JUSTIÇA EM MODELOS DE INTELIGÊNCIA ARTIFICIAL: UM ESTUDO COMPARATIVO DE TÉCNICAS DE MITIGAÇÃO E SUAS IMPLEMENTAÇÕES NO CONTEXTO EDUCACIONAL

Fairness in Artificial Intelligence Models: A Comparative Study of Mitigation Techniques and Their Implementations in the Educational Context

Pedro Henrique dos Santos Pereira
phsp@discente.ifpe.edu.br
Luciano de Souza Cabral
luciano.cabral@jaboatao.ifpe.edu.br

RESUMO

Cada vez mais, modelos de machine learning vem sendo utilizados para influenciar tomadas de decisão que impactam direta e indiretamente a vida das pessoas. Especificamente tomadas de decisão relacionadas à educação podem influenciar profundamente, tanto no momento presente, como no futuro com a repercussão dos resultados gerados por esses modelos. Nesse contexto, tem crescido a preocupação quanto ao *fairness* (justiça) dos modelos, ou seja, se possuem vieses incompatíveis com a equidade de grupos socialmente estigmatizados, como por exemplo: gênero, raça, idade, renda e etc. O objetivo deste trabalho é comparar as principais bibliotecas que implementam métricas e técnicas de mitigação de fairness, e fornecer dados que ajudem a escolher a implementação de mitigação mais adequada para casos semelhantes, além de demonstrar o impacto das variáveis sensíveis sobre a saída dos modelos.

Palavras-chave: fairness; equidade; educação; machine learning; explicabilidade;

ABSTRACT

Machine learning models are increasingly being used to influence decision-making that directly and indirectly impacts people's lives. Specifically, decision-making related to education can have a profound influence, both in the present and in the future, on the repercussions of the results generated by these models. In this context, there has been growing concern about the fairness of models, that is, whether they have biases that are incompatible with the equity of socially stigmatized groups, such

Instituto Federal de Pernambuco. Campus Jaboatão. Curso de tecnologia em análise e desenvolvimento de sistemas. 01 de julho de 2025.

as gender, race, age, income, etc. The objective of this work is to compare the main libraries that implement fairness metrics and mitigation techniques, and to provide data that help choose the most appropriate mitigation implementation for similar cases, in addition to demonstrating the impact of sensitive variables on the output of the models.

Keywords: fairness; equity; education; machine learning; explainability;

1 INTRODUÇÃO

O uso de análises estatísticas e algoritmos para inferir conhecimento a partir de um conjunto de dados é uma tendência crescente nos dias atuais. Cada vez mais, vem crescendo a viabilização e a coleta de dados em todos os aspectos de nossas vidas, através de dispositivos IoT, sensores em smartphones, métricas de uso e visualização em redes sociais e em outros cenários (**MAYER-SCHONBERGER** et al., 2014). Diariamente uma enorme massa de dados é gerada e coletada. Esses dados, porém, não tem muito valor em si, mas no potencial que tem para inferência, para isso são necessários algoritmos que processem esses dados. Dentre os principais algoritmos usados para isso estão os algoritmos de *Machine Learning*.

Machine Learning é uma área da inteligência artificial e da computação, que estuda técnicas e algoritmos que são capazes de identificar padrões (**JORDAN** et al., 2015). Modelos de *ML* são treinados com grandes quantidades de dados e "aprendem" a classificar dados de estrutura semelhante a partir do conhecimento que inferiram dos dados de treinamento. Dessa forma, alvos posteriores podem ser classificados em atributos predefinidos, atributos não nomeados ou realizando regressões ("predições").

O padrão aprendido por um modelo é totalmente dependente e baseado na inferência feita nos dados de treinamento, assim sendo, tudo o que estiver implícito nos dados, vai ser aprendido pelo modelo e tratado como verdadeiro. É de extrema importância e de competência de quem criou o modelo, garantir que os dados estão livres de vieses, tendências parciais que conduzem a erros de aprendizagem que podem induzir o modelo a resultados tendenciosos.

Um modelo pode aprender preconceitos presentes nos dados e reforçá-los se forem usados para tomadas de decisões que impactem indivíduos ou grupos alvo dessas vieses, fortalecendo o viés em um ciclo chamado por Mehrabi et al. (2021) e Barocas et al., (2023) de ***feedback loop***, onde o modelo é treinado sobre o mundo real, gera predições que são usadas para modificação do mesmo, essas modificações geram dados que irão reforçar o viés quando ocorrer um novo treinamento do modelo e novamente, as predições do modelo irão reforçar os preconceitos. Impedir a realização deste ciclo é uma tarefa essencial para assegurar equidade ao tratar com grupos sensíveis e historicamente prejudicados. Sob essa ótica, a ética de inteligência artificial tem figurado como uma área de debate ético e filosófico extremamente relevante.

Esse ciclo de reforço de preconceitos se torna ainda mais preocupante quando vemos que decisões institucionais tomadas por modelos de machine learning já vem ocorrendo, principalmente relacionadas à educação (**ANDERS** et al., 2020). Em países com altos índices de desigualdade social, a educação é vista como a única oportunidade de ascensão econômica, possibilitando melhorias significativas na qualidade de vida, alimentação, saúde, bem-estar físico e mental.

É essencial para assegurar a *ética da inteligência artificial*, medir o enviesamento do modelo e tomar medidas para assegurar que o modelo não prejudique grupos ao gerar respostas enviesadas que reforcem ou criem preconceitos. Evitar e mitigar esses preconceitos são tarefas muitos debatidas na esfera ética e legal, e a ciência de identificar e corrigir esses vieses é chamada por Su Cong et al. (2022) de ***Machine Learning Fairness***, ***justiça de inteligência artificial*** em português.

Inspirado pelos trabalhos produzidos por Le Quy et al. (2023) e Mehrabi et al. (2021), o presente artigo tem como objetivo final avaliar duas populares bibliotecas de código que implementam as técnicas, métricas e ferramentas citadas nos referidos estudos, a fim de indicar a mais adequada para os parâmetros definidos. Ao longo da seção 2 deste trabalho é tratado acerca das definições dos conceitos abordados; na seção 3 é descrito a metodologia de pesquisa e o experimento realizado na mesma; na seção 4 é feita uma análise dos resultados obtidos; por fim, na seção 5 são apresentadas as conclusões realizadas.

2 DESENVOLVIMENTO

No contexto educacional, ferramentas de inteligência artificial já vem sendo usadas para as mais diversas tarefas e tem demonstrado enorme potencial ao potencializar o desempenho dos alunos, através de *insights* sobre personalização do método de ensino, apoio ao professor, e de forma mais abrangente no apoio a decisões institucionais, entre outras formas (**ZAWACKI-RICHTER** et al., 2019). Apesar de ser uma ferramenta útil e poderosa, limitações e questões éticas surgem. Um algoritmo toma decisões melhor que humanos? Pode ser mais rápido e eficaz delegar a tarefa de decisão a um algoritmo, mas o risco que modelos mal treinados ou enviesados trazem deve ser cuidadosamente avaliado.

Em modelos criados para decisões educacionais é crítico identificar e mitigar vieses. Um modelo enviesado pode limitar, segregar e prejudicar profundamente a vida escolar e por consequência a vida pessoal, profissional e capacidade cultural dos alunos (**BAROCAS** et al., 2019; **ZAWACKI-RICHTER** et al., 2019; **LE QUY** et al., 2023). Tendo como exemplo anedótico: uma escola com número limitado de vagas para um curso de iniciação à robótica pode usar um *algoritmo enviesado* para selecionar os alunos que podem apresentar o melhor desempenho e afinidade com o assunto, excluindo alunos a critério do preconceito embutido no modelo, desconsiderando o potencial individual de cada aluno.

2.1 Fairness

Mehrabi et al. (2021) definem *fairness* como “ausência de qualquer preconceito ou favoritismo em relação a um indivíduo ou grupo com base em suas características inerentes ou adquiridas”. Isso se traduz no contexto de inteligência artificial como a justiça relativa à equidade de classificação ou predição para diferentes grupos em relação a grupos conhecidamente discriminados. Em âmbitos gerais, hoje essa discriminação começa a ser reconhecida, e políticas públicas estão sendo criadas, como por exemplo com a criação de leis que estabelecem porcentagem de vaga reservadas a grupos discriminados, as *cotas*, que buscam nivelar a desigualdade enfrentada por grupos sociais, étnicos e outros. As medidas de mitigação de viés de injustiça, visam promover a igualdade a esses grupos discriminados, medindo e corrigindo tendências discriminantes em algoritmos.

2.1.1 Atributos protegidos

Atributos protegidos são subconjuntos de dados definidos por atributos sensíveis, como raça, gênero, idade, orientação sexual, deficiência, religião ou nacionalidade como cita (**MEHRABI** et al., 2021), cuja proteção contra discriminação é garantida por princípios legais, regulatórios ou éticos. Em sistemas de aprendizado de máquina, garantir tratamento justo entre diferentes atributos protegidos é fundamental para evitar que modelos perpetuem ou amplifiquem disparidades históricas. Métricas de equidade e algoritmos de mitigação são frequentemente aplicados para identificar e corrigir vieses nos resultados preditivos entre esses atributos.

2.2 Métricas

Para avaliar os modelos tanto em termos de desempenho quanto de justiça, utilizam-se diversas métricas tradicionais de performance — dentre as quais destaco *accuracy*, *precision*, recall, *taxas de falso positivo* (FPR) e *falso negativo* (FNR) — e métricas específicas de *fairness*, como *Demographic Parity*, *Equalized Odds* e *Equal Opportunity* (**FAWCETT**, 2006; **MEHRABI** et al., 2021). Essas métricas permitem identificar não apenas a eficácia geral dos modelos, mas também possíveis disparidades no tratamento entre diferentes grupos demográficos, fornecendo uma análise mais completa dos resultados e sua conformidade ética.

2.2.1 Métricas de Performance dos Modelos

As métricas de performance são fundamentais para avaliar a capacidade preditiva dos modelos (**FAWCETT**, 2006). Destaco:

- **Accuracy (Acurácia):** representa a proporção de previsões corretas em relação ao total de exemplos avaliados. Embora útil como métrica geral, pode ser enganosa em conjuntos de dados desbalanceados. Varia em frações entre 0 e 1, $\mathrm{acc} \in [0, 1]$, que representam a porcentagem de acertos.
- **Precision (Precisão):** indica a proporção de verdadeiros positivos entre todas as previsões positivas, sendo relevante em cenários onde falsos positivos devem ser minimizados.
- **Recall:** mensura a proporção de verdadeiros positivos entre todos os exemplos que são realmente positivos. É especialmente importante em contextos nos quais falsos negativos representam riscos significativos, como diagnósticos médicos ou triagens sociais.
- **False Positive Rate (FPR):** mede a proporção de exemplos negativos incorretamente classificados como positivos. Está diretamente relacionada ao risco de superestimar casos positivos.
- **False Negative Rate (FNR):** calcula a proporção de exemplos positivos que foram erroneamente classificados como negativos, sendo crítica em aplicações onde a omissão de casos reais pode ter consequências graves.

Essas métricas fornecem uma visão abrangente sobre o comportamento dos modelos em termos de acerto, erro e impacto prático das decisões automáticas.

2.2.2 Demographic Parity

Demographic Parity é uma métrica de fairness que exige que a taxa de decisões positivas previstas por um modelo seja igual entre todos os grupos definidos por um atributo sensível, (**DWORK** et al., 2012; **LE QUY** et al., 2023). Formalmente, requer que:

$$P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = a') \quad \forall a, a' \in \mathcal{A}$$

onde $\hat{Y}$ é a predição do modelo e A é o atributo sensível binário. Esse critério ignora o verdadeiro rótulo Y, focando apenas na igualdade da taxa de predições positivas, o que pode resultar em um dilema de otimização com a acurácia.

Assume valores reais no intervalo $\mathrm{DP} \in [0, 1]$, representando a diferença (ou razão) entre as taxas de previsão positiva entre diferentes grupos sensíveis. Com 0 sendo o resultado desejado, onde não há parcialidade na classificação em relação aos grupos sensíveis e 1 representando a máxima parcialidade beneficiando de forma desigual o grupo sensível privilegiado. A paridade demográfica é atingida quando a probabilidade de um indivíduo receber um resultado positivo é a mesma, independentemente do grupo ao qual pertence.

2.2.3 Equalized odds

Equalized Odds exige que as taxas de verdadeiro positivo (TPR) e falso positivo (FPR) sejam iguais entre os grupos sensíveis, garantindo que o modelo tenha o mesmo comportamento de erro para diferentes grupos (**HARDT** et al., 2016; **LE QUY** et al., 2023). Formalmente:

$$P(\hat{Y} = 1 \mid Y = y, A = a) = P(\hat{Y} = 1 \mid Y = y, A = a') \quad \forall a, a' \in \mathcal{A}, y \in \{0, 1\}$$

Isso significa que tanto os positivos reais quanto os negativos reais devem ser tratados de maneira similar em todos os grupos. Assume valores reais no intervalo [0,1], $\mathrm{EOdds} \in [0, 1]$, com valores próximos a 0 representando maior imparcialidade em relação aos atributos sensíveis. Equalized odds representa a igualdade nas **taxas de verdadeiro positivo e de falso positivo** entre os diferentes grupos sensíveis. A métrica avalia se o modelo comete erros e acertos na mesma proporção entre os grupos (**MEHRABI** et al., 2021).

2.2.4 Equal opportunity

Equal Opportunity é uma versão simplificada de Equalized Odds que considera apenas a taxa de verdadeiros positivos. Ela exige que indivíduos que pertencem à atributo positivo recebam tratamento igual, independentemente do grupo sensível, (**HARDT** et al., 2016; **LE QUY** et al., 2023):

$$P(\hat{Y} = 1 \mid Y = 1, A = a) = P(\hat{Y} = 1 \mid Y = 1, A = a') \quad \forall a, a' \in \mathcal{A}$$

Esse critério é especialmente relevante em contextos em que se deseja garantir acesso equitativo a oportunidades sem necessariamente exigir simetria nos erros para o atributo negativo.vlws

Assume valores reais no intervalo [0,1], $\mathrm{EOp} \in [0, 1]$, quando mais próximo de 0 mais justas são as predições feitas pelo modelo medido. Em linhas gerais, *Equal opportunity* é um caso específico de *Equalized Odds* que considera apenas a igualdade na **taxa de verdadeiro positivo** entre os grupos. Garante que indivíduos que pertencem ao atributo positivo tenham a mesma chance de serem corretamente classificados, independentemente do grupo.

2.3 Técnicas de mitigação

Diversos métodos têm sido propostos para mitigar o viés em sistemas de inteligência artificial, com o objetivo de promover justiça algorítmica. De forma geral, essas técnicas se organizam em três categorias principais (**MEHRABI** et al., 2021):

- Pré-processamento: modificações nos dados antes do treinamento, com o intuito de remover discriminações subjacentes. Essa abordagem é aplicável quando é possível alterar o conjunto de dados original.
- Durante o processamento (in-processing): adaptações no algoritmo de aprendizado durante o treinamento, por meio de mudanças na função objetivo ou imposição de restrições. Esse tipo de mitigação é viável quando se tem controle sobre o processo de aprendizagem do modelo.
- Pós-processamento: ajustes aplicados após o treinamento, especialmente úteis quando o modelo é uma “caixa-preta”. Nesse caso, os rótulos previstos podem ser atribuídos novamente com base em funções corretivas.

Esse panorama evidencia que o viés pode emergir de diversas formas e em diferentes pontos do sistema, exigindo atenção crítica de pesquisadores quanto a potenciais fontes de discriminação.

Neste trabalho, foram comparadas três técnicas de mitigação que fazem ajustes nos modelos durante o processamento. Segundo Mehrabi et al. (2021), “se for permitido alterar o procedimento de aprendizagem de um modelo de aprendizagem de máquina, a mitigação durante o processamento pode ser usada durante o treinamento de um modelo”. Por ser o cenário que mais se adequa ao experimento desenvolvido posteriormente nesse artigo, dentre os descritos por Mehrabi, a mitigação durante o processamento foi escolhida para a avaliação, apesar de não ser explicitamente recomendada pela autora.

2.3.1 Exponentiated Gradient Reduction

Exponentiated Gradient Reduction é um algoritmo de *mitigação de viés durante o processamento*. Ele reformula o problema de aprendizado como um jogo entre um otimizador que visa minimizar a perda preditiva e um adversário que impõe restrições de equidade (por exemplo, igualdade de oportunidades) (**AGARWAL** et al., 2018). O método utiliza uma técnica de atualização multiplicativa (gradiente exponencial) para encontrar uma distribuição mista de classificadores que otimize a acurácia do modelo enquanto respeita limites pré-definidos de disparidade entre grupos sensíveis.

2.3.2 Grid Search Reduction

O **Grid Search Reduction** é outra abordagem de *mitigação durante o processamento*. Essa técnica busca uma combinação de modelos que equilibram desempenho e equidade ao explorar sistematicamente um espaço de soluções (grid) de forma explícita (**AGARWAL** et al., 2018). O algoritmo resolve múltiplos problemas de otimização, cada um com diferentes pesos atribuídos a precisão e justiça, gerando uma *fronteira de Pareto*, um equilíbrio ótimo, entre essas duas dimensões. O resultado é uma distribuição mista de modelos com desempenho controlado em termos de viés.

2.3.3 Adversarial Debiasing

O **Adversarial Debiasing** é um método de *mitigação durante o processamento* baseado em *aprendizado adversarial*, uma técnica de *Deep Learning* onde duas redes neurais artificiais são construídas para tentar falsear uma a outra, a primeira concentra-se na tarefa principal e a segunda testa a qualidade da predição da primeira. No caso do Adversarial Debiasing, o modelo consiste em um preditor principal que aprende a tarefa alvo e um adversário que tenta inferir atributos sensíveis a partir das saídas do preditor (**ZHANG** et al., 2018). O objetivo é treinar o modelo preditor de forma que suas previsões maximizem a acurácia, mas que ao mesmo tempo não revelem os atributos sensíveis para o modelo adversário, reduzindo assim a correlação entre os atributos sensíveis e as previsões. Essa técnica promove a equidade ao minimizar o viés diretamente no processo de treinamento.

2.4 Ferramentas de mitigação de injustiça

2.4.1 Fairlearn

Fairlearn é uma biblioteca de código para Python que fornece ferramentas para avaliação e mitigação de desigualdades algorítmicas em sistemas de aprendizado de máquina. Seu principal objetivo é ajudar desenvolvedores e pesquisadores a identificar e reduzir impactos discriminatórios em modelos preditivos, promovendo decisões mais justas com base em métricas de equidade. A biblioteca oferece funcionalidades para análise de desempenho por subgrupos sensíveis, bem como algoritmos de mitigação que ajustam modelos para equilibrar justiça e acurácia, alinhando-se a princípios éticos e regulatórios de uso responsável da inteligência artificial (**BIRD** et al., [s.d.]). Dentre outras inúmeras técnicas e ferramentas para mitigação que Fairlearn implementa, em destaque estão as técnicas e métricas de mitigação de injustiça, descritas anteriormente.

2.4.2 AI Fairness 360

AI Fairness 360 (AIF360) é uma biblioteca de código aberto desenvolvida pela IBM, voltada para a detecção, mensuração e mitigação de vieses em modelos de aprendizado de máquina. Ela fornece um conjunto abrangente de métricas para avaliação de equidade em dados e previsões, além de algoritmos de pré-processamento, durante o processamento e pós-processamento destinados a reduzir disparidades entre grupos sensíveis. Com suporte a múltiplos tipos de dados e modelos, o AIF360 é uma ferramenta robusta para promover o desenvolvimento ético e responsável de sistemas de inteligência artificial (**IBM**, [s.d.]). Semelhante a Fairlearn, AIF360 implementa inúmeras técnicas e fornece diversas ferramentas auxiliares, entre elas, se destacam estão as técnicas e métricas de mitigação de injustiça, descritas nas seções anteriores.

3 METODOLOGIA

A metodologia utilizada foi a pesquisa experimental de natureza quantitativa, cujo objetivo é avaliar o viés de fairness em um modelo de aprendizado de máquina e utilizar diferentes algoritmos de mitigação durante o processamento para corrigir vieses, para complementar a análise de enviesamento, foram utilizadas técnicas de explicação de modelos de IA para descrever a influência de variáveis sensíveis sobre o resultado. O experimento é reprodutível, e os scripts, jupyter notebooks e relatórios parciais e finais podem ser encontrados em repositório: https://github.com/SantosPereira/tcc.

3.1 O conjunto de dados

O conjunto de dados utilizado é o UCI Student Performance (**CORTEZ** et al., 2008), um conjunto de dados com 32 atributos e 382 registros sobre os resultados finais de um ano letivo na disciplina de língua portuguesa, incluindo informações demográficas, como: se o aluno mora área urbana ou rural, tamanho da família, se os pais são separados, nível de educação dos pais, entre outros possíveis indicadores de qualidade de vida e socioeconômicos, além das notas preliminares e média final dos alunos. Segue descrição detalhada do conjunto de dados:

Atributo	Tipo	Descrição Completa
school	Binário	Escola do estudante: "GP" (Escola Gabriel Pereira) ou "MS" (Escola Mousinho da Silveira)
sex	Binário	Sexo do estudante: "F" (feminino) ou "M" (masculino)

age	Numérico	Idade do estudante (de 15 a 22 anos)
address	Binário	Tipo de endereço residencial do estudante: "U" (urbano) ou "R" (rural)
famsize	Binário	Tamanho da família: "LE3" (menor ou igual a 3 membros) ou "GT3" (maior que 3 membros)
Pstatus	Binário	Situação de coabitação dos pais: "T" (morando juntos) ou "A" (separados)
Medu	Numérico	Nível de educação da mãe: 0 (nenhum), 1 (educação primária - 4ª série), 2 (5ª a 9ª série), 3 (ensino secundário), 4 (ensino superior)
Fedu	Numérico	Nível de educação do pai: 0 (nenhum), 1 (educação primária - 4ª série), 2 (5ª a 9ª série), 3 (ensino secundário), 4 (ensino superior)
Mjob	Nominal	Profissão da mãe: "teacher" (professor), "health" (saúde), "services" (serviços civis), "at_home" (em casa), "other" (outros)
Fjob	Nominal	Profissão do pai: "teacher" (professor), "health" (saúde), "services" (serviços civis), "at_home" (em casa), "other" (outros)
reason	Nominal	Razão para escolher a escola: "home" (perto de casa), "reputation" (reputação da escola), "course" (preferência de curso), "other" (outra)
guardian	Nominal	Guardião do estudante: "mother" (mãe), "father" (pai) ou "other" (outros)
traveltime	Numérico	Tempo de viagem de casa para a escola: 1 (<15 min), 2 (15 a 30 min), 3 (30 min a 1 hora), 4 (>1 hora)
studytime	Numérico	Tempo de estudo semanal: 1 (<2 horas), 2 (2 a 5 horas), 3 (5 a 10 horas), 4 (>10 horas)
failures	Numérico	Número de reprovações anteriores: n se 1≤n<3, senão 4
schoolsup	Binário	Suporte educacional extra (sim ou não)
famsup	Binário	Suporte educacional familiar (sim ou não)
paid	Binário	Aulas extras pagas na disciplina (Português) (sim ou não)
activities	Binário	Atividades extracurriculares (sim ou não)
nursery	Binário	Frequentou o jardim de infância (sim ou não)
higher	Binário	Deseja cursar ensino superior (sim ou não)
internet	Binário	Acesso à internet em casa (sim ou não)
romantic	Binário	Em um relacionamento romântico (sim ou não)
famrel	Numérico	Qualidade das relações familiares: 1 (muito ruim) a 5 (excelente)
freetime	Numérico	Tempo livre após a escola: 1 (muito baixo) a 5 (muito alto)
goout	Numérico	Saídas com amigos: 1 (muito baixo) a 5 (muito alto)
Dalc	Numérico	Consumo de álcool em dias de semana: 1 (muito baixo) a 5 (muito alto)
Walc	Numérico	Consumo de álcool no fim de semana: 1 (muito baixo) a 5 (muito

		alto)
health	Numérico	Status atual de saúde: 1 (muito ruim) a 5 (muito bom)
absences	Numérico	Número de faltas escolares (de 0 a 93)
G1	Numérico	Nota do primeiro período (de 0 a 20)
G2	Numérico	Nota do segundo período (de 0 a 20)
G3	Numérico	Nota final (de 0 a 20, variável alvo de saída)

Tabela 1 - Descrição do conjunto de dados

Historicamente, o sexo tem sido um dos principais fatores de discriminação em diversos contextos sociais. Na sociedade ocidental, mulheres são frequentemente desfavorecidas tanto cultural quanto institucionalmente. Elas recebem, em média, salários menores que os homens para a realização das mesmas funções, enfrentam barreiras à ascensão profissional e são frequentemente consideradas inadequadas para determinadas áreas, especialmente aquelas associadas à técnica, engenharia ou tecnologia (**WORLD ECONOMIC FORUM**, 2025; **UNESCO**, 2018). Além disso, estigmas de gênero desencorajam, desde cedo, o interesse de muitas meninas por carreiras tradicionalmente vistas como masculinas, contribuindo para a perpetuação da desigualdade de gênero no mercado de trabalho e na produção científica e tecnológica. Por esse motivo, "*sexo*" foi escolhido como atributo sensivel e "*feminino*" como categoria protegida. O modelo foi mitigado de forma que o valor do atributo "sexo" não influencie a taxa de falso positivo e falso negativos, garantindo que o modelo não tenha um viés erroneo que venha a prejudicar o grupo protegido.

O *target*, ou seja, o atributo que guarda o conjunto de categorias que o modelo deve classificar, foi a média final dos alunos, atributo "*G3*" do conjunto de dados. Inicialmente esse atributo possuia um intervalo de valores inteiros, entre 0 e 20, sendo 20 média final máxima e 0 a média final mínima, para a avaliação desenvolvida nesse trabalho essa granularidade não é relevante, considerando isso e na intenção de simplificar o modelo, foi definido um *limiar* a partir do qual as notas seriam binarizadas, receberiam valores 0 ou 1, 0 representando todos os valores abaixo do limiar, e 1 representando todos os valores acima desse limiar. O limiar definido foi 15, que numa escala de 0 a 10, como a usada para atribuir notas em muitas instituições de ensino, equivale a *média escolar* 7,5, estando os alunos com nota superior ao limiar considerados "aprovados" e os demais "reprovados". A escolha desse valor como limiar se deu na tentativa de definir um limiar rigoroso, que demonstrasse que o aluno não apenas atingiu uma média razoável mas também que tem um conhecimento consolidado.

3.2 Preparação e transformação dos dados

A importação e transformação dos dados é uma parte essencial do processo de treinamento. Geralmente, nessa etapa os dados são trazidos a partir da sua fonte em arquivos de dados, limpos e transformados para maior conveniência de uso. No caso tratado neste artigo, foram feitos alguns processos de codificação dos dados para que fossem processados adequadamente pelo modelo classificador, como ilustra a Figura 1. O modelo-base usado, a implementação do Random Forest Classifier da biblioteca scikit-learn aceita somente dados numéricos, então foi necessário converter os atributos categóricos em atributos numéricos.

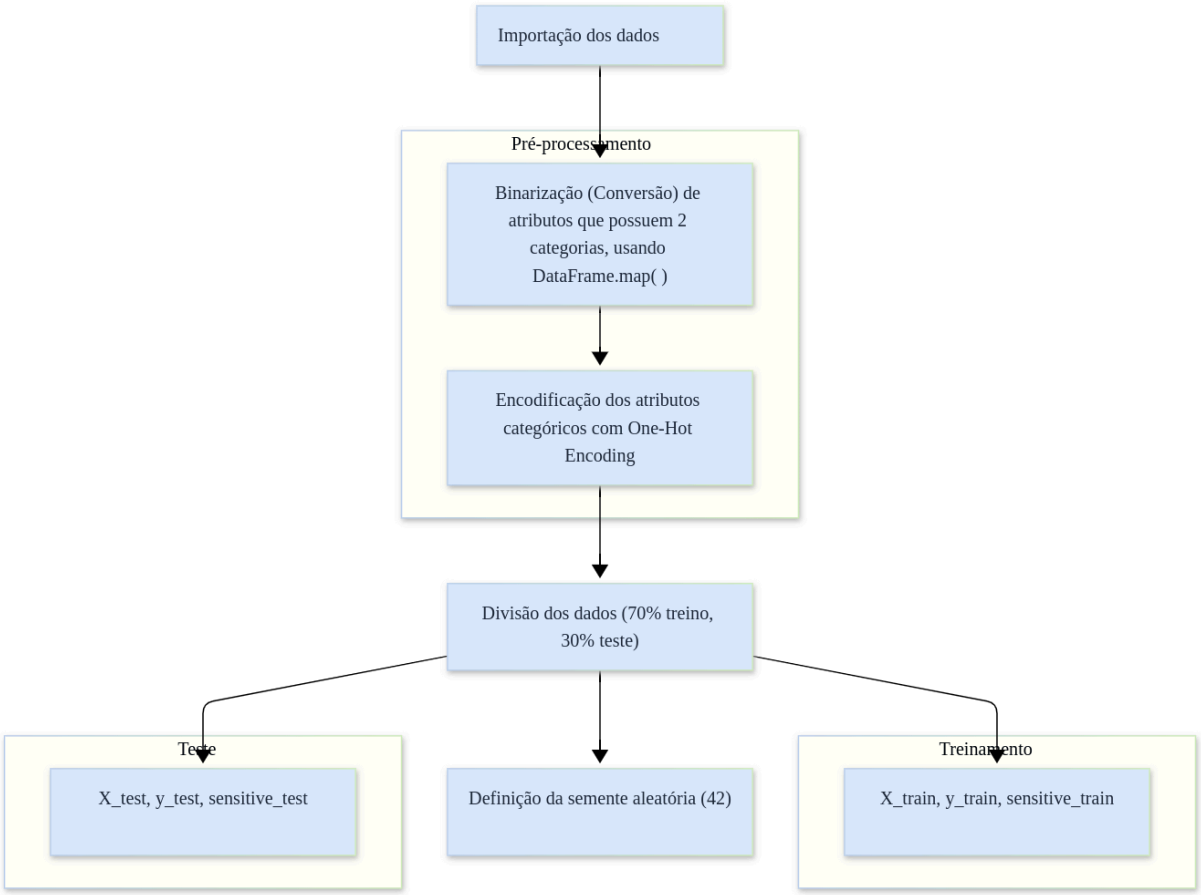

Figura 1 - Fluxograma descrevendo o processo de preparação dos dados

Para isso usou-se a técnica de One Hot Encoder (GÉRON, 2019), uma técnica de codificação utilizada para transformar variáveis categóricas em vetores binários. Cada categoria é representada por um vetor com valor 1 na posição correspondente e 0 nas demais, assim, podemos por exemplo representar o atributo *sexo* como um vetor de duas posições [1,0] para a categoria “masculino” e [0,1] para a categoria “feminino”, com a posição do valor 1 representando a categoria. Essa abordagem é

amplamente adotada em algoritmos de aprendizado de máquina que requerem entradas numéricas.

Com os dados devidamente tratados e convertidos para a forma mais adequada, foram divididos em 6 subgrupos: ***X_train, X_test, y_train, y_test, sensitive_train, sensitive_test***. Grupos com sufixo "***_train***", foram usados no processo de treinamento dos modelos, e grupos com sufixo "***_test***" no processo de teste e coletas de métricas. Os grupos com raiz "***X***" contém os atributos descritores. Os grupos "***y***" contém os rótulos, as "respostas" das classificações que o modelo deve fazer em seu processo de treinamento e teste. Os grupos "***sensitive***" contém os valores do atributo sensível. A divisão desses grupos teve proporção de 70% dos dados para o processo de treinamento e 30% para o processo de teste, com embaralhamento das amostras para evitar vieses presentes na ordenação dos dados e com uma semente aleatória fixa para garantir reprodutibilidade. A semente aleatória utilizada foi ***42*** e foi a utilizada em todos métodos e algoritmos onde uma semente aleatória foi necessária.

3.3 Mitigação de injustiça

A princípio foi construído um modelo controle, onde não foi realizada nenhuma otimização pensada para fairness ou técnica de mitigação, ele foi usado como referência de comparação das métricas, útil para saber como o modelo se comportaria sem intervenção. Sem isso, não seria possível afirmar se uma técnica de fairness melhorou ou piorou o desempenho ou a justiça do modelo. Esse modelo teve como base a implementação do algoritmo ***Random Forest Classifier*** da biblioteca ***sci-kit learn***.

As três técnicas de mitigação durante o processamento descritas anteriormente foram usadas: ***Exponentiated Gradient Reduction, Grid Search Reduction e Adversarial Debiasing***. Em todas, exceto em Adversarial Debiasing que utiliza redes neurais, o modelo-base foi o mesmo: o algoritmo ***Random Forest Classifier*** usado no modelo controle, o mesmo é amplamente usado no desenvolvimento de modelos de propósito geral. Tanto a implementação de cada um das técnicas feita pela biblioteca Fairlearn, quanto pela biblioteca AI Fairness 360 foram avaliados, e os resultados foram compilados em duas tabelas, uma para as métricas de performance do modelo e uma para as métricas de fairness.

Adversarial Debiasing, diferente dos outros algoritmos, requereu um algoritmo base diferente. No caso do AIF360, isso é transparente e não completamente configurável, apenas através de hiperparâmetros, já com o Fairlearn foi necessário criar duas redes neurais, uma preditora e adversária. Usando a API Keras do Tensorflow, tanto a rede neural preditiva quanto a rede neural adversária foram definidas como um modelo sequencial de duas camadas densas: uma camada oculta com 10 neurônios e função de ativação ReLU, e uma camada de saída com

um único neurônio e função de ativação Sigmoide, adequada para tarefas de classificação binária.

3.4 Explicação do modelo

Por fim, foi feita a explicação da influência das características no resultado final, utilizando o explicador **LIME**. O resultado produzido pelo **LIME** (*Local Interpretable Model-agnostic Explanations*) consiste em uma lista ordenada de características acompanhadas de seus respectivos pesos, que representam a contribuição local de cada característica para a predição de um determinado atributo. (**XIMENES**, 2020; **RIBEIRO** et al., 2016) O *LIME* constrói uma aproximação linear interpretável do modelo original em torno de uma instância específica, e os pesos resultantes indicam o grau e a direção (positiva ou negativa) da influência de cada *feature* sobre a decisão do modelo para aquela instância. No caso do modelo contruido nesse artigo, se a influência é positiva maiores as chances de um aluno do sexo feminino receber nota aceitável ao limiar definido, sendo considerada "**aprovada**", é se a influência é negativa maiores as chances de um aluno do sexo feminino receber nota inferior ao limiar, sendo considerada "**reprovada**". Esse resultado do explicador é essencial para nossa análise, pois permite analisar de forma transparente e localizada quais atributos foram mais determinantes para a predição feita pelos modelos.

Todos os atributos do conjunto de dados foram considerados na explicação, exceto, é claro, o atributo alvo (G3).

4 RESULTADOS E ANÁLISE

A partir dos experimentos realizados foi feita uma análise, tanto para métricas tradicionais de performance, quanto de fairness. Com a eficácia em mitigação de injustiça determinada pela menor média de valores de *paridade demográfica*, *equalized odd* e *equal opportunity* próximos a 0, e eficiência determinada por média de valores próximos a 1 para *precisão* e *recall*.

4.1 Análise Comparativa

Como visível na Figura 2, as métricas coletadas indicam uma tendência leve de ganho de acurácia e precisão em modelos mitigados se comparados ao modelo controle. Apesar dos modelos construídos com implementações fornecidas pelo AIF360 terem tido melhor desempenho nas maioria das métricas de fairness, como mostra a Figura 3, eles obtiveram valores de recall ligeiramente menores, até mesmo que o modelo controle.

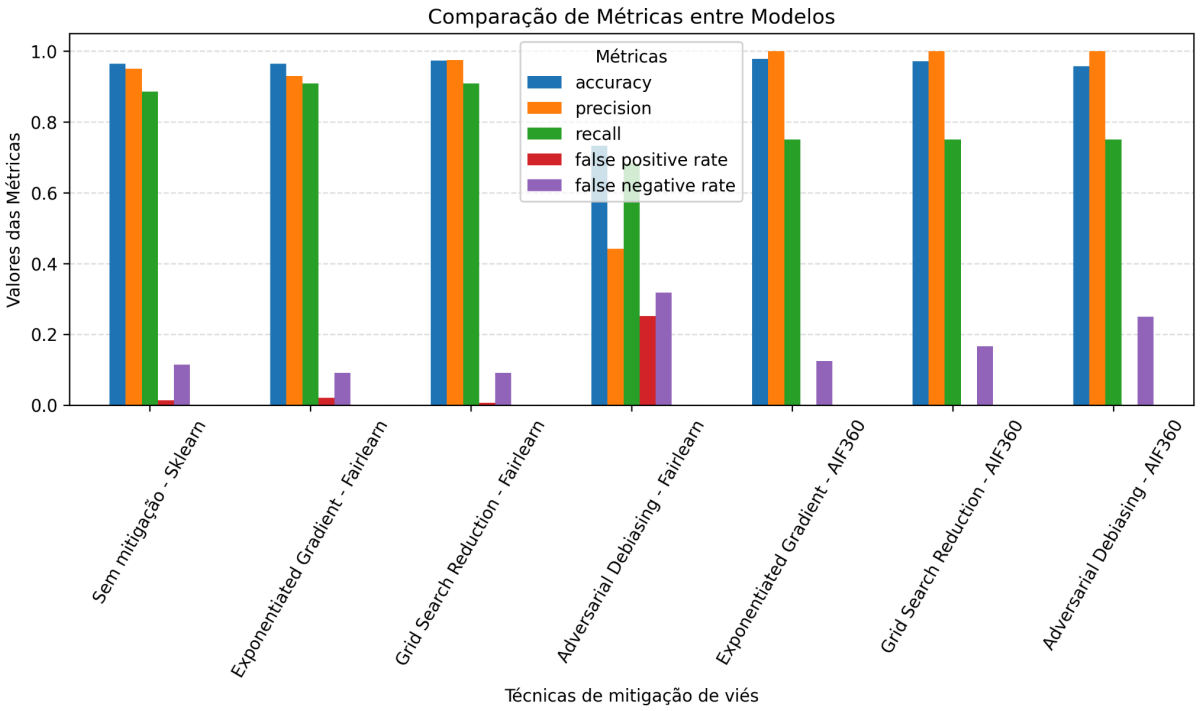

Figura 2 - Gráfico de métricas de performance dos modelos

As métricas de justiça avaliadas foram utilizadas para mensurar disparidades no comportamento do modelo entre os grupos sensíveis, tendo os modelos mitigados resultados consideravelmente melhores, principalmente os modelos mitigados com as implementações da biblioteca ***AI Fairness 360***, destacando-se: ***Grid Search Reduction***, que possue a menor média entre as métricas de fairness avaliadas, como mostra a Figura 3.

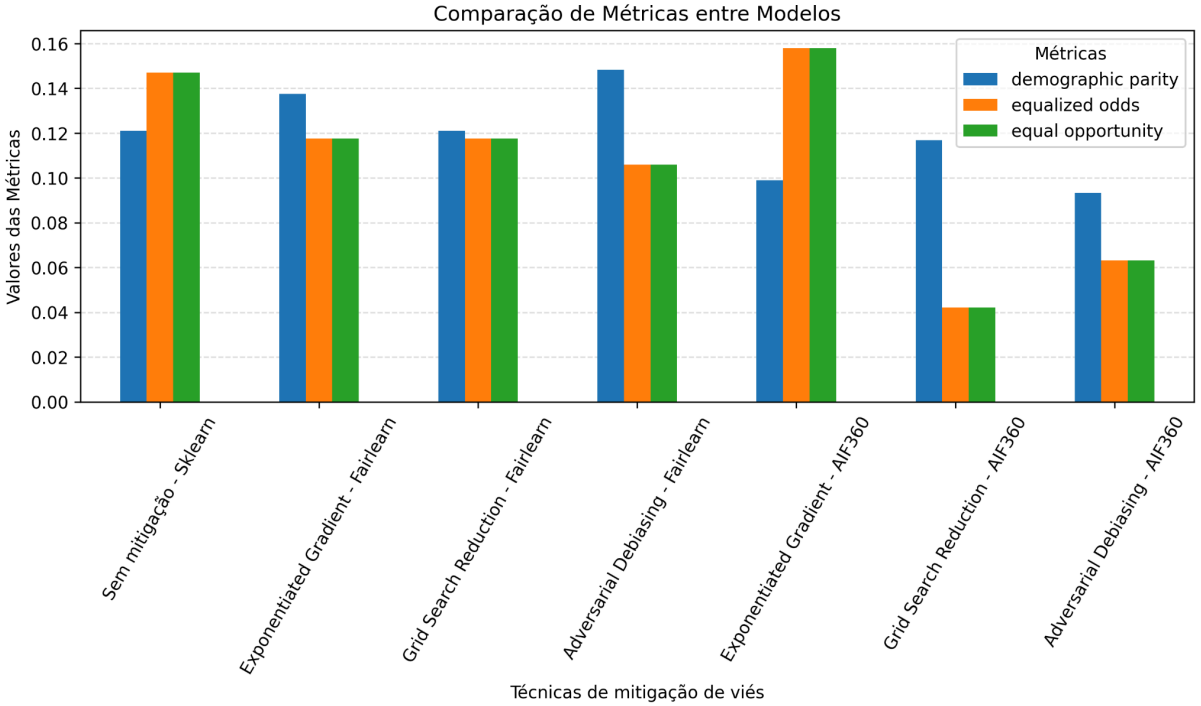

Figura 3 - Gráfico de métricas de fairness dos modelos

4.2 Influências dos atributos

A explicação gerada para uma amostra do conjunto de classificações indica de forma geral a influência de cada característica sobre o resultado da amostra avaliada, dando indícios sobre combinação de fatores que guiavam a resposta do modelo enviesado e dos modelos corrigidos.

Para padronização das explicações foi utilizada a mesma amostra presente no conjunto de predições realizadas pelos modelos, a amostra de índice **15** do conjunto de classificações.

A explicação (Figura 4) para a amostra do modelo controle sugere, que para a dada amostra onde G3 foi predito como 0 e sexo presente no conjunto de treinamento igual a 1, outras variáveis sensíveis, que não a mitigada, têm mais influência sobre a predição, nesse caso principalmente a área de atuação profissional da mãe do aluno e o status de relacionamento dos pais (casados ou separados).

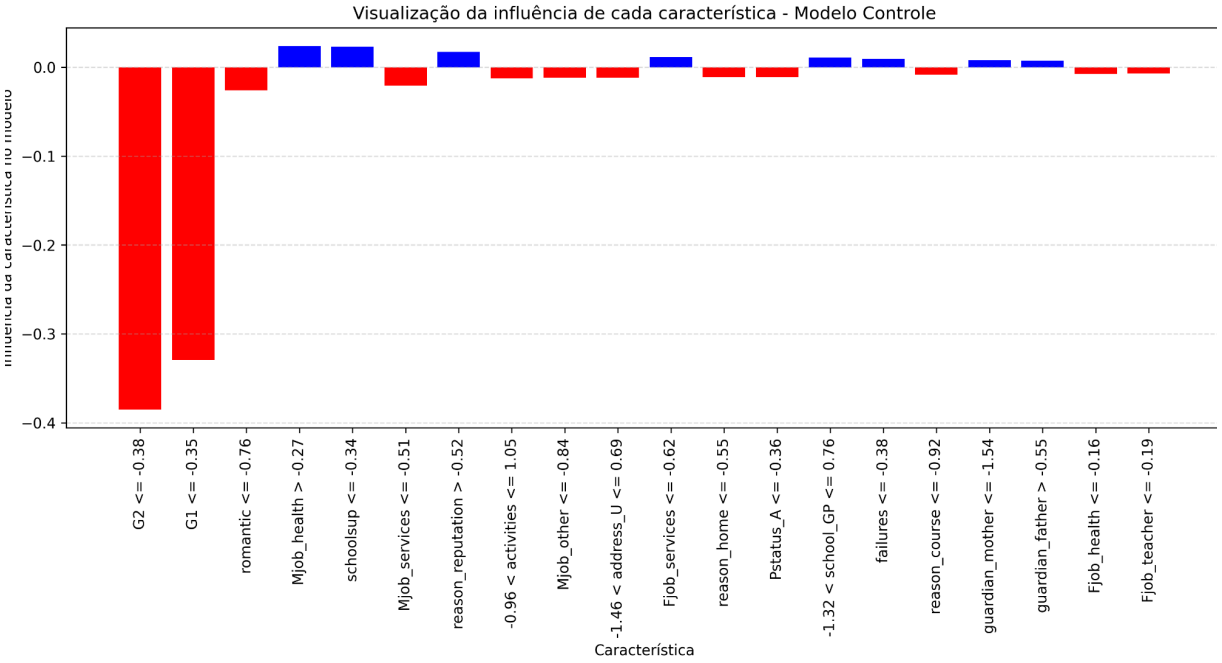

Figura 4 - Gráfico de influência das características para o modelo controle

A explicação (Figura 5) para a amostra do modelo com mitigação por Exponentiated Gradient Reduction implementada pelo *Fairlearn*, onde G3 foi predito como 0 e sexo presente no conjunto de treinamento igual a 1, demonstra mais uma vez, que para a dada amostra, a variável sensível não foi tão relevante sobre a predição, não estando nem entre as vinte características mais influentes.

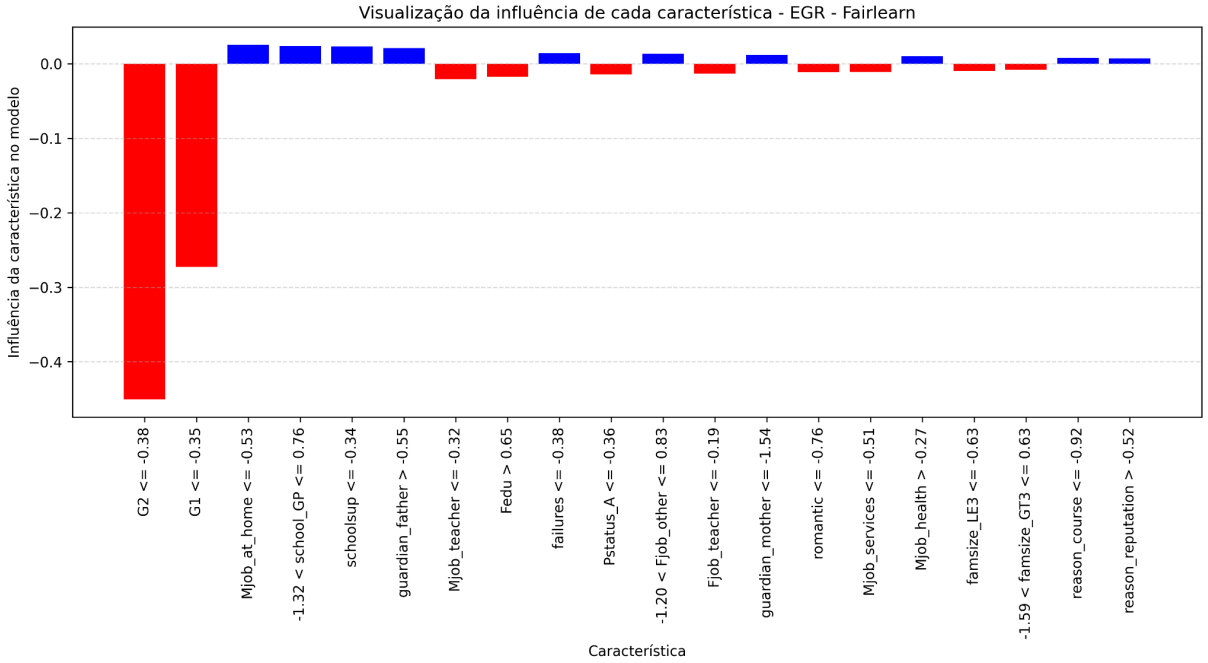

Figura 5 - Gráfico de influência das características para a implementação de EGR do Fairlearn

A explicação (Figura 6) para a amostra do modelo com mitigação por Grid Search Reduction implementada pelo *Fairlearn* sugere, que para a dada amostra onde G3 foi predito como 0 e sexo presente no conjunto de treinamento igual a 1, mais uma vez, outras variáveis sensíveis que não a mitigada pelo modelo têm mais influência sobre a predição, e novamente a área de atuação profissional dos pais e se são separados ou não, foram fatores sensíveis mais decisivos.

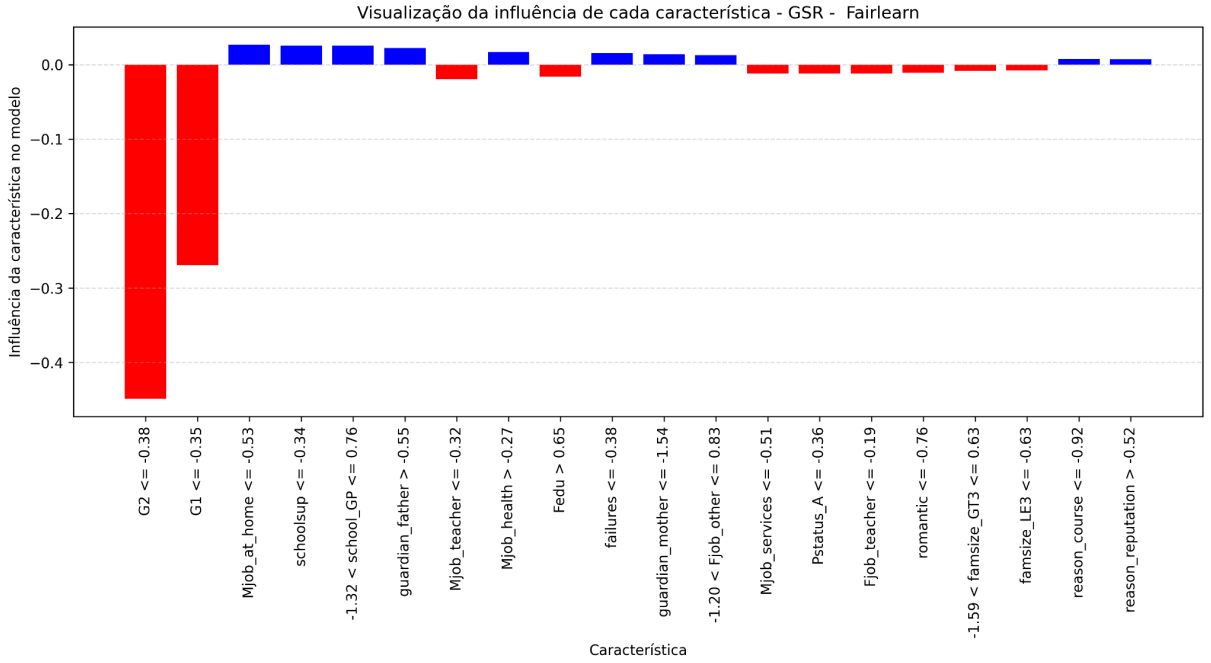

Figura 6 - Gráfico de influência das características para a implementação de GSR do Fairlearn

A explicação (Figura 7) para a amostra do modelo com mitigação por Adversarial Debiasing implementada pelo *Fairlearn* sugere, que para a dada amostra onde G3 foi predito como 1 e sexo presente no conjunto de treinamento igual a 1, mais uma vez a variável sensível avaliada não é tão relevante, e a profissão dos pais, principalmente da mãe, se mostra mais decisiva.

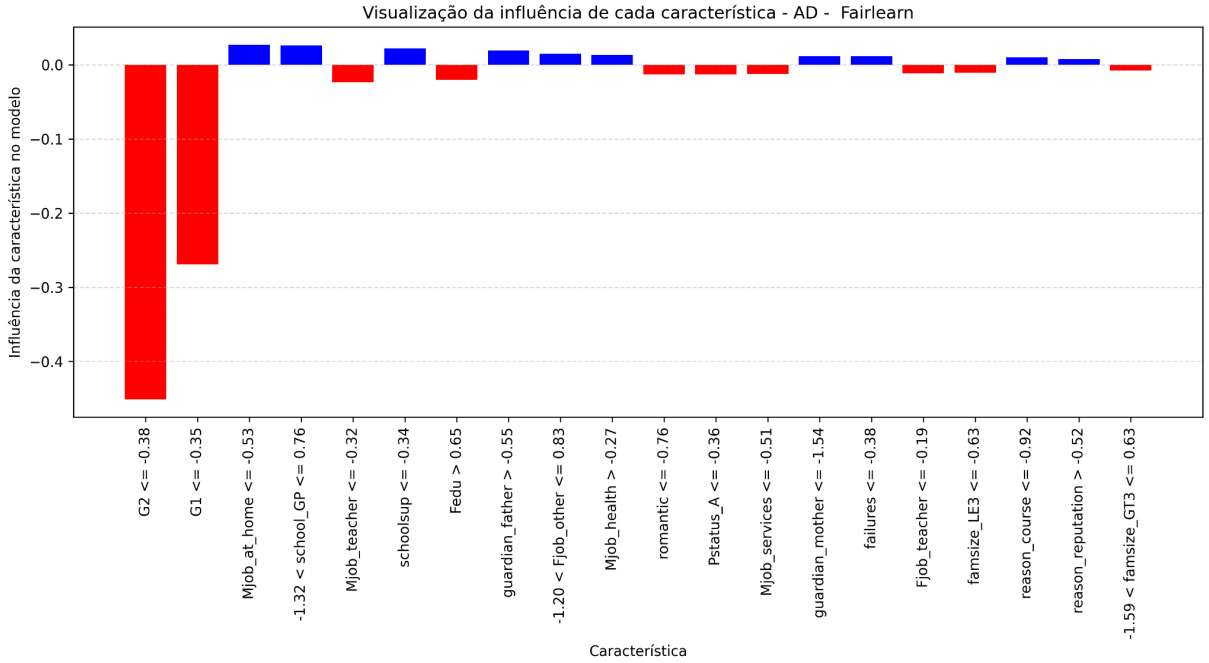

Figura 7 - Gráfico de influência das características para a implementação de AD do Fairlearn

A explicação (Figura 8) para a amostra do modelo com mitigação por Exponentiated Gradiente Reduction implementada pelo *AI Fairness 360* sugere, que para a dada amostra onde G3 foi predito como 0 e sexo presente no conjunto de treinamento igual a 1, diversos outros fatores foram mais relevantes, e mais uma vez a profissão da mãe figura como uma característica relevante no resultado final.

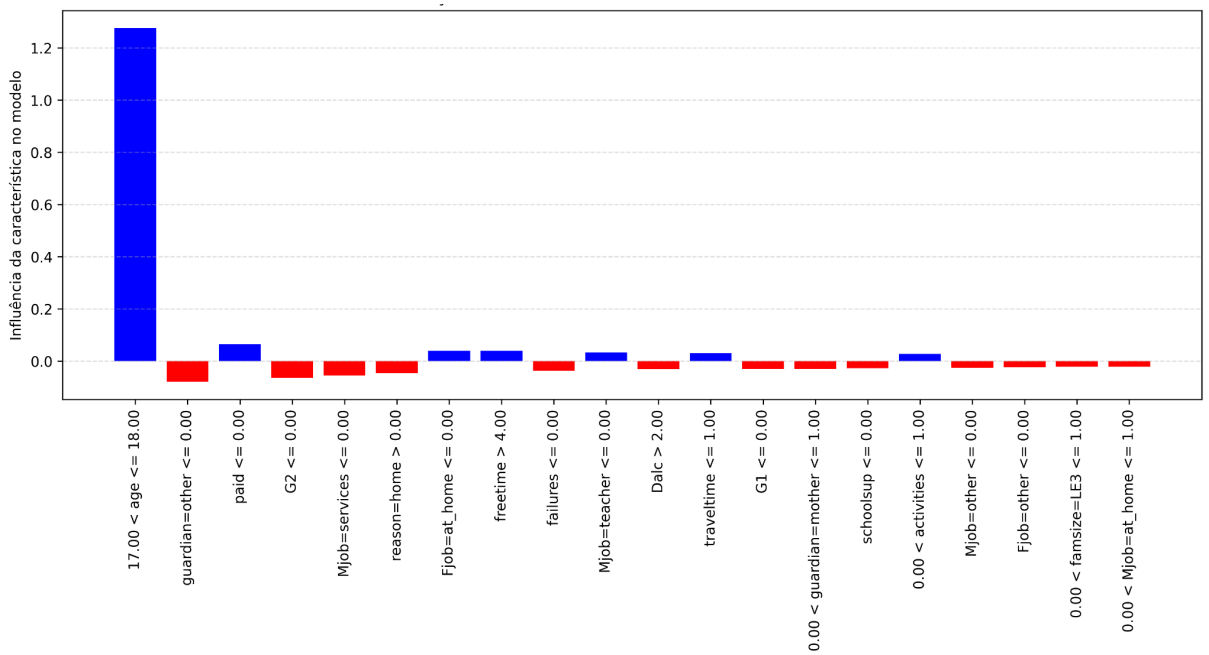

Figura 8 - Gráfico de influência das características para a implementação de EGR do AIF360

A explicação (Figura 9) para a amostra do modelo com mitigação por Grid Search Reduction implementada pelo *AI Fairness 360*, onde G3 foi predito como 0 e sexo presente no conjunto de treinamento igual a 1, revela que entre os modelos avaliados essa mitigação foi a que mais reduziu a influência da variável sensível, não sendo nem a vigésima mais influente no resultado final dessa amostra.

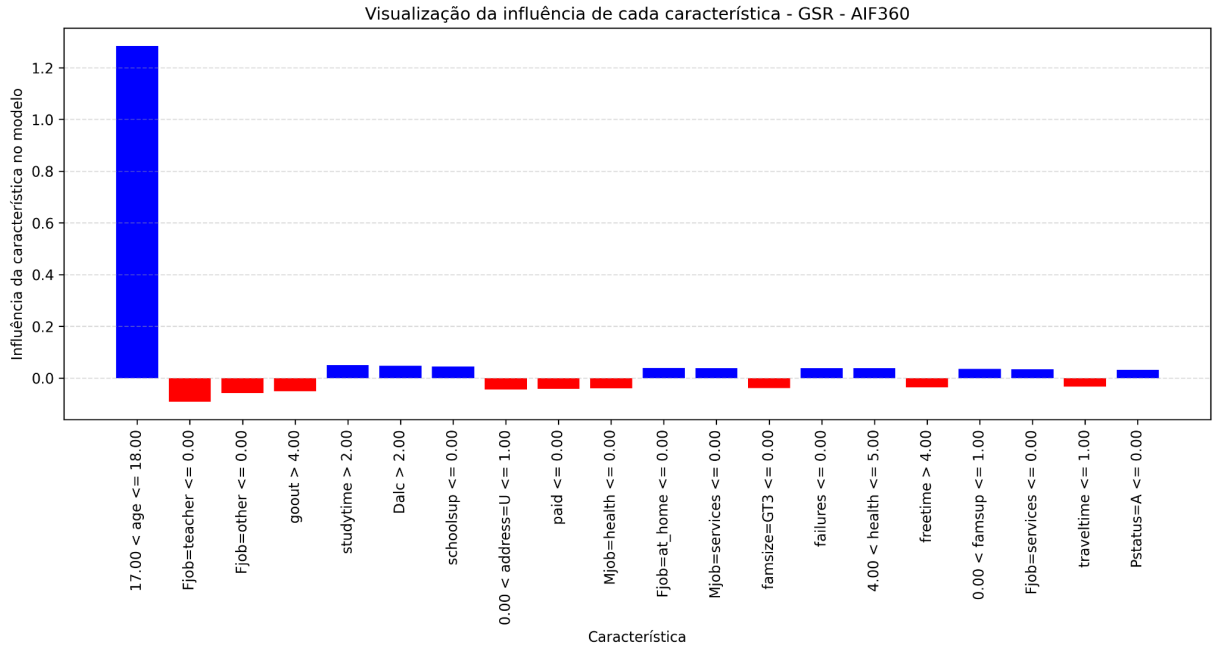

Figura 9 - Gráfico de influência das características para a implementação de GSR do AIF360

A explicação (Figura 10) para a amostra do modelo com mitigação por Adversarial Debiasing implementada pelo *AI Fairness 360* sugere, que para a dada amostra onde G3 foi predito como 0 e sexo presente no conjunto de treinamento igual a 1, a variável sensível não teve influência significativa na predição.

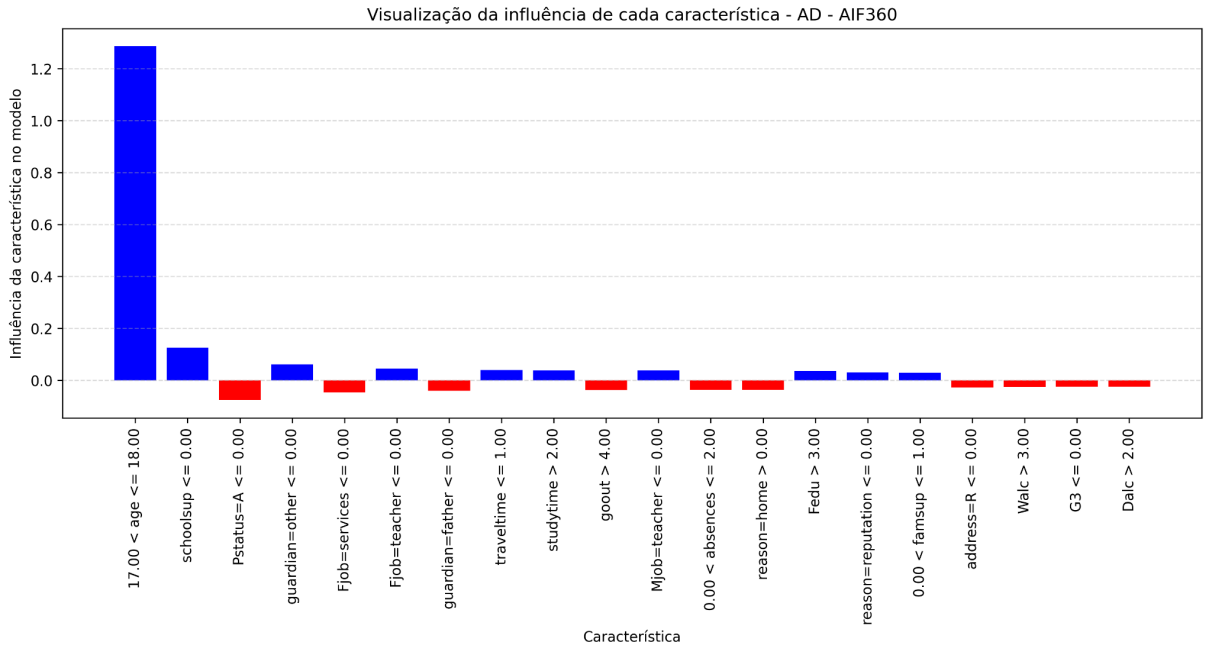

Figura 10 - Gráfico de influência das características para a implementação de AD do AIF360

5 CONSIDERAÇÕES FINAIS E CONCLUSÕES

Com base nos resultados, observa-se que a técnica de ***Grid Search Reduction*** implementada pela biblioteca ***AI Fairness 360***, foi capaz de reduzir o viés de forma mais equilibrada, com impacto controlado na acurácia e muito positivo na precisão, **se mostrando a opção mais eficiente**, segundo parâmetros definidos na seção 4 deste trabalho, são eles: menor média de valores de *paridade demográfica*, *equalized odd* e *equal opportunity* entre as implementações avaliadas e maior média de precisão e recall.

No contexto de fairness desenvolvido no experimento, apesar de as explicações locais fornecidas pelo LIME indicarem que a variável "*sexo*" **não teve influência direta significativa nas predições individuais do modelo controle**, as métricas de equidade apontam para disparidades relevantes entre os grupos sensíveis. Isso indica que o modelo pode estar **reproduzindo desigualdades** presentes nos dados **de forma indireta**, por meio de variáveis correlacionadas, o que é detectado pelas métricas estatísticas de fairness, mas não necessariamente captado em explicações locais.

5.1 Trabalhos futuros

Como trabalho futuro, propõe-se a reprodução do modelo desenvolvido neste estudo, aplicando a mitigação de *fairness* **antes do treinamento**, ou seja, na fase de preparação dos dados. A principal ideia é investigar o impacto da **remoção ou transformação de atributos com alta correlação com a variável sensível** (neste

caso, o sexo), que possam contribuir para o viés algorítmico. O objetivo é reduzir o enviesamento diretamente nos dados, **preservando a precisão do modelo** e, idealmente, **dispensando a necessidade de aplicação de técnicas de correção posteriores**. Essa abordagem permitiria verificar a efetividade de estratégias de mitigação baseadas em engenharia de atributos e análise estatística prévia.

REFERÊNCIAS

AGARWAL, A. et al. **A Reductions Approach to Fair Classification**. 6 mar. 2018.

ANDERS, J. et al. **Grade Expectations: How well can we predict future grades based on past performance?** [s.l: s.n.]. Disponível em: <https://EconPapers.repec.org/RePEc:ucl:cepeow:20-14>.

BAROCAS, S.; HARDT, M.; NARAYANAN, A. **FAIRNESS AND MACHINE LEARNING Limitations and Opportunities**. [s.l: s.n.]. Disponível em: <https://fairmlbook.org/>. 2023.

BIRD, S. et al. **Fairlearn: A toolkit for assessing and improving fairness in AI**. [s.l: s.n.]. Disponível em: <https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/>. Acesso em: 3 jun. 2025.

CORTEZ, P.; SILVA, A. M. G. **Using data mining to predict secondary school student performance**. 2008.

DWORK, C. et al. **Fairness through awareness**. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. **Anais**...New York, NY, USA: ACM, 8 jan. 2012.

FAWCETT, T. **An introduction to ROC analysis**. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, jun. 2006.

GÉRON, A. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow Conceitos, Ferramentas e Técnicas para a Construção de Sistemas Inteligentes**. [s.l: s.n.]. p. 65–66. 2019.

HARDT, M.; PRICE, E.; SREBRO, N. Equality of opportunity in supervised learning. **Advances in neural information processing systems**, v. 29, 2016.

IBM. ***AI Fairness 360 (AIF360): An open-source toolkit for detecting, understanding, and mitigating unwanted bias in machine learning models*.** Disponível em: https://aif360.readthedocs.io/. Acesso em: 3 jun. 2025.

JORDAN, M. I.; MITCHELL, T. M. **Machine learning: Trends, perspectives, and prospects**. **Science**, v. 349, n. 6245, p. 255–260, 2015.

LE QUY, T. et al. **Evaluation of Group Fairness Measures in Student Performance Prediction Problems.** In: [s.l: s.n.]. p. 119–136. 2023.

MAYER-SCHONBERGER, V.; CUKIER, K. **Big data: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana**. [s.l.] Elsevier Brasil, 2014. v. 1

MEHRABI, N. et al. **A survey on bias and fairness in machine learning. ACM computing surveys (CSUR)**, v. 54, n. 6, p. 1–35, 2021.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. **"Why Should I Trust You?": Explaining the Predictions of Any Classifier.** Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. **Anais**...2016.

SU, C.; YU, G; WANG, J; YAN, Z; CUI, L. **A review of causality-based fairness machine learning**. Intell Robot. 2022.

UNESCO. **Decifrar o código:educação de meninas e mulheres em ciências, tecnologia, engenharia e matemática (STEM)**. [s.l.]: UNESCO, 2018. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000264691. Acesso em: 11 jul. 2025.

WORLD ECONOMIC FORUM. ***Global Gender Gap Report 2025: Digest*. Genebra; Cologny: World Economic Forum**, 11 jun. 2025. Disponível em: https://www.weforum.org/publications/global-gender-gap-report-2025/digest/ . Acesso em: 11 jul. 2025.

XIMENES, Bianca. **Explicabilidade em Machine Learning: isso existe?** Campinas: CDG Campinas, 2020. 1 vídeo (26min). Publicado por: InfoQ Brasil. Disponível em: https://www.youtube.com/watch?v=SkojnqoKzPg. Acesso em: 25 nov. 2024.

ZAWACKI-RICHTER, O. et al. **Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education**Springer Netherlands, , 1 dez. 2019.

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. **Anais**...New York, NY, USA: ACM, 27 dez. 2018.