UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

LUIS VINICIUS LAURIANO DE FRANÇA

Balanceamento de dados para mitigar vieses amostrais e algorítmicos: um estudo comparativo

Recife

2025

LUIS VINICIUS LAURIANO DE FRANÇA

Balanceamento de dados para mitigar vieses amostrais e algorítmicos: um estudo comparativo

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Aprendizado de Máquina

Orientador (a): Ricardo Bastos C. Prudencio

Recife

2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

França, Luis Vinicius Lauriano de.
Balanceamento de dados para mitigar vieses amostrais e algorítmicos: um estudo comparativo / Luis Vinicius Lauriano de França. - Recife, 2025.
89f.: il.

Dissertação (Mestrado)- Universidade Federal de Pernambuco, Centro de Informática, Programa de Pós-Graduação em Ciência da Computação, 2025.
Orientação: Ricardo Bastos Cavalcante Prudêncio.

1. Aprendizado de máquina; 2. Justiça algorítmica; 3. Balanceamento de dados. I. Prudêncio, Ricardo Bastos Cavalcante. II. Título.

UFPE-Biblioteca Central

Luis Vinicius Lauriano de França

“Balanceamento de dados para mitigar vieses amostrais e algorítmicos: um estudo comparativo”

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional

Aprovado em: 29/07/2025.

BANCA EXAMINADORA

_____
Profa. Dra. Juscimara Gomes Avelino
Centro de Informática / UFPE

_____
Prof. Dr. Péricles Barbosa Cunha de Miranda
Departamento de Computação /UFRPE

_____
Prof. Dr. Ricardo Bastos Cavalcante Prudêncio
Centro de Informática / UFPE
(orientador)

Dedico este trabalho a minha mãe, Márcia, meus tios, Antônio e Mônica, e a todos os outros familiares e amigos, por todo apoio para que minha formação pudesse ser obtida com êxito.

AGRADECIMENTOS

Primeiramente, agradeço a minha mãe Márcia e aos meus tios Antônio e Mônica, por apoiar e se dedicar para que eu pudesse sempre ir além com os estudos. Não menos importante, também agradeço aos meus primos Emmanuel, Laryssa e Vitória, e aos amigos leais Ricardo, Lucas e Renan, pela amizade, acreditarem e estenderem as mãos em meus projetos e estudos. Agradeço também a todos os outros familiares que não foram mencionados, mas sempre estão apoiando e não deixam de ser importantes. Dedico aqui a vocês meu amor e admiração.

Agradeço ao meu orientador, contribuindo para que eu pudesse sempre ir além, desafiar-me e atingir os objetivos da melhor forma possível, muito obrigado.

Agradeço aos extremamente capacitados professores das disciplinas cursadas e o Centro de Informática (Cin), que me forneceu uma excelente oportunidade com uma ótima estrutura e um adequado ambiente de aprendizado.

RESUMO

A crescente aplicação de modelos de aprendizado de máquina em decisões de alto impacto social exige uma análise rigorosa de seus potenciais vieses. A justiça algorítmica é um campo de pesquisa fundamental, que frequentemente lida com desafios técnicos como o desbalanceamento de grupos sociais, onde a sub-representação de grupos pode levar a resultados discriminatórios. Técnicas de balanceamento de dados são amplamente utilizadas para melhorar a performance preditiva nesses cenários, mas seu impacto sobre a equidade do modelo é pouco compreendido. O objetivo deste trabalho foi, portanto, investigar empiricamente o trade-off entre performance preditiva e justiça algorítmica ao aplicar um conjunto de dez técnicas de balanceamento de dados. Para tal, foi conduzido um estudo comparativo de larga escala, avaliando dez abordagens de balanceamento de dados sobre oito bases de dados distintas, com onze algoritmos de classificação. A análise foi conduzida sob uma ótica dupla, avaliando-se tanto a performance preditiva, medida principalmente pelo F1-Score, quanto a justiça algorítmica, quantificada por meio de índices de paridade de grupo. Os resultados demonstram que a eficácia de cada técnica é altamente dependente do contexto da base de dados. Enquanto técnicas de sobreamostragem, como o SMOTE, frequentemente ofereceram um bom equilíbrio entre ganho de performance e mitigação de viés, abordagens de subamostragem agressiva mostraram-se capazes de degradar a equidade em cenários de desbalanceamento severo, evidenciando um trade-off crítico. Conclui-se que não existe uma técnica de balanceamento universalmente superior e que a construção de modelos de aprendizado de máquina justos exige uma avaliação conjunta e contextual de múltiplas métricas. Este trabalho contribui com um mapeamento empírico dos efeitos dessas técnicas, oferecendo um guia prático para a seleção de estratégias de mitigação de viés de forma mais consciente e responsável.

Palavras-chaves: Aprendizado de Máquina. Justiça Algorítmica. Viés. Balanceamento de Dados. Justiça.

ABSTRACT

The increasing use of machine learning models in high-stakes decision-making requires a rigorous analysis of their potential biases. Algorithmic fairness is a fundamental research field that often deals with technical challenges such as the imbalance of social groups, where the underrepresentation of certain groups can lead to discriminatory outcomes. Data balancing techniques are widely used to improve predictive performance in such scenarios, but their impact on model fairness is still poorly understood. This study aims to empirically investigate the trade-off between predictive performance and algorithmic fairness by applying ten different data balancing techniques. To this end, a large-scale comparative study was conducted, evaluating ten balancing approaches across eight different datasets using eleven classification algorithms. The analysis adopted a dual perspective, employing standard metrics to measure predictive performance and five fairness metrics, computed using the DALEX library, to assess algorithmic fairness. The results show that the effectiveness of each technique is highly dependent on the dataset context. While oversampling methods such as SMOTE often provided a good balance between performance gains and bias mitigation, aggressive undersampling approaches were found to degrade fairness in highly imbalanced scenarios, revealing a critical trade-off. The study concludes that there is no universally superior balancing technique, and that building fair machine learning models requires a joint, context-aware evaluation of multiple metrics. This work contributes an empirical mapping of the effects of these techniques, offering a practical guide for more conscious and responsible bias mitigation strategy selection.

Keywords: Machine Learning. Algorithmic Fairness. Bias. Data Balancing. Justice.

LISTA DE FIGURAS

Figura 1 – A arquitetura da rede adversária 42
Figura 2 – Exemplo de gráfico gerado após aplicação do método fairness_check. . . . 60
Figura 3 – Distribuição do atributo sensível sexo por risco de crédito 87

LISTA DE TABELAS

Tabela 1 – Resumo das características das 8 bases de dados selecionadas para os experimentos. 51

Tabela 2 – Relação dos 11 algoritmos de classificação selecionados para os experimentos, agrupados por família. 52

Tabela 3 – Descrição dos 11 algoritmos de classificação utilizados. 53

Tabela 4 – As 10 abordagens ténicas de balanceamento de dados utilizadas nos experimentos. 54

Tabela 5 – Descrição das métricas de justiça 57

Tabela 6 – Resultados consolidados de performance e justiça para o Experimento 1 (COMPAS). As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações. 65

Tabela 7 – Resultados consolidados de performance e justiça para o Experimento 2 (German Credit). As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações. 66

Tabela 8 – Resultados consolidados de performance e justiça para o Experimento 3 (Adult Income), avaliando o trade-off entre as abordagens. As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações. 68

Tabela 9 – Resultados consolidados de performance e justiça para o Experimento 4 (Default Credit), destacando os efeitos contrastantes das abordagens. As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações. 69

Tabela 10 – Resultados consolidados de performance e justiça para o Experimento 5 (Heart Disease). As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações. 70

Tabela 11 – Resultados consolidados de performance e justiça para o Experimento 6 (CDC Diabetes), ilustrando o ganho duplo em um cenário de desbalanceamento severo. As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações. 72
Tabela 12 – Resultados consolidados de performance e justiça para o Experimento 7 (Credit Card Approval), ilustrando a manutenção da equidade em um cenário de baixo viés inicial. As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações. 73
Tabela 13 – Resultados consolidados de performance e justiça para o Experimento 8 (LSAC), ilustrando o efeito adverso da subamostragem agressiva. As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações. 74
Tabela 14 – Desempenho médio (F1-Score) dos 11 algoritmos de classificação nos diferentes experimentos, considerando o cenário *baseline* (sem intervenção). . 77
Tabela 15 – Desempenho médio (F1-Score) dos 11 algoritmos de classificação nos diferentes experimentos, após a aplicação das técnicas de intervenção. 77

LISTA DE ABREVIATURAS E SIGLAS

ACC	Accuracy
ADASYN	Adaptive Synthetic Sampling
AIF360	AI Fairness 360 Toolkit
AM	Aprendizado de Máquina
AUC-PR	Area Under the Precision - Recall Curve
AUC-ROC	Area Under the Curve - ROC
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
DALEX	Descriptive Machine Learning Explanations
FPR	False Positive Rate
GANs	Generative Adversarial Networks
IA	Inteligência Artificial
KNN	K-Nearest Neighbors
PCA	Análise de Componentes Principais
PN	Negativo Privilegiado
PP	Positivo Privilegiado
PPV	Positive Predictive Value
ROS	Random Oversampling
RUS	Random Undersampling
SMOTE	Synthetic Minority Over-sampling Technique
STP	Statistical Parity
SVC	Support Vector Classifier
SVM	Support Vector Machines
TPR	True Positive Rate
UN	Negativo Não Privilegiado
UP	Positivo Não Privilegiado

VAEs	Variational Autoencoders
WQRF	Weighted Quadratic Random Forest

SUMÁRIO

1 INTRODUÇÃO 16
1.1 OBJETIVOS 19
1.2 CONTRIBUIÇÕES 20
1.3 ESTRUTURA DA DISSERTAÇÃO 20
2 REFERENCIAL TEÓRICO 22
2.1 JUSTIÇA 22
2.1.1 Justiça em aprendizado de máquina 24
2.1.2 Critérios de justiça 26
2.1.3 Métricas de justiça 28
2.2 BALANCEAMENTO DE DADOS 30
2.2.1 Abordagens principais 32
2.2.1.1 Abordagens de niveis de dados 34
2.2.1.2 Abordagens híbridas 36
2.3 BALANCEAMENTO DE DADOS PARA JUSTIÇA 37
2.3.1 Novas técnicas e ferramentas para justiça algorítmica 39
2.3.1.1 Mixup para equidade 39
2.3.1.2 Reponderação de instância 40
2.3.1.3 Abordagens de balanceamento baseadas em redes generativas adversariais . 41
2.3.1.4 Bibliotecas especializadas em justiça 41
2.3.1.5 Remoção de viés adversarial 42
2.3.2 Técnicas de balanceamento de dados aplicadas 45
2.3.2.1 Embaralhamento do Atributo Sensível (Shuffling) 45
2.3.2.2 Duplicação com Classe Oposta 46
2.3.2.3 Subamostragem Aleatória (Random Under-Sampling) 46
2.3.2.4 SMOTE 46
2.3.2.5 FairSMOTE 47
2.3.2.6 ADASYN 47
2.3.2.7 Sobreamostragem Aleatória (Random Over-Sampling) 47
2.3.2.8 NearMiss 47
2.3.2.9 BorderlineSMOTE 48

2.3.2.10 Reponderação sensível de classe . . . 48
3 TRABALHO DESENVOLVIDO . . . 49
3.1 METODOLOGIA GERAL . . . 49
3.2 SELEÇÃO DAS BASES DE DADOS . . . 50
3.3 ALGORITMOS DE CLASSIFICAÇÃO SELECIONADOS . . . 51
3.4 DESCRIÇÃO DAS TÉCNICAS DE BALANCEAMENTO DE DADOS APLICADAS . . . 53
3.5 DIVISÃO DOS CONJUNTOS DE DADOS . . . 54
3.6 MEDIDAS DE AVALIAÇÃO DE PERFORMANCE PREDITIVA . . . 54
3.6.1 Acurácia (Accuracy) . . . 55
3.6.2 Recall (Revocação ou Sensibilidade) . . . 55
3.6.3 Precisão (Precision) . . . 55
3.6.4 F1-Score . . . 56
3.6.5 AUC (Area Under the ROC Curve) . . . 56
3.7 MÉTRICAS DE AVALIAÇÃO DE JUSTIÇA . . . 56
3.7.1 Paridade Estatística . . . 57
3.7.2 Igualdade de Oportunidade . . . 57
3.7.3 Igualdade Preditiva . . . 58
3.7.4 Paridade Preditiva . . . 58
3.7.5 Igualdade de Acurácia . . . 58
3.8 BIBLIOTECA DALEX . . . 58
3.9 EXECUÇÃO DA ETAPA 1: ANÁLISE DA BASELINE (BASELINE) . . . 60
3.10 EXECUÇÃO DA ETAPA 2: ANÁLISE PÓS-INTERVENÇÃO . . . 61
3.11 ANÁLISE COMPARATIVA DOS RESULTADOS . . . 61
4 RESULTADOS E DISCUSSÕES . . . 63
4.1 RESULTADOS OBTIDOS . . . 64
4.1.1 Compas . . . 64
4.1.2 German Credit . . . 65
4.1.3 Adult Census . . . 66
4.1.4 Default of Credit Card Clients . . . 68
4.1.5 Heart Disease . . . 69
4.1.6 Diabetes Health Indicators . . . 71
4.1.7 Credit Card Approval . . . 72

4.1.8 Law School **73**
4.2 DISCUSSÃO GERAL E SÍNTESE DOS RESULTADOS 74
4.2.1 Análise geral do desempenho dos algoritmos **76**
4.2.2 Análise geral das métricas **78**
5 CONCLUSÕES **79**
REFERÊNCIAS **81**

APÊNDICES 85
APÊNDICE A – DESCRIÇÃO DETALHADA DAS BASES DE DADOS **86**

1 INTRODUÇÃO

Diante do mundo globalizado do qual vivemos, com uma transformação digital constante, o Aprendizado de Máquina (AM) surge como uma das principais tecnologias de impacto na sociedade. Sejam sistemas de recomendação que moldam nosso consumo cultural a soluções médicas que podem salvar vidas, os algoritmos de AM estão cada vez mais integrados e comuns na sociedade. Apesar de muitas das vezes acabar passando despercebido, estes algoritmos tornam-se úteis por serem capazes de otimizar processos, gerar insights a partir de vastos conjuntos de dados e automatizar decisões em uma escala sem precedentes (JORDAN; MITCHELL, 2015). A capacidade desses modelos de AM em aprender padrões a partir de dados históricos e fazer previsões acuradas sobre eventos futuros deu a eles um papel central em setores críticos como finanças, saúde, segurança e transporte.

O potencial do AM para impulsionar e impactar a sociedade como um todo é inegável. Entretanto, é preciso discutir também os desafios complexos e de alta relevância ética e social que os modelos de AM, apesar de sua matemática sofisticada, podem trazer. Os modelos de AM são um reflexo direto dos dados com os quais são treinados, porém, se os dados de treinamento contêm vieses históricos, desigualdades sistêmicas ou representações distorcidas da realidade, o modelo resultante não apenas aprenderá, mas poderá amplificar essas falhas, levando a consequências prejudiciais e à perpetuação de injustiças (O'NEIL, 2016). A mitigação de viés algorítmico depende crucialmente da adoção de estratégias conscientes e da responsabilização dos profissionais que constroem os modelos, cujo papel é central na determinação da equidade do sistema final.

Um dos problemas técnicos mais prevalentes e estudados que afeta diretamente a performance e a equidade dos modelos de AM é o desbalanceamento de grupos sociais. Este fenômeno ocorre quando a distribuição de exemplos entre as diferentes classes em um conjunto de dados é significativamente desigual. Em problemas de classificação binária, por exemplo, é comum que a classe de interesse (a classe minoritária), como a detecção de uma transação fraudulenta, o diagnóstico de uma doença rara ou a previsão de inadimplência de crédito, seja muito menos frequente que a classe majoritária (eventos normais). Algoritmos de AM padrão, que são geralmente otimizados para maximizar a acurácia geral, tendem a desenvolver um viés preditivo em favor da classe majoritária, simplesmente por ser a estratégia que minimiza o erro global. Como consequência, eles exibem um desempenho sofrível na identificação da

classe minoritária, que é, ironicamente, muitas vezes a mais crucial do ponto de vista prático (HE; GARCIA, 2009).

Outro ponto importante relacionado ao desafio técnico do desbalanceamento, é uma preocupação ética ainda mais profunda: o tratamento de atributos sensíveis ou protegidos, como gênero e raça. Em muitos contextos, especialmente em áreas de alto impacto na sociedade como justiça criminal ou concessão de crédito, os dados históricos estão repletos de vieses sociais e discriminações estruturais. Um modelo treinado nesses dados pode aprender a associar resultados negativos a determinados grupos demográficos, mesmo que essa correlação seja injusta. A criação de modelos de AM exige, portanto, uma abordagem que não apenas busque a precisão preditiva, mas que também incorpore ativamente os princípios de justiça, equidade e transparência. Desconsiderar a forma como um modelo impacta diferentes subpopulações é arriscar a automação da discriminação em larga escala.

A relação entre o desbalanceamento de classes e a presença de atributos sensíveis é particularmente perigosa. Grupos minoritários em termos demográficos são, frequentemente, também minoritários nos conjuntos de dados, o que agrava o problema. Um modelo treinado em dados de grupos sociais desbalanceados pode não apenas falhar em prever corretamente para a classe minoritária em geral, mas pode falhar de forma desproporcional para indivíduos que pertencem a grupos demográficos protegidos dentro dessa classe minoritária. Isso cria um duplo viés: o viés amostral, oriundo da baixa representatividade, e o viés algorítmico, que penaliza ainda mais subgrupos específicos, resultando em um modelo que é simultaneamente impreciso e discriminatório.

Durante muito tempo, uma estratégia amplamente utilizada para evitar discriminação nos modelos era simplesmente remover os atributos sensíveis dos dados de entrada, assumindo que, ao não considerar informações como sexo ou raça, os algoritmos não seriam capazes de discriminar os indivíduos. A premissa era que, se o modelo não "visse"esses atributos sensíveis, ele não poderia usá-los para tomar decisões discriminatórias. Esta prática era vista como suficiente para garantir a neutralidade e evitar possíveis injustiças (GAJANE; PECHENIZKIY, 2017). Contudo, com o avanço dos estudos na área de justiça em AM, notou-se que tal abordagem não só é inadequada, como pode ser contraproducente. A remoção de atributos protegidos não impede que o modelo aprenda vieses de forma indireta. Outros atributos aparentemente neutros, como CEP, histórico de compras, instituição de ensino ou mesmo a frequência de certas palavras em um texto, podem funcionar como proxies (variáveis substitutas) para os atributos sensíveis. Por exemplo, o CEP pode estar altamente correlacionado com a composição racial

e socioeconômica de um bairro. Assim, mesmo sem o atributo raça ou gênero, o modelo pode inferir e discriminar com base na localização, perpetuando o mesmo viés que se pretendia eliminar. Ignorar os atributos protegidos impede, na verdade, que possamos medir e mitigar ativamente o viés, tornando o problema invisível, mas não inexistente. A abordagem moderna defende que esses atributos devem ser mantidos e tratados com sua devida importância, não como preditores, mas como variáveis de auditoria para garantir que os resultados do modelo sejam equitativos entre os diferentes grupos.

Os problemas causados por desbalanceamento de classes são, felizmente, objeto de intensas pesquisas em busca de soluções para eliminar, ou pelo menos reduzir, os efeitos negativos causados durante o treinamento. A literatura apresenta um vasto leque de métodos e técnicas de balanceamento, que podem ser categorizados em três tipos de técnicas principais. A primeira são as técnicas a nível de dados, que modificam a distribuição de classes no conjunto de treinamento, incluindo métodos de sobreamostragem (oversampling), como o popular Synthetic Minority Over-sampling Technique (SMOTE) (CHAWLA et al., 2002), que cria exemplos sintéticos da classe minoritária, e métodos de subamostragem (undersampling), que removem exemplos da classe majoritária. A segunda são técnicas que ocorrem a nível de algoritmo, modificando o próprio processo de aprendizado para dar mais peso aos erros na classe minoritária, uma técnica conhecida como aprendizado sensível ao custo (cost-sensitive learning). Por fim, as técnicas de conjunto (ensemble methods) combinam múltiplos modelos treinados em subconjuntos balanceados dos dados para melhorar a robustez geral da predição (KRAWCZYK, 2016)

A escolha da estratégia de mitigação, tanto para o desbalanceamento quanto para o viés, não é trivial e seus efeitos precisam ser cuidadosamente avaliados. A aplicação de uma técnica de balanceamento pode melhorar a performance na classe minoritária, mas, se não for feita com cautela, pode acabar aumentando a disparidade de performance entre diferentes grupos demográficos, trocando um problema por outro. Portanto, é preciso que seja feito uma análise de diferentes técnicas de tratamento de dados e que seja conduzida sob uma análise de performance preditiva e da justiça algorítmica para entender se realmente está sendo possível obter um resultado positivo.

Tendo em vista a problemática do desbalanceamento de classes em bases de dados das mais diversas áreas (Finanças, Saúde, Justiça Criminal, Educação e entre outras), o objetivo central deste trabalho é investigar como as técnicas de balanceamento impactam a performance preditiva e de justiça para as classes minoritárias. Para tal, será realizado um estudo

empírico robusto, envolvendo 8 experimentos com bases de dados públicas e distintas, que contêm tanto desbalanceamento de classes quanto atributos sensíveis. A análise foi dividida em duas etapas fundamentais: a primeira etapa consistirá no treinamento e avaliação de modelos sem a utilização de quaisquer técnicas de balanceamento, estabelecendo uma linha de base (baseline) para a performance e o viés. A segunda etapa replicará os experimentos, mas desta vez aplicando um conjunto de 10 abordagens de técnicas de balanceamento de dados. Em ambas as etapas, serão treinados 11 algoritmos de classificação distintos para garantir a generalidade dos resultados. Para avaliar o impacto no viés e na justiça, será utilizada a biblioteca Descriptive Machine Learning Explanations (DALEX) (MI2.AI, 2020), uma ferramenta poderosa para explicabilidade e auditoria de modelos AM. Por meio dela, serão utilizadas métricas de justiça True Positive Rate (TPR), False Positive Rate (FPR), Statistical Parity (STP), Positive Predictive Value (PPV) e Accuracy (ACC) para avaliar se, ao final do processo e após a aplicação das técnicas, os modelos podem ser considerados, de fato, justos ou não.

1.1 OBJETIVOS

O presente trabalho tem como objetivo geral investigar o impacto da aplicação de técnicas de balanceamento de dados na mitigação de vieses amostrais (decorrentes do desbalanceamento de classes) e algorítmicos (relacionados a atributos protegidos), sob a perspectiva de desempenho preditivo e de justiça algorítmica. Para alcançar o objetivo geral proposto, os seguintes objetivos específicos foram definidos:

1. Levantamento de bases de dados com diferentes características e domínios.

2. Investigar a presença de viés algorítmico relacionado ao gênero em modelos de classificação treinados com conjuntos de dados desbalanceados, sem aplicação de técnicas de mitigação.

3. Avaliar o desempenho de múltiplos modelos de classificação de aprendizado de máquina sob diferentes estratégias de balanceamento.

4. Utilizar métricas de justiça para mensurar a equidade dos modelos treinados, considerando o grupo privilegiado e o grupo não privilegiado.

5. Analisar, por meio da biblioteca DALEX, a eficácia de cada técnica de balanceamento

na mitigação de viés, comparando os resultados obtidos nas etapas com e sem balanceamento.

6. Comparar quais técnicas de balanceamento são mais adequadas para diferentes cenários e quais os cuidados necessários para garantir que a busca por performance não comprometa a equidade do modelo.

1.2 CONTRIBUIÇÕES

As principais contribuições deste trabalho concentram-se na análise sistemática da mitigação de viés algorítmico de gênero por meio da aplicação de técnicas de balanceamento de dados em modelos de AM. Sendo conduzido oito experimentos com conjuntos de dados de diferentes domínios, permitindo avaliar a generalização das abordagens testadas. O estudo implementou dez técnicas distintas de balanceamento, desde métodos clássicos como undersampling e oversampling até variações modernas como SMOTE, FairSMOTE, Adaptive Synthetic Sampling (ADASYN) e técnicas de reponderação por pesos sensíveis. Além disso, foram utilizados onze algoritmos de classificação de diferentes paradigmas, possibilitando uma análise abrangente da interação entre técnicas de balanceamento, modelos preditivos e métricas de justiça. A biblioteca DALEX foi adotada como principal ferramenta para auditoria de viés, fornecendo uma base rigorosa para a interpretação de métricas de justiça. A originalidade desta dissertação está na proposição de um arcabouço metodológico para a avaliação empírica de técnicas de balanceamento. O acabouço metodológico foi desenhado para avaliar sistematicamente o impacto de múltiplas técnicas de balanceamento sobre os eixos de performance e justiça, através de um conjunto diversificado de algoritmos e bases de dados. Os resultados obtidos fornecem subsídios importantes para a construção de modelos mais justos, éticos e alinhados com princípios de equidade em aplicações sensíveis do aprendizado de máquina.

1.3 ESTRUTURA DA DISSERTAÇÃO

Este trabalho está organizado em sete capítulos, cuja estrutura é descrita a seguir, desde a fundamentação teórica até as conclusões finais:

- **Capítulo 2:** É apresentado a fundamentação teórica que serve de base para a pesquisa. Inicialmente, são explorados os conceitos de justiça em aprendizado de máquina, deta-

lhando os critérios e as principais métricas de justiça utilizadas para quantificar o viés. Este capítulo também está focado no trabalho desenvolvido, apresentando ferramentas e abordagens alternativas. São analisadas bibliotecas proeminentes para a avaliação de justiça de modelos de AM, como o AI Fairness 360 Toolkit (AIF360) e o Fairlearn. Além disso, são discutidas técnicas de mitigação de viés e aumento de dados que diferentes das exploradas neste trabalho, como o balanceamento baseado em redes generativas adversariais, o uso do mixup para equidade e a remoção de viés adversarial.

- **Capítulo 3:** Este capítulo detalha a metodologia de análise comparativa criada para este trabalho. A proposta consiste em um experimento de duas etapas: primeiro, modelos são treinados e avaliados em seu estado original (baseline), em seguida, são retreinados após a aplicação de dez diferentes técnicas de balanceamento. O objetivo é analisar sistematicamente o impacto de cada técnica na performance preditiva e na justiça algorítmica, utilizando a biblioteca DALEX para identificar os trade-offs resultantes.

- **Capítulo 4:** Este capítulo é dedicado à apresentação e análise dos resultados. Os dados coletados nas duas etapas experimentais são apresentados de forma comparativa, utilizando tabelas, gráficos e visualizações para destacar o impacto de cada técnica de intervenção na performance preditiva e nas métricas de justiça dos modelos.

- **Capítulo 5:** O capítulo 5 encerra a dissertação com a conclusão. É realizada uma sumarização dos principais achados, retomando os objetivos propostos e discutindo as implicações dos resultados.

2 REFERENCIAL TEÓRICO

Esta seção apresenta a fundamentação teórica que serviu de base para o desenvolvimento deste trabalho. Primeiramente, o campo de justiça de modo geral será introduzido na Seção 2.1, contextualizando a origem dos vieses algorítmicos e a necessidade de modelos mais equitativos. Em seguida, na Seção 2.1.1, serão discutidos os principais critérios conceituais de justiça em AM. O foco, no entanto, será direcionado para as métricas de justiça quantificáveis, detalhadas na Seção 2.1.2 e 2.1.3, uma vez que são elas que permitem uma avaliação objetiva dos modelos neste trabalho, se são justos ou não. Posteriormente, na Seção 2.2, será abordada a problemática do balanceamento de dados, com uma descrição das suas abordagens principais, como sobreamostragem e subamostragem. Por fim, a Seção 2.3 conectará os dois campos, explorando o uso do balanceamento de dados para promover a equidade e justiça, apresentando as novas técnicas da literatura que tratam desses desafios de forma conjunta e as técnicas que foram aplicadas a este trabalho, estabelecendo o contexto para o modelo experimental proposto.

A literatura atual apresenta uma vasta gama de estudos e novas técnicas no balanceamento de dados para otimização de performance, ou na mitigação de viés através de métodos específicos de justiça, como é possível observar em 2.3.1. No entanto, poucos trabalhos realizaram uma análise comparativa de larga escala para investigar sistematicamente o trade-off entre esses dois objetivos. Frequentemente, as técnicas de balanceamento são aplicadas sem uma auditoria rigorosa de seu impacto sobre a equidade, e as técnicas de justiça são avaliadas sem considerar seu efeito sobre a performance em cenários diversificados de dados desbalanceados. Por tanto, o diferencial desta dissertação está exatamente nesta lacuna. Nesta seção será possível observar o referencial teórico utilizado para criar uma ponte entre esses dois campos, fornecendo um mapeamento empírico abrangente do comportamento de dez técnicas de balanceamento sobre os eixos de performance e justiça simultaneamente.

2.1 JUSTIÇA

Ao decorrer dos últimos anos, nota-se que está se tornando cada vez mais comum e habitual a utilização de algoritmos capazes de tomar decisões. De modo que, os modelos de AM estão sendo inseridos em diversas atividades cotidianas e também em tomadas de decisões críticas

para o aspecto humano. Estas ferramentas, por exemplo, podem ir de uma atividade simples como sugerir palavras ou responder clientes de e-commerce, até mesmo uma atividade mais complexa como decidir se um cidadão está apto a receber um empréstimo.

A confiabilidade em permitir que modelos de aprendizado de máquina sejam construídos e tomem tais decisões pode ser explicada, em partes, tendo em vista que os avanços do campo de aprendizado de máquina e também da inteligência artificial têm levado a resultados impressionantes. Por outro lado, assim como decisões humanas precisam ser tomadas e asseguradas por razões éticas e sociais, quando os modelos de AM tomam decisões que envolvem dilemas éticos e consequentemente irão impactar na vida de pessoas, é fundamental e necessário lidar com as implicações sociais e de justiça. (MEHRABI et al., 2021)

Os beneficios trazidos por soluções que envolvem aprendizado de máquina são essenciais para a sociedade, mas dependendo de como o modelo for construído, seus resultados podem mostrar discriminação contra certos grupos de pessoas que não estão sendo levados em conta na contrução do modelo e levar a questões de equidade na prática (MEHRABI et al., 2021). Dessa forma, têm surgido estudos na área de justiça, responsáveis por desenvolver uma nova área de pesquisa focada em problemas sócio-algorítmicos em soluções de Inteligência Artificial (IA), como justiça, transparência, responsabilidade, explicabilidade e privacidade (KEARNS; ROTH, 2019).

É possível evidenciar alguns casos de algoritmos que refletem preconceitos sociais que perduram anos, em sua maioria relacionados à sexismo, machismo e racismo. Como exemplo, um programa de reconhecimento de imagens do google photos classificou os rostos de dois amigos negros como gorilas (GIBBS, 2015), em outro exemplo os algoritmos ao analisarem fotos de pessoas nas mais diversas situações, classificaram erroneamente homens como se fossem mulheres quando eles estavam na cozinha (ZHAO et al., 2017). Logo, é de suma importância além da utilização de medidas de imparcialidade para quantificar o grau de viés nos modelos de AM, também a busca por abordagens de balanceamento de dados, uma estratégia de mitigação consolidada na literatura como proposto em (CATON; HAAS, 2020) para garantir que os modelos de aprendizado de máquina não exibam vieses em suas previsões e recomendações.

A busca por justiça, portanto, está relacionada além do que apenas garantir precisão nos modelos: exige o compromisso de que esses sistemas sejam justos, transparentes e responsáveis, considerando seus efeitos sociais, especialmente sobre grupos historicamente marginalizados. Isso envolve a adoção de métricas específicas, como métricas de justiça para avaliar disparidades de desempenho entre grupo, bem como o uso de técnicas para mitigação de viés, sendo o

balanceamento de dados uma das abordagens centrais. (HARDT; PRICE; SREBRO, 2016)

Introduzir equidade no desenvolvimento de sistemas de aprendizado de máquina não é apenas uma exigência técnica, mas uma responsabilidade ética que deve partir, principalmente, das pessoas responsáveis pelo seu desenvolvimento. Mais do que corrigir modelos, trata-se de refletir criticamente sobre os dados, os contextos em que são utilizados e os impactos que as decisões automatizadas podem gerar. A equidade algorítmica se apresenta, portanto, como uma dimensão fundamental da confiabilidade e da legitimidade dos sistemas inteligentes em uma sociedade que precisa ser mais justa e inclusiva.

2.1.1 Justiça em aprendizado de máquina

No contexto do aprendizado de máquina, justiça, trata-se dos vieses que estão naturalmente presentes nos modelos de AM, quando não são construídos ou tratados com o devido cuidado. Essa discussão engloba não só os aspectos técnicos, como as diferenças entre correlação e causalidade, mas também discussões de cunho profundamente ético, como a própria definição de justiça. De modo que, a justiça pode ser entendida como um conceito social e subjetivo que descreve a adequação de como uma construção social é medida, e sua definição pode sofrer alterações de acordo com as diferentes culturas e sociedades.(SELBST et al., 2019)

No contexto do aprendizado de máquina, a injustiça algorítmica se manifesta quando um modelo produz resultados sistematicamente desfavoráveis para indivíduos com base em seus atributos protegidos, como raça e gênero. É crucial, no entanto, distinguir entre o viés no sentido social e o viés no sentido puramente estatístico. Um modelo pode apresentar um viés estatístico, por exemplo, uma tendência a superestimar consistentemente um valor em 5 porcentos para todos os indivíduos, sem que isso de fato constitua uma injustiça social. A injustiça surge quando o desempenho ou o erro do modelo é correlacionado com atributos sensíveis, afetando de forma desigual diferentes grupos demográficos.

Inicialmente uma abordagem para evitar modelos de aprendizado de máquina tendenciosos e acreditava-se ser o ideal estava relacionada a desconsiderar os atributos sensíveis presentes, como raça e gênero, ao treinar o modelo e isto já era visto como algo suficiente para evitar possíveis discriminações que os resultados do modelo poderia ter, abordagem que ficou conhecida como "justiça por desconhecimento" (CHEN et al., 2019). Com o avanço dos estudos, notou-se que tal abordagem acaba sendo inadequada, tendo em vista que os atributos sensíveis têm correlação considerável com outros atributos que são muito úteis para o modelo. Dessa

forma, ocorre a discriminação através de tais atributos correlacionados, um fenômeno conhecido como "proxy discrimination"(BAROCAS; SELBST, 2016). Um clássico exemplo trata-se do código postal, que é frequentemente correlacionado com a raça devido a padrões históricos de segregação residencial.

A partir disso, algumas estratégias precisaram ser pensadas e ainda estão sendo desenvolvidas para lidar com este problema de discriminação gerada por modelos de AM e inteligências artificiais, consequentemente afetando até mesmo pessoas, levando em consideração os seus atributos sensíveis. Segundo Mehrabi et al. (2021), a discriminação no contexto do AM pode ser entendida, como o fato de haver prejuízo contra um indivíduo ou um grupo de individuos na tomada de decisão.

Como caminho viável para evitar o desenvolvimento de modelos de aprendizado de máquinas tendenciosos algumas medidas de justiça foram introduzidas na literatura. Como exemplo, paridade estatística condicional (CORBETT-DAVIES et al., 2017), oportunidade igual (HARDT; PRICE; SREBRO, 2016), justiça contrafactual (CHIAPPA, 2019) e probabilidades equalizadas (HARDT; PRICE; SREBRO, 2016).

Uma das conclusões mais contundentes da literatura sobre equidade em AM é que o maior desafio não está na implementação de uma métrica, mas na própria definição do que é justo em um determinado contexto. De fato, argumenta-se que a dificuldade em garantir a justiça algorítmica é mais um problema de filosofia e de escolha de valores do que um desafio puramente técnico (CORBETT-DAVIES; GOEL, 2018). Essa complexidade emerge porque qualquer definição matemática de justiça é, na verdade, uma simplificação de um conceito social multifacetado, um processo de abstração que inevitavelmente perde nuances cruciais (SELBST et al., 2019). Portanto, não existe uma métrica de justiça universalmente superior; a adequação de qualquer medida é estritamente dependente do contexto de aplicação, dos valores sociais em jogo e dos danos que se pretende evitar, exigindo uma seleção cuidadosa (MEHRABI et al., 2021).

A busca por equidade no pré-processamento de dados tem sido explorada sob diversas perspectivas na literatura recente. Trabalhos como o de (WANG et al., 2018) oferecem uma visão crítica, argumentando que o balanceamento demográfico por si só é insuficiente para corrigir vieses em representações profundas. Outros, como (QUARESMINI; PRIMIERO, 2023), ampliam o escopo, defendendo que a justiça algorítmica depende fundamentalmente de múltiplas dimensões da qualidade dos dados. Esses estudos se somam a investigações seminais, como a de (BUOLAMWINI; GEBRU, 2018) sobre o viés em classificação facial, que estabeleceram a urgência

do tema. Apesar da relevância dessas contribuições, elas frequentemente negligenciam a avaliação do impacto dessas técnicas sob a ótica de métricas de performance robustas para classes desbalanceadas. O presente trabalho se diferencia ao abordar esta lacuna específica, propondo uma metodologia que avalia sinergicamente o efeito das técnicas de balanceamento tanto em métricas de justiça quanto em métricas de desempenho apropriadas para o desbalanceamento.

2.1.2 Critérios de justiça

Modelos de aprendizado de máquina, assim como decisões humanas, podem ser enviesados e produzir resultados sistematicamente desfavoráveis para grupos de pessoas com base em atributos sensíveis, como gênero ou raça. Para diagnosticar e mitigar tais vieses, a literatura de justiça algorítmica desenvolveu um conjunto de critérios formais que traduzem noções de equidade em declarações matemáticas.

Por isso torna-se valioso abordar sobre métricas de justiça, entretanto, primeiramente é importante entender do que se trata os critérios de justiça. Seguindo a notação e a estrutura propostas por (BAROCAS; HARDT; NARAYANAN, 2019), os critérios de justiça podem ser definidos a partir das seguintes variáveis:

Atributo Protegido (A): Uma variável que denota pertencimento a um grupo (ex: $A = a$ para o grupo privilegiado e $A = b$ para o não privilegiado).

Rótulo Real (Y): O resultado verdadeiro, onde $Y = 1$ representa o resultado favorável (ex: ser qualificado para o empréstimo).

Predição do Modelo ($\hat{Y}$): A decisão binária do modelo, baseada em um score de risco (R), onde $\hat{Y} = 1$ é a predição favorável.

A partir desta notação, três das mais importantes definições de justiça de grupo podem ser expressas como declarações de independência condicional entre a predição do modelo (R), o atributo protegido (A) e o resultado real (Y).

Independência (ou Paridade Estatística): $R \perp A$.

Conceito: Exige que as predições do modelo sejam estatisticamente independentes do atributo protegido. Na prática, isso significa que a probabilidade de receber um resultado

positivo (a "taxa de seleção") deve ser a mesma para todos os grupos, independentemente de suas características. Por exemplo, a porcentagem de aprovações de crédito deve ser a mesma para homens e mulheres.

Separação (ou Igualdade de Chances/Oportunidades): $R \perp A \mid Y$.

Conceito: Exige que as predições do modelo sejam independentes do atributo protegido, mas condicionadas ao resultado real. Isso se desdobra em duas condições principais:

1. **Igualdade de Oportunidades:** A taxa de verdadeiros positivos deve ser igual para todos os grupos. Ou seja, indivíduos qualificados de todos os grupos devem ter a mesma chance de serem classificados corretamente.
2. **Igualdade de Chances (Equalized Odds):** Exige tanto a igualdade na taxa de verdadeiros positivos quanto na taxa de falsos positivos.

Suficiência: $Y \perp A \mid R$.

Conceito: Exige que o resultado real seja independente do atributo protegido, **condicionado ao score de risco do modelo**. Isso significa que, para indivíduos com o mesmo score de risco, a probabilidade de eles realmente pertencerem à classe positiva deve ser a mesma, não importando o grupo. Esta noção está ligada à **paridade de valor preditivo**.

Os critérios teóricos de Independência, Separação e Suficiência dão origem a um conjunto de métricas práticas, que são utilizadas para auditar o comportamento de um modelo. Essas métricas geralmente comparam as taxas de erro ou de sucesso entre o grupo privilegiado ($A = a$) e o não privilegiado ($A = b$), buscando a paridade entre eles. Os cinco critérios de paridade de grupo centrais avaliados neste trabalho são: Paridade Estística, igualdade de oportunidades, igualdade preditiva, paridade preditiva e igualdade de acurácia. Dessa forma, garante que os classificadores tenham a mesma taxa de falsos positivos para cada subgrupo.

A questão fundamental, contudo, é que nem todas as métricas de justiça são igualmente importantes em todos os cenários. A adequação de uma medida de equidade é estritamente dependente do contexto de aplicação, dos valores sociais e dos danos que se busca evitar, por isso a importância de entender a problemática e o contexto dos dados.

Profissionais informados no domínio devem, portanto, selecionar as métricas que são mais relevantes para seu problema específico. Por exemplo, em um estudo de caso sobre pontuação de crédito justa, (KOZODOI; JACOB; LESSMANN, 2021) concluem que o critério de Separação

(que inclui a Igualdade de Oportunidade) é o mais apropriado para evitar a penalização de grupos que já são financeiramente desfavorecidos. Essa visão é corroborada por guias práticos, como os da Agência da União Europeia para os Direitos Fundamentais (European Union Agency for Fundamental Rights, 2018), que orientam sobre a seleção de atributos protegidos e a avaliação de impacto. Além disso, é crucial reconhecer que as métricas de justiça de grupo não descobrem todos os tipos de injustiça, como a discriminação a nível individual (DWORK et al., 2012), e que, por vezes, soluções não-técnicas para problemas de equidade são não apenas benéficas, mas necessárias para uma mudança social efetiva (ABEBE et al., 2021).

Por mais tentador que seja aspirar a um modelo que satisfaça todos os critérios de justiça simultaneamente, trabalhos seminais como o de (KLEINBERG; MULLAINATHAN; RAGHAVAN, 2016) provaram matematicamente que, para qualquer modelo imperfeito, as principais noções de justiça são mutuamente excludentes. Demonstrando que isso é, em geral, impossível.

Como detalhado por Barocas, Hardt e Narayanan (2019), com exceção de cenários triviais (como um classificador perfeito ou taxas de base idênticas entre os grupos), nenhum par dos três critérios fundamentais: Independência, Separação e Suficiência, pode ser cumprido ao mesmo tempo. Consequentemente, a busca por justiça em aprendizado de máquina não é sobre alcançar um ideal matemático perfeito, mas sim sobre navegar em um campo de trade-offs, decidindo conscientemente qual critério priorizar ou qual grau de desequilíbrio é aceitável para um determinado problema.

2.1.3 Métricas de justiça

As métricas de justiça têm como objetivo quantificar e mensurar a presença de viés algorítmico em modelos de aprendizado de máquina. Enquanto os critérios de justiça definem as condições ideais que um modelo deveria atender para ser considerado justo, as métricas oferecem formas concretas de avaliar se essas condições estão sendo, de fato, satisfeitas na prática. Tais métricas são fundamentais para permitir o diagnóstico sistemático de disparidades de tratamento entre diferentes subgrupos definidos por atributos sensíveis, como raça, gênero, orientação sexual, entre outros.

Em geral, as métricas de justiça são calculadas com base na comparação de indicadores estatísticos entre um grupo considerado privilegiado e um ou mais grupos não privilegiados. A matriz de confusão composta por verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos desempenha papel central nesse processo, pois a partir dela derivam-se

medidas como TPR, FPR e ACC. O foco, nesse caso, não é apenas o desempenho geral do modelo, mas a busca pela paridade dessas taxas entre os grupos. Assim, a equidade é avaliada pela comparação direta desses indicadores, e um catálogo extenso dessas métricas e de suas inter-relações pode ser encontrado em trabalhos de revisão como os de (VERMA; RUBIN, 2018) e (MEHRABI et al., 2021).

Como apresentado na seção anterior, existem diferentes perspectivas para avaliar justiça: paridade estatística, igualdade de oportunidade, paridade preditiva, igualdade preditiva e igualdade de precisão. Cada uma dessas perspectivas está associada a uma métrica específica, e nenhuma delas, isoladamente, é capaz de capturar a totalidade do que se entende por justiça em contextos sociais complexos. Por isso, é interessante que sejam avaliadas múltiplas métricas simultaneamente, permitindo uma análise mais abrangente do comportamento do modelo, além da utilização de diferentes técnicas e abordagens.

A escolha de uma métrica de justiça não é um ato neutro, ela representa uma decisão normativa sobre qual tipo de equidade é mais importante para um determinado contexto social. Cada métrica ilumina um tipo diferente de disparidade, e a compreensão de suas nuances é crucial para uma avaliação de viés responsável. Por exemplo, a paridade estatística busca garantir que a taxa de previsões positivas seja igual entre os grupos. Embora útil para detectar tratamento desigual direto, sua aplicação pode ser enganosa se as taxas de base (a prevalência real do resultado positivo) diferirem entre os grupos, mascarando injustiças ao invés de corrigi-las. Em contraste, a igualdade de oportunidade foca apenas nos indivíduos que de fato merecem o resultado positivo ($Y=1$), exigindo que a taxa de verdadeiros positivos seja a mesma para todos. Esta métrica torna-se especialmente relevante em cenários como admissões ou contratações, onde o objetivo principal é não privar candidatos qualificados de uma oportunidade.

Por outro lado, a paridade preditiva e a igualdade preditiva analisam o erro do modelo de formas distintas. A paridade preditiva garante que uma previsão positiva tenha o mesmo significado para todos os grupos, exigindo que a precisão seja a mesma. Já a igualdade preditiva foca na taxa de falsos positivos, sendo essencial em contextos como o sistema de justiça criminal, onde uma previsão positiva incorreta (um "falso alarme") pode ter consequências devastadoras. É importante notar que, essas duas métricas são frequentemente incompatíveis quando as taxas de base diferem. Por fim, a igualdade de precisão oferece uma visão mais geral, propondo que a acurácia global do modelo seja equivalente entre os subgrupos. A dificuldade em satisfazer múltiplas métricas simultaneamente torna a análise comparativa de

seus trade-offs, como proposto por (FRIEDLER et al., 2019), um passo indispensável na prática.

É importante ressaltar que essas métricas nem sempre podem ser otimizadas simultaneamente. Segundo Barocas, Hardt e Narayanan (2019), existem incompatibilidades fundamentais entre os critérios de independência, separação e suficiência, o que implica que, na maioria dos contextos práticos, é necessário fazer escolhas entre as métricas com base nos objetivos sociais e éticos do sistema. Em outras palavras, a busca por justiça algorítmica não é apenas uma tarefa técnica, mas também envolve decisões normativas e contextuais.

Além disso, o uso de métricas de justiça deve considerar a natureza dos dados e a distribuição das classes. Métricas calculadas em conjuntos de dados altamente desbalanceados podem ser enganosas, tendo em vista que as disparidades podem surgir de diferentes fontes, incluindo o próprio viés nos rótulos do conjunto de treinamento, conhecido como label bias (CHEN; JOHANSSON; SONTAG, 2018). Essa constatação reforça a importância de técnicas de balanceamento, como sobreamostragem, subamostragem ou geração de dados sintéticos, como medidas complementares na mitigação de viés.

Dessa forma, o uso consciente e crítico de métricas de justiça é essencial para a construção de sistemas de aprendizado de máquina mais éticos, transparentes e responsáveis. Elas oferecem não apenas um diagnóstico técnico da equidade do modelo, mas também subsidiam a tomada de decisões sobre ajustes necessários no pipeline de dados e no design do sistema preditivo.

2.2 BALANCEAMENTO DE DADOS

O balanceamento de dados é um dos desafios fundamentais no desenvolvimento de modelos de aprendizado de máquina, especialmente em cenários onde a distribuição das classes dentro do conjunto de dados é desproporcional. Esse fenômeno, conhecido como desbalanceamento de classes, ocorre quando uma ou mais classes estão sub-representadas em relação a outras, o que pode levar a modelos enviesados e a um desempenho reduzido na predição das classes minoritárias. Esse problema é amplamente observado em diversas aplicações reais, como na detecção de fraudes financeiras, diagnóstico médico, segurança cibernética e reconhecimento de padrões em imagens.

A principal consequência do desbalanceamento de classes é que os modelos de aprendizado de máquina, especialmente aqueles treinados com algoritmos supervisionados, tendem a otimizar métricas globais que podem mascarar o mau desempenho para a classe minoritária

(HE; GARCIA, 2009). Em um cenário altamente desbalanceado, um modelo pode atingir uma acurácia global elevada, simplesmente prevendo a classe majoritária na maioria dos casos, enquanto a classe de menor ocorrência é frequentemente ignorada. Esse problema se agrava quando métricas como acurácia são utilizadas como principal critério de avaliação, já que não refletem adequadamente o desempenho do modelo para todas as classes. Para contornar essa limitação, métricas como F1-score, Area Under the Curve - ROC (AUC-ROC) e Area Under the Precision - Recall Curve (AUC-PR) são mais indicadas para avaliar o impacto do desbalanceamento (SAITO; REHMSMEIER, 2015).

Para mitigar os impactos do desbalanceamento de classes, diversas estratégias podem ser empregadas, podendo ser divididas em três principais categorias: Técnicas baseadas em reamostragem, ajustes na modelagem e técnicas baseadas na geração de dados sintéticos. (HE; GARCIA, 2009)

As técnicas baseadas em reamostragem visam modificar a distribuição do conjunto de dados para reduzir o desbalanceamento, seja ampliando a representação da classe minoritária (oversampling) ou reduzindo a quantidade de amostras da classe majoritária (undersampling).

Outra abordagem para lidar com o desbalanceamento de classes é modificar a forma como os modelos de aprendizado de máquina processam as diferentes classes, sem alterar diretamente a distribuição dos dados. Como a técnica de ponderação de classes (Class Weighting), esta técnica atribui pesos diferentes às classes durante o treinamento, forçando o modelo a prestar mais atenção aos erros cometidos nas instâncias da classe minoritária. Muitos algoritmos, como árvores de decisão e redes neurais, permitem a configuração de hiperparâmetros que ajustam a importância de cada classe na função de perda (ELKAN, 2001). Ou a técnica de alteração na função de perda: Em modelos de redes neurais profundas, é possível utilizar funções de perda especializadas, como a Focal Loss. Esta função foi projetada para reduzir o impacto de exemplos fáceis e bem classificados durante o treinamento, fazendo com que o modelo foque na aprendizagem dos exemplos mais difíceis, que frequentemente pertencem à classe minoritária (LIN et al., 2017)

Além das técnicas tradicionais de oversampling, abordagens mais avançadas utilizam modelos generativos para criar dados sintéticos de alta qualidade, aumentando a diversidade dos exemplos da classe minoritária. Como por exemplo a ténica de redes adversárias generativas: Modelos baseados em Generative Adversarial Networks (GANs) podem gerar exemplos realistas da classe minoritária ao aprender a distribuição dos dados originais. Isso é particularmente útil em domínios como visão computacional e geração de texto (SAMPATH et al., 2021). Outra

técnica generativa baseada em modelos probabilísticos que pode ser usada para criar amostras sintéticas representativas da classe minoritária é a Variational Autoencoders (VAEs)

O desbalanceamento de classes não apenas compromete o desempenho dos modelos, mas também pode introduzir viés algorítmico, especialmente em aplicações sensíveis, como seleção de candidatos, concessão de crédito e diagnósticos médicos. Modelos treinados com dados desbalanceados podem perpetuar desigualdades, prejudicando sistematicamente grupos sub-representados. Por isso a equidade na predição deve ser avaliada com métricas de justiça, que analisam a diferença na performance do modelo entre diferentes grupos populacionais.

A escolha da estratégia de mitigação, portanto, deve ser feita com cautela. Nem sempre aumentar a quantidade de exemplos da classe minoritária garante um modelo mais justo ou com melhor capacidade de generalização. O oversampling pode introduzir redundâncias e sobreajuste, enquanto o undersampling pode levar à perda de informações valiosas. É essencial que a aplicação de técnicas como reamostragem e ponderação de classes seja combinada com uma avaliação rigorosa, utilizando métricas apropriadas tanto para o desbalanceamento quanto para a justiça, a fim de garantir que o modelo final seja equilibrado, robusto e equitativo.

2.2.1 Abordagens principais

O balanceamento de dados é um desafio significativo em AM e análise estatística, e há diversas maneiras de lidar com essa questão. Uma das conclusões centrais desta dissertação é que a abordagem ideal depende de fatores como o tamanho do conjunto de dados, a severidade do desbalanceamento e a natureza do problema a ser resolvido. Algumas técnicas visam reduzir o impacto da classe majoritária, enquanto outras buscam aumentar a representatividade da classe minoritária. Além disso, há métodos que incorporam ajustes diretamente nos algoritmos de aprendizado, bem como estratégias mais avançadas que utilizam inteligência artificial para gerar novos dados sintéticos (JOHNSON; KHOSHGOFTAAR, 2019). O balanceamento de dados pode ser abordado de diferentes maneiras, dependendo das características do conjunto de dados e do problema em questão.

Por outro lado, o desbalanceamento de dados ocorre quando as classes de um conjunto de dados não estão distribuídas de forma equitativa, resultando em uma disparidade na quantidade de exemplos entre as diferentes categorias. Esse fenômeno pode comprometer a eficácia dos modelos de aprendizado de máquina, pois muitos algoritmos convencionais tendem a favorecer a classe majoritária, ignorando ou classificando incorretamente a classe minoritária.

Para mitigar o problema dos dados desbalanceados, existem diversas estratégias de tratamento, que podem ser divididas em três categorias principais. A primeira, a nível de dados, engloba as técnicas de balanceamento propriamente ditas (como reamostragem), que visam modificar a distribuição das classes. A segunda, a nível de algoritmo, adapta o modelo para que seja mais sensível à classe minoritária, sem alterar os dados. Por fim, as estratégias híbridas combinam ambas as abordagens. A escolha da abordagem ideal depende de fatores como o tamanho do conjunto de dados e a severidade do desbalanceamento (JAFARIGOLA; TRAFALIS, 2023).

Para as abordagens no nível dos dados, os métodos aplicados diretamente sobre os dados visam modificar a distribuição das classes por meio de técnicas de reamostragem. Esses métodos podem ser divididos em duas categorias principais. Undersampling: Consiste na redução do número de amostras da classe majoritária para equilibrar a proporção entre as classes. O método mais simples dessa abordagem é o RandomUnderSampler, que remove aleatoriamente exemplos da classe dominante. Apesar de ser eficiente na correção do desbalanceamento, pode levar à perda de informações relevantes. E oversampling: Responsável por aumentar artificialmente a quantidade de exemplos da classe minoritária, inserindo novas amostras para garantir uma distribuição mais equilibrada. Uma técnica básica é o RandomOverSampler, que replica aleatoriamente amostras da classe minoritária. Contudo, a repetição de instâncias pode levar a problemas de sobreajuste do modelo.

Técnicas mais avançadas foram desenvolvidas para gerar amostras sintéticas em vez de simplesmente replicar dados existentes. Um dos métodos mais utilizados é o SMOTE, que cria novas amostras ao interpolar pontos entre as instâncias existentes da classe minoritária e seus vizinhos mais próximos. Essa abordagem evita a simples duplicação de exemplos e melhora a generalização do modelo.

Outras variações do SMOTE foram propostas para lidar com desafios específicos, como BorderlineSMOTE focado em amostras próximas à fronteira de decisão, onde a distinção entre classes é mais difícil. ADASYN responsável por gerar amostras sintéticas com base na complexidade da região de decisão, atribuindo mais exemplos sintéticos a áreas com maior dificuldade de separação. SVMSMOTE que utiliza Support Vector Machines (SVM) para identificar a região de decisão e gerar amostras sintéticas próximas a essa fronteira. E KMeansSMOTE que aplica clustering para identificar padrões dentro da classe minoritária e criar novas amostras de maneira mais controlada.

Outra estratégia para lidar com conjuntos de dados desbalanceados são as abordagens

no nível do algoritmo, que tem como objetivo modificar o funcionamento dos algoritmos de aprendizado de máquina, atribuindo um peso maior às classes menos representadas. Essa técnica garante que os modelos não sejam enviesados para a classe majoritária. Métodos comuns incluem modificação da função de custo, onde algoritmos podem ser ajustados para penalizar erros na classe minoritária com maior intensidade, reduzindo o impacto do desbalanceamento. Ou ajuste de probabilidades, alguns classificadores probabilísticos podem ser configurados para compensar o desbalanceamento, ajustando os limiares de decisão.

As abordagens híbridas combinam técnicas de reamostragem e ajustes nos algoritmos para obter um melhor equilíbrio entre representatividade e desempenho. Por exemplo, um modelo pode aplicar o SMOTE para gerar novas amostras e, em seguida, utilizar um classificador que penalize erros na classe minoritária.

O impacto do desbalanceamento de dados é um dos desafios mais críticos na área de aprendizado de máquina e análise de dados. A escolha da técnica mais adequada depende da natureza do problema, do volume de dados disponível e da complexidade da tarefa de classificação. Estratégias eficazes devem considerar tanto a representatividade dos dados quanto a capacidade dos modelos de generalizar corretamente para novas amostras.

2.2.1.1 Abordagens de niveis de dados

As abordagens de níveis de dados para mitigação de viés e balanceamento de conjuntos de dados são estratégias fundamentais para melhorar a equidade e a qualidade de modelos de aprendizado de máquina. Essas técnicas atuam diretamente sobre os dados antes do treinamento do modelo, com o objetivo de corrigir distribuições desbalanceadas, reduzir vieses e melhorar a representatividade de todas as classes ou grupos sensíveis.

Neste nível, a abordagem mais comum é a reamostragem de dados, que modifica a distribuição das classes. O undersampling reduz a quantidade de instâncias da classe majoritária, removendo exemplos redundantes ou menos informativos, de forma a igualar a proporção entre as classes. Já o oversampling aumenta artificialmente o número de instâncias da classe minoritária, seja por duplicação aleatória de amostras ou por técnicas mais avançadas, como o BorderlineSMOTE, que cria novos exemplos sintéticos a partir da interpolação de exemplos reais, enriquecendo a representação da classe minoritária (HAN; WANG; MAO, 2005). Ambas as estratégias visam evitar que o modelo aprenda a associar padrões apenas à classe dominante, promovendo uma melhor generalização.

Além das técnicas tradicionais de balanceamento de classes, outra abordagem relevante no nível de dados é a embaralhamento de atributos sensíveis (shuffling). Essa técnica consiste em randomizar o valor de um atributo sensível, como gênero ou etnia, dentro do conjunto de dados, impedindo que o modelo aprenda padrões discriminatórios associados a esses atributos. Dessa forma, busca-se reduzir o viés algorítmico e garantir que as decisões sejam baseadas em informações relevantes para a tarefa predita, e não em características sensíveis. Outras técnicas mais complexas envolvem a modificação dos dados para que satisfaçam critérios de justiça antes mesmo do treinamento (KAMIRAN; CALDERS, 2012). É importante notar que, embora a reponderação de instâncias possa ser vista como uma técnica de pré-processamento, ela é mais precisamente categorizada como uma abordagem a nível de algoritmo, pois altera a função de perda durante o treinamento do modelo (ZADROZNY; LANGFORD; ABE, 2003).

Outra estratégia importante é a subamostragem da classe majoritária, que reduz a quantidade de exemplos da classe predominante para tornar a distribuição mais uniforme. Isso pode ser feito de maneira aleatória ou utilizando técnicas mais elaboradas, como a seleção de exemplos mais representativos da classe majoritária para manter a diversidade e evitar perda excessiva de informações úteis.

Além disso, a duplicação da classe oposta é uma técnica frequentemente utilizada para balanceamento. Nesse método, os exemplos da classe minoritária são replicados dentro do conjunto de treinamento para garantir uma representação mais equitativa das classes. No entanto, essa abordagem pode aumentar o risco de overfitting, pois o modelo pode memorizar as instâncias repetidas em vez de aprender padrões generalizáveis.

Outra técnica amplamente utilizada no nível de dados é a geração de dados sintéticos, que busca criar novos exemplos realistas para enriquecer a classe minoritária. Esse tipo de técnica é especialmente útil em contextos onde a obtenção de novos dados reais é difícil, como na área da saúde ou em visão computacional (SHORTEN; KHOSHGOFTAAR, 2019). Além do SMOTE, abordagens baseadas em aprendizado profundo, como modelos GANs, podem ser empregadas para criar amostras sintéticas mais complexas e variadas. Esse tipo de técnica é útil principalmente em contextos onde a obtenção de novos dados reais é difícil, como na área da saúde ou na detecção de fraudes.

Além das estratégias voltadas para balanceamento, existem métodos de pré-processamento de dados que ajudam a mitigar vieses antes do treinamento do modelo. Esses métodos incluem a remoção de atributos sensíveis do conjunto de dados ou a normalização de distribuições para garantir que todas as classes ou grupos sensíveis tenham características estatísticas semelhan-

tes. Outra técnica relevante é a reponderação de instâncias, na qual diferentes pesos são atribuídos às amostras de acordo com sua representatividade ou importância para a tarefa predita. Isso ajuda o modelo a focar mais nos exemplos da classe minoritária, sem precisar alterar a quantidade de instâncias no conjunto de dados.

A escolha da abordagem mais adequada depende do contexto da aplicação e das características do conjunto de dados. Enquanto o oversampling e a geração de dados sintéticos são úteis quando há uma quantidade muito pequena de exemplos na classe minoritária, o undersampling pode ser preferível quando há grande redundância de exemplos na classe majoritária. Já a remoção de atributos sensíveis e o embaralhamento de classes são mais apropriados em situações onde o objetivo principal é mitigar viés discriminatório.

Em resumo, as abordagens de níveis de dados desempenham um papel fundamental na construção de modelos de aprendizado de máquina mais justos, equilibrados e robustos. Elas ajudam a corrigir problemas estruturais nos conjuntos de dados, garantindo que todas as classes sejam representadas de forma justa e que o modelo não aprenda padrões enviesados. Quando bem aplicadas, essas estratégias contribuem para o desenvolvimento de soluções mais confiáveis e responsáveis em diversas áreas, como saúde, finanças, segurança e inteligência artificial ética.

2.2.1.2 Abordagens híbridas

As abordagens híbridas representam uma estratégia sofisticada para o tratamento de dados, combinando diferentes técnicas para maximizar a equidade e a representatividade das classes, especialmente em cenários de desbalanceamento extremo. Onde uma única técnica pode ser insuficiente, a combinação de métodos oferece uma solução mais robusta e eficaz.

Uma das estratégias híbridas mais consolidadas é a combinação de oversampling e undersampling. Nesse caso, o oversampling é utilizado para aumentar a representatividade da classe minoritária, enquanto o undersampling reduz a cardinalidade da classe majoritária. O objetivo é evitar tanto a perda de informação que pode ocorrer com o undersampling isolado, quanto o risco de overfitting associado ao oversampling simples. Uma variação eficaz é a aplicação do SMOTE para gerar exemplos sintéticos, seguida da remoção de ruídos ou instâncias ambíguas com técnicas como proposto em (BATISTA; PRATI; MONARD, 2004).

Outra opção de hibridização consiste em unir intervenções nos dados com a mitigação de viés. Por exemplo, pode-se aplicar o SMOTE para corrigir o desbalanceamento de classes e,

simultaneamente, embaralhar um atributo sensível para impedir que o modelo aprenda correlações espúrias. Essa abordagem busca tratar os problemas de representação e de discriminação em conjunto.

Adicionalmente, as abordagens híbridas podem combinar manipulação de dados com ajustes a nível de algoritmo. Um exemplo clássico é o SMOTEBoost, que integra a geração de dados sintéticos do SMOTE dentro do processo de treinamento de um algoritmo de ensemble como o AdaBoost. Outra possibilidade é aplicar oversampling nos dados e, em seguida, treinar um modelo com ponderação de classes (class weighting), forçando o algoritmo a focar duplamente na classe minoritária (CHAWLA et al., 2003).

Em cenários com dados de alta dimensionalidade, como imagens ou texto, as abordagens híbridas podem incluir o uso de modelos generativos avançados. Técnicas baseadas em GANs podem ser empregadas para criar amostras sintéticas realistas para a classe minoritária, combinadas com métodos de subamostragem inteligente para limpar a classe majoritária.

Além disso, uma estratégia híbrida que otimiza o esforço de rotulagem é a combinação de aprendizado ativo com técnicas de balanceamento. Nesses casos, o modelo seleciona ativamente as instâncias mais informativas de um conjunto de dados não rotulados para serem rotuladas por um especialista, focando em exemplos que ajudem a balancear as classes de maneira mais eficiente e inteligente (AGGARWAL; POPESCU; HUDELOT, 2020).

As abordagens híbridas são particularmente úteis em aplicações sensíveis onde o viés pode ter consequências graves, como em sistemas de recrutamento, diagnósticos médicos e análise de crédito. A combinação estratégica de técnicas permite alcançar um compromisso entre representatividade, mitigação de viés e preservação da qualidade dos dados, resultando em modelos mais justos, generalizáveis e robustos.

2.3 BALANCEAMENTO DE DADOS PARA JUSTIÇA

O balanceamento de dados é uma etapa crucial na construção de sistemas de aprendizado de máquina justos, especialmente quando se busca mitigar o viés algorítmico presente em conjuntos de dados com atributos sensíveis, como gênero, raça ou idade. Embora o balanceamento de classes seja tradicionalmente utilizado para lidar com a assimetria entre categorias da variável de saída, o balanceamento orientado à justiça vai além, ele visa redistribuir ou reestruturar os dados de forma a reduzir disparidades injustas no desempenho do modelo entre diferentes grupos sociais (KAMIRAN; CALDERS, 2012).

Em muitos cenários reais, o viés algorítmico não surge apenas da desigualdade entre rótulos, mas da sub-representação de certos grupos nos dados. Por exemplo, se mulheres são minoria em um dataset de concessão de crédito, o modelo poderá aprender padrões enviesados que resultam em uma menor taxa de aprovação para esse grupo. Nesses casos, técnicas de balanceamento aplicadas de maneira consciente ao atributo sensível e não apenas à variável alvo tornam-se ferramentas essenciais para promover justiça algorítmica.

Ao tratar justiça como um objetivo explícito, o balanceamento de dados pode ser aplicado diretamente sobre os subgrupos definidos pela combinação entre o atributo sensível e a variável de saída. Por exemplo, em vez de apenas balancear entre classes "aprovado" e "reprovado", pode-se equilibrar as quantidades de amostras entre "mulher-aprovada", "mulher-reprovada", "homem-aprovado" e "homem-reprovado". Essa abordagem garante que o modelo veja, durante o treinamento, distribuições mais equitativas entre os subgrupos de interesse, evitando a natural tendência dos algoritmos de favorecer padrões mais frequentes (LAMMERS; VAQUET; HAMMER, 2025).

Entre as técnicas mais utilizadas para balanceamento com foco em justiça estão o oversampling e o undersampling aplicados especificamente aos grupos sub-representados. Técnicas como o SMOTE, originalmente desenvolvidas para lidar com desbalanceamento de classes, vêm sendo adaptadas para gerar exemplos sintéticos de grupos minorizados, como no caso do Fair-SMOTE, com o objetivo de suavizar disparidades nas taxas de erro entre subgrupos (KHAKUREL; ABDELMOUMIN; RAWAT, 2025).

A utilização de estratégias de balanceamento para fairness, no entanto, requer cuidados metodológicos. O simples aumento da representação de um grupo pode, por exemplo, introduzir padrões irreais se não houver diversidade suficiente nos dados originais. Além disso, quando se cria ou duplica exemplos para minorias sociais historicamente marginalizadas, há o risco de reforçar estereótipos ou de o modelo focar em correlações espúrias, um lembrete de que apenas corrigir os dados não é uma panaceia para o problema do viés (HOOKER, 2021). Por isso, as técnicas de balanceamento devem ser acompanhadas de uma análise crítica sobre os dados disponíveis e os possíveis efeitos colaterais da intervenção.

De forma geral, o balanceamento de dados para justiça vai além da correção estatística de classes, trata-se de uma estratégia crítica e intencional voltada à construção de modelos mais justos. Ao reestruturar os dados para garantir que grupos sub-representados estejam adequadamente refletidos no processo de aprendizagem, busca-se não apenas melhorar métricas quantitativas, mas também contribuir para um uso mais ético e responsável do aprendizado

de máquina. A efetividade dessas técnicas depende tanto da qualidade da intervenção quanto da sensibilidade analítica dos profissionais envolvidos, pois justiça não é apenas uma questão técnica, mas também social e contextual.

2.3.1 Novas técnicas e ferramentas para justiça algorítmica

Entre as técnicas mais recentes voltadas ao balanceamento de dados com objetivo de promover justiça, destacam-se as abordagens de aumento de dados (data augmentation) que foram adaptadas para o contexto da justiça algorítmica. Uma abordagem recente de aumento de dados focada em equidade é a mixagem de subgrupos (subgroup mixup). Proposta por (NAVARRO et al., 2023), a técnica consiste em gerar amostras sintéticas a partir da interpolação de exemplos de diferentes grupos sensíveis, promovendo um aprendizado mais robusto e justo. Esta técnica se inspira no princípio do mixup tradicional, uma estratégia consagrada na literatura para a regularização e o aumento da robustez de modelos, que cria novos dados a partir da interpolação linear entre pares de exemplos.

2.3.1.1 Mixup para equidade

A mixagem de subgrupos funciona criando novas amostras sintéticas a partir de combinações convexas de pares de exemplos rotulados pertencentes a diferentes subgrupos sensíveis. A intuição é que, ao interpolar exemplos de grupos com diferentes frequências ou tratamentos históricos no modelo, o algoritmo se torna mais exposto a regiões "intermediárias" do espaço de atributos que geralmente são ignoradas quando os dados são desbalanceados. Essas novas amostras são incorporadas ao conjunto de treinamento, sem a necessidade de impor restrições adicionais ao modelo ou utilizar termos explícitos de regularização o que torna a técnica simples de implementar e altamente escalável.

A abordagem de mixagem de subgrupos diferencia-se fundamentalmente do uso tradicional da técnica de Mixup na literatura. Originalmente, o Mixup foi introduzido como um mecanismo de regularização para melhorar a generalização e a robustez de redes neurais, operando através da interpolação de pares aleatórios de amostras de todo o conjunto de dados (ZHANG et al., 2017).

Em contraste, o trabalho de (NAVARRO et al., 2023) redireciona essa estratégia, propondo o uso do Mixup como uma técnica de aumento de dados direcionada explicitamente para a equi-

dade. Os autores demonstram que, ao gerar amostras sintéticas na fronteira entre subgrupos demográficos, é possível mitigar o viés algorítmico de forma eficaz. Seus resultados indicam uma melhoria em métricas de justiça como a TPR e a STP, frequentemente sem um sacrifício significativo no desempenho global do modelo.

Essa técnica representa uma importante evolução conceitual na mitigação de viés. Diferentemente de abordagens que apenas regularizam o modelo ou alteram o treinamento com pesos (a nível de algoritmo), o subgroup mixup atua diretamente na fonte do problema: a distribuição dos dados. Ao promover uma exposição mais balanceada entre os subgrupos durante a fase de aprendizado, a técnica busca preencher lacunas de representação que são uma causa raiz do viés algorítmico. Essa estratégia está em total alinhamento com o movimento de IA Centrada em Dados (Data-Centric AI), que defende que o aprimoramento da qualidade e da estrutura dos dados é um dos caminhos mais eficazes para o desenvolvimento de sistemas mais justos e inclusivos.

2.3.1.2 Reponderação de instância

Uma abordagem clássica e amplamente explorada para promover justiça no AM é a reponderação de instâncias com base em atributos sensíveis. No trabalho de (HEIDARPOUR; LOUGHRAN; MCDAID, 2023), os autores propõem um método de reponderação no qual pesos diferenciados são atribuídos a cada exemplo de treinamento, considerando a interação entre o atributo sensível (como gênero) e o rótulo de classe. Para o caso do atributo "sexo", os exemplos são categorizados em quatro grupos: Positivo Privilegiado (PP), Positivo Não Privilegiado (UP), Negativo Privilegiado (PN) e Negativo Não Privilegiado (UN). Os exemplos nos grupos UP e PN recebem pesos maiores durante o treinamento, a fim de corrigir a sub-representação desses perfis e forçar o modelo a prestar mais atenção a esses casos potencialmente marginalizados. Já os grupos PP e UN, considerados super-representados ou com menor risco de injustiça, recebem pesos menores.

Como nem todos os algoritmos de classificação suportam vetores de peso diretamente, os autores também aplicam técnicas de amostragem para refletir esses pesos no conjunto de treinamento. O método consiste em subamostrar os grupos com pesos menores PP, UN e superamostrar aqueles com pesos maiores PN e UP, modificando assim a distribuição dos dados de entrada. O objetivo é simular o efeito da reponderação, mesmo quando o classificador não oferece suporte nativo à ponderação.

2.3.1.3 Abordagens de balanceamento baseadas em redes generativas adversariais

Além das técnicas clássicas de balanceamento como SMOTE e undersampling, a literatura mais recente tem explorado o uso de modelos de deep learning para a geração de dados sintéticos. As GANs surgem como uma alternativa sofisticada. Uma GAN consiste em duas redes neurais, um gerador e um discriminador, que competem entre si. O gerador aprende a criar novos exemplos (neste caso, da classe minoritária) que sejam indistinguíveis dos dados reais, enquanto o discriminador aprende a diferenciar os exemplos reais dos sintéticos.

Essa abordagem tem o potencial de capturar distribuições de dados muito mais complexas e gerar amostras sintéticas de alta qualidade e diversidade, superando as limitações da interpolação linear do SMOTE (GOODFELLOW et al., 2014). Métodos como o MedGAN, por exemplo, foram propostos para gerar dados médicos sintéticos realistas. Embora promissoras, essas técnicas não foram incluídas no escopo experimental desta dissertação devido à sua alta complexidade computacional, à necessidade de grandes volumes de dados para um treinamento eficaz e à dificuldade de ajuste de hiperparâmetros, o que as tornaria inviáveis para uma análise de larga escala envolvendo 8 bases de dados e 11 algoritmos distintos, onde a reprodutibilidade e a comparação sistemática eram prioridades.

2.3.1.4 Bibliotecas especializadas em justiça

A complexidade de implementar e avaliar o impacto dessas diversas técnicas de mitigação de viés levou ao desenvolvimento de frameworks especializados. A existência dessas ferramentas reforça a importância de avaliar a equidade com métricas de justiça específicas, para além das métricas de performance. Duas das bibliotecas mais proeminentes na área são:

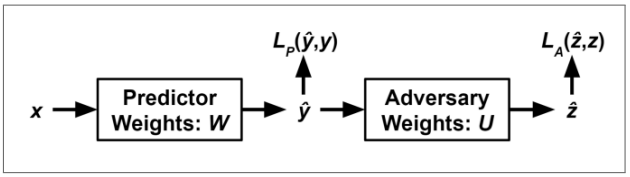
- A AIF360, desenvolvida pela IBM Research, é uma das bibliotecas mais completas para detecção e mitigação de viés em modelos de aprendizado de máquina. Ela fornece mais de 70 métricas para avaliar a justiça de modelos e implementa diversas técnicas de mitigação, organizadas em três categorias: pré-processamento, in-processamento e pós-processamento. Dentre suas funcionalidades, destaca-se a compatibilidade com scikit-learn e sua capacidade de integrar-se com pipelines existentes de aprendizado de máquina. AIF360 permite comparar rapidamente os efeitos de diferentes métodos de mitigação de viés, facilitando a avaliação quantitativa da equidade algorítmica.

- O fairlearn é uma biblioteca desenvolvida em Python que oferece métricas e algoritmos para avaliação e mitigação de viés. Ela se destaca por oferecer uma interface simples para implementar restrições de equidade diretamente no processo de treinamento dos modelos, utilizando técnicas como *Exponentiated Gradient Reduction* e *Grid Search Reduction*. Além disso, Fairlearn apresenta ferramentas gráficas que auxiliam na análise de trade-offs entre precisão e justiça, oferecendo suporte para métricas como TPR e STP. Sua integração com o ecossistema scikit-learn facilita sua adoção em projetos reais e acadêmicos.

2.3.1.5 Remoção de viés adversarial

Inspiradas no sucesso das Redes Generativas Adversariais (GANs), as técnicas de remoção de viés adversarial representam uma abordagem in-processamento sofisticada para a construção de modelos mais justos. Proposta em trabalhos seminais como o de (ZHANG; LEMOINE; MITCHELL, 2018), essa estratégia enquadra a mitigação de viés como um jogo entre dois componentes de uma rede neural: um preditor e um adversário. O Preditor tem a tarefa principal de prever o resultado desejado (ex: aprovação de crédito) com a maior acurácia possível, utilizando os dados de entrada. O Adversário, por sua vez, não tem acesso ao resultado real, mas observa as predições (ou as representações internas) geradas pelo Preditor e tem um único objetivo: tentar adivinhar o atributo sensível do indivíduo (ex: gênero, raça) a partir dessa informação. Observe a figura abaixo:

Figura 1 – A arquitetura da rede adversária



Fonte: ZHANG; LEMOINE; MITCHELL (2018)

Essas técnicas ilustram como o pré-processamento dos dados pode ser uma poderosa ferramenta para mitigar viés algorítmico antes mesmo do treinamento do modelo, evitando intervenções posteriores mais complexas ou restrições no algoritmo de aprendizado. Além disso, os resultados demonstram que tais intervenções podem melhorar significativamente a equidade entre os subgrupos, com impacto direto nas métricas de justiça como TPR e STP, sem necessariamente comprometer o desempenho global dos modelos.

A necessidade de aplicar técnicas de balanceamento para promover a equidade não é apenas uma questão teórica, mas uma área de intensa investigação na literatura de IA responsável, motivando estudos em domínios de alto impacto. Um exemplo notável é o trabalho de (AL-ALAWI; ALBUAINAIN, 2024). Nesta pesquisa, os autores abordam o problema da classificação de promoções em recursos humanos, um cenário inerentemente sensível a vieses de gênero e outros atributos, além de avaliarem comparativamente o impacto de diferentes técnicas de balanceamento de dados. No estudo em questão, um conjunto diversificado de sete técnicas de preparação de dados e modelagem foi analisado para avaliar seu impacto em um problema de classificação com classes desbalanceadas. Essas abordagens podem ser organizadas em três categorias distintas: pré-processamento padrão, balanceamento por sobreamostragem e uma abordagem a nível de algoritmo.

Inicialmente, foram aplicadas técnicas de pré-processamento padrão, como a Seleção de Atributos (Feature Selection), a normalização de dados e a Análise de Componentes Principais (PCA). O objetivo dessas etapas é otimizar o conjunto de dados, reduzindo a dimensionalidade e garantindo que as escalas das variáveis não influenciem indevidamente o treinamento do modelo. Em seguida, foram exploradas três técnicas de balanceamento por sobreamostragem, que atuam diretamente na distribuição das classes: Random Oversampling (ROS), SMOTE e ADASYN. Conforme discutido anteriormente, essas estratégias visam aumentar a representatividade da classe minoritária, seja por replicação ou pela geração de exemplos sintéticos, demonstrando grande potencial para melhorar o desempenho dos modelos em classes sub-representadas.

A abordagem mais proeminente analisada foi o Weighted Quadratic Random Forest (WQRF). O WQRF representa uma evolução do Random Forest tradicional, introduzindo um sofisticado mecanismo de ponderação. Nele, cada árvore do ensemble tem seu voto ponderado pela F-measure que ela alcança, e os atributos utilizados nos nós da árvore também recebem pesos quadráticos. Essa dupla ponderação torna o algoritmo inerentemente mais robusto a dados desbalanceados. No estudo citado, essa técnica se mostrou particularmente eficaz, com o classificador alcançando uma acurácia notável de 92,80 porcentos, indicando seu grande potencial em tarefas onde desempenho e equilíbrio são cruciais. A técnica de ROS, também utilizada no estudo, consiste na simples replicação de amostras da classe minoritária até igualar sua proporção à da classe majoritária. Apesar de seu risco conhecido de overfitting, a técnica mostrou-se eficaz para melhorar o desempenho em situações onde a classe minoritária sofre de baixa representatividade uma condição comum em problemas de viés.

No esforço para mitigar o viés algorítmico em contextos reais, o estudo de (AL-ALAWI; ALBUAINAIN, 2024) oferece uma visão abrangente de como diferentes técnicas de balanceamento de dados podem ser aplicadas de forma estratégica para melhorar a performance e a justiça preditiva em cenários sensíveis, como o de recursos humanos. Enquanto o estudo de (AL-ALAWI; ALBUAINAIN, 2024) oferece uma valiosa contribuição ao demonstrar a aplicação estratégica de técnicas de balanceamento no domínio específico de Recursos Humanos, a presente dissertação expande essa análise de forma significativa. O diferencial deste trabalho reside em sua abordagem multi-domínio e na escala da investigação empírica, que avalia um conjunto muito mais amplo de algoritmos e técnicas em oito cenários distintos. O foco aqui não é a aplicação em um único setor, mas sim a criação de um mapeamento generalizável dos trade-offs fundamentais entre performance e justiça.

Em sua análise, a técnica SMOTE foi explorada como uma forma sofisticada de gerar exemplos sintéticos da classe minoritária por meio de interpolação com seus vizinhos mais próximos (CHAWLA et al., 2002). Essa abordagem visa aumentar a diversidade dos dados e reduzir o risco de sobreajuste em comparação com a duplicação aleatória, sendo responsável por ganhos significativos de desempenho nos experimentos.

A variante ADASYN também foi investigada, atuando como uma extensão do SMOTE com foco em gerar mais amostras sintéticas nas regiões onde a classe minoritária é mais difícil de aprender. Conforme detalhado por (HE et al., 2008), essa estratégia adaptativa promove um melhor equilíbrio entre a sensibilidade e a especificidade do modelo, sendo particularmente útil em situações onde o viés ocorre próximo às fronteiras de decisão.

Além dessas técnicas de reamostragem, o estudo também investigou abordagens que indiretamente contribuem para o balanceamento e a justiça do modelo, como feature selection, normalização de dados e PCA. A seleção de atributos ajuda a remover redundâncias e a complexidade, evitando que o modelo aprenda padrões espúrios. A normalização e o PCA, por sua vez, atuam na padronização e na redução de dimensionalidade, o que indiretamente contribui para uma modelagem mais estável e menos tendenciosa.

Em seu conjunto, os resultados ressaltam que a escolha da técnica de balanceamento deve considerar não apenas a acurácia, mas também os impactos sobre a equidade entre grupos, especialmente quando atributos sensíveis estão envolvidos, alinhando-se com as melhores práticas para o desenvolvimento de modelos de aprendizado de máquina mais responsáveis.

De mogo geral, a evolução das técnicas de balanceamento para a promoção da justiça algorítmica demonstra uma clara trajetória em direção a abordagens mais integradas e contextuais.

As pesquisas mais recentes estão se movendo para além de simples intervenções nos dados, explorando métodos que combinam geração de dados sintéticos com modelos generativos, como GANs, e estratégias de aprendizado ativo para otimizar o processo. Essa sofisticação crescente reflete o reconhecimento de que a justiça em aprendizado de máquina não é um problema com uma solução única, mas um desafio dinâmico que exige uma combinação de avanços técnicos, avaliação rigorosa e uma profunda compreensão do domínio de aplicação.

2.3.2 Técnicas de balanceamento de dados aplicadas

Para a execução dos experimentos desta dissertação, foi selecionado um conjunto de 10 abordagens técnicas distintas descritas na tabela 4. A escolha deste conjunto foi estratégica, visando criar um panorama abrangente e representativo das principais estratégias discutidas na literatura para o tratamento de desbalanceamento de dados e mitigação de viés. Foram incluídas desde técnicas clássicas de reamostragem, como ROS e Random Undersampling (RUS), até variantes sintéticas mais sofisticadas, como ADASYN e BorderlineSMOTE. Além disso, o conjunto incorpora métodos que atuam a nível de algoritmo, como a Reponderação (Reweighting), e abordagens desenvolvidas explicitamente para o contexto da justiça, como o FairSMOTE. Para garantir uma análise robusta e isolar o impacto de certas variáveis. Técnicas de controle, como o embaralhamento de atributos, também foram implementadas. A seguir, cada uma dessas 10 abordagens utilizadas nos experimentos será descrita em detalhe.

2.3.2.1 Embaralhamento do Atributo Sensível (Shuffling)

Esta abordagem foi utilizada como um método de controle para a análise de justiça. A técnica consiste em embaralhar aleatoriamente os valores da coluna do atributo sensível (neste caso, "sexo") no conjunto de treinamento. O objetivo não é melhorar o modelo, mas sim quebrar a correlação direta entre o atributo sensível e a variável alvo, permitindo avaliar o quanto o viés do modelo original era dependente dessa informação explícita versus outras correlações latentes nos dados.

2.3.2.2 Duplicação com Classe Oposta

Esta é uma técnica de aumento de dados focada em equidade. Para cada amostra no conjunto de treinamento, uma cópia contrafactual é criada, na qual o valor do atributo sensível é invertido (ex: 'homem' torna-se 'mulher' e vice-versa), enquanto todos os outros atributos e o rótulo da classe são mantidos. A intenção é ensinar ao modelo que o atributo sensível não deve ser um fator decisivo na predição, forçando-o a aprender padrões mais robustos e menos discriminatórios.

2.3.2.3 Subamostragem Aleatória (Random Under-Sampling)

Esta é uma das técnicas mais diretas para lidar com o desbalanceamento. Ela consiste em remover aleatoriamente instâncias da classe majoritária até que a distribuição de classes no conjunto de treinamento atinja um equilíbrio desejado. Embora eficaz para reduzir o viés do modelo em direção à classe majoritária, sua principal desvantagem é o risco de descartar informações potencialmente úteis.

2.3.2.4 SMOTE

A quarta abordagem, SMOTE, é uma das técnicas de sobreamostragem mais influentes e amplamente utilizadas para o tratamento de dados desbalanceados. Proposta para superar a principal limitação da sobreamostragem aleatória, o risco de sobreajuste (overfitting), o SMOTE não apenas duplica instâncias existentes, mas cria novos exemplos sintéticos para a classe minoritária. O algoritmo opera selecionando uma instância da classe minoritária e identificando seus k-vizinhos mais próximos que também pertencem a essa classe. Em seguida, um desses vizinhos é escolhido aleatoriamente, e um novo ponto de dados sintético é gerado por meio da interpolação linear no espaço de atributos, em um ponto aleatório ao longo do segmento de linha que une a instância original e o vizinho selecionado. Este processo enriquece o conjunto de treinamento com exemplos novos e plausíveis, ajudando a criar regiões de decisão mais bem definidas e generalizáveis para a classe minoritária (CHAWLA et al., 2002)

2.3.2.5 FairSMOTE

Esta técnica é uma adaptação do SMOTE projetada explicitamente com o objetivo de promover a justiça. O algoritmo primeiro identifica e remove instâncias consideradas fontes de viés e, em seguida, utiliza um processo de sobreamostragem que gera dados sintéticos de forma a balancear não apenas a classe alvo, mas também a representação dos grupos sensíveis (ex: homens e mulheres) dentro de cada classe. O objetivo é alcançar equidade tanto na distribuição de classes quanto na demográfica.

2.3.2.6 ADASYN

O ADASYN é uma técnica de sobreamostragem adaptativa e uma evolução do SMOTE. Sua principal característica é gerar mais exemplos sintéticos para as instâncias da classe minoritária que são mais difíceis de aprender (ou seja, aquelas que têm mais vizinhos da classe majoritária). Ao focar nessas "regiões difíceis", o ADASYN ajuda o modelo a prestar mais atenção às fronteiras de decisão complexas.

2.3.2.7 Sobreamostragem Aleatória (Random Over-Sampling)

Esta é a abordagem mais fundamental de sobreamostragem. Ela consiste em duplicar aleatoriamente instâncias da classe minoritária até que o conjunto de dados esteja balanceado. É uma técnica simples e rápida, mas seu uso excessivo pode levar ao sobreajuste (overfitting), pois o modelo pode acabar memorizando as instâncias duplicadas em vez de aprender padrões generalizáveis.

2.3.2.8 NearMiss

NearMiss é uma família de algoritmos de subamostragem mais sofisticados que o método aleatório. Em vez de descartar instâncias da classe majoritária ao acaso, ele seleciona quais manter com base em sua distância em relação às instâncias da classe minoritária. O objetivo é reter as amostras da classe majoritária que são mais importantes para definir a fronteira de decisão, reduzindo a perda de informação útil.

2.3.2.9 BorderlineSMOTE

Esta é outra variante inteligente do SMOTE. O algoritmo primeiro identifica as instâncias da classe minoritária que estão na "fronteira"da decisão, ou seja, aquelas que são mais prováveis de serem classificadas incorretamente. Em seguida, ele gera exemplos sintéticos exclusivamente a partir dessas instâncias da fronteira, reforçando as regiões mais críticas para o aprendizado do classificador.

2.3.2.10 Reponderação sensível de classe

Diferente das outras técnicas, esta é uma abordagem a nível de algoritmo que não modifica a quantidade de dados, mas sim o processo de aprendizado. Ela atribui um peso maior às instâncias da classe minoritária ou de um grupo demográfico sub-representado durante o treinamento do modelo. Ao fazer isso, a função de perda do algoritmo penaliza mais severamente os erros cometidos nesses casos, forçando o modelo a dar mais atenção a esses exemplos sem alterar a distribuição original dos dados. O cálculo dos pesos para este trabalho seguiu o método da ponderação inversa à frequência do grupo sensível. Para cada conjunto de treinamento, foi calculada a proporção de cada grupo definido pelo atributo "sexo"(ex: 60% masculino, 40% feminino). Em seguida, o peso atribuído a cada instância foi definido como o inverso da frequência de seu respectivo grupo.

Este método aumenta matematicamente a importância de cada exemplo do grupo sub-representado durante o cálculo da função de perda, incentivando o modelo a aprender padrões desse grupo com maior atenção, sem o risco de sobreajuste que a duplicação de amostras pode gerar.

Esta técnica foi aplicada utilizando o parâmetro sample weight, disponível em diversas implementações da biblioteca scikit-learn. No entanto, nem todos os 11 classificadores selecionados para este estudo suportam este parâmetro de forma nativa. Por essa razão técnica, a análise para a abordagem de Reponderação foi conduzida em um subconjunto de 7 algoritmos compatíveis: Random Forest, Logistic Regression, Gradient Boosting, XGBoost, CatBoost, Ridge Classifier e SGD Classifier. Os algoritmos KNeighbors, SVC, GaussianNB e MLPClassifier foram excluídos dos experimentos para esta abordagem específica.

3 TRABALHO DESENVOLVIDO

Este capítulo apresenta o arcabouço metodológico desenvolvido para investigar o impacto de técnicas de manipulação de dados na performance preditiva e na equidade de modelos de aprendizado de máquina. A questão central desta dissertação é: como as intervenções nos dados afetam os vieses algorítmicos relacionados a atributos protegidos?

Para responder a esta questão, a abordagem metodológica consistiu em utilizar técnicas tradicionalmente usadas para o balanceamento de classes para a promoção de equidade de gênero. Em vez de aplicar as técnicas diretamente sobre a variável alvo (ex: 'aprovado' vs. 'reprovado'), o balanceamento foi realizado sobre os subgrupos interseccionais, formados pela combinação do atributo sensível sexo com o rótulo da classe. Ao fazer isso, o conjunto de treinamento foi reestruturado para apresentar ao modelo uma distribuição onde homens e mulheres possuem representações equitativas em ambos os resultados (positivo e negativo). A eficácia desta estratégia foi então avaliada por meio de um estudo comparativo de larga escala, cuja execução e métricas são detalhadas nas seções subsequentes.

3.1 METODOLOGIA GERAL

O presente trabalho propõe uma metodologia experimental que reside em uma análise comparativa robusta, estruturada em duas etapas fundamentais. O objetivo é contrastar o comportamento de múltiplos algoritmos de classificação quando treinados em cenários distintos: um cenário de linha de base (baseline), sem qualquer tratamento dos dados, e um cenário de intervenção, onde os dados de treinamento são pré-processados por um portfólio de técnicas de balanceamento.

- **Etapa 1: Análise da baseline (Baseline).** Nesta fase inicial, os modelos são treinados utilizando os conjuntos de dados em seu estado original, com o desbalanceamento de classes e os vieses presentes. O desempenho e a equidade de cada modelo são mensurados e armazenados através da biblioteca DALEX, servindo como um ponto de referência para todas as comparações futuras. Esta etapa busca responder à pergunta: "Qual é o nível de viés e de performance dos modelos sem qualquer intervenção?".

- **Etapa 2: Análise Pós-Intervenção.** Na segunda fase, os mesmos conjuntos de dados de treinamento são submetidos a 10 abordagens diferentes de técnicas de balanceamento.

Para cada uma das versões balanceadas dos dados, os 11 algoritmos de classificação são novamente treinados e avaliados. O desempenho e a equidade são mensurados da mesma forma que na Etapa 1. Esta etapa busca responder à pergunta: "Como cada técnica de balanceamento altera o viés e a performance dos modelos?".

A análise final consiste na comparação direta dos resultados obtidos nas duas etapas, permitindo uma avaliação aprofundada dos trade-offs entre a correção do desbalanceamento e a promoção da justiça algorítmica.

3.2 SELEÇÃO DAS BASES DE DADOS

A seleção de um conjunto de dados diversificado é um pilar fundamental para garantir a validade e a generalização dos resultados deste estudo. Para este fim, foram escolhidas oito bases de dados públicas, amplamente reconhecidas e utilizadas como benchmarks na literatura de aprendizado de máquina e, em especial, nos estudos sobre justiça algorítmica (MEHRABI et al., 2021).

Esses conjuntos de dados abrangem domínios críticos e socialmente sensíveis, como justiça criminal, finanças, censo e saúde. O critério de seleção principal, além da presença de desbalanceamento do grupo social, foi a disponibilidade de atributos sensíveis que permitissem uma análise robusta do viés algorítmico, com um foco particular no atributo gênero. A tabela 1 sumariza as características das bases de dados escolhidas. Uma descrição detalhada de cada base de dados, incluindo seu contexto, a tarefa de classificação e os atributos de interesse, pode ser encontrada no apêndice A.

Tabela 1 – Resumo das características das 8 bases de dados selecionadas para os experimentos.

Nome da Base	Domínio	Inst.	Atr.	Atributo Alvo	Atributo Sensível	Desbal. (%)
Adult Census	Censo	32.561	14	Renda Anual > 50k	Sexo	33.8
German Credit	Finanças	1.000	26	Risco de Crédito	Sexo	38.0
COMPAS	Justiça Criminal	7.214	11	Predição de reincidência	Sexo	61.3
Credit Card Approval	Finanças	30.459	52	Bom ou Mau pagador	Sexo	34.6
Default Credit	Finanças	30.000	25	Inadimplência	Sexo	20.7
Heart Disease	Saúde	1.025	14	Presença Doença Cardíaca	Sexo	39.12
Diabetes Health Indicators	Saúde	23.638	13	Presença Diabete	Sexo	6.2
Law School	Educação	20.798	12	Aprovação no exame da ordem	Sexo	12.22

3.3 ALGORITMOS DE CLASSIFICAÇÃO SELECIONADOS

A fim de garantir que as conclusões deste trabalho sobre o impacto das técnicas de balanceamento sejam robustas e generalizáveis, foi selecionado um conjunto de 11 algoritmos de classificação distintos. A escolha desses modelos foi estratégica para abranger uma ampla gama de paradigmas de aprendizado, incluindo modelos lineares (como Regressão Logística e Ridge Classifier), métodos baseados em instâncias como K-Nearest Neighbors (KNN), máquinas de vetores de suporte como Support Vector Classifier (SVC), modelos probabilísticos (Gaussian Naive Bayes), redes neurais (MLPClassifier) e, com especial atenção, um forte conjunto de métodos de ensemble baseados em árvores (Random Forest, Gradient Boosting, XGBoost e CatBoost), que representam o estado da arte em muitos problemas de classificação tabular. Os algoritmos podem ser visualizados na tabela 2

Tabela 2 – Relação dos 11 algoritmos de classificação selecionados para os experimentos, agrupados por família.

Família do Algoritmo	Algoritmo Selecionado
Modelos Lineares	Regressão Logística
	Ridge Classifier
Support Vector Machines (SVM)	Support Vector Classifier (SVC)
Baseado em Instâncias	K-Vizinhos Mais Próximos (KNN)
Modelos Probabilísticos	Naive Bayes Gaussiano
Baseado em Árvores	Decision Tree
Métodos de Ensemble	Random Forest
	Gradient Boosting Machines
	XGBoost
	CatBoost
Redes Neurais	MLPClassifier

Para todos os algoritmos como é possível observar na tabela 3, foram utilizadas as implementações padrões da biblioteca `sklearn 1.7.1` e de bibliotecas compatíveis como `xgboost` e `lightgbm`. A fim de garantir a consistência no pré-processamento, cada classificador foi integrado a um pipeline, que aplicou as mesmas etapas de transformação de dados antes do treinamento. A reprodutibilidade dos experimentos foi assegurada pela fixação do random_state=123 tanto na divisão dos dados (70% para treino, 30% para teste) quanto na inicialização de todos os algoritmos. Os modelos foram treinados com seus hiperparâmetros padrões para garantir uma base de comparação justa. As exceções foram os modelos DecisionTreeClassifier e RandomForestClassifier, que tiveram sua complexidade controlada com a definição de max_depth=7, e n_estimators=200, valores definidos a partir de uma análise exploratória inicial para evitar sobreajuste.

Tabela 3 – Descrição dos 11 algoritmos de classificação utilizados.

Algoritmo	Descrição
Logistic Regression	Modelo linear que utiliza a função logística para modelar a probabilidade de uma ocorrência de classe binária.
Ridge Classifier	Classificador linear que aplica a regularização L2 (Ridge) para penalizar coeficientes de grande magnitude, prevenindo o sobreajuste.
SGD Classifier	Implementa modelos lineares (como SVM e Regressão Logística) com treinamento via Gradiente Descendente Estocástico, ideal para grandes volumes de dados.
SVC (Support Vector)	Encontra o hiperplano ótimo que melhor separa as classes no espaço de atributos, maximizando a margem entre elas.
KNN	Algoritmo não-paramétrico que classifica um novo ponto com base na classe da maioria de seus "k"vizinhos mais próximos no conjunto de treino.
GaussianNB	Classificador probabilístico baseado no Teorema de Bayes com a suposição "ingênua"de independência condicional entre os atributos.
Random Forest	Método de ensemble que opera combinando múltiplas árvores de decisão para reduzir a variância e o sobreajuste.
Gradient Boosting	Método de ensemble que constrói árvores de forma sequencial, onde cada nova árvore é treinada para corrigir os erros residuais da anterior.
XGBoost	Implementação otimizada e de alta performance do Gradient Boosting, conhecida por sua velocidade, regularização e precisão.
CatBoost	Variante de Gradient Boosting com tratamento nativo e eficiente para atributos categóricos, evitando a necessidade de pré-processamento manual.
MLPClassifier	Uma rede neural artificial do tipo Perceptron Multicamadas (Multi-layer Perceptron), capaz de aprender padrões complexos e não-lineares nos dados.

3.4 DESCRIÇÃO DAS TÉCNICAS DE BALANCEAMENTO DE DADOS APLICADAS

O cerne da intervenção experimental reside na aplicação de um grupo de 10 abordagens técnicas de balanceamento de dados, implementadas através da biblioteca imblearn 0.14.0 para as técnicas de reamostragem mais consolidadas (como SMOTE e suas variantes), e scripts customizados com pandas `2.3.2` e `sklearn 1.7.1` para as abordagens conceituais, como o embaralhamento de atributo e a reponderação. As técnicas que foram escolhidas para representar as principais famílias de abordagens a nível de dados podem ser observadas na tabela 4:

Tabela 4 – As 10 abordagens ténicas de balanceamento de dados utilizadas nos experimentos.

Categoria da Abordagem	Técnica / Descrição
Baseline (Justiça)	**1.** Embaralhamento do atributo sensível (Shuffling)
Sobreamostragem (Oversampling)	**2.** RandomOverSampler (Duplicação da classe minoritária)
	3. BorderlineSMOTE
	4. SMOTE
	5. ADASYN (Adaptive Synthetic Sampling)
Subamostragem (Undersampling)	**6.** RandomUnderSampler (Subamostragem da classe majoritária)
	7. NearMiss
Focada em Equidade (Fairness)	**8.** Fair-SMOTE (Balanceamento com remoção de viés nos dados)
Nível de Algoritmo	**9.** Reponderação sensível de classes
Híbrida (Combinação)	**10.** Duplicação com classe oposta

3.5 DIVISÃO DOS CONJUNTOS DE DADOS

Uma etapa fundamental para a avaliação imparcial de modelos de aprendizado de máquina é o particionamento do conjunto de dados em subconjuntos de treinamento e teste. Para todos os experimentos realizados nesta dissertação, foi adotada uma divisão na proporção de 70% para o conjunto de treinamento e 30% para o conjunto de teste. É crucial ressaltar que todas as técnicas de balanceamento e pré-processamento foram aplicadas exclusivamente ao conjunto de treinamento. O conjunto de teste, mantido intacto, foi utilizado somente na etapa final para avaliar a capacidade de generalização do classificador em dados não vistos, permitindo assim uma avaliação justa e indireta da eficácia de cada abordagem de intervenção.

3.6 MEDIDAS DE AVALIAÇÃO DE PERFORMANCE PREDITIVA

A avaliação da performance de um modelo de classificação é uma etapa crítica para determinar a eficácia e capacidade de generalização do modelo. No entanto, em cenários de dados desbalanceados, que é o foco deste estudo, a escolha da métrica de avaliação é particularmente crucial, pois medidas tradicionais como a acurácia podem ser enganosas. Elas tendem

a favorecer modelos que ignoram a classe minoritária, resultando em uma falsa sensação de bom desempenho enquanto falham em identificar os casos de maior interesse prático. Para contornar essa limitação e obter uma visão multidimensional e fidedigna do comportamento dos classificadores, foi utilizado um conjunto de métricas complementares. Para a avaliação da performance preditiva, foi utilizado um conjunto abrangente de métricas. Este conjunto inclui a Acurácia como ponto de referência, mas também a Precisão, o Recall, o F1-Score e a AUC-ROC, que, juntos, oferecem uma análise robusta dos trade-offs do modelo. Embora todas as métricas tenham sido coletadas, a análise de resultados dará foco principal ao F1-Score, por ser o indicador mais equilibrado para cenários com desbalanceamento de classes. A seguir, cada uma das cinco métricas será brevemente definida.

3.6.1 Acurácia (Accuracy)

A acurácia é a métrica mais intuitiva e representa a proporção geral de previsões corretas que o modelo fez em relação ao total de amostras. Ela responde à pergunta: "De todas as classificações, qual a porcentagem de acertos?". Embora seja de fácil interpretação, a acurácia é uma medida enganosa em conjuntos de dados desbalanceados, pois um modelo pode alcançar um valor alto simplesmente ao prever a classe majoritária. Neste trabalho, ela é utilizada principalmente como um ponto de referência comparativo.

3.6.2 Recall (Revocação ou Sensibilidade)

O recall mede a capacidade do modelo de identificar corretamente todas as instâncias positivas. Ele responde à pergunta: "De todos os casos que eram realmente positivos, quantos o modelo conseguiu encontrar?". Esta métrica é crucial em cenários onde os falsos negativos (não encontrar um caso positivo) são muito custosos, como em diagnósticos médicos ou detecção de fraudes, sendo fundamental para avaliar o desempenho na classe minoritária.

3.6.3 Precisão (Precision)

A precisão avalia a confiabilidade das previsões positivas feitas pelo modelo. Ela responde à pergunta: "Das vezes que o modelo previu um resultado positivo, quantas vezes ele realmente acertou?". Uma alta precisão indica uma baixa taxa de falsos positivos. Esta métrica é impor-

tante em cenários onde os falsos positivos geram custos, como em campanhas de marketing direcionadas ou na concessão de benefícios.

3.6.4 F1-Score

O F1-Score é a média harmônica entre a precisão e o recall, combinando ambas as métricas em um único valor. Ele busca um equilíbrio entre as duas, sendo particularmente útil quando há um trade-off e ambas são igualmente importantes. Por ser robusto em cenários de desbalanceamento de classes, o F1-Score é uma das principais métricas de performance para avaliar a eficácia das técnicas de balanceamento nesta dissertação.

3.6.5 AUC (Area Under the ROC Curve)

A AUC-ROC, mede a capacidade geral de um classificador de distinguir entre as classes positiva e negativa em todos os os limiares de classificação. Uma AUC-ROC de 1.0 representa um modelo perfeito, enquanto 0.5 representa um desempenho aleatório. Ela responde à pergunta: "Qual a probabilidade de o modelo atribuir uma pontuação maior a um caso positivo aleatório do que a um caso negativo aleatório?". É uma métrica robusta e agregada do desempenho geral do modelo.

3.7 MÉTRICAS DE AVALIAÇÃO DE JUSTIÇA

Além da performance preditiva, um pilar central desta dissertação é a avaliação da equidade algorítmica. Para mover a discussão sobre justiça de um plano puramente conceitual para uma análise quantitativa e objetiva, é necessário adotar um conjunto de métricas de justiça. Essas métricas são projetadas especificamente para detectar e quantificar a presença de vieses, medindo as disparidades no tratamento que o modelo oferece a diferentes subgrupos populacionais, definidos por atributos sensíveis.

A maioria dessas métricas, e todas as utilizadas neste trabalho, baseiam-se no princípio da paridade de grupo, onde o comportamento do modelo é comparado entre um grupo historicamente privilegiado e um ou mais grupos não privilegiados. Conforme detalhado no protocolo experimental, essas métricas são calculadas como rácios entre os indicadores de desempenho de cada grupo (como a Taxa de Verdadeiros Positivos ou a Acurácia), permitindo uma ava-

liação clara de quão equitativo é o modelo. As subseções seguintes descrevem as principais métricas de justiça empregadas nesta pesquisa para auditar os modelos.

Tabela 5 – Descrição das métricas de justiça

Nome de justiça entre subgrupos	Métrica	Fórmula
Paridade Estatística	Taxa de Seleção (STP)	$\frac{VP + FP}{VP + FN + FP + VN}$
Igualdade de Oportunidade	Taxa de Verdadeiros Positivos (TPR)	$\frac{VP}{VP + FN}$
Igualdade Preditiva	Taxa de Falsos Positivos (FPR)	$\frac{FP}{FP + VN}$
Paridade Preditiva	Valor Preditivo Positivo (PPV)	$\frac{VP}{VP + FP}$
Igualdade de Acurácia	Acurácia (ACC)	$\frac{VP + VN}{VP + FN + FP + VN}$

Onde: VP = Verdadeiro Positivo; VN = Verdadeiro Negativo; FP = Falso Positivo; FN = Falso Negativo.

3.7.1 Paridade Estatística

A paridade estatística é a implementação direta do critério de **Independência**. Ela exige que a fração de predições positivas seja a mesma para todos os subgrupos, independentemente de seus resultados reais (DWORK et al., 2012).

$$P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b)$$

3.7.2 Igualdade de Oportunidade

Esta métrica corresponde a uma das condições do critério de **Separação**. Ela verifica se o classificador possui a mesma Taxa de Verdadeiros Positivos TPR para cada subgrupo, garantindo que indivíduos qualificados tenham a mesma chance de receber uma predição positiva (HARDT; PRICE; SREBRO, 2016).

$$P(\hat{Y} = 1 \mid A = a, Y = 1) = P(\hat{Y} = 1 \mid A = b, Y = 1)$$

3.7.3 Igualdade Preditiva

Sendo uma das condições para satisfazer o critério de **Separação**, esta métrica garante que o classificador tenha a mesma Taxa de Falsos Positivos FPR para cada subgrupo (CORBETT-DAVIES et al., 2017). Quando a Igualdade de Oportunidade e a Igualdade Preditiva são satisfeitas em conjunto, o critério de Separação é alcançado.

$$P(\hat{Y} = 1 \mid A = a, Y = 0) = P(\hat{Y} = 1 \mid A = b, Y = 0)$$

3.7.4 Paridade Preditiva

Esta métrica é uma relaxação do critério de **Suficiência**. Ela mede se um modelo possui o mesmo Valor Preditivo Positivo PPV para cada subgrupo, ou seja, se a probabilidade de um indivíduo ser verdadeiramente positivo, dado que recebeu uma predição positiva, é a mesma para todos os grupos (CHOULDECHOVA, 2017).

$$P(Y = 1 \mid A = a, \hat{Y} = 1) = P(Y = 1 \mid A = b, \hat{Y} = 1)$$

3.7.5 Igualdade de Acurácia

Esta métrica de justiça avalia se a acurácia geral do modelo, definida como a probabilidade de uma predição correta ($P(\hat{Y} = Y)$), é a mesma entre os diferentes subgrupos.

$$P(\hat{Y} = Y \mid A = a) = P(\hat{Y} = Y \mid A = b)$$

3.8 BIBLIOTECA DALEX

Para garantir uma comparação rigorosa e consistente entre os 11 algoritmos de classificação, que possuem implementações distintas, foi adotado o framework de auditoria agnóstica DALEX. Dentre as ferramentas disponíveis na literatura, como AIF360 e Fairlearn, optou-se pela biblioteca DALEX por ser uma ferramenta de desenvolvimento mais recente, por suas poderosas capacidades de visualização gráfica e pela clareza com que seus relatórios permitem interpretar e comparar os níveis de justiça entre múltiplos modelos. A metodologia de uso consistiu em, para cada modelo treinado, encapsulá-lo em um objeto Explainer do DALEX, que padroniza a interface de avaliação. Este objeto foi então submetido a duas funções de

análise principais: (1) a função model performance(), para extrair métricas de desempenho preditivo como F1-Score; e (2) a função fairness check(), para calcular os índices de paridade para as cinco métricas de justiça. A combinação dessas funções permitiu, de forma sistemática e reprodutível, avaliar cada um dos modelos gerados nos eixos de performance e equidade, fornecendo a base quantitativa para todas as análises apresentadas neste trabalho.

A criação de um objeto Explainer é o primeiro e mais fundamental passo. Ele requer o modelo treinado, os dados de validação ou teste (X e y) e, opcionalmente, um rótulo para o modelo. Uma vez criado, este objeto Explainer se torna a porta de entrada para todas as análises de performance, importância de variáveis e, crucialmente para esta pesquisa, de justiça.

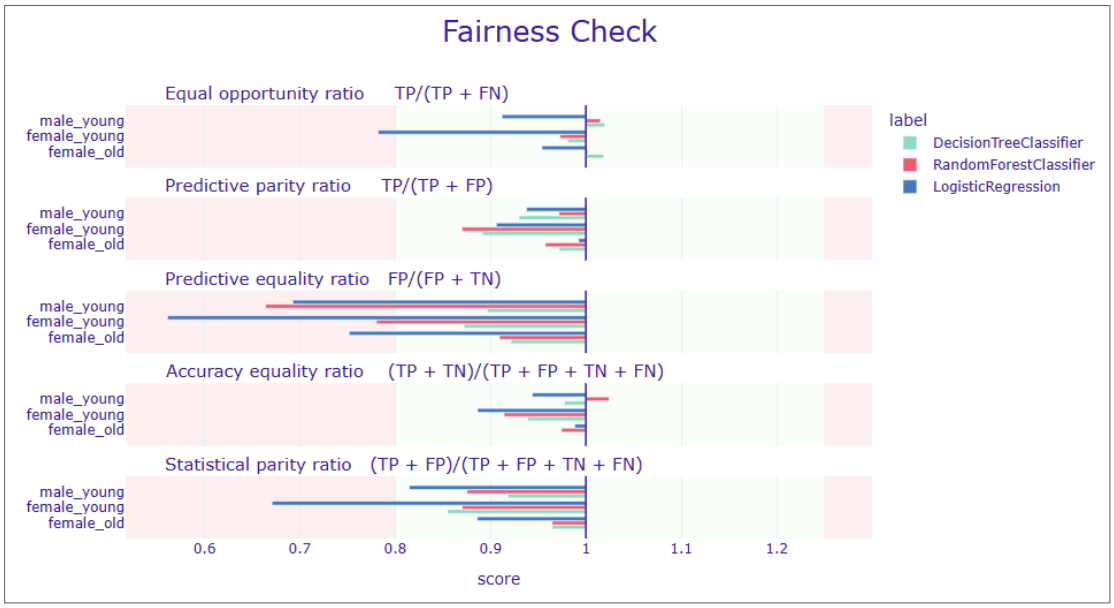
O cálculo de justiça da função fairness check() baseia-se no ratio de paridade. Para cada métrica de avaliação (ex: TPR, Acurácia), a função calcula seu valor para o subgrupo não privilegiado e o divide pelo valor da mesma métrica para o subgrupo privilegiado. Um ratio de 1.0 indica uma paridade perfeita, significando que o modelo se comporta de maneira idêntica para ambos os grupos naquela métrica específica. Para determinar se um desvio de 1.0 é significativo, a ferramenta adota a "regra dos quatro quintos"(four-fifths rule), um princípio prático utilizado em contextos legais antidiscriminatórios. Esta regra estabelece que um ratio abaixo de 0.8 é um indicador de impacto adverso. Por simetria, isso define um intervalo de equidade de [0.8, 1.25]. Qualquer ratio fora desta faixa é sinalizado pela ferramenta como um viés detectado. Por exemplo, se a Taxa de Verdadeiros Positivos (TPR) para o grupo privilegiado for de 90% e para o não privilegiado for de 63%, o ratio de 0.7 (63/90) cairia fora do intervalo, indicando a presença de viés.

Essa abordagem garante que a comparação entre os 11 algoritmos diferentes seja justa e consistente, pois todos são avaliados exatamente sob as mesmas condições e com as mesmas ferramentas de medição, eliminando vieses que poderiam surgir do uso de funções de avaliação específicas de cada biblioteca de modelo.

Além da análise numérica, a ferramenta oferece uma poderosa visualização do viés através da função plot(). Com um gráfico que exibe os resultados de paridade para todas as métricas simultaneamente. O protocolo experimental desta dissertação consistiu em aplicar este processo de verificação a cada um dos classificadores, tanto no cenário de linha de base quanto após a aplicação de cada uma das 10 técnicas de balanceamento de dados. A análise de justiça foi realizada com a biblioteca DALEX. Como podemos ver no exemplo disponibilizado pela própria biblioteca na Figura 2, a ferramenta gera um gráfico que permite uma interpretação

visual e imediata do viés do modelo.

Figura 2 – Exemplo de gráfico gerado após aplicação do método fairness_check.



Fonte: Documentação oficial da biblioteca Dalex (MI2.AI, 2020).

3.9 EXECUÇÃO DA ETAPA 1: ANÁLISE DA BASELINE (BASELINE)

A primeira etapa do experimento consiste em estabelecer a linha de baseline de performance e equidade para cada combinação de base de dados e algoritmo, sem qualquer intervenção de balanceamento. O processo é executado de forma sistemática e automatizada.

O procedimento segue um pipeline rigoroso: para cada uma das 8 bases de dados, é realizada uma partição estratificada para dividir os dados em conjuntos de treinamento (70) e teste (30), garantindo que a proporção de classes seja mantida em ambas as partições. Em seguida, um loop é executado sobre os 11 algoritmos de classificação selecionados. Dentro deste loop, cada algoritmo é treinado utilizando exclusivamente os dados de treinamento originais. Após o treinamento, um explicador dalex é instanciado com o modelo treinado e os dados de teste. A função fairness check() é então invocada, passando o explicador e o atributo protegido relevante para aquela base de dados. Ao final de cada execução, foram extraídas e armazenadas cinco métricas de performance preditiva (Acurácia, Recall, Precisão, AUC-ROC e F1 Score) e às cinco métricas de justiça TPR, ACC, PPV, FPR e STP.

3.10 EXECUÇÃO DA ETAPA 2: ANÁLISE PÓS-INTERVENÇÃO

A segunda etapa experimental é onde o efeito das técnicas de balanceamento é investigado. O processo é uma expansão da etapa 1. O loop para os 11 algoritmos de classificação selecionados é executado para cada uma das 10 técnicas de balanceamento de dados selecionadas.

Para cada base de dados e para cada técnica de balanceamento, a técnica é aplicada apenas ao conjunto de treinamento. Isso é crucial para evitar o vazamento de dados, pois o conjunto de teste deve permanecer em seu estado original para refletir o ambiente real de implantação. Após a aplicação da técnica, uma nova versão balanceada dos dados de treinamento é gerada. É sobre este novo conjunto de dados que os 11 algoritmos de classificação são treinados. A partir daí, o processo segue de forma idêntica à Etapa 1: um explicador dalex é criado para cada modelo treinado, e a função fairness check() é utilizada para extrair e armazenar as métricas de performance e justiça.

3.11 ANÁLISE COMPARATIVA DOS RESULTADOS

A fase final da metodologia proposta é a definição dos resultados coletados nas etapas 1 e 2 e suas análises e comparações. Esta análise, cujos resultados detalhados serão apresentados no capítulo 6, será conduzida sob duas perspectivas principais:

- **Perspectiva de Performance Preditiva:** O primeiro objetivo é verificar se as técnicas de balanceamento cumprem sua promessa primária de melhorar o desempenho do modelo na classe minoritária. A principal métrica para esta análise será o F1-Score, que representa um equilíbrio entre precisão e recall, sendo particularmente informativo em cenários desbalanceados. Será analisada a variação percentual do F1-Score para cada técnica em comparação com a linha de base.
- **Perspectiva de Justiça Algorítmica:** O segundo e mais crítico objetivo é avaliar como as técnicas de balanceamento impactam a equidade do modelo. Para cada métrica de justiça, será analisada a variação em relação à baseline. A análise buscará responder a perguntas como: A técnica melhorou a equidade, aproximando o ratio de 1.0? A técnica piorou a equidade, afastando o rácio de 1.0? Existe um trade-off evidente, onde a melhoria no F1-Score veio ao custo de uma piora na equidade?

Serão geradas tabelas para comparar o efeito médio de cada técnica de balanceamento sobre as métricas de performance e justiça em todos os cenários experimentais. Esta análise dupla permitirá a extração de conclusões robustas sobre a adequação e os riscos de cada abordagem de balanceamento.

4 RESULTADOS E DISCUSSÕES

Este capítulo apresenta e discute os resultados obtidos a partir dos experimentos conduzidos para avaliar o impacto de técnicas de balanceamento de dados aplicadas a modelos de AM. A análise foi estruturada para investigar diferentes cenários, incluindo aqueles com alto viés intrínseco, desbalanceamento de classe severo e baixo viés, através das métricas comuns e de justiça.

O objetivo central desta pesquisa é analisar como diferentes estratégias de balanceamento de dados afetam a performance preditiva e a justiça algorítmica dos modelos de AM, não apenas em sua capacidade de predição, mas também em sua conformidade com métricas de justiça consolidadas na literatura. Os resultados são apresentados de forma comparativa, contrastando o desempenho dos modelos em um cenário baseline (sem intervenção) com os resultados obtidos após a aplicação de dez abordagens distintas de balanceamento de dados. A discussão busca extrair insights sobre a eficácia de cada técnica, os desafios inerentes a cada tipo de problema e as implicações práticas para o desenvolvimento de modelos de aprendizado de máquina mais justos e robustos.

Dentre as técnicas avaliadas, a abordagem 5 (FairSmote), não apresentou bons resultados nos experimentos 4.1.2, 4.1.5 e 4.1.6. Nesses casos, a estratégia de remover rótulos considerados tendenciosos e aplicar um rebalanceamento interno com base no atributo sensível não foi suficiente para garantir justiça nos modelos. Um dos principais problemas foi a distribuição altamente desbalanceada entre os subgrupos sensíveis, o que dificultou a criação de exemplos equilibrados entre as classes positiva e negativa. Além disso, a baixa quantidade de dados em certos grupos minoritários impediu a geração de amostras sintéticas representativas, limitando a eficácia do balanceamento. Outro fator que prejudicou o desempenho foi a eliminação de exemplos reais durante o processo de filtragem de rótulos, o que pode ter removido informações relevantes para o aprendizado. Esses fatores combinados explicam o insucesso da abordagem 5 nos experimentos mencionados, destacando a necessidade de considerar a qualidade e a representatividade dos dados sensíveis ao aplicar estratégias de balanceamento mais agressivas.

4.1 RESULTADOS OBTIDOS

4.1.1 Compas

A análise consolidada do Experimento 1, apresentada na tabela 6, permite uma visualização direta do impacto de cada abordagem de intervenção sobre os eixos de performance preditiva e justiça algorítmica. Os resultados foram agregados calculando-se a média das métricas para os 11 algoritmos de classificação, permitindo uma avaliação geral da eficácia de cada técnica.

Em termos de performance, quase todas as técnicas de intervenção promoveram uma melhoria no F1-Score em comparação com o baseline (0.587). As abordagens de reamostragem, de forma geral, foram as mais eficazes. O destaque foi para a Abordagem 4 (SMOTE), a Abordagem 3 (Undersampling) e a Abordagem 9 (BorderlineSMOTE), que registraram os maiores ganhos, com aumentos de 11.9%, 11.5% e 11.1%, respectivamente. Isso confirma que as estratégias de reamostragem são robustas para melhorar a capacidade do modelo em identificar a classe minoritária.

Na dimensão da justiça, no entanto, a história é mais complexa. O cenário baseline era severamente enviesado, com uma média de 3.0 vieses detectados por modelo. Notavelmente, muitas técnicas que melhoraram a performance, como a Subamostragem Aleatória e o SMOTE padrão, não conseguiram reduzir o número de vieses, mantendo a média em 3.0. A abordagem que mais se destacou na mitigação de viés foi a Abordagem 5, que reduziu o número médio de vieses em 20%. Contudo, foi a Abordagem 9 que obteve o resultado de equidade mais promissor, trazendo o índice de paridade da TPR para 1.20, muito próximo do limiar de justiça.

O trade-off entre os dois objetivos fica evidente ao comparar, por exemplo, SMOTE e BorderlineSMOTE. Embora o SMOTE tenha proporcionado o maior ganho de F1-Score (+11.9%), seu impacto na justiça foi nulo. Já o BorderlineSMOTE, com um ganho de performance quase idêntico (+11.1%), foi a abordagem mais eficaz na mitigação do viés de Igualdade de Oportunidade (TPR Ratio de 1.20) e também reduziu o número geral de vieses. Este resultado ilustra de forma contundente a tese central deste trabalho: a escolha de uma técnica de balanceamento implica uma decisão consciente sobre priorizar performance, justiça, ou o delicado equilíbrio entre ambos.

Tabela 6 – Resultados consolidados de performance e justiça para o Experimento 1 (COMPAS). As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações.

	Performance		Justiça		
Abordagem	**F1-Score (Avg)**	**Δ% vs. Baseline**	**Avg. Vieses Det.**	**Δ Vieses**	**TPR Ratio (Avg)**
Baseline	0.5873	–	3.0	–	2.32
1. Shuffle	0.6143	+4.6%	2.9	-3%	1.36
2. Duplicate Opposite	0.5960	+1.5%	3.0	+0%	2.21
3. Undersampling	0.6554	+11.5%	3.0	+0%	2.24
4. SMOTE	0.6570	+11.9%	3.0	+0%	2.07
5. Fair-SMOTE	0.6378	+8.5%	2.4	-20%	1.30
6. ADASYN	0.6243	+6.3%	5.0	+67%	1.84
7. Random Oversampling	0.6490	+10.5%	3.0	+0%	1.33
8. NearMiss	0.6128	+4.3%	3.0	+0%	1.43
9. BorderlineSMOTE	0.6525	+11.1%	2.7	-10%	1.20
10. Reweighting	0.6138	+4.5%	3.0	+0%	1.74

4.1.2 German Credit

Em forte contraste com o cenário de alto viés do Experimento 1, a análise do Experimento 2 (German Credit) revelou um comportamento distinto e igualmente revelador. Conforme apresentado na Tabela 7, os modelos treinados no cenário baseline já se mostraram majoritariamente justos, com uma média de apenas 0.27 vieses detectados e um excelente índice de paridade da TPR de 0.95. O desafio, portanto, não era criar equidade, mas sim otimizar a performance preditiva (F1-Score de 0.761) sem degradar a justiça.

Notavelmente, a aplicação das técnicas de reamostragem mais comuns resultou em um trade-off severamente negativo. Abordagens de subamostragem como NearMiss (-21.0%) e de sobreamostragem como SMOTE (-6.3%) não apenas degradaram a performance preditiva, mas também destruíram a equidade do sistema. O número médio de vieses para essas técnicas saltou de quase zero para mais de 3.5, um aumento de mais de 1200%, indicando que a manipulação da distribuição dos dados introduziu um viés significativo que não existia anteriormente.

Em contrapartida, as abordagens que não alteram a distribuição de forma agressiva mostraram-se as mais eficazes. A Abordagem 10 (Reweighting) emergiu como a melhor estratégia, proporcionando o maior ganho de performance (+5.0% no F1-Score) e mantendo um excelente nível de justiça. Similarmente, as técnicas de controle, como Abordagem 1 (Shuffle) e Abordagem 2

(Duplicate Opposite), também melhoraram a performance enquanto reduziam ou mantinham o número de vieses em praticamente zero.

Este resultado é uma evidência crítica de que, em cenários de baixo viés inicial, as técnicas de reamostragem podem ser desestabilizadoras e prejudiciais. A análise sugere que intervenções mais sutis, como a reponderação, ou aquelas focadas em dissociar o atributo sensível da predição, são mais seguras e eficazes para otimizar modelos em datasets já equitativos.

Tabela 7 – Resultados consolidados de performance e justiça para o Experimento 2 (German Credit). As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações.

	Análise de Performance		**Análise de Justiça**		
Abordagem	**F1-Score Médio**	**Δ% vs. Baseline**	**Avg. Vieses Detectados***	**Δ Vieses**	**TPR Ratio (Avg)**
Baseline	0.761	-	0.27	-	0.95
Abordagens de Intervenção					
1. Shuffle	0.785	+3.2%	0.00	-100%	1.04
2. Duplicate Opposite	0.784	+3.0%	0.18	-33.3%	1.02
3. Undersampling	0.678	-10.9%	3.09	+1044%	0.63
4. SMOTE	0.713	-6.3%	3.73	+1281%	0.43
5. Fair-SMOTE	N/A	N/A	N/A	N/A	N/A
6. ADASYN	0.741	-2.6%	3.82	+1315%	0.42
7. Random Oversampling	0.711	-6.6%	1.00	+270%	1.01
8. NearMiss	0.601	-21.0%	1.73	+541%	1.35
9. BorderlineSMOTE	0.718	-5.6%	0.27	0%	0.97
10. Reweighting	0.799	+5.0%	0.29	+7.4%	0.95

** Média do número de métricas com índice fora do intervalo [0.8, 1.25]. Valores 'NaN' ou 'None' foram contados como 0 vieses detectados.*

4.1.3 Adult Census

Enquanto alguns cenários experimentais demonstraram um "ganho duplo", o Experimento 3 (Adult Income) revela a complexidade e os perigos dos trade-offs entre a otimização da performance e a garantia da equidade. A análise do baseline para esta base de dados, um benchmark clássico para estudos de justiça, mostrou um duplo desafio: baixo desempenho preditivo, com um F1-Score médio de apenas 0.55, e um viés algorítmico severo e sistêmico, com uma média de 3.0 vieses detectados para cada modelo. Os índices de paridade, especialmente para Igualdade de Oportunidade (TPR Ratio de 0.43), estavam extremamente baixos,

indicando uma forte discriminação na capacidade do modelo de identificar corretamente os casos positivos entre os grupos.

A aplicação das técnicas de intervenção expôs um resultado crucial, conforme sumarizado na Tabela 8. A Abordagem 3 (Subamostragem Aleatória), embora tenha sido uma das mais eficazes em melhorar a performance, elevando o F1-Score médio em 10.9%, falhou drasticamente em mitigar o viés. Pelo contrário, esta técnica aumentou o número médio de vieses detectados em 18% e manteve o TPR Ratio em um nível inaceitável (0.58). Este é um exemplo claro de um trade-off negativo, onde um pesquisador focado apenas em performance poderia erroneamente concluir que a técnica foi um sucesso, enquanto na verdade ela agravou a injustiça do sistema.

Em forte contraste, a Abordagem 4 (SMOTE) demonstrou um equilíbrio muito mais desejável. Além de promover um ganho de performance comparável (aumento de 9.1% no F1-Score), esta técnica foi uma das mais eficazes em promover a justiça, reduzindo o número de vieses detectados e elevando o índice de paridade da TPR para 0.83, um valor considerado justo pelo limiar de [0.8, 1.25]. O comparativo direto entre Subamostragem e SMOTE nesta base de dados ilustra a tese central deste trabalho: a escolha da técnica de balanceamento é uma decisão crítica com implicações profundas e, por vezes, divergentes para a equidade e a performance do modelo final.

Tabela 8 – Resultados consolidados de performance e justiça para o Experimento 3 (Adult Income), avaliando o trade-off entre as abordagens. As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações.

Abordagem	Análise de Performance		Análise de Justiça		
	F1-Score Médio	Δ% vs. Baseline	Avg. Vieses Detectados*	Δ Vieses	TPR Ratio (Avg)**
Baseline	0.55	-	3.00	-	0.43
Abordagens de Intervenção					
1. Shuffle	0.50	-10.0%	3.0	0%	0.85
2. Duplicate Opposite	0.50	-10.0%	2.91	-3.0%	0.86
3. Undersampling	0.612	+10.91%	3.55	+18%	0.58
4. SMOTE	0.605	+9.09%	2.91	-3.0%	0.83
5. Fair-SMOTE	0.574	+3.51%	2.91	-3.0%	0.53
6. ADASYN	0.59	+6.78%	2.00	-33%	0.82
7. Random Oversampling	0.58	+5.1%	2.36	-21%	0.85
8. NearMiss	0.53	-3.6%	2.00	-33%	0.84
9. BorderlineSMOTE	0.595	+6.78%	2.09	-30%	0.80
10. Reweighting	0.524	-5.45%	2.71	-10%	0.54

** Média do número de métricas com índice fora do intervalo [0.8, 1.25].*
*** Média calculada desconsiderando o outlier do classificador GaussianNB.*

4.1.4 Default of Credit Card Clients

A análise do Experimento 4, conduzida na base de dados Default of Credit Card Clients, revelou os riscos inerentes às técnicas de subamostragem e reforçou a necessidade de uma escolha contextual das estratégias de intervenção. O cenário baseline para este dataset apresentou um desafio duplo: baixa performance preditiva (F1-Score médio de 0.428) e um viés moderado, com uma média de 2.0 vieses detectados por modelo.

Conforme demonstrado na Tabela 9, a aplicação de técnicas de sobreamostragem, como o Random Oversampling (Abordagem 7), mostrou-se uma estratégia viável, proporcionando o maior ganho de performance com um aumento de 23.4% no F1-Score médio, enquanto mantinha os níveis de justiça relativamente estáveis.

Em total contraste, as técnicas de subamostragem tiveram um efeito fortemente negativo na performance. A Abordagem 8 (NearMiss), em particular, causou uma queda drástica no

F1-Score médio para 0.344, uma perda de quase 20% em relação ao já baixo desempenho do baseline. Este resultado sugere que, para este dataset, a remoção de instâncias da classe majoritária, mesmo que bem-intencionada, acarreta uma perda de informação tão significativa que o modelo se torna incapaz de aprender padrões preditivos úteis. Este achado evidencia que, embora o desbalanceamento seja um problema, a solução não pode vir ao custo da eliminação de dados cruciais para a tarefa de classificação, destacando a superioridade das abordagens de sobreamostragem neste cenário específico.

Tabela 9 – Resultados consolidados de performance e justiça para o Experimento 4 (Default Credit), destacando os efeitos contrastantes das abordagens. As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações.

	Análise de Performance		**Análise de Justiça**	
Abordagem	**F1-Score Médio**	**Δ% vs. Baseline**	**Avg. Vieses Detectados***	**Avg. TPR Ratio****
Baseline	0.428	-	2.08	0.86
Abordagens de Intervenção				
1. Shuffle	0.424	-0.9%	1.91	1.01
2. Duplicate Opposite	0.421	-1.6%	1.91	0.93
3. Undersampling	0.485	+13.3%	3.17	0.65
4. SMOTE	0.457	+6.8%	0.00	1.01
5. Fair-SMOTE	0.573	+33.9%	3.00	1.21
6. ADASYN	0.486	+13.6%	5.00	1.04
7. Random Oversampling	0.531	+24.1%	0.00	1.04
8. NearMiss	0.344	-19.6%	2.42	1.10
9. BorderlineSMOTE	0.463	+8.2%	0.00	1.07
10. Reweighting	0.425	-0.7%	2.00	0.86

** Média do número de métricas com índice fora do intervalo [0.8, 1.25]. Valores 'NaN' ou 'None' foram contados como 0.*
*** Média do Índice de Paridade da Taxa de Verdadeiros Positivos (Igualdade de Oportunidade).*

4.1.5 Heart Disease

A análise do Experimento 5 (Heart Disease) oferece uma perspectiva distinta sobre a utilidade das técnicas de balanceamento. Este conseguiu que os modelos baseline alcançassem um alto desempenho preditivo, com um F1-Score médio de 0.886, conforme detalhado na

Tabela 10.

Neste cenário, a aplicação das técnicas de reamostragem, como a Subamostragem Aleatória (Abordagem 3) e o SMOTE (Abordagem 4), não resultou em ganhos de performance significativos; em alguns casos, houve até uma leve degradação. Este resultado era esperado, uma vez que não havia uma forte assimetria de classes para ser corrigida.

O achado mais relevante, no entanto, surgiu na análise de justiça. Embora o baseline apresentasse poucos vieses detectáveis, em parte devido ao pequeno tamanho da amostra que resultou em várias métricas NaN, a aplicação de praticamente todas as técnicas de reamostragem foi capaz de eliminar os vieses restantes, resultando em modelos perfeitamente justos segundo os critérios da ferramenta DALEX. A Tabela 10 ilustra como, mesmo com uma variação mínima no F1-Score, o número médio de vieses detectados foi consistentemente reduzido a zero pelas intervenções.

Tabela 10 – Resultados consolidados de performance e justiça para o Experimento 5 (Heart Disease). As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações.

Abordagem	**Análise de Performance**		**Análise de Justiça**		
	F1-Score Médio	**Δ% vs. Baseline**	**Avg. Vieses Detectados***	**Δ Vieses**	**TPR Ratio (Avg)**
Baseline	0.886	-	2.40	-	1.19
Abordagens de Intervenção					
1. Shuffle	0.879	-0.8%	1.64	+46%	1.01
2. Duplicate Opposite	0.874	-1.4%	0.00	-100%	1.00
3. Undersampling	0.849	-4.2%	0.18	-83%	1.08
4. SMOTE	0.871	-1.7%	0.27	-75%	1.01
6. ADASYN	0.857	-3.3%	0.18	-83%	1.00
7. Random Oversampling	0.866	-2.3%	0.00	-100%	1.00
8. NearMiss	0.803	-9.4%	0.00	-100%	0.99
9. BorderlineSMOTE	0.870	-1.8%	0.27	-75%	0.99
10. Reweighting	0.871	-1.7%	0.00	-100%	1.00

** Média do número de métricas com índice fora do intervalo [0.8, 1.25]. Valores 'NaN' ou 'None' foram contados como 0 vieses detectados.*

4.1.6 Diabetes Health Indicators

A análise do baseline confirmou o impacto do desbalanceamento da classe-alvo: os modelos foram incapazes de aprender a identificar a classe minoritária, resultando em valores de F1-Score extremamente baixos (média de 0.255), o que os torna funcionalmente inúteis. No entanto, na dimensão da justiça, devido ao equilíbrio mais balanceado do atributo "Sexo"comparado aos outros experimentos, os mesmos modelos baseline já se mostraram majoritariamente justos, com poucos ou nenhum viés de grupo detectado, criando um cenário de um modelo justo, porém ineficaz.

Neste contexto, a aplicação das técnicas de reamostragem, como a Subamostragem (Abordagem 3) e o SMOTE (Abordagem 4), teve um efeito notável. Conforme os resultados, essas técnicas promoveram um aumento massivo e necessário na performance preditiva, com o F1-Score médio mais do que dobrando. Crucialmente, essa melhoria de performance foi alcançada sem degradar a equidade do modelo. Como o atributo sensível já era balanceado, as técnicas de reamostragem da classe-alvo não introduziram novo viés, mantendo os excelentes índices de paridade do baseline.

Tabela 11 – Resultados consolidados de performance e justiça para o Experimento 6 (CDC Diabetes), ilustrando o ganho duplo em um cenário de desbalanceamento severo. As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações.

Abordagem	Análise de Performance		Análise de Justiça		
	F1-Score Médio	Δ% vs. Baseline	Avg. Vieses Detectados*	Δ Vieses	TPR Ratio (Avg)
Baseline	0.255	-	1.09	-	0.99
Abordagens de Intervenção					
1. Shuffle	0.258	+1.2%	0.18	-83%	1.17
2. Duplicate Opposite	0.253	-0.8%	0.00	-100%	1.14
3. Undersampling	0.457	+79.2%	0.09	-92%	0.95
4. SMOTE	0.463	+81.6%	0.00	-100%	1.01
6. ADASYN	0.595	+133.3%	2.09	+92%	1.15
7. Random Oversampling	0.461	+80.8%	0.00	-100%	1.01
8. NearMiss	0.334	+31.0%	0.00	-100%	1.02
9. BorderlineSMOTE	0.454	+78.0%	0.00	-100%	1.06
10. Reweighting	0.276	+8.2%	1.00	-8%	0.96

** Média do número de métricas com índice fora do intervalo [0.8, 1.25]. Valores 'NaN' ou 'None' foram contados como 0 vieses detectados.*

4.1.7 Credit Card Approval

Em forte contraste com os experimentos anteriores, a análise do Experimento 7 (Credit Card Approval) investiga um cenário com desbalanceamento de classes razoável, crucialmente, baixo viés demográfico inicial. Uma característica notável desta base de dados é a inversão do padrão de privilégio, onde o grupo feminino constituiu a maioria, sendo, portanto, tratado como o grupo privilegiado na análise. Conforme detalhado, os modelos treinados no cenário baseline já se mostraram excepcionalmente um pouco mais justos e com baixo viés detectado pelo DALEX para a grande maioria dos classificadores.

Neste contexto, o desafio para as técnicas de intervenção não era criar de fato justiça, mas sim manter a justiça enquanto se buscava otimizar a performance. Os resultados indicam que a maioria das abordagens foi bem-sucedida neste duplo objetivo. As técnicas de reamostragem, como a Subamostragem (Abordagem 3) e a Sobreamostragem (Abordagem 7), proporciona-

ram ganhos incrementais, porém consistentes, no F1-Score médio, ao mesmo tempo em que preservaram os excelentes índices de paridade do baseline.

Este tipo de análise também é de grande importância prática, pois sugere que em base de dados com uma representação demográfica um pouco mais equilibrada, o risco de que as técnicas de balanceamento de classes introduzam viés é significativamente menor. A aplicação dessas técnicas pode, portanto, ser considerada uma estratégia segura e eficaz para otimizar a performance preditiva sem comprometer a equidade do sistema, contanto que a auditoria de justiça seja mantida como uma etapa de verificação.

Tabela 12 – Resultados consolidados de performance e justiça para o Experimento 7 (Credit Card Approval), ilustrando a manutenção da equidade em um cenário de baixo viés inicial. As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações.

Abordagem	**Análise de Performance**		**Análise de Justiça**		
	F1-Score Médio	**Δ% vs. Baseline**	**Avg. Vieses Detectados***	**Δ Vieses**	**TPR Ratio (Avg)**
Baseline	0.684	-	0.00	-	0.99
Abordagens de Intervenção					
1. Shuffle	0.684	0.0%	0.00	0%	1.02
2. Duplicate Opposite	0.690	+0.9%	0.00	0%	1.01
3. Undersampling	0.703	+2.8%	0.00	0%	1.02
4. SMOTE	0.703	+2.8%	0.00	0%	1.01
5. Fair-SMOTE	0.677	-1.0%	0.00	0%	0.98
6. ADASYN	0.718	+5.0%	0.00	0%	0.99
7. Random Oversampling	0.717	+4.8%	0.00	0%	0.99
8. NearMiss	0.654	-4.4%	0.00	0%	1.03
9. BorderlineSMOTE	0.704	+2.9%	0.00	0%	1.00
10. Reweighting	0.686	+0.3%	0.14	N/A	1.00

** Média do número de métricas com índice fora do intervalo [0.8, 1.25]. 'None' foi contado como 0 vieses detectados.*

4.1.8 Law School

A análise do Experimento 8 (LSAC), caracterizado por um desbalanceamento de classe de 12%, forneceu o exemplo mais contundente dos riscos associados à aplicação indiscrimi-

nada de técnicas de balanceamento. Surpreendentemente, apesar do desequilíbrio na variável alvo, os modelos treinados no cenário baseline se mostraram excepcionalmente justos, com praticamente nenhum viés detectado em todos os 11 classificadores. Neste cenário de "justiça natural", a aplicação da maioria das técnicas de sobreamostragem (como SMOTE e Random Oversampling) conseguiu manter a equidade dos modelos. No entanto, a Abordagem 8 (NearMiss), uma técnica de subamostragem agressiva, teve um efeito adverso e catastrófico na justiça dos modelos, como pode ser visto na Tabela 13

Tabela 13 – Resultados consolidados de performance e justiça para o Experimento 8 (LSAC), ilustrando o efeito adverso da subamostragem agressiva. As células são coloridas para indicar o impacto da técnica em relação ao baseline: verde para melhorias, amarelo para resultados neutros e vermelho para degradações.

Abordagem	**Análise de Performance**		**Análise de Justiça**		
	F1-Score Médio	**Δ% vs. Baseline**	**Avg. Vieses Detectados***	**Δ Vieses**	**TPR Ratio (Avg)**
Baseline	0.933	-	0.00	-	1.03
Abordagens de Intervenção					
1. Shuffle	0.932	-0.1%	0.00	0%	1.03
2. Duplicate Opposite	0.931	-0.2%	0.00	0%	1.03
3. Undersampling	0.835	-10.5%	0.00	0%	1.03
4. SMOTE	0.865	-7.3%	0.00	0%	1.00
5. Fair-SMOTE	0.914	-2.0%	0.00	0%	1.03
6. ADASYN	0.850	-8.9%	0.00	0%	1.00
7. Random Oversampling	0.865	-7.3%	0.00	0%	1.03
8. NearMiss	0.537	-42.4%	2.67	N/A	1.30
9. BorderlineSMOTE	0.863	-7.5%	0.00	0%	1.01
10. Reweighting	0.934	+0.1%	0.14	N/A	1.03

** Média do número de métricas com índice fora do intervalo [0.8, 1.25]. 'None' foi contado como 0 vieses detectados.*

4.2 DISCUSSÃO GERAL E SÍNTESE DOS RESULTADOS

A jornada através dos oito experimentos apresentados neste capítulo revelou a natureza multifacetada e dependente de contexto da relação entre performance preditiva e justiça algorítmica. O objetivo central desta pesquisa foi investigar como diferentes técnicas de pré-

processamento, especificamente o balanceamento de dados, impactam este delicado equilíbrio sob variadas condições de desbalanceamento de classe e viés demográfico. Esta seção final sintetiza os principais achados, oferecendo uma visão holística sobre o comportamento dos algoritmos, a sensibilidade das métricas e as implicações práticas para o desenvolvimento de modelos de aprendizado de máquina mais justos e eficazes.

O achado mais proeminente e recorrente em toda a análise é a confirmação de que não existe uma solução universal para o duplo desafio de performance e justiça, com a eficácia de cada técnica de intervenção sendo intrinsecamente ligada às características específicas do conjunto de dados. Isso ficou especialmente evidente no Experimento 1 (COMPAS), que representou um cenário de alto viés inicial, com um F1-Score médio de apenas 0.587 e uma média de 3.0 vieses detectados por modelo no baseline. Neste cenário desafiador, embora a maioria das técnicas de reamostragem tenha obtido sucesso em aumentar a performance preditiva, observou-se uma clara divergência no impacto sobre a equidade. Abordagens como o SMOTE padrão (Abordagem 4) e a Subamostragem Aleatória (Abordagem 3), apesar de elevarem o F1-Score em mais de 11%, não foram capazes de reduzir o número de vieses detectados, ilustrando um trade-off onde a otimização da performance não garantiu a melhoria da justiça.

O Experimento 3 (Adult) exemplificou o clássico cenário de trade-off. Com um desbalanceamento moderado e um viés inicial presente, mas não extremo, as intervenções que otimizavam a performance (aumento do F1-Score) por vezes causavam uma leve degradação em algumas métricas de justiça, e vice-versa. Este cenário sublinha a importância da seleção cuidadosa do modelo e da técnica, onde o analista deve ponderar qual dimensão, performance ou equidade, é prioritária para o caso de uso específico.

Os Experimentos 7 (Credit Approval) e o baseline do Experimento 8 (Law School) mostraram que, quando um dataset já possui um equilíbrio demográfico razoável, o principal objetivo da intervenção muda de "criar"para "manter"a justiça ou avaliar se está sendo prejudical. Nesses casos, a maioria das técnicas de balanceamento se mostrou segura, permitindo otimizar a performance sem introduzir novo viés. Isso sugere que o risco de uma intervenção causar danos à equidade é significativamente menor quando a representação dos grupos já é justa.

O Experimento 8 (Law School) serviu como um conto de advertência fundamental. A aplicação da técnica de subamostragem agressiva NearMiss em um cenário de desbalanceamento extremo foi catastrófica para a justiça, introduzindo vieses severos em modelos que eram originalmente justos. A hipótese é que a remoção massiva de dados da classe majoritária destruiu

informações contextuais vitais, forçando os modelos a se apoiarem em correlações espúrias e nos próprios atributos sensíveis para fazer predições. Este resultado prova que uma intervenção mal aplicada pode ser pior do que nenhuma intervenção.

4.2.1 Análise geral do desempenho dos algoritmos

Ao longo dos experimentos, observou-se um padrão consistente no desempenho relativo dos diferentes algoritmos, como é possível observar na tabela 14 sem intervenção e na tabela 15 após a aplicação das técnicas de intervenção.

- Modelos de Ensemble Baseados em Árvores (Random Forest, XGBoost, CatBoost, Gradient Boosting): Estes algoritmos consistentemente se posicionaram no topo do ranking de performance em termos de F1-Score e AUC-ROC. Sua robustez para lidar com dados tabulares e sua capacidade de capturar interações complexas os tornaram a escolha mais confiável na maioria dos cenários, tanto antes quanto depois das intervenções.

- Modelos Lineares (Logistic Regression, Ridge Classifier): Frequentemente subestimados, estes modelos provaram ser baselines fortes e estáveis. Embora raramente tenham alcançado o pico de performance dos modelos de ensemble, sua simplicidade e interpretabilidade os tornam valiosos. Em muitos casos, após o balanceamento dos dados, seu desempenho se tornou altamente competitivo.

- Modelos Baseados em Instância e Probabilidade (KNN, GaussianNB): Estes modelos exibiram a maior variabilidade. O KNeighbors mostrou-se sensível à "maldição da dimensionalidade"e à escala dos dados, enquanto o GaussianNB foi fortemente influenciado pela distribuição dos atributos. O GaussianNB, em particular, tendeu a se beneficiar imensamente das técnicas de sobreamostragem que ajudam a normalizar a distribuição das classes.

- Redes Neurais (MLPClassifier): O Perceptron de Múltiplas Camadas demonstrou um desempenho sólido e consistente, geralmente figurando entre os melhores modelos. No entanto, ele não superou de forma definitiva os melhores modelos de ensemble, sugerindo que, para os dados tabulares estruturados utilizados nesta pesquisa, o custo computacional e a complexidade de uma rede neural podem não se traduzir em um ganho de

performance que justifique sua preferência sobre algoritmos como XGBoost ou CatBoost sem uma otimização extensiva de hiperparâmetros.

Tabela 14 – Desempenho médio (F1-Score) dos 11 algoritmos de classificação nos diferentes experimentos, considerando o cenário *baseline* (sem intervenção).

Modelo	**Exp. 1**	**Exp. 2**	**Exp. 3**	**Exp. 4**	**Exp. 5**	**Exp. 6**	**Exp. 7**	**Exp. 8**
Random Forest	0.587	0.796	0.590	0.471	0.964	0.146	0.691	0.946
Logistic Regression	0.583	0.797	0.545	0.356	0.889	0.261	0.662	0.945
Gradient Boosting	0.591	0.777	0.610	0.473	0.944	0.282	0.698	0.945
XGBoost	0.586	0.776	0.624	0.459	0.970	0.296	0.736	0.941
CatBoost	0.592	0.810	0.621	0.479	0.980	0.279	0.722	0.943
Ridge Classifier	0.571	0.796	0.427	0.249	0.875	0.135	0.654	0.943
SGD Classifier	0.615	0.640	0.513	0.244	0.803	0.194	0.656	0.944
SVC	0.598	0.804	0.567	0.452	0.920	0.178	0.691	0.945
KNeighbors	0.562	0.782	0.570	0.446	0.858	0.305	0.640	0.937
GaussianNB	0.523	0.782	0.413	0.463	0.834	0.415	0.619	0.886
MLPClassifier	0.605	0.810	0.595	0.461	0.951	0.303	0.697	0.944

Tabela 15 – Desempenho médio (F1-Score) dos 11 algoritmos de classificação nos diferentes experimentos, após a aplicação das técnicas de intervenção.

Modelo	**Exp. 1**	**Exp. 2**	**Exp. 3**	**Exp. 4**	**Exp. 5**	**Exp. 6**	**Exp. 7**	**Exp. 8**
Random Forest	0.616	0.750	0.612	0.493	0.950	0.418	0.702	0.835
Logistic Regression	0.607	0.777	0.584	0.441	0.844	0.428	0.688	0.868
Gradient Boosting	0.628	0.758	0.619	0.512	0.938	0.422	0.709	0.863
XGBoost	0.613	0.749	0.593	0.490	0.941	0.389	0.727	0.897
CatBoost	0.625	0.761	0.618	0.492	0.944	0.413	0.723	0.902
Ridge Classifier	0.592	0.776	0.553	0.398	0.850	0.410	0.680	0.855
SGD Classifier	0.620	0.697	0.560	0.421	0.790	0.447	0.673	0.842
SVC	0.611	0.770	0.591	0.480	0.871	0.435	0.703	0.856
KNeighbors	0.559	0.724	0.553	0.455	0.836	0.382	0.636	0.806
GaussianNB	0.504	0.760	0.452	0.447	0.810	0.452	0.650	0.824
MLPClassifier	0.630	0.781	0.598	0.511	0.926	0.442	0.698	0.873

4.2.2 Análise geral das métricas

- Métricas de Performance: A escolha do F1-Score como métrica primária de performance foi validada repetidamente. Em cenários de desbalanceamento, a acurácia se mostrou uma métrica enganosa, enquanto o F1-Score, ao harmonizar precision e recall, ofereceu uma avaliação muito mais fiel da capacidade do modelo em identificar corretamente a classe minoritária, que é frequentemente a classe de maior interesse.

- Métricas de Justiça: O estudo demonstrou que alcançar uma paridade perfeita em todas as métricas de justiça é um objetivo irrealista e, por vezes, indesejável. A abordagem de aceitar um intervalo de tolerância (ex: [0.8, 1.25]) do DALEX é pragmática e eficaz. A métrica de STP emergiu como a mais difícil de ser satisfeita, indicando que as taxas de seleção geral entre grupos são altamente sensíveis a vieses de representação. O fato de que a maioria das intervenções de balanceamento conseguiu aproximar as razões de TPR, ACC e PPV de 1.0 reforça seu valor como ferramentas de mitigação de viés.

De modo geral, esta pesquisa abrangente fornece uma contribuição clara que a busca por modelos de aprendizado de máquina justos e precisos não reside na descoberta de um único algoritmo ou técnica superior, mas sim no desenvolvimento de um framework metodológico rigoroso de avaliação, intervenção contextual e verificação contínua. O trabalho aqui apresentado serve como um guia prático, demonstrando os potenciais benefícios, os trade-offs inerentes e os riscos latentes no uso de técnicas de balanceamento, capacitando cientistas a tomar decisões mais informadas e responsáveis na construção de modelos de aprendizado de máquina.

5 CONCLUSÕES

A crescente automação de decisões por meio de sistemas de aprendizado de máquina impõe à comunidade científica e à sociedade o desafio de garantir que essas tecnologias sejam não apenas eficientes, mas fundamentalmente justas. Esta dissertação abordou a interseção crítica de dois problemas centrais nesse contexto: o viés algorítmico, que ameaça a equidade, e o desbalanceamento de classes, que compromete a performance preditiva. O objetivo principal foi investigar empiricamente se, e como, as técnicas de balanceamento de dados, tradicionalmente focadas em performance, poderiam ser empregadas como ferramentas para a mitigação de vieses, e quais seriam os trade-offs inerentes a essa abordagem. Para tal, foi executado um estudo comparativo de larga escala, avaliando dez técnicas de intervenção sobre oito bases de dados distintas, com onze algoritmos de classificação, sob a ótica dupla da performance e da justiça.

A análise dos modelos gerados permitiu extrair conclusões robustas e importantes. O achado mais importante deste trabalho é a confirmação de que não existe uma técnica de balanceamento universalmente superior. A eficácia de cada abordagem mostrou-se altamente dependente do contexto da base de dados, incluindo o grau do desbalanceamento e, de forma crucial, a distribuição do atributo sensível. Em cenários de desbalanceamento severo, as técnicas de reamostragem promoveram um "ganho duplo", melhorando drasticamente tanto a performance quanto a justiça. Em contrapartida, em cenários onde o viés era um pouco mais baixo, o principal benefício das técnicas foi manter a equidade enquanto otimizavam a performance. A pesquisa também revelou trade-offs claros, onde a subamostragem melhorou a performance, mas falhou em mitigar o viés, e casos de resultados adversos, onde a técnica NearMiss degradou ativamente a justiça de modelos previamente equitativos.

A principal contribuição desta dissertação é, portanto, um mapeamento empírico detalhado dos efeitos e interações entre as técnicas de balanceamento de dados e a justiça algorítmica. O trabalho fornece um guia prático para cientistas, demonstrando que a otimização da performance não garante a equidade e que a seleção de uma técnica de pré-processamento deve ser uma decisão informada e contextual. Reforça-se a tese de que a auditoria de justiça, por meio de um framework de avaliação duplo como o aqui proposto, é um passo indispensável no ciclo de vida de desenvolvimento de modelos de aprendizado de máquina responsáveis.

As implicações práticas deste estudo são diretas: a mitigação de viés deve começar com

uma análise cuidadosa dos dados de origem. Em bases de dados com bom equilíbrio demográfico, o foco pode ser a otimização da performance com menor risco para a equidade. Em cenários de alto viés ou desbalanceamento, a escolha da técnica torna-se crítica e deve ser validada por múltiplas métricas. A pesquisa desmistifica a ideia de que o balanceamento é uma solução milagrosa, enquadrando-o como uma ferramenta poderosa, mas que exige uma aplicação consciente e crítica.

Apesar da abrangência, este trabalho possui limitações. A análise de justiça se concentrou em métricas de paridade de grupo, não explorando a justiça individual. Os hiperparâmetros dos modelos foram mantidos em seus valores padrão para garantir a comparabilidade, e o conjunto de técnicas, embora diverso, não é exaustivo. Essas limitações abrem caminhos para trabalhos futuros. Sugere-se a expansão desta metodologia para investigar o impacto do balanceamento em noções de justiça mais complexas, como a contrafactual e a interseccional. Além disso, a criação de novas técnicas híbridas, otimizadas para navegar o trade-off performance-justiça, e a aplicação deste framework a outros domínios, como processamento de linguagem natural e séries temporais, representam fronteiras de pesquisa promissoras para o avanço de uma inteligência artificial mais confiável e equitativa.

REFERÊNCIAS

ABEBE, R.; BAROCAS, S.; KLEINBERG, J.; LEVY, K.; RAGHAVAN, M.; ROBINSON, D. G. Roles for computing in social change. In: . [S.l.]: ACM, 2021.

AGGARWAL, U.; POPESCU, A.; HUDELOT, C. Active learning for imbalanced datasets. In: . [S.l.]: IEEE, 2020.

AL-ALAWI, A.; ALBUAINAIN, H. Machine learning in human resource analytics: Promotion classification using data balancing techniques. In: . [S.l.]: IEEE, 2024.

BAROCAS, S.; HARDT, M.; NARAYANAN, A. *Fairness and machine learning*. 2019. Disponível em: <https://fairmlbook.org/>. Acesso em: 22 Mai. 2025.

BAROCAS, S.; SELBST, A. D. Big data's disparate impact. *California Law Review*, v. 104, 2016.

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, ACM, v. 6, n. 1, p. 20–29, 2004.

BUOLAMWINI, J.; GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: . [S.l.]: PMLR, 2018.

CATON, S.; HAAS, C. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, 2002.

CHAWLA, N. V.; LAZAREVIC, A.; HALL, L. O.; BOWYER, K. W. SMOTEBoost: Improving prediction of the minority class in boosting. In: . [S.l.]: Springer-Verlag, 2003.

CHEN, I.; JOHANSSON, F. D.; SONTAG, D. Why is my classifier discriminatory? In: *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*. [S.l.]: Curran Associates, Inc., 2018. p. 3544–3555.

CHEN, J.; KALLUS, N.; MAO, X.; SVACHA, G.; UDELL, M. Fairness under unawareness: Assessing disparity when protected attributes are not observed. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. [S.l.]: ACM, 2019.

CHIAPPA, S. Path-specific counterfactual fairness. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.]: AAAI Press, 2019. v. 33, n. 01, p. 7801–7808.

CHOULDECHOVA, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, Mary Ann Liebert, Inc., publishers, v. 5, n. 2, p. 153–163, 2017.

CORBETT-DAVIES, S.; GOEL, S. The measure and mismeasure of fairness: A critical review of fair machine learning. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. [S.l.]: ACM, 2018. (AIES '18), p. 97–102.

CORBETT-DAVIES, S.; PIERSON, E.; FELLER, A.; GOEL, S.; HUQ, A. Algorithmic decision making and the cost of fairness. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.]: ACM, 2017. (KDD '17).

DWORK, C.; HARDT, M.; PITASSI, T.; REINGOLD, O.; ZEMEL, R. Fairness through awareness. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. [S.l.]: ACM, 2012. p. 214–226.

ELKAN, C. The foundations of cost-sensitive learning. In: *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*. [S.l.]: Morgan Kaufmann Publishers Inc., 2001. p. 973–978.

European Union Agency for Fundamental Rights. *Preventing unlawful profiling today and in the future: a guide*. Luxembourg, 2018.

FRIEDLER, S. A.; SCHEIDEGGER, C.; VENKATASUBRAMANIAN, S.; CHOUDHARY, S.; HAMILTON, E. P.; ROTH, D. A comparative study of fairness-enhancing interventions in machine learning. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. [S.l.]: ACM, 2019. p. 329–338.

GAJANE, P.; PECHENIZKIY, M. On formalizing fairness in prediction with machine learning. In: . [S.l.: s.n.], 2017.

GIBBS, S. *Google Photos apologises for labelling black people as 'gorillas'*. 2015. The Guardian. Disponível em: <https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app>. Acesso em: 9 jul. 2025.

GOODFELLOW, I. J.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative adversarial networks. In: *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. [S.l.]: Curran Associates, Inc., 2014. p. 2672–2680.

HAN, H.; WANG, W.-Y.; MAO, B.-H. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In: *Proceedings of the International Conference on Intelligent Computing (ICIC 2005)*. [S.l.]: Springer-Verlag, 2005. p. 878–887.

HARDT, M.; PRICE, E.; SREBRO, N. Equality of opportunity in supervised learning. In: *Advances in Neural Information Processing Systems 29 (NIPS 2016)*. [S.l.]: Curran Associates, Inc., 2016. p. 3315–3323.

HE, H.; BAI, Y.; GARCIA, E. A.; LI, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: . [S.l.]: IEEE, 2008.

HE, H.; GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, v. 21, n. 9, p. 1263–1284, 2009.

HEIDARPOUR, M.; LOUGHRAN, R.; MCDAID, K. Pre-processing techniques to mitigate against algorithmic bias. In: . [S.l.: s.n.], 2023.

HOOKER, S. Moving beyond "algorithmic bias is a data problem". In: . [S.l.: s.n.], 2021.

JAFARIGOLA, E.; TRAFALIS, T. B. A review of machine learning techniques in imbalanced data and future trends. In: . [S.l.]: IEEE, 2023.

JOHNSON, J. M.; KHOSHGOFTAAR, T. M. A survey on deep learning with imbalanced data. *Journal of Big Data*, Springer, v. 6, n. 1, p. 27, 2019.

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

KAMIRAN, F.; CALDERS, T. Data preprocessing techniques for classification without discrimination. In: . [S.l.]: Springer, 2012. v. 33, n. 1, p. 1–33.

KEARNS, M.; ROTH, A. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. New York: Oxford University Press, 2019.

KHAKUREL, U.; ABDELMOUMIN, G.; RAWAT, D. B. Performance evaluation for detecting and alleviating biases in predictive machine learning models. *ACM Transactions on Probabilistic Machine Learning*, ACM, v. 1, n. 2, p. 1–34, jun 2025.

KLEINBERG, J.; MULLAINATHAN, S.; RAGHAVAN, M. Inherent trade-offs in the fair determination of risk scores. In: *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. [S.l.: s.n.], 2016.

KOZODOI, N.; JACOB, J.; LESSMANN, S. Fairness in credit scoring: Assessment, implementation and profit implications. 2021.

KRAWCZYK, B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, Springer, v. 5, n. 4, p. 221–257, 2016.

LAMMERS, K.; VAQUET, V.; HAMMER, B. Continuous fair smote – fairness-aware stream learning from imbalanced data. In: . [S.l.: s.n.], 2025.

LIN, T.-Y.; GOYAL, P.; GIRSHICK, R.; HE, K.; DOLLáR, P. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2017.

MEHRABI, N.; MORSTATTER, F.; SAXENA, N.; LERMAN, K.; GALSTYAN, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, ACM, v. 54, n. 6, p. 1–35, 2021.

MI2.AI. *DALEX: Fairness module in dalex*. 2020. <https://dalex.drwhy.ai/python-dalex-fairness.html>. Developed by MI2.AI. License: CC-BY-SA-4.0. Accessed: 10 jul. 2025.

NAVARRO, M.; LITTLE, C.; ALLEN, G. I.; SEGARRA, S. Data augmentation via subgroup mixup for improving fairness. In: . [S.l.: s.n.], 2023.

O'NEIL, C. *Algoritmos de destruição em massa*. New York: Crown, 2016.

QUARESMINI, C.; PRIMIERO, G. Data quality dimensions for fair ai. Springer, 2023.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, Public Library of Science, v. 10, n. 3, p. e0118432, 2015.

SAMPATH, V.; MAURTUA, I.; AGUILAR-MARTIN, J. H.; RODRIGUEZ-AYERBE, P. A survey on generative adversarial networks for imbalance data. *Journal of Big Data*, Springer, v. 8, n. 1, p. 8, 2021.

SELBST, A. D.; BOYD, D.; FRIEDLER, S. A.; VENKATASUBRAMANIAN, S.; VERTESI, J. Fairness and abstraction in sociotechnical systems. *Communications of the ACM*, ACM, v. 62, n. 1, 2019.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data*, Springer, v. 6, n. 1, p. 60, 2019.

VERMA, S.; RUBIN, J. Fairness definitions explained. In: *Proceedings of the International Workshop on Software Fairness (FairWare '18)*. [S.l.]: ACM, 2018. p. 1–7.

WANG, T.; ZHAO, J.; YATSKAR, M.; CHANG, K.-W.; ORDONEZ, V. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2018.

ZADROZNY, B.; LANGFORD, J.; ABE, N. Cost-sensitive learning by cost-proportionate example weighting. In: *Proceedings of the Third IEEE International Conference on Data Mining (ICDM '03)*. [S.l.]: IEEE Computer Society, 2003. p. 435–442.

ZHANG, B. H.; LEMOINE, B.; MITCHELL, M. Mitigating unwanted biases with adversarial learning. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. [S.l.]: ACM, 2018. (AIES '18), p. 335–340.

ZHANG, H.; CISSE, M.; DAUPHIN, Y. N.; LOPEZ-PAZ, D. mixup: Beyond empirical risk minimization. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2017.

ZHAO, J.; WANG, T.; YATSKAR, M.; ORDONEZ, V.; CHANG, K.-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.]: Association for Computational Linguistics, 2017. p. 2979–2989.

Apêndices

APÊNDICE A – DESCRIÇÃO DETALHADA DAS BASES DE DADOS

Nesta seção, cada uma das oito bases de dados utilizadas nos experimentos é descrita em detalhe.

A.1 COMPAS

O conjunto de dados Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) é amplamente utilizado em pesquisas sobre a justiça algorítmica. Ele contém informações detalhadas de mais de dez mil réus criminais do condado de Broward, Flórida, Estados Unidos, e tem como objetivo principal prever a probabilidade de reincidência criminal em um período de dois anos. A tarefa principal desse conjunto de dados é uma classificação binária: determinar se um indivíduo será ou não reincidente. Os atributos disponíveis no conjunto de dados incluem características demográficas, como sexo, idade e etnia, além de informações sobre o histórico criminal e o status socioeconômico dos réus. Essas variáveis são utilizadas pelo algoritmo COMPAS para atribuir uma pontuação de risco que orienta decisões judiciais, como concessões de liberdade condicional ou penas alternativas. No entanto, estudos anteriores demonstraram que o sistema do COMPAS apresenta vieses significativos, embora grande parte da atenção tenha sido voltada para vieses raciais, também foi constatado que o algoritmo pode tratar homens e mulheres de maneira diferente, o que gera disparidades de gênero nas previsões.

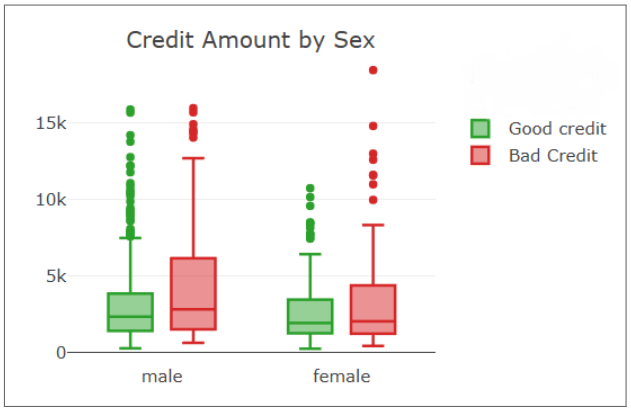
A.2 GERMAN CREDIT DATA

A base de dados German Credit é amplamente utilizado em estudos sobre aprendizado de máquina, especialmente em tarefas relacionadas à concessão de crédito. Ele contém informações de clientes de instituições financeiras na Alemanha e a tarefa é classificar cada solicitante como tendo um bom ou mau risco de crédito. Essa tarefa de classificação binária é essencial para a tomada de decisões no setor financeiro, tornando o conjunto de dados uma ferramenta valiosa para estudos sobre equidade em algoritmos de classificação.

Os atributos do conjunto de dados incluem variáveis como idade, sexo, ocupação, histórico financeiro, informações sobre contas bancárias e crédito. Esses fatores são usados para calcular

a probabilidade de inadimplência de cada cliente. No entanto, o German Credit é conhecido por conter desbalanceamentos em suas variáveis demográficas, o que pode resultar em vieses nos modelos de aprendizado de máquina. Em particular, o sexo é um atributo sensível que tem sido associado a disparidades nas decisões de crédito, com mulheres frequentemente enfrentando maiores taxas de recusa em comparação aos homens.

Figura 3 – Distribuição do atributo sensível sexo por risco de crédito



A.3 ADULT CENSUS INCOME

A base de dados Adult Income, também conhecida como "Census Income", é um dos conjuntos de dados mais canônicos e amplamente utilizados para tarefas de classificação e, especialmente, para estudos de justiça em aprendizado de máquina. Sendo um benchmark clássico para a avaliação de viés algorítmico. A tarefa de classificação consiste em prever se um indivíduo possui uma renda anual que excede 50.000 (classe positiva) ou se é igual ou inferior a este valor (classe negativa).

Sua utilização na literatura de equidade algorítmica tem relação com à presença de atributos socialmente sensíveis, como gênero e raça. Estes atributos permitem a avaliação de viés, investigando se o modelo tende a favorecer ou desfavorecer certos grupos demográficos, independentemente de suas outras qualificações.

A.4 DEFAULT OF CREDIT CARD CLIENTS

Default of Credit Card Clients trata-se de uma base de dados muito popular para a modelagem de risco de crédito. Este dataset contém informações de 30.000 clientes de um grande banco em Taiwan, coletadas em 2005, e serve como um cenário financeiro realista para a aná-

lise de inadimplência. A tarefa de classificação binária consiste em prever se um cliente irá ou não cumprir com o pagamento de sua fatura no mês subsequente. O conjunto de dados possui informações sobre o limite de crédito, dados demográficos como gênero, nível de educação e estado civil, e um rico histórico de status de pagamento e valores de faturas dos seis meses anteriores.

A.5 HEART DISEASE

O quinto experimento aborda a área da saúde com o clássico dataset Heart Disease, um dos mais tradicionais para tarefas de diagnóstico preditivo, originário do repositório de aprendizado de máquina. Este conjunto de dados é, na verdade, uma compilação de quatro bases distintas (Cleveland, Hungria, Suíça e Long Beach), com dados coletados a partir de 1988. Para garantir a consistência com a maioria dos trabalhos publicados na literatura, este estudo utiliza a versão mais processada e conhecida, a "Cleveland". A tarefa consiste em uma classificação binária para prever a presença (valor 1) ou ausência (valor 0) de doença cardíaca em um paciente, com base em 13 atributos preditores que incluem informações demográficas como idade e sexo, bem como dados clínicos importantes como tipo de dor no peito, nível de colesterol e resultados de eletrocardiograma.

A.6 DIABETES HEALTH INDICATORS

Diabetes Health Indicators é uma base de dados de grande escala derivado do Behavioral Risk Factor Surveillance System, uma pesquisa anual de saúde conduzida pelo centro de controle e prevenção de doenças dos Estados Unidos. A tarefa de classificação consiste em prever o status de diabetes de um indivíduo com base em uma ampla gama de indicadores de saúde e respostas sobre estilo de vida. A variável alvo é binária, onde o valor 1 representa a presença de diabetes ou pré-diabetes, e 0 indica um indivíduo saudável. Por compor mais de 250.000 linhas, neste estudo foi utilizado 10 porcentos da base total.

A.7 CREDIT CARD APPROVAL

O sétimo experimento utiliza a base de dados Credit Card Approval disponibilizada no site público Kaggle, um conjuntos de dados para problemas de classificação de crédito, que

consiste em prever se um candidato é um cliente "bom"ou "ruim". As pontuações de crédito são um método comum de controle de risco no setor financeiro. Eles utilizam informações pessoais e dados enviados pelos solicitantes de cartão de crédito para prever a probabilidade de inadimplências e empréstimos futuros. O banco pode decidir se deve ou não emitir um cartão de crédito para o solicitante. As pontuações de crédito podem quantificar objetivamente a magnitude do risco.

A.8 LAW SCHOOL (LSAC)

A base de dados do Law School Admission Council é proveniente do National Longitudinal Bar Passage Study. Este é um dos conjuntos de dados importante por ser socialmente relevante para o estudo da justiça algorítmica, pois lida com o acesso à profissão jurídica nos Estados Unidos. A tarefa de classificação consiste em prever se um estudante de direito será aprovado no exame da ordem (bar exam) em sua primeira tentativa. A importância desta base de dados paro o estudo em justiça está relacionado aos atributos sensíveis de gênero, o que permite uma análise aprofundada de vieses interseccionais na aprovação dos estudantes.