mitigação de viés, devido à simplicidade para treinamento do modelo e interpretação dos resultados, conforme estratégias de mitigação.

4.2. Protocolo Experimental

A decisão pela escolha do algoritmo de aprendizado para treinamento do modelo levou em questão sua aderência ao fator *explicabilidade*, o qual refere-se à capacidade de entender e interpretar as decisões tomadas por um modelo de AM [Silva et al. 2024]. Sendo assim, foi selecionado o algoritmo clássico de *Árvore de Decisão* (AD) como base para os experimentos, devido à sua capacidade de oferecer uma representação visual e lógica do processo de inferência. Essa característica permite compreender como o viés se manifesta em relação ao atributo sensível considerado (gênero), ao longo das predições efetuadas pelo modelo.

Para estimar o desempenho do modelo preditivo e sua capacidade de generalização, aplicou-se a técnica de validação cruzada (*cross-validation*) com 5 *folds* e *random_state=42*. O melhor valor de *k-folds* foi estimado a partir da execução de experimentos preliminares com *k* variando de 1 a 10 *folds*. O algoritmo de AD é instanciado com seguintes hiperparâmetros ajustados após execução de uma busca exaustiva dos parâmetros: *max_depth=5*, *criterion=gini* e *min_samples_split=2*. O desempenho do modelo é avaliado utilizando a tabela de matriz de confusão, com o intuito de visualizar, em termos de números absolutos, os erros e acertos do classificador. As métricas de justiça Paridade Demográfica (DP) e Igualdade de Oportunidades (EO), definidas na Seção 2.4, são empregadas para avaliar a equidade entre os grupos do atributo “gênero” no resultado produzido pelo modelo.

4.3. Métricas de Avaliação de Desempenho

Além de investigar os efeitos das técnicas de mitigação de vieses sob a perspectiva das métricas *DP* e *EO*, é relevante analisar paralelamente como a busca por justiça algorítmica pode impactar o desempenho preditivo global do modelo. As tradicionais métricas *accuracy* (acurácia), *precision* (precisão) e *recall* (revocação) permitem obter diferentes perspectivas acerca da qualidade das previsões efetuadas pelo modelo. Considere VP = Verdadeiros Positivos, VN = Verdadeiros Negativos, FP = Falsos Positivos e FN = Falsos Negativos, as métricas são definidas como segue:

- **Accuracy**: um percentual estatístico que mede a proporção de previsões corretas de todas as classes em relação ao total de previsões realizada pelo modelo.

$$\textbf{Accuracy} = \frac{VP + VN}{VP + VN + FP + FN} \tag{3}$$

- **Precision**: é a proporção de predições positivas corretas em relação ao total de predições positivas (corretas ou não) determinadas pelo modelo.

$$\textbf{Precision} = \frac{VP}{VP + FP} \tag{4}$$

- **Recall**: mede a proporção de exemplos positivos classificados corretamente em relação ao total de exemplos que são efetivamente positivos.

$$\textbf{Recall} = \frac{VP}{VP + FN} \tag{5}$$

- **F1-score**: é a média harmônica entre *precision* e *recall*, equilibrando os dois indicadores em uma única métrica, especialmente útil quando há desbalanceamento entre as classes e desempenhos opostos em relação ao *precision* e *recall*.

$$\textbf{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision + Recall}} \tag{6}$$

5. Resultados

O desempenho do modelo preditivo foi avaliado considerando cinco cenários experimentais distintos, apresentados a seguir.

- **Cenário Experimental 1 (CE1)** - *Baseline*: modelo treinado com os dados originais, sem calibração no (pré) pós-processamento ou uso de bibliotecas externas;
- **Cenário Experimental 2 (CE2)**: modelo treinado com aplicação do método *EqOddsPostprocessing* da biblioteca *AI Fairness 360*;
- **Cenário Experimental 3 (CE3)**: modelo treinado ajustando pesos no atributo sensível;
- **Cenário Experimental 4 (CE4)**: modelo treinado com aleatorização do atributo sensível "gênero" apenas nos dados de treinamento
- **Cenário Experimental 5 (CE5)**: modelo treinado com aplicação do método *Reweighing* da biblioteca *AI Fairness 360*.

Ressalta-se que no conjunto de dados há majoritariamente instâncias da classe positiva que são do gênero feminino, logo, instâncias do gênero masculino estão sub-representadas.

A análise foi conduzida com base nos resultados de saída produzidos pelo modelo, discutido por cenário de experimentação conforme métricas de avaliação previamente apresentadas. A Tabela 2 ilustra uma representação genérica da matriz de confusão em razão das classes reais e preditas. Também, são indicadas a localização de VP, FP, FN e VN.

Valores Previstos	**Valores Reais**	
	Positivo (1)	**Negativo (0)**
Positivo (1)	VP	FP
Negativo (0)	FN	VN

Tabela 2. Abstração da Matriz de Confusão.

A Tabela 3 apresenta, para fins de ilustração, a matriz de confusão referente ao desempenho do modelo em uma iteração do *Cross-Validation* referente ao CE1. Nela, é possível observar a distribuição dos valores previstos e reais por grupo sensível (masculino e feminino), bem como o resultado dos cálculos das métricas de justiça *Equal Opportunity* (EO) e *Demographic Parity* (DP).

	Masculino		**Feminino**	
	Real		Real	
Valores Previstos	Concluído	Abandono	Concluído	Abandono
Concluído	106	29	315	25
Abandono	12	108	16	115
Total de Exemplos	118	137	331	140
EO	106/118 = 90%		315/331 = 95%	
DP	(106+29)/(118+137) = 53%		(315+25)/(331+140) = 72%	

Tabela 3. Desempenho do modelo treinado com os dados originais.

A Tabela 4 apresenta de forma condensada os resultados das métricas de justiça (atributo “gênero") e desempenho reveladas por cada cenário de experimentação. Conforme descrito anteriormente, o objetivo é avaliar o comportamento preditivo do modelo para cada valor do atributo “gênero” (masculino e feminino), bem como examinar de que modo as configurações dispostas em cada cenário impactam a equidade e a eficácia do modelo. O conjunto de teste é constituído por 726 instâncias, sendo 255 do gênero masculino (35%) e 471 do gênero feminino (65%). Em todos os cenários experimentais, observou-se um desvio padrão baixo (≤ 0.01) para as métricas de desempenho (Accuracy, Precision, Recall e F1-score) referentes às médias das duas classes ao longo dos folds da validação cruzada.

	CE1	**CE2**	**CE3**	**CE4**	**CE5**
Accuracy	89%	91%	89%	89%	90%
Precision	89%	91%	90%	89%	91%
Recall	89%	91%	89%	89%	90%
F1-score	89%	91%	89%	89%	90%
DP - Masculino	48%	49%	50%	48%	51%
DP - Feminino	73%	73%	74%	73%	71%
EO - Masculino	90%	93%	92%	90%	95%
EO - Feminino	95%	97%	95%	95%	96%

Tabela 4. Resultados dos experimentos com diferentes abordagens de justiça algorítmica.

5.1. Cenário Experimental 1

A Tabela 4 mostra os resultados para os gêneros masculino (*EO=90%*, *DP=48%*) e feminino (*EO=95%*, *DP=73%*). Este cenário representa o aprendizado do modelo sem nenhuma intervenção nos dados ou algoritmo. Embora a métrica EO apresente uma maior precisão na identificação de casos positivos em ambos os gêneros (90% para homens e

95% para mulheres), o modelo demonstra um desempenho desigual em termos de DP, com valores de 48% para o gênero masculino e 73% para o feminino. Ainda que as métricas de desempenho do modelo sugiram um resultado teoricamente aceitável, essa disparidade (DP) sugere que o modelo não está distribuindo de forma equitativa as classificações positivas (corretas ou não) entre os grupos analisados, caracterizando um viés para classificações positivas ao gênero feminino.

5.2. Cenário Experimental 2

No Cenário 2, a biblioteca *AI Fairness 360* da IBM foi configurada para utilizar o método *EqOddsPostprocessing*. Esse método atua após o treinamento do modelo, reajustando a distribuição das previsões positivas entre os grupos favorecidos e desfavorecidos do atributo "gênero", de forma a equilibrar a Taxa de Verdadeiros Positivos (TVP) entre os diferentes grupos.

Os resultados indicam que o uso do método *EqOddsPostprocessing* promoveu uma leve aproximação das métricas EO (masculino = *93%*, feminino = *97%*) e DP (masculino = *49%*, feminino = *73%*) entre os grupos, ao garantir que a TVP fosse aproximada para ambos os gêneros. Neste cenário, observa-se uma sutil melhora (cerca de 1% em termos de EO e DP em relação ao CE1) nas predições finais do modelo, mitigando ligeiramente a disparidade entre os grupos, sem alterar os dados de entrada e o modelo subjacente.

5.3. Cenário Experimental 3

Neste cenário, foram aplicados pesos aos exemplos com base no atributo sensível, utilizando o argumento *class_weight='balanced'* da biblioteca *scikit-learn*. Esse argumento ajusta os pesos do atributo gênero de acordo com a sua frequência no conjunto de dados. Haja vista que o grupo masculino é menos frequente em comparação com o grupo feminino, este recebe um peso maior. A Tabela 4 apresenta os resultados para os gêneros masculino (*EO=92%*, *DP=50%*) e feminino (*EO=95%*, *DP=74%*).

Nota-se, portanto, a indicação de uma leve melhoria em relação às métricas EO, enquanto a DP manteve a mesma distância de 24%, demonstrando valores equiparados ao do **CE1** e **CE2**. Embora perceba-se um sutil decréscimo da disparidade entre os grupos na EO e DP, torna-se necessário realizar um estudo mais aprofundado sobre essa estratégia de aplicação dos pesos, objetivando verificar fatores que expliquem tal comportamento em relação a atributos sensíveis.

5.4. Cenário Experimental 4

Neste cenário de experimento, foi mantido o esquema de particionamento de dados utilizando o *Cross Validation (k=5)* com *random_state=42*, assegurando que as amostras dos conjuntos de teste de cada fold permanecessem idênticas às utilizadas nos experimentos anteriores. No entanto, o conteúdo de cada *fold* correspondente ao conjunto de treinamento foi modificado intencionalmente aplicando-se uma aleatorização dos valores do atributo sensível "gênero", de maneira a tentar reduzir uma correlação entre o gênero e o status do estudante (Graduou ou Abandonou). Essa estratégia tem a finalidade de avaliar como a remoção parcial do viés (desconsiderando o valor real do atributo sensível) afeta as métricas de justiça sem comprometer a consistência estrutural dos dados.

Percebe-se que a aleatorização do atributo "gênero" no conjunto de treinamento não indicou nenhuma variação no resultado das métricas EO (masculino=*90%*, feminino=*95%*) e DP (masculino=*48%*, feminino=*73%*) em relação ao **CE1** (ver Tabela 4), mantendo-se as mesmas margens de desigualdade. Entende-se que há influência ao preservar os valores originais do atributo sensível no conjunto de teste, fazendo com que o modelo não consiga adotar critérios de equidade para os grupos estudados.

5.5. Cenário Experimental 5

O quinto experimento aplicou a técnica de *Reweighing*, disponível na biblioteca AI Fairness 360 (IBM). Essa abordagem atribuiu pesos distintos às amostras do conjunto de treinamento, equilibrando a distribuição conjunta do grupo sensível e da classe alvo, sem alterar os valores originais das variáveis. A aplicação do *Reweighing* possibilitou que o modelo aprendesse de forma mais equitativa, aumentando a representatividade do grupo minoritário (masculino) e reduzindo disparidades nas previsões.

A Tabela 4 apresenta os resultados para o gênero masculino (*EO=95%*, *DP=51%*) e feminino (*EO=96%*, *DP=71%*). Observa-se uma melhoria nas métricas de justiça, com DP apresentando a menor diferença e, por consequência, a menor disparidade (20%) entre as classes. De forma semelhante, EO registrou a maior aproximação entre os grupos dentre os cenários avaliados, com uma diferença de apenas 1%. Além disso, em comparação com o **CE1**, verificou-se uma pequena melhora no desempenho do classificador, indicando um equilibrio mais consistente entre justiça algorítmica e desempenho preditivo do modelo. Esses resultados corroboram a eficácia da técnica de pré-processamento empregada na promoção de decisões algorítmicas mais justas.

6. Discussão

Os experimentos realizados permitiram examinar de forma aplicada a manifestação de viés na saída do modelo preditivo (inspeção da matriz de confusão) e como grupos de um atributo sensível são afetados em termos de equidade no processo decisório. Essa perspectiva é particularmente relevante quando se deseja compreender o viés e (in)justiça acontecendo na prática, haja vista que possibilita uma análise visível dos impactos distintos entre grupos, indo além do uso e interpretação de métricas de desempenho. Alguns aspectos puderam ser observados na presente investigação, os quais são discutidos como segue.

No que concerne ao comportamento do modelo preditivo em termo das métricas de desempenho, a Tabela 4 revela que as estratégias de promoção da justiça e equidade algorítimica, em relação ao atributo sensível, apresenta para todos os cenários avaliados um resultado muito próximo do cenário *baseline* **CE1**. Entretanto, a melhoria é observada em termos de aumento do nível de equidade entre os grupos do atributo sensível.

O **CE5** demonstrou capacidade de promover a justiça algorítmica considerando diferentes grupos associados a um atributo sensível, conforme evidenciado pela análise das métricas de justiça adotadas. Esse resultado corrobora com a possível estratégia de ajustar pesos entre grupos de um atributo sensível e a variável-alvo, mitigando o viés que possa ser induzido em favorecimento ao grupo majoritário. Como consequência, observou-se um maior equilíbrio entre grupos do atributo "gênero" em relação às métricas *EO* e *DP*, estabelecendo, então, um tratamento de equidade e manifestando um sutil acréscimo no valor das métricas de desempenho do modelo preditivo.

O uso da biblioteca *AI Fairness 360* no **CE2** mostrou percepções positivas. O método *EqOddsProcessing* e o *Reweighing* evidenciaram uma melhoria no valor de EO para os grupos, ao mesmo tempo que registra um avanço em direção ao equilíbrio da métrica na diferença entre os grupos. Em menor intensidade, percebeu-se também uma leve redução da disparidade entre os grupos em termos da métrica DP. Os achados sugerem que o método *EqOddsProcessing* contribui para regular o equilíbrio de oportunidades entre os grupos.

Nos cenários **CE3** e **CE5**, foram aplicadas duas estratégias distintas de pesos para lidar com desequilíbrios e vieses de gênero na árvore de decisão. No **CE3**, utilizou-se o parâmetro *class_weight='balanced'* baseada na distribuição de frequência do atributo sensível (gênero). Isso aumentou a representatividade numérica do grupo minoritário (homens), mas manteve a correlação original entre o gênero e o desfecho (abandono), razão pela qual as métricas de justiça não apresentaram melhora significativa. Já no **CE5**, foi empregada a técnica de *Reweighing*, disponível na biblioteca AI Fairness 360. Enquanto o parâmetro *class_weight='balanced'* busca compensar o desbalanceamento do atributo sensível, o método *Reweighing* atua de maneira mais direcionada à mitigação de vieses associados à interseção entre o atributo sensível e a variável alvo. Como resultado, observou-se melhorias mais consistentes nas métricas de justiça **Demographic Parity (DP)** e **Equal Opportunity (EO)**, sem comprometer as métricas de desempenho do modelo.

Tomando o cenário **CE1** como referência, nota-se que as estratégias de mitigação estabelecidas nos cenários **CE3** e **CE4** não produziram efeitos suficientes que fizessem o modelo reduzir a diferença entre grupos em relação às métricas EO e DP. Embora a aplicação de pesos estabelecida em **CE3** tenha estimulado uma suave equidade do modelo, ainda assim não foi o suficiente para caracterizar uma mitigação efetiva do viés algorítimico.

Sob uma perspectiva de implicações práticas no domínio educacional, a superioridade do método Reweighing **CE5** frente às abordagens de linha de base (*baseline*) **CE1** e de ajuste de pesos em relação às classes **CE3**, transmite um indicativo de melhoria em termos de equidade. A disparidade observada no modelo original, que favorecia sistematicamente o gênero feminino nas previsões de conclusão, poderia resultar, em um ambiente real de tomada de decisão, na invisibilidade de estudantes do gênero masculino que demandam intervenções pedagógicas. Ao mitigar esse viés e garantir que EO atingisse paridade próxima a 96% para ambos os grupos, o sistema deixa de reproduzir padrões históricos de sub-representação presentes nos dados de treinamento. Isso sugere que a adoção de estratégias de pré-processamento é um fator crítico para evitar que sistemas de apoio à decisão perpetuem desigualdades estruturais, assegurando que a alocação de recursos voltados à retenção estudantil seja pautada pelo risco real de evasão, e não por distorções aprendidas pelo algoritmo.

Não obstante os benefícios práticos observados, é necessário ressaltar que a avaliação da eficácia das estratégias, em especial do método *Reweighing* **CE5**, deve ser interpretada sob a ótica das delimitações experimentais deste estudo. Os padrões de melhoria observados nas métricas de justiça e de desempenho, ainda que consistentes ao longo dos folds da validação cruzada (cross-validation), refletem o viés indutivo específico da Árvore de Decisão. Além disso, as diferenças percentuais reportadas entre os cenários (e.g.

entre **CE1** e **E5**) constituem indicativos de tendência baseados nas métricas de avaliação adotadas, não tendo sido submetidas, nesta etapa exploratória, a testes de hipótese para validar estatisticamente as evidências.

Em frente às discussões expostas, para responder à questão de pesquisa definida na Seção 1: *Como identificar e mitigar vieses associados ao atributo "gênero" no treinamento de um modelo preditivo, com base na aplicação de métricas de justiça?* evidencia-se a importância de reconhecer e adotar estratégias sistemáticas para atenuar vieses propagados por algoritmos preditivos, com o objetivo de conceber modelos mais justos. Ressalta-se que tais vieses, ao emergirem do próprio funcionamento do algoritmo, podem favorecer de forma sistemática grupos majoritários e prejudicar outros, mesmo quando os dados de treinamento se apresentem em um estado "neutro". Nesse sentido, a mitigação de vieses associados a atributos sensíveis requer uma abordagem abrangente, que compreende desde uma análise crítica dos dados até a aplicação de estratégias para correção de decisões potencialmente injustas. A adoção de técnicas de mitigação de vieses em diferentes etapas do pipeline de AM deve ser cuidadosamente analisada e testada, seja no estágio de pré-processamento ou pós-processamento. De forma crucial, os achados deste estudo refutam a premissa de que haveria um *trade-off* negativo necessário entre justiça e eficácia. A mitigação de viés no **CE5** não apenas promoveu maior equidade entre os grupos analisados, como também manteve (e marginalmente superou) o desempenho global do modelo (Acurácia e F1-score de 90%). Esses resultados demonstram a viabilidade técnica de construir soluções que sejam, simultaneamente, precisas e éticamente orientadas.

7. Conclusão e Trabalhos Futuros

Este trabalho realizou uma série de experimentos para avaliação de justiça algorítmica em um modelo preditivo, utilizando um conjunto de dados previamente pré-processado, com o objetivo de examinar os efeitos da aplicação de diferentes estratégias de identificação e mitigação de vieses associados a um atributo sensível. Os resultados e análises postas permitiram compreender os padrões de injustiças reproduzidas pelo modelo e mitigar as classificações tendenciosas em favor da equidade preditiva. A aplicação e interpretação das métricas de justiça EO e DP, associadas ao uso de *frameworks* e estratégias de mitigação de viés, demonstraram eficácia na redução do viés algorítmico e na promoção de decisões mais justas entre grupos associados a um atributo sensível.

A presente pesquisa forneceu uma análise aplicada a um problema preditivo originalmente enviesado, evidenciando o comportamento do modelo frente às estratégias de mitigação de vieses adotadas. Nesse contexto, enfatiza-se a importância da incorporação de diretrizes éticas no desenvolvimento de sistemas de AM, bem como a adoção de metodologias que fomentem a equidade como princípio fundamental do processo de modelagem.

Em síntese, a principal contribuição deste trabalho para a área de Sistemas de Informação é de natureza metodológica e empírica, ao demonstrar que a justiça algorítmica não emerge espontaneamente do simples balanceamento de classes (como observado no **C3**), mas requer intervenções deliberadas na fase de pré-processamento dos dados. Os resultados indicam que a técnica de *Reweighing* se apresenta como uma abordagem promissora para o desenvolvimento de sistemas éticos, como visto no domínio educacional

com dados desbalanceados, permitindo conciliar a precisão preditiva com a equidade social. Dessa forma, o estudo oferece diretrizes práticas para o desenvolvimento de modelos de AM socialmente responsáveis, reduzindo assim a amplificação de discriminações contra grupos minoritários, conforme preconizado pelas teorias de viés sócio-técnico.

Como trabalhos futuros, planeja-se expandir os experimentos no cenário **CE5** para investigar a justiça sob a perspectiva do viés interseccional, considerando a sobreposição de múltiplos atributos sensíveis como gênero, raça e sexualidade, além de incorporar outras métricas e técnicas de mitigação de justiça algorítmica. Por fim, planeja-se expandir os experimentos por ora limitados ao algoritmo de Árvore de Decisão, acrescentando outros modelos de aprendizado de maior complexidade e robustez, como o Random Forest e Gradient Boosting. O objetivo é examinar se a eficácia da técnica de *Reweighing* se mantém e se o trade-off entre desempenho e equidade apresenta comportamento análogo ao observado com o modelo de Árvore de Decisão.

Referências

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3):671–732.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943.

Binns, R. (2018). What can political philosophy teach us about algorithmic fairness? *IEEE Security & Privacy*, 16(3):73–80.

Brasil (2018). Lei nº 13.709, de 14 de agosto de 2018. `https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm`. Accessed: 2024-12-18.

Castaneda, J., Jover, A., Calvet, L., Yanes, S., Juan, A., and Sainz, M. (2022). Dealing with gender bias issues in data-algorithmic processes: A social-statistical perspective. *Algorithms*, 15(9):1–16.

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(166):1–38.

Dutenhefner, P., Lemos, G., Rezende, T., Fernandes, J., Tuler, D., Pappa, G., Paixão, G., Ribeiro, A., and Jr., W. M. (2024). Ecg-resnext: Age prediction in pediatric electrocardiograms and its correlations with comorbidities. In *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional*, pages 49–60, Porto Alegre, RS, Brasil. SBC.

Favaretto, M., De Clercq, E., and Elger, B. S. (2019). Big data and discrimination: perils, promises and solutions. a systematic review. *J Big Data*, 6:12.

Fernandes, D. Y. d. S. and Rêgo, A. S. d. C. (2023). Viés, ética e responsabilidade social em modelos preditivos. *Computação Brasil*, (51):19–23.

Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(3). Accessed: 2024-04-28.

Freitas, C. C. (2024). Análise preditiva e equitativa com ai fairness 360. Orientador: Clerivaldo José Roccia. Contribuidores: Thiago Salhab Alves, Luciene Maria Garbuio Castello Branco.

Hastings, C., Qian, L., Gibson, C., Obiomon, P., and Dong, X. (2024). Comprehensive validation on reweighting samples for bias mitigation via aif360. *Applied Sciences*, 14(9):3826. C. Hastings, L. Qian, P. Obiomon and X. Dong are with the Department of Electrical and Computer Engineering, Prairie View A&M University, Texas A&M University System, Prairie View, TX 77446, USA. C. Gibson is with the College of Juvenile Justice, Executive Director of Texas Juvenile Crime Prevention Center, Prairie View A&M University, Texas A&M University System, Prairie View, TX 77446, USA. Email: {chastings, liqian, cbgibson, phobiomon, xidong}@pvamu.edu.

Kordzadeh, N. and Ghasemaghaei, M. (2021). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*. Accessed: 2024-11-20.

Martini, V. and Berton, L. (2024). Fairness analysis in ai algorithms in healthcare: A study on post-processing approaches. In *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional*, pages 553–564, Porto Alegre, RS, Brasil. SBC.

Martins, M. V. e. a. (2021). Early prediction of student's performance in higher education: a case study. In *Trends and Applications in Information Systems and Technologies*, volume 1 of *Advances in Intelligent Systems and Computing*. Springer.

Mehrabi, N. e. a. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35. Accessed: 2024-11-20.

Menezes, H. F. e. a. (2021). Bias and fairness in face detection. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, Online. Sociedade Brasileira de Computação.

Pagano, T. P. e. a. (2023). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7:15.

Ribeiro, H. B., da Silva, L. O., and de Andrade Lira Rabêlo, R. (2024). Estratégias para lidar com desbalanceamento de dados em aprendizado de máquina. In *Anais Estendidos do X Simpósio Brasileiro de Computação Aplicada à Saúde*. Sociedade Brasileira de Computação.

Ruback, L., Avila, S., and Cantero, L. (2021). Vieses no aprendizado de máquina e suas implicações sociais: Um estudo de caso no reconhecimento facial. In *Anais do 2º Workshop sobre as Implicações da Computação na Sociedade (WICS)*, pages 90–101, Evento Online. Sociedade Brasileira de Computação.

Silva, F., Feitosa, R., Batista, L., and Santana, A. (2024). Análise comparativa de métodos de explicabilidade da inteligência artificial no cenário educacional: um estudo de caso sobre evasão. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, pages 2968–2977, Porto Alegre, RS, Brasil. SBC.

Uddin, S., Lu, H., Rahman, A., et al. (2024). A novel approach for assessing fairness in deployed machine learning algorithms. *Scientific Reports*, 14:17753.

Wong, P.-H. (2019). The socio-technical dimensions of fairness in machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Honolulu, HI, USA.

Yang, K., Huang, B., Stoyanovich, J., and Schelter, S. (2020). Fairness-aware instrumentation of preprocessing pipelines for machine learning. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics (HILDA '20)*, page 4, New York, NY, USA. ACM.