UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS DE COMPUTAÇÃO
COORDENAÇÃO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

RODOLFO BOLCONTE DONATO

ANÁLISE DE TRANSFORMAÇÕES METAMÓRFICAS EM CONJUNTO DE DADOS PARA GARANTIA DE *FAIRNESS* EM MODELOS DE CLASSIFICAÇÃO DE INFORMAÇÕES

Campina Grande, Paraíba, Brasil
2024

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Análise de Transformações Metamórficas em Conjunto de Dados para Garantia de *Fairness* em Modelos de Classificação de Informações

Rodolfo Bolconte Donato

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Aprendizagem de Máquina

Patrícia Duarte de Lima Machado
(Orientadora)

Campina Grande, Paraíba, Brasil

D677a Donato, Rodolfo Bolconte.
Análise de transformações metamórficas em conjunto de dados para garantia de Fairness em modelos de classificação de informações / Rodolfo Bolconte Donato – Campina Grande, 2024.
139 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2024.
"Orientação: Profa. Dra. Patrícia Duarte de Lima Machado."
Referências.

1. Ciência da Computação. 2. Aprendizagem de Máquina. 3. Classificação de Informações. 4. Fairness. 5. Transformações. I. Machado, Patrícia Duarte de Lima. II. Título.

CDU 004(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA ITAPUANA SOARES DIAS GONÇALVES CRB-15/093

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO EM CIENCIA DA COMPUTACAO
Rua Aprígio Veloso, 882, Edifício Telmo Silva de Araújo, Bloco CG1, - Bairro Universitário, Campina Grande/PB, CEP 58429-900
Telefone: 2101-1122 - (83) 2101-1123 - (83) 2101-1124
Site: http://computacao.ufcg.edu.br - E-mail: secretaria-copin@computacao.ufcg.edu.br / copin@copin.ufcg.edu.br

FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

RODOLFO BOLCONTE DONATO

APLICAÇÃO DE TRANSFORMAÇÕES METAMÓRFICAS EM CONJUNTO DE DADOS PARA GARANTIA DE FAIRNESS EM MODELOS DE CLASSIFICAÇÃO DE DADOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Mestre em Ciência da Computação.

Aprovada em: 08/02/2024

Profa. Dra. PATRICIA DUARTE DE LIMA MACHADO, UFCG, Orientadora

Prof. Dr. EVERTON LEANDRO GALDINO ALVES, UFCG, Examinador Interno

Profa. Dra. NATASHA CORREIA QUEIROZ LINO, UFPB, Examinadora Externa

---

Documento assinado eletronicamente por **EVERTON LEANDRO GALDINO ALVES**, **PROFESSOR 3 GRAU**, em 15/02/2024, às 09:55, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **PATRICIA DUARTE DE LIMA MACHADO**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 19/02/2024, às 10:11, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Natasha Correia Queiroz Lino**, **Usuário Externo**, em 21/02/2024, às 22:37, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Resumo

O viés de classificação é um problema recorrente em sistemas de aprendizagem, sendo causado também pela presença de preconceitos e injustiças do mundo real inseridas nos dados digitais. O estudo de tal assunto se concentra na área intitulada *Fairness*, que embora não tenha uma definição unificada pela literatura, representa a garantia de que decisões tomadas por sistemas sejam realizadas de forma imparcial, evitando a presença de preconceitos e discriminações contra minorias de grupos sociais, sejam por gênero, raça, poder aquisitivo, etc. Porém, tais sistemas são muitas vezes intitulados como "não testáveis", visto que não está claro quais são os resultados corretos para as classificações dos dados e, com isso, surge a discussão do quão os modelos podem generalizar seus resultados. Visando à imparcialidade a partir da generalização dos modelos, surgem as técnicas de Transformações Metamórficas, sendo alterações realizadas nos conjuntos de dados, tanto em suas estruturas como também em seus valores, empregadas no presente trabalho com o propósito de que as classificações dos modelos executados com os dados alterados possam gerar os resultados mais imparciais possíveis independentemente dos grupos sociais presentes nos dados. As Transformações Metamórficas empregadas são idealizadas e executadas em quatro cenários distintos, sejam alterando informações em partes de valores de cada atributo do conjunto de dados utilizado como também a alteração de todas as informações de um único grupo de amostras, como informações de pessoas apenas do sexo masculino ou feminino. É com este contexto que o presente trabalho se desenvolve, através da análise de diferentes aplicações e combinações de transformações metamórficas em dados de amostras que tem o gênero feminino como grupo suscetível ao preconceito por parte dos modelos de classificação, com seus valores postos em evidência através do cálculo de métricas voltadas para as áreas de *Fairness*,

fazendo a análise seja somente com os valores classificados como também levando em consideração os valores reais, além do cálculo de métricas consolidadas em tarefas de Classificação de informações. Em termos de resultados foi possível alcançar valores que indicam melhorias de até 20% quando utilizando instâncias de modelos treinadas com os dados transformados. Ao analisar as transformações em diferentes modelos de aprendizagem foi possível embasar a discussão sobre se realmente é possível melhorar os índices de *Fairness* nas classificações, com alguns cenários se mostrando promissores em responder esta questão e também no engrandecimento da literatura sobre o assunto, que até então tem se mostrado escassa.

Abstract

Classification bias is a recurring problem in learning systems and is also caused by the presence of real-world prejudices and injustices embedded in digital data. The study of this subject focuses on the area known as Fairness, which, although it does not have a unified definition in the literature, represents the assurance that decisions made by systems are carried out impartially, avoiding the presence of prejudice and discrimination against minorities of social groups, whether by gender, race, purchasing power, etc. However, such systems are often labeled as 'untestable', since it is not clear what the correct results for data classifications are, and this raises the discussion about how models can generalize their results. Aiming for impartiality from the generalization of the models, Metamorphic Transformations techniques emerge, with changes made to the datasets, both in their structures and also in their values. These techniques are used in this work with the aim that the classifications of the models executed with the altered data can generate the most impartial results possible, regardless of the social groups present in the data. The Metamorphic Transformations employed are conceptualized and executed in four different scenarios, either by changing information in parts of the values of each attribute of the dataset used or by changing all information from a single group of samples, such as information about people of just one sex, male or female. It is within this context that this work is developed, through the analysis of different applications and combinations of metamorphic transformations in sample data that have the female gender as a group susceptible to prejudice on the part of classification models. Their values are highlighted through the calculation of metrics focused on the areas of Fairness, carrying out the analysis not only with the classified values but also taking into account the real values, in addition to the calculation of consolidated metrics

in Information Classification tasks. In terms of results, it was possible to achieve values that indicate improvements of up to 20% when using model instances trained with the transformed data. By analyzing the transformations in different learning models, it was possible to support the discussion on whether it is really possible to improve Fairness indices in classifications, with some scenarios showing promise in answering this question and also in expanding the literature on the subject, which until then has proven to be scarce.

"Don't hang your head in sorrow

And, please, don't cry"

Don't Cry - Guns N' Roses

Agradecimentos

O Primeiro é direcionado aos professores Everton Galdino, Herman Martins e Natasha Queiroz, que como membros das Bancas Avaliadoras da Qualificação e da Defesa Final puderam me ajudar na construção do trabalho, repassando suas experiências e expertises sempre priorizando o engrandecimento do estudo, muito aprendi com eles e levarei seus ensinamentos comigo na vida.

Ao LaCInA dedico o Segundo agradecimento, que foi uma verdadeira casa para mim durante o meu período de mestrado, principalmente pelo professor Cláudio Campelo tendo me acolhido no laboratório ainda no início do curso, com quem pude aprender muito tanto profissionalmente quanto pessoalmente, assim como com os professores Carlos Eduardo e Ricardo Oliveira também. Além dos professores, principal e indispensavelmente sou grato aos meus colegas diários do laboratório, com quem construí fortes vínculos e me ajudaram a amenizar as tensões e pressões acadêmicas, obrigado Bryan Khelven, David Eduardo, Diego Ribeiro, Eleonilia Rodrigues, Francicláudio Dantas, Gabriel Medeiros, Igor Pereira, João Antônio, Lucas Ribeiro, Luis Thiago, Matheus Maciel, Rafael Guerra e Tiago Brasileiro, destes especialmente ao grupo CSI - Insalubre, a melhor equipe de desenvolvimento que já integrei até então, as pausas pro café falam por si!

O Terceiro agradecimento é especialmente para o grupo dos Mascotes de Telemática, amigos que fiz ainda na graduação e que continuam comigo até então. Apesar da distância, sempre se fizeram presente a mim quando possível, ajudando no que fosse necessário e compartilhando momentos únicos da vida. Obrigado Maxsuel Medeiros, Miqueas Galdino e Nathalya Leite, que possamos ter muitos outros bons momentos em vida.

O Quarto, eu direciono para os amigos que me ajudaram mesmo que indiretamente a superar as dificuldades acadêmicas, sem eles eu não conseguiria ter uma boa saúde mental no desenvolvimento do estudo e muito menos finalizar ele. Sou grato a Caio Luna, Daniel Maribond, Deborah Hemmely, Enrique Eliardo, Esther Sousa, Fernando Ribeiro, João Victor Barros, Laís Vitória, Rubem Ribeiro, Samya Amado, Sáskya Lima e Vanessa Diniz, independente do vínculo que temos ou não neste momento, foram pessoas especiais para mim.

O Quinto vai para todos aqueles que não foram citados, mas que de alguma forma me ajudaram no decorrer do estudo, sou grato por todo o apoio e pelos momentos que vivemos.

O Sexto e segundo agradecimento mais importante vai para as duas pessoas mais importantes na minha vida, meus pais, Vilma Bolconte e Roberto Donato, que sempre me deram forças para continuar no que chamamos de vida, me ensinando, me acolhendo, me repreendendo, me escutando, me mostrando, me várias coisas sempre com o propósito de que eu seja a melhor pessoa possível tanto para mim quanto para os outros ao meu redor. Hoje eu sou o que sou por eles e sou eternamente grato por isso. Junto dos meus pais quero citar meu filhote Dante, Dantinho, Dantão, Dantchovski, o Gato, que me ajudou em muitas crises de ansiedade de madrugada com sua fofura. Ainda neste agradecimento se faz presente a minha segunda família, meus tios Vailson Bolconte e Gesilene Lima, e meus primos Arthur, Júlia e Raquel, eles que no começo da minha trajetória em pós-graduação receberam a mim e meu filhote como membros da família, e desde então me incentivaram e me divertiram sempre.

O Sétimo, último e mais importante agradecimento é destinado a minha orientadora Patricia Machado, que sempre quando alguém pergunta como é minha orientadora, eu costumo dizer que ela é uma fofa. Devido um problema pessoal que tive no último ano do mestrado, me encontrei bastante desmotivado na tentativa de concluir o estudo, foi uma verdadeira guinada pra mim e em vários momentos achei que não fosse conseguir, porém Patricia se mostrou uma digna orientadora, com sua calma e empatia entendeu meu problema e permaneceu comigo até que fosse possível eu finalizar o trabalho, se doando ao máximo na melhora do mesmo, até em pleno réveillon lendo e enviando correções, jamais esquecerei o seu empenho. Peço desculpas a ela por todo esse tempo

que levei e eventuais aperreios. Serei eternamente grato por toda a sua ajuda durante esses três anos e espero que o Oklahoma City Thunder possa lhe dar muitas alegrias nos campeonatos de basquete!

Por fim, dedico a realização deste trabalho ao meu avô, Raimundo Bolconte, grande mestre de obras vindo do sertão da Paraíba, que inclusive foi responsável por obras na própria Universidade Federal de Campina Grande. Uma das pessoas mais calmas e gentis que conheci até então, sempre me apoiou nos estudos e sempre perguntou como eu estava neles, obrigado vô, eu consegui, e o senhor também!

Conteúdo

Lista de Símbolos

AI	*Artificial Intelligence*
AMD	*Advanced Micro Devices*
COMPAS	*Correctional Offender Management Profiling for Alternative Sanctions*
CSI	*Crime Scene Investigation*
CSV	*Comma Separated Value*
FN	Falso Negativo
FP	Falso Positivo
FPR	*False Positive Rate*
GB	*Gigabyte*
KNN	*K-Nearest Neighbors*
LaCInA	Laboratório de Computação Inteligente Aplicada
NB	*Naive Bayes*
QP	Questão de Pesquisa
SVM	*Support Vector Machine*
TPR	*True Positive Rate*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

Lista de Figuras

Lista de Tabelas

1 Introdução

A Aprendizagem de Máquina é uma área de pesquisa da Inteligência Artificial que visa o desenvolvimento de sistemas computacionais com a capacidade de aprender a realizar diferentes atividades com base na extração de conhecimento a partir de determinados dados. Isto quer dizer que os sistemas de aprendizagem são capazes de aprender por si só ao utilizar informações que possam representar experiências passadas. Além da Inteligência Artificial, a Aprendizagem acaba englobando também probabilidade e estatística, teoria da complexidade computacional e da informação, filosofia, psicologia, entre outras, variando de acordo com o cenário em que ela é trabalhada também (CERRI; CARVALHO, 2017).

A aprendizagem pode ser aplicada em diversas áreas para diferentes finalidades. É possível citar a sua utilização em diversos campos, como: 1) visão computacional, como para os veículos autônomos sendo possível compreender e reagir aos ambientes em que se encontram de forma semelhante à capacidade humana, com a possibilidade de mapear e analisar inúmeros objetos ao seu redor (GHIRARDELLO, 2023); 2) prevenção de falhas, em que é possível prever a ocorrência de certas falhas em sistemas a partir de determinados sinais que precedem a ocorrência destas, como alterações em métricas de monitoramento; 3) detecção de fraude, com a compreensão de comportamentos fora de padrão em acessos de contas, por exemplo, ou ainda transações financeiras atípicas; 4) análise de currículo, ao verificar de forma automática as informações de candidatos a vagas para definir aqueles de maior potencial; 5) ofertas de produtos, com o reconhecimento de padrões de compras de usuários possibilitando o anúncio de produtos de forma direcionada em que novas vendas sejam garantidas (SEROKELL,

2020).

Outro forte exemplo que tem-se atualmente relacionado a aprendizagem de máquina é o ChatGPT, uma ferramenta de processamento de linguagem natural desenvolvida pela empresa OpenAI e utilizado para diversas finalidades, tais como *chat bots*, geração automática de conteúdo, tradução automática de textos, entre outras, se apresentando como uma das ferramentas de Inteligência Artificial mais avançadas e disponíveis no mercado atual, aprimorada constantemente pelos seus desenvolvedores com relação a sua precisão e também a capacidade de compreensão em diversos idiomas (ROSSONI; CHAT, 2022). A ferramenta se mostrou revolucionária no âmbito tecnológico, pressionando até as famosas *Big Techs*, como Google, Meta e Microsoft a correrem contra o tempo no desenvolvimento e/ou integração de suas próprias tecnologias semelhante ao que o ChatGPT proporciona (METZ et al., 2023).

Tendo cada vez mais destaque e com uma crescente utilização seja no âmbito da pesquisa acadêmica como também no profissional, a aprendizagem se tornou, em partes, um fator chave para o desenvolvimento tecnológico por proporcionar a classificação de dados, esta que se orienta como um método de aprendizagem capaz de atribuir classes a um conjunto de amostras com base em suas características. Para isto, geralmente é necessário um conjunto de dados para a realização de uma atividade de treino, em que as informações passadas já possuem uma atribuição de classe, a fim de que modelos de classificação de informações possam aprender padrões e definir classes para informações futuras (JULIAN, 2016).

Além das áreas já citadas, a classificação automática de dados através de modelos de aprendizagem também é bastante disseminada nos seguintes domínios: 1) detecção de *spam*, em que algoritmos são essenciais na filtragem de mensagens indesejadas que contém certas características como padrões textuais e links suspeitos, variando desde e-mails até comentários em sites e redes sociais; 2) previsão de *churn*, referente a perda de clientes por parte das empresas a partir da percepção de dados e comportamentos de compras dos próprios clientes, para que medidas de retenção possam ser tomadas; e 3) análise de sentimento, permitindo assimilar o estado emocional de indivíduos em textos ou falas, sendo útil para análises de opiniões e também de *feedbacks* impactando na melhora de produtos/serviços por parte das empresas (CRUZ, 2023).

Porém, com o uso da classificação em grandes quantidades de dados de diversas áreas, é perceptível a presença de erros nas execuções, dentre eles o viés de classificação que está ligado a generalização de um algoritmo, ou seja, o quão bem ele é capaz de prever novos casos. O viés pode ser causado também através da presença de preconceitos e injustiças do mundo real, que como forma de conscientização do problema é possível a elaboração de certos questionamentos durante o desenvolvimento de sistemas, sendo eles: como a base de dados é construída? os dados são tendenciosos? como fazer uma previsão justa? o que poderia ser aceito como justo? (CLEGER; PESSOA; LIMA, 2019). Tais questionamentos com relação à preconceitos reais sendo adaptados em sistemas inteligentes se concentram na área de estudo intitulada *Fairness*.

1.1 Formulação do Problema

Apesar das vantagens ao utilizar a Aprendizagem de Máquina para prever ou classificar objetos, como a análise de grandes volumes de dados e a realização de processos automáticos de forma mais rápida em relação à atividades manuais, é importante salientar os problemas recorrentes nesta área, que podem, em certos casos, induzir os desenvolvedores a tomar decisões equivocadas com base nos resultados obtidos. Dentre os problemas que podem ocorrer, se destacam:

1. *Overfiting*: a possibilidade de que cálculos dos modelos sejam específicos para o conjunto de treinamento utilizado, ou seja, como os dados de treino são apenas uma amostra do domínio trabalhado, ao induzir hipóteses que melhorem o desempenho dos modelos para o conjunto de treinamento, o desempenho para dados além dos presentes no treinamento se mostra pior, evidenciando um ajuste somente para determinados dados (REZENDE et al., 2003);

2. Prevalência de Classe: relacionado ao desbalanceamento de classes em um conjunto de dados, quando a quantidade de objetos de uma ou mais classes se mostra bem maior em relação as outras, por exemplo um conjunto que possui 90 objetos da classe 1 e 5 objetos das classes 2 e 3, não assimilando bem quando as classes minoritárias possuem uma informação que seja importante para o cenário

trabalhado (REZENDE et al., 2003).

Além dos problemas citados acima, quando decisões são tomadas sobre indivíduos com base nos resultados obtidos pelos modelos, inevitavelmente surgem novos tipos de preocupações que vem ganhando cada vez mais destaque nas pesquisas de Aprendizagem, sendo a discriminação e/ou a parcialidade. E se modelos influenciarem decisões sistematicamente tendenciosas contra pessoas pertencentes a grupos minoritários da sociedade de acordo com sua raça, gênero ou religião? Através deste problema, um paradigma da computação tem ganhado cada vez mais notoriedade na literatura: *Fairness* (BINNS, 2018).

O estudo de *Fairness*, embora não tenha uma definição única, está relacionado na garantia de que decisões tomadas por um sistema de aprendizagem sejam realizadas da maneira certa e também pela razão certa, evitando a presença de preconceitos e discriminações que violem leis de acordo com os direitos humanos (ZHANG et al., 2019). A fim de contextualizar os preconceitos encontrados na área de *fairness* é possível citar: 1) uma pesquisa realizada na The University of Virginia demonstrando como algoritmos treinados com bases de dados de fotos reproduziam preconceitos ao classificarem fotos de homens como se fossem mulheres quando eles estavam na cozinha (VIEIRA, 2019); 2) a *Amazon*, empresa tecnológica multinacional, descontinuou o uso de uma ferramenta de recrutamento que estava penalizando candidatas mulheres dando preferência à candidatos homens (DASTIN, 2018); 3) uma função de reconhecimento de imagens do Google Photos que classificou os rostos de dois amigos negros como gorilas (HOWLEY, 2015); e 4) o sistema de cálculo de probabilidade de reincidência criminal dos Estados Unidos, *COMPAS – Correctional Offender Management Profiling for Alternative Sanctions*, prevendo de forma errônea que uma pessoa negra cometeria reincidência de crime a uma taxa de 44,9%, esta duas vezes maior do que a taxa para pessoas brancas, cerca de 23,5% (DRESSEL; FARID, 2018).

A partir dos problemas relacionados ao viés de classificação evidenciando preconceitos do mundo real presentes em sistemas de aprendizagem, é notória a necessidade de um monitoramento de tais sistemas a fim de não somente avaliar, mas também buscar melhorar os índices de *fairness* nas classificações. Uma tentativa de solução para tal problema é a utilização de Transformações Metamórficas em conjuntos de dados,

estas sendo alterações nos próprios valores do conjunto, sejam através de operações matemáticas como a multiplicação por constantes numéricas como também a remoção de informações do conjunto, para que ocorra uma maior generalização nas classificações dos sistemas de aprendizagem, a fim de torná-las as mais imparciais possíveis com relação aos atributos sensíveis.

Para que seja possível o combate ao viés de classificação, é importante a realização de testes na fase de aprendizagem de um modelo para avaliação de desempenho do mesmo, que é exatamente onde o presente trabalho se concentra, através do estudo de diferentes execuções de transformações metamórficas em diferentes modelos de aprendizagem a fim de definir os cenários que melhor garantem classificações justas, independente de características sensíveis de indivíduos ou grupos.

1.2 Objetivos

1.2.1 Objetivo Geral

Com o contexto melhor apresentado nas seções anteriores, o objetivo geral deste trabalho se faz presente, o de analisar o impacto de diferentes transformações metamórficas de dados sobre classificações realizadas por modelos de aprendizagem, visando o aumento de índices de *fairness* e, consequentemente, a diminuição da parcialidade e do preconceito nos resultados em relação a indivíduos e grupos de classes minoritárias na sociedade.

1.2.2 Objetivos Específicos

Expandindo o objetivo geral do trabalho, tem-se os seguintes objetivos específicos:

- Consolidar a utilização de transformações metamórficas na área de classificação de informações com foco em *fairness*;
- Estudar e identificar transformações metamórficas que possam ser capazes de causar impacto em classificações de modelos;
- Identificar métricas da literatura que compreendam de fato o problema da parcialidade contra informações de indivíduos e grupos na sociedade propícios a

preconceito;

- Verificar o comportamento de transformações metamórficas na fase de treino de modelos de diferentes tipos de aprendizagem;
- Analisar de forma exaustiva o impacto de transformações metamórficas de dados nas classificações de modelos de aprendizagem;
- Definir os métodos de transformações metamórficas que apresentam resultados favoráveis na diminuição da parcialidade nas classificações;
- Indicar panoramas para a utilização de transformações metamórficas para melhores resultados em ambientes finais de desenvolvimento.

1.3 Justificativa e Relevância

Preocupações com o comportamento ético de sistemas de aprendizagem tem motivado discussões que são cada vez mais presentes em entidades governamentais, no âmbito acadêmico e também no industrial (CESARO, 2021). Entidades governamentais a partir de diversos países tem incentivado discussões relacionadas aos impactos éticos da Inteligência Artificial na sociedade, tendo certo destaque a atuação do parlamento Europeu com a idealização de um guia de regras éticas recomendadas para qualquer concepção, desenvolvimento, implementação ou utilização de produtos e serviços de Inteligência Artificial na União Europeia (MADIEGA, 2019).

No setor empresarial, casos de modelos com viés, como alguns citados anteriormente, vem causando certa aversão por parte do público, fazendo com que empresas se preocupem mais com os possíveis impactos éticos nos seus respectivos projetos que lidam com Aprendizagem de Máquina. Corroborando tais preocupações, gigantes tecnológicas tem disponibilizado ferramentas de acesso livre ao público para auxiliar no desenvolvimento de projetos de aprendizagem de forma ética. Um exemplo de tais ferramentas pode ser a "What-If Tool", da Google, que permite gerar gráficos diversos para interpretação de variáveis de modelos e avaliação de viés (CESARO, 2021; WEXLER et al., 2019).

É notório na literatura o estudo de *fairness* em contextos diferentes, a fim de analisar o quão imparciais são os sistemas de aprendizagem e, da mesma forma, a

utilização de testes e transformações metamórficas principalmente como forma de análise de generalização de modelos, porém estudos utilizando transformações metamórficas voltadas para o contexto de *fairness* sejam apenas para análise como também para a mitigação de preconceitos, são escassos, embora a aplicação de tais técnicas se mostrem grandes aliadas no monitoramento de *fairness* nos sistemas. Sendo assim, mostra-se a necessidade de uma maior quantidade de estudos que conciliem as duas áreas a fim de definir que os sistemas de aprendizagem sejam cada vez mais imparciais, evitando preconceitos do mundo real.

1.4 Contribuições

Concentrando-se no problema da ausência de *fairness* contra indivíduos ou grupos suscetíveis a discriminação por parte de modelos de classificação de dados, foi possível a realização de um estudo exaustivo utilizando transformações metamórficas de dados a fim de analisar o impacto destas na melhora dos índices de *fairness* nas classificações. Como resultado, espera-se as seguintes contribuições para este trabalho:

- A criação de um conjunto adequado de transformações metamórficas de dados para a melhoria de resultados;
- Uma análise consistente do comportamento de modelos de classificação quando treinados com dados originais e transformados;
- A indicação do custo/benefício da utilização ou não de transformações metamórficas de dados em atividades de classificação;
- Contribuição à literatura referente a utilização de transformações metamórficas na melhoria de índices de *fairness*.

1.5 Organização do Trabalho

A estrutura da presente dissertação se encontra da seguinte forma: no Capítulo 2 é apresentado todo o referencial teórico necessário para compreender o trabalho, estruturando a atividade de Classificação de Dados através de modelos de aprendizagem de

informações, a área de *Fairness* voltada para a diminuição da parcialidade e preconceito na classificação de informações e as Transformações Metamórficas de valores em conjuntos de dados, como a possível solução no problema exposto. O Capítulo 3 fornece uma visão geral do que a literatura atual dispõe sobre as áreas de Transformações Metamórficas e *Fairness* em Classificações, expondo também diferenças entre as pesquisas já realizadas em relação ao presente trabalho. No Capítulo 4 todo o estudo é descrito, elucidando as Questões de Pesquisas que guiam o trabalho e também apresentando os componentes utilizados, como os dados, modelos de classificação e transformações metamórficas, além do método de comparação para validar as execuções. O Capítulo 5 trás os resultados mais importantes para cada um dos cenários de aplicação e análise das transformações metamórficas. Por fim, o Capítulo 6 apresenta as conclusões obtidas e limitações percebidas durante a execução do estudo, além de pensamentos futuros para o engrandecimento do trabalho. O trabalho conta ainda com dois Apêndices: o A com uma breve análise exploratória dos dados utilizados e o B mostrando todos os valores obtidos em cada uma das execuções realizadas.

2 Fundamentação Teórica

Neste capítulo são apresentados tópicos necessários para a compreensão da presente pesquisa. A Seção 2.1 traz uma discussão sobre o método de Classificação de Dados, expondo diferentes técnicas de como realizar tal tarefa e métodos para mensuração dos resultados obtidos, a fim de saber o quão bons eles de fato podem ser. Na Seção 2.2 a área de *Fairness* é então apresentada, passando pelos seus conceitos, exemplos de aplicações e, assim como o processo de classificação, as formas de mensuração de *fairness* nos valores previstos. Encerrando os tópicos da fundamentação, a Seção 2.3 descreve as Transformações Metamórficas, mostrando como são aplicadas nos conjuntos de dados e a diferença da utilização destas na presente pesquisa em relação aos tradicionais Testes Metamórficos.

2.1 Classificação de Dados

Cientistas de dados geralmente encaram problemas que para suas soluções são necessárias técnicas de decisões automatizadas. Um *e-mail* recebido é uma mensagem séria ou uma tentativa de *pishing*? Tal problema pode ser resolvido através da Classificação de Dados, esta sendo uma tarefa da aprendizagem de máquina que consiste em associar um objeto a uma classe pré-estabelecida: se o objeto é um 0 (não *pishing*) ou um 1 (*pishing*), ou, em alguns casos, estabelecer uma classe dentro de muitas outras categorias, por exemplo, a filtragem da caixa de entrada do Gmail, que conta com as categorias “Principal”, “Social”, “Atualizações”, “Fóruns” e “Promoções” (BRUCE; BRUCE, 2019).

Citada no Capítulo anterior, a Aprendizagem de Máquina trata-se de uma extração

de conhecimento a partir de determinados dados, sendo um campo de pesquisa que une a Estatística, a Inteligência Artificial e a Ciência da Computação (MÜLLER; GUIDO, 2016). Com ela é possível então apresentar características sobre um determinado conjunto de dados sem que seja necessário produzir um código personalizado para o problema, apenas com uma implementação inicial de um modelo genérico que seja capaz de construir sua própria lógica de funcionamento, a partir da obtenção de determinados dados (GEITGEY, 2014).

Para exemplificar o funcionamento da Aprendizagem de Máquina, pode ser citado um problema de definição de preço de um imóvel. Uma das possíveis soluções para o problema é contar com a ajuda de um corretor de imóveis, que utiliza da sua experiência com outros imóveis já vendidos para estimar um valor justo ao objeto. Tal experiência do corretor é possível através da análise de algumas informações que o objeto dispõe, como: tipo de imóvel, área total, localização, cômodos, entre outros. Porém, para o problema exemplificado, não é possível solicitar a ajuda de um corretor, tendo à disposição apenas as informações de imóveis que já foram vendidos. Neste caso, é possível encontrar padrões dos dados dispostos ou então definir quais das informações são as mais relevantes para a obtenção do preço do imóvel com base nos anteriores. É aí que a aprendizagem se sobressai, com a capacidade de encontrar padrões que definem o preço mais real possível de forma automática, através do treinamento de modelos com informações de imóveis já vendidos.

Visando a resolução de um determinado problema utilizando a aprendizagem de máquina, é necessário ponderar sobre que tipo de dado é desejado como resultado, para que seja possível então saber qual das categorias de aprendizagem pode vir a resolver o problema de forma eficiente e eficaz. Na literatura é comum a divisão da Aprendizagem de Máquina em quatro tipos, que embora tenham como base inferir informações sobre determinados objetos, se distinguem na forma como exercem tal atividade.

Sendo uma das mais utilizadas e bem-sucedidas, a Aprendizagem Supervisionada é utilizada sempre que seja necessário prever um determinado resultado para uma determinada entrada quando há exemplos de pares entrada-saída. Seu funcionamento se dá através do recebimento de um conjunto de dados rotulados para que o modelo possa aprender o que exatamente é cada objeto passado. É possível resolver problemas

de dois tipos: 1) de Regressão, em que o objetivo é prever um valor contínuo, como o rendimento de uma fazenda, a partir de dados anteriores como clima, número de funcionários, animais, etc; e 2) de Classificação, para prever um rótulo de classe dentro de um intervalo predefinido (MÜLLER; GUIDO, 2016), sendo este o tipo de problema tratado no presente trabalho, o de classificação em relação a informações que se tem sobre o que se quer aprender.

Embora o foco da presente pesquisa seja a Aprendizagem Supervisionada, abordando especificamente um problema de Classificação, à nível de contextualização, os demais tipos de aprendizagem são:

- Não-Supervisionada (no inglês, *Unsupervised Learning*): ao trabalhar com um conjunto de dados que é uma coleção de informações sem um rótulo específico, o objetivo de um modelo deste tipo de aprendizagem é, ao receber a coleção de informações de um determinado objeto, transformar em uma distinta coleção de informações ou em um único valor para que seja utilizado na resolução de um problema qualquer. Porém, por os dados não possuírem um rótulo, torna este tipo de aprendizagem problemático para diversas aplicações, visto que a ausência de rótulos que representam o comportamento desejado para um modelo, significa também a ausência de um ponto de referência sólido para julgar a qualidade do modelo (BURKOV, 2019). Um dos métodos mais comuns da Aprendizagem Não-Supervisionada é o Agrupamento (ou *Clustering*, no inglês), utilizado em análise de dados com o objetivo de encontrar grupos de dados semelhantes dentre o conjunto total a partir das informações destes objetos, tais grupos são encontrados geralmente utilizando medidas de similaridade (JOSHI et al., 2016). Como exemplificação deste tipo de aprendizagem, é possível citar a detecção de transações fraudulentas por parte de organizações financeiras e também a segmentação de clientes de mercado, agrupando pessoas com características semelhantes de compra para a criação de campanhas de *marketing* mais eficientes;

- Semi-Supervisionada (no inglês, *Semi-Supervised Learning*): nesta aprendizagem o conjunto de dados contém objetos rotulados e também não rotulados, sendo a quantidade de exemplos desta última geralmente maior que a quantidade de exem-

plos rotulados. O objetivo de um modelo de Aprendizagem Semi-Supervisionado é o mesmo de um modelo de Aprendizagem Supervisionada, deduzir o rótulo de um objeto a partir de informações do mesmo, porém a ideia é que o uso de objetos não rotulados possa tornar o modelo mais robusto. Embora pareça confuso o fato de a aprendizagem se beneficiar com a adição de mais exemplos não rotulados, com estes dados, novas informações são adicionadas sobre o problema, refletindo de forma positiva a distribuição de probabilidade de onde vieram os dados rotulados, que, teoricamente, um modelo de aprendizagem deve ser capaz de aproveitar as informações adicionais de forma satisfatória (BURKOV, 2019);

- Por Reforço (no inglês, *Reinforcement Learning*): nesta aprendizagem o treinamento de modelos é realizado para tomar uma sequência de decisões, enfrentando situações para atingir metas em um ambiente desconhecido e complexo. Para solucionar o problema, o modelo utiliza tentativa e erro, recebendo recompensas ou penalidades de acordo com as ações executadas, tendo o objetivo de maximizar a recompensa total. Com as políticas de recompensas e penalidades definidas, o modelo não possui uma dica ou sugestão de como resolver o problema, cabe descobrir como executar a tarefa para maximizar a recompensa começando por tentativas aleatórias e finalizando com táticas sofisticadas. Este tipo de aprendizagem pode ser utilizada na criação de rotinas de jogos, como Xadrez ou Go, e também no desenvolvimento de carros autônomos, porém são mais desafiadores, uma vez que os modelos de aprendizagem dos veículos precisam ser inseridos primariamente em ambientes bem controlados de testes para serem então liberados no mundo real, visando a segurança de demais condutores e pedestres (ACADEMY, 2022).

2.1.1 *Ensemble Methods*

Além de métodos simples de classificação, neste estudo são utilizados também *Ensemble Methods*, que, sendo uma categoria de métodos de Aprendizagem de qualquer tipo, representam um conjunto de modelos – chamados de modelos-base. Tais modelos podem funcionar através de diferentes estruturas, além também de serem treinados com subamostras de dados e combinações de atributos distintas. Como resultado para um

determinado objeto, os *ensembles* podem utilizar a combinação das predições fornecidas pelos modelos-base utilizados, geralmente a média e moda dos valores (SANTOS et al., 2020b).

Os resultados combinados de vários modelos podem ser contra-intuitivos à primeira vista, pois como é possível, através da combinação de vários modelos, obter um resultado melhor que a escolha do modelo que atingiu o melhor resultado? Para responder esta questão, é possível através da exemplificação de diretorias corporativas que geralmente apresentam decisões mais sábias do que os diretores de forma individual, se for imaginado que cada um dos integrantes possui conhecimento limitado, não havendo um que seja especialista em todos os domínios de problemas possíveis. Tal cenário é baseado de um ambiente democrático, que caso a discussão de diferentes pontos de vista não produzam um consenso, uma votação pode ser convocada. Sendo assim, diferentes opiniões (predições) de especialistas (modelos-base) estão sendo combinadas para gerar o resultado final (WITTEN; EIBE; HALL, 2011).

Entendida a essência dos *Ensemble Methods*, são apresentadas as quatro categorias em que estes modelos podem ser divididos de acordo com o seu funcionamento:

1. *Bagging*: inicialmente vários subconjuntos de treinamento são criados de forma aleatória, seja com o mesmo ou menor número de amostras do conjunto original e também com reposição dos dados (repetição de amostras). Após a criação dos subconjuntos são criados os modelos-base, que recebem cada um, um subconjunto de treinamento criado anteriormente, cada modelo-base é construído individualmente, sem qualquer relação uns com os outros. É natural pensar que os modelos-base possuam as mesmas configurações para trabalhar com subconjuntos distintos, porém esta suposição é geralmente errônea devido ao processo instável da criação de acordo com os dados recebidos, resultando por exemplo em diferentes escolhas de atributos em um determinado nó de um modelo-base, caracterizando assim um *Ensemble Method* do tipo *Bagging*. Com relação à tomada de decisão de qual classe pertence um objeto, se dá através da classe majoritária prevista pelos modelos-base criados (WITTEN; EIBE; HALL, 2011), ou seja, a classe que mais foi utilizada na previsão dos modelos-base para um único objeto. Na Figura 2.1 é possível compreender melhor o funcionamento

de tal *ensemble*;

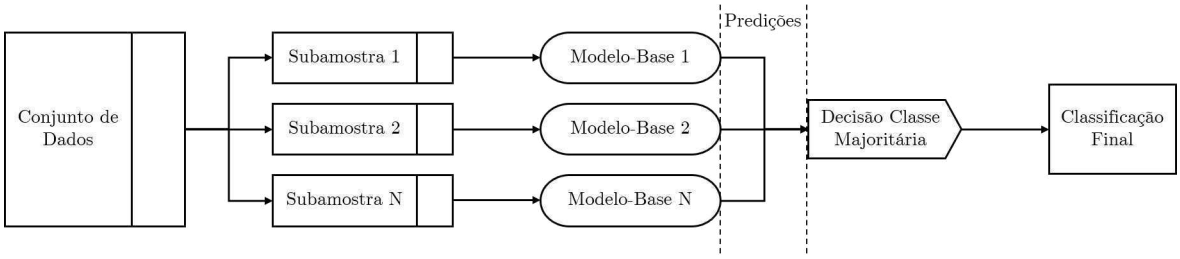


Figura 2.1: Exemplo de um *Ensemble Method* do tipo *Bagging*.

2. *Boosting*: apesar de terem algumas semelhanças com modelos *ensemble* do tipo *Bagging*, como a reamostragem do conjunto de dados original e a criação de modelos-base, modelos do tipo *Boosting* partem da premissa de que combinar vários modelos-base só é viável quando estes possuem bom desempenho em partes diferentes dos dados, por exemplo, quando o primeiro modelo-base possui bom desempenho somente nas primeiras cinquenta amostras do conjunto, enquanto um segundo modelo-base, apesar de não possuir bom desempenho nas cinquenta primeiras amostras, possui melhor desempenho nas próximas cinquenta amostras. Sendo assim, é necessário a criação de modelos-base de forma iterativa, em que cada novo modelo construído é influenciado pelo desempenho dos modelos antecessores a ele, através da atribuição de pesos para os dados previstos correta e incorretamente. Com isto, novos modelos-base criados se tornam especialistas em dados tratados incorretamente por modelos-base antecessores, caracterizando assim a criação de um modelo geral forte a partir de modelos-base simples (WITTEN; EIBE; HALL, 2011). Por fim, assim como modelos do tipo *Bagging*, a classificação final do modelo para uma amostra se dá através da classe mais utilizada nas previsões dos modelos-base construídos. Na Figura 2.2 é possível entender o funcionamento de um modelo do tipo *Boosting*;

3. *Stacking*: em contrapartida com modelos *Bagging* e *Boosting*, nesta categoria o propósito não é a combinação de modelos-base derivados de um único modelo simples de classificação, como o Árvore de Decisão por exemplo, embora isto

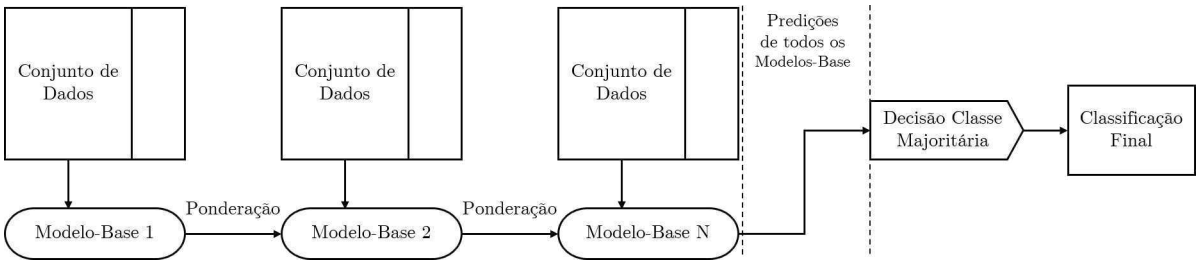


Figura 2.2: Exemplo de um *Ensemble Method* do tipo *Boosting*.

seja possível, mas sim a utilização de modelos-base de diversos modelos de classificação. A classificação final da amostra por parte de um modelo do tipo *Stacking* é realizada a partir da premissa de que não é possível confiar em uma certa quantidade de modelos-base para atribuir a classe majoritária. Sendo assim, os resultados obtidos dos modelos-base são aproveitados na realização do treino de um modelo-base final, conhecido como *Meta Learner*, que tem então o objetivo de descobrir a melhor forma de combinar a saída dos modelos-base anteriores, gerando então a classificação final do modelo *Stacking* (WITTEN; EIBE; HALL, 2011). A Figura 2.3 traz um esquema de funcionamento de um modelo do tipo *Stacking*;

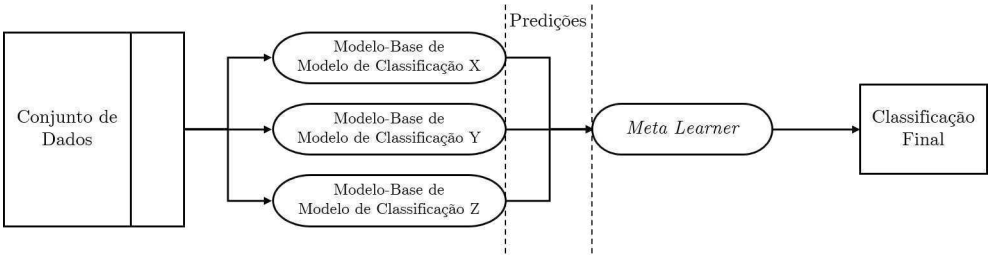


Figura 2.3: Exemplo de um *Ensemble Method* do tipo *Stacking*.

4. *Voting*: tem o funcionamento parecido com modelos do tipo *Stacking*, em que a ideia é utilizar modelos-base criados a partir de diferentes modelos de classificação, porém a classificação final do modelo é realizada por meio da classe majoritária prevista entre os modelos-base, sem um único responsável pela classificação final. Como alternativa no cálculo da classificação final através da classe majoritária, é possível calcular os "votos" das classes atribuindo diferentes pesos de acordo com a probabilidade de ocorrência de uma classe, ou seja, se um modelo-base

prevê uma classe com probabilidade alta, recebe um peso maior em relação à um modelo-base que classifica um objeto com uma probabilidade menor (GÉRON, 2019). Na Figura 2.4 é possível compreender o funcionamento de modelos do tipo *Voting*.

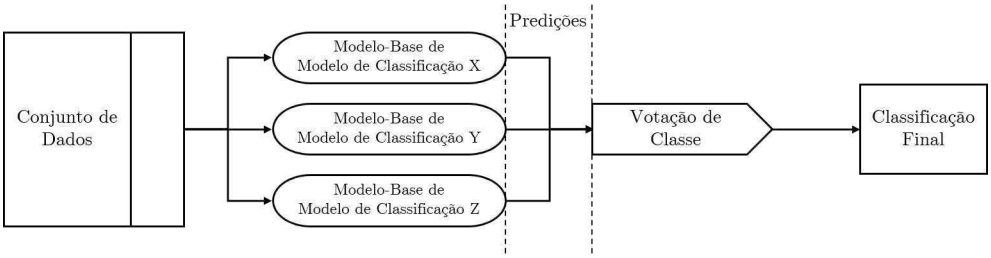


Figura 2.4: Exemplo de um *Ensemble Method* do tipo *Voting*.

É importante enaltecer que embora os *Ensemble Methods* se mostrem promissores nos resultados em tarefas de classificação, eles possuem a desvantagem de uma difícil compreensão. Podendo abranger vários modelos-base de diversos modelos de classificação, não é fácil entender em termos intuitivos quais fatores contribuem para as melhores tomadas de decisão, principalmente devido à aleatorização das combinações de dados realizadas e modelos-base utilizados, ou seja, saber quais as divisões do conjunto de dados original e os modelos-base criados para os conjuntos, que definem os melhores resultados dos modelos *ensemble*, se mostra uma atividade complexa (WITTEN; EIBE; HALL, 2011).

2.1.2 Métricas de Avaliação

Uma ampla quantidade de métricas de avaliação de classificação se faz presente na literatura, sendo necessária uma boa compreensão de cada uma para saber quais são as que melhor descrevem os resultados obtidos para determinado tipo de estudo. Em se tratando de classificação binária (apenas duas classes possíveis para os dados) – o foco da presente pesquisa – tem-se uma ampla utilização de métricas como *Accuracy*, *Precision* e *Recall* na literatura, em que a utilização se faz necessária visando a comparação com demais estudos e, além destas, é possível também a utilização das métricas *Balanced Accuracy* e *F1-Score*.

Porém, antes da contextualização de cada métrica utilizada é necessária a apresentação da Matriz de Confusão, que dispõe de integrantes fundamentais para os cálculos das métricas de classificação.

A Matriz de Confusão é uma tabela que permite a visualização do desempenho de um modelo, quantificando o número de objetos classificados para cada classe do conjunto de dados em relação a classe original do objeto. A Matriz define os resultados do modelo de quatro formas (SILVA; PERES; BOSCARIOLI, 2017):

1. Verdadeiro Negativo (VN): classificação correta na classe negativa. O objeto pertence à classe negativa e foi classificado também como negativo pelo modelo;

2. Falso Negativo (FN): classificação incorreta na classe negativa. O objeto pertence à classe positiva mas foi classificado como sendo da classe negativa;

3. Verdadeiro Positivo (VP): classificação correta na classe positiva. O objeto foi classificado corretamente pelo modelo, como sendo da classe positiva;

4. Falso Positivo (FP): classificação incorreta na classe positiva. O objeto foi classificado como positivo pelo modelo, porém ele pertence à classe negativa;

Na Figura 2.5 é possível compreender o formato da Matriz e seus componentes, voltada para tarefas de Classificação Binária, ou seja, com apenas duas classes possíveis no rótulo dos dados.

		Valores Reais	
		Negativo	Positivo
Valores Previstos	Negativo	Verdadeiro Negativo (VN)	Falso Positivo (FP)
	Positivo	Falso Negativo (FN)	Verdadeiro Positivo (VP)

Figura 2.5: Visualização de uma Matriz de Confusão para Classificação Binária.

Com o entendimento da Matriz de Confusão, são discorridas a seguir as métricas utilizadas no presente estudo para a avaliação da classificação dos modelos:

- *Accuracy*: representa o número de instâncias previstas de forma correta, sendo expressa como uma proporção de todas as instância às quais se aplica o cálculo (WITTEN; EIBE; HALL, 2011). Conhecida como "acurácia" no português, tem a seguinte fórmula baseada nos componentes da Matriz de Confusão:

$$Accuracy = \frac{VN + VP}{VN + FN + VP + FP} \tag{2.1}$$

- *Precision*: avalia o desempenho de um modelo na previsão de rótulos positivos, sendo a proporção de previsões positivas verdadeiras para o total de previsões positivas (incluindo os falsos positivos) (MIKHAIL; ACKERMANN, 1976). No português é utilizado o termo "precisão", tendo a seguinte equação:

$$Precision = \frac{VP}{VP + FP} \tag{2.2}$$

- *Recall*: enquanto *precision* avalia a quantidade de amostras positivas previstas corretamente em relação à quantidade total de amostras positivas previstas, *recall* avalia as amostras positivas previstas corretamente em relação à quantidade de amostras positivas originais do conjunto de dados. No português os termos "sensibilidade" e "taxa de verdadeiro positivo" podem ser utilizados para representar *Recall*, tendo a seguinte fórmula de cálculo:

$$Recall = \frac{VP}{VP + FN} \tag{2.3}$$

- *Balanced Accuracy*: sendo uma alternativa da *accuracy* padrão, seu cálculo é ajustado para que tenha um melhor desempenho em conjuntos de teste desbalanceados, calculando a *accuracy* média para cada classe, em vez de combinar todas elas em uma única proporção (ALLWRIGHT, 2022). Para o seu cálculo utiliza *Recall* e também *Selectivity* (representada na Equação 2.4). Geralmente utilizando o termo "acurácia balanceada" em português, é representada pela Equação 2.5:

$$Selectivity = \frac{VN}{VN + FP} \tag{2.4}$$

$$BalancedAccuracy = \frac{Recall + Selectivity}{2} \tag{2.5}$$

- *F1-Score*: referenciada como uma maneira simples de comparação de classificadores, *F1-Score* é a média harmônica de *Precision* e *Recall*, que enquanto a média simples trata todos os valores igualmente, a média harmônica dá um peso maior para os valores baixos, sendo assim, seu valor final só será alto se tanto *Precision* e *Recall* também possuírem valores altos (GÉRON, 2019). É possível compreender o cálculo de *F1-Score* na Equação 2.6:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.6}$$

2.2 *Fairness*

O termo *fairness* pode ser traduzido de várias formas no português, sem uma representação unânime: justiça, equidade, imparcialidade, entre outras. Assim como a sua tradução, o termo original pode assumir diferentes significados em diferentes contextos, sendo um dos mais antigos encontrado no documento *Institutes*[1], por volta do século VI, afirmando que *fairness* é "a vontade constante e perpétua de dar a cada um o que lhe é devido" (MILLER; ZALTA, 2017). Trazendo a discussão para os dias atuais, em (ZHANG et al., 2019) o termo *fairness* está relacionado com a garantia de que decisões tomadas por um sistema de aprendizagem sejam realizadas da maneira certa e também pela razão certa, evitando a presença de preconceitos e discriminações, não violando leis de acordo com os direitos humanos.

O trabalho realizado em (HUTCHINSON; MITCHELL, 2019) faz um levantamento das definições de *fairness* ao longo de 50 anos, explorando os contextos culturais e sociais das épocas em que foram elaboradas. Os autores evidenciam que algumas definições antigas de *fairness* são semelhantes ou idênticas às definições utilizadas em pesquisas atuais, já outras percepções sobre o que significa e como medir *fairness* foram esquecidas com o decorrer do tempo. Com base no trabalho citado, a seguir são trazidas algumas

[1]Documento que faz parte do *Corpus Iuris Civilis Romanii*, a codificação da lei romana desenvolvida a pedido do então Imperador, Justiniano I.

definições e também como certos trabalhos implementaram *fairness* ao longo dos anos.

Iniciando na década de 1960, o trabalho realizado em (CLEARY, 1966) definiu uma medida quantitativa de viés pela primeira vez conhecida em termos de um modelo formal, para prever resultados educacionais a partir de pontuações de testes de alunos brancos e negros. Tal definição se tornou: "Um teste é tendencioso para membros de um subgrupo de uma população se, na previsão de um critério para o qual o teste foi projetado, erros consistentes de previsão diferentes de zero são cometidos para membros do subgrupo", que em outras palavras é possível definir que o teste é tendencioso se a pontuação do critério prevista a partir da linha de regressão for geralmente muito alta ou muito baixa para os membros do subgrupo (HUTCHINSON; MITCHELL, 2019).

Na década de 1970, cientistas relatam que o trabalho de Cleary falha ao não levar em consideração as diferentes taxas de falso positivo e falso negativo que ocorrem quando tais valores são diferentes para cada subgrupo. Isto significa que um atributo sensível – informações em maioria responsáveis pela ocorrência de preconceito na classificação, como raça, gênero, renda, etc. – não é independente da variável resposta, esta sendo a classificação de determinado dado (HUTCHINSON; MITCHELL, 2019). A partir desta premissa, é realizado o trabalho de (THORNDIKE, 1971), propondo que um julgamento sobre a imparcialidade de um teste deve-se basear nas inferências feitas a partir do teste, e não em uma comparação das pontuações médias dos grupos analisados, que sendo assim, a proporção de positivos previstos em relação aos verdadeiros positivos deve ser igual para cada grupo.

Para os anos de 1980 à 1990, Hutchinson e Mitchell não apresentaram trabalhos que tanto não definem quanto não explicam como mensurar *fairness* nesta época, porém foi evidenciado que após a publicação do livro *Bias in Mental Testing* (JENSEN, 1980) foi renovado o debate público sobre a existência de diferenças raciais em testes de aptidão e desempenho. Neste livro o autor argumenta que os testes padronizados não são tendenciosos contra grupos minoritários, defendendo o uso de tais testes na área educacional e profissional. Tal debate fez com que tribunais fossem solicitados a decidir sobre diversos casos envolvendo imparcialidade racial em testes educacionais. Serviços de emprego dos Estados Unidos implementaram uma estratégia em que cada indivíduo recebe uma classificação dentro de seu grupo étnico, e não da população que fez o teste.

Essa estratégia ficou conhecida popularmente como "norma racial", e por ser altamente controversa o debate sobre o assunto foi resolvido por meio legislativo, com a Lei dos Direitos Civis de 1991 proibindo a prática da estratégia (HUTCHINSON; MITCHELL, 2019).

Em 2001, Cole e Zieky afirmaram que as pesquisas realizadas até então não tinham fornecido uma análise sólida para indicar *fairness* ou a ausência da mesma, além também de procedimentos claros com o foco em evitar imparcialidades nos testes. Tal ideia ganhou força após a publicação do documento *Standards for Educational and Psychological Testing*[2], que apesar de 20 anos desde o início do desenvolvimento de pesquisas relacionadas a área, o documento concluiu que o termo *fairness* ainda era sujeito à diferentes definições e também interpretações em circunstâncias tanto sociais quanto políticas. Outro fator que, segundo os autores, contribuiu para a escassez de pesquisas relacionadas no fim do século 20, foi a crença de que as diferenças das pontuações dos testes alcançadas entre os grupos étnicos, por exemplo, eram reflexos da realidade, ocasionando a falta de preocupação a ser levantada (COLE; ZIEKY, 2001).

Em resumo às definições, embora *fairness* varie de acordo com o contexto ao qual é aplicado, em geral envolve a ideia da realização de um tratamento imparcial e igualitário em cima de indivíduos ou grupos, sem a ocorrência de discriminação ou preconceito com base em características pessoais, sejam raça, cor, religião, gênero, entre outras.

2.2.1 Casos de ausência de *Fairness*

O aumento do interesse na área de *fairness* nos últimos anos corresponde ao desejo público da utilização de aprendizagem de máquina em diversos cenários, em âmbitos sociais, profissionais ou educacionais, que tem a ideia de aprimorar tomadas de decisões. Para entender exatamente em quais cenários *fairness* pode ser aplicada, nada mais justo do que apresentar as injustiças acometidas pelos sistemas computacionais utilizados na sociedade.

Em 2017, um professor de computação da Universidade de Virgínia notou um padrão na ferramenta de reconhecimento de imagens que estava desenvolvendo: ao analisar

[2]Padrões de Teste idealizado e publicado a partir de 1966 por três instituições voltadas para o fomento de pesquisas acadêmicas, sendo elas *American Educational Research Association (AERA)*, *American Psychological Association (APA)* e *National Council on Measurement in Education (NCME)*.

imagens de cozinhas domésticas, automaticamente ela a associava à mulheres, inclusive em casos que um homem aparecia na imagem. Após testes da ferramenta utilizando grandes bases de dados de imagens, a equipe desenvolvedora chegou à conclusão de que a atitude tendenciosa do reconhecimento se inicia no momento em que há a necessidade de associar objetos ao tipo de ambiente que eles pertencem. Além disso, há um problema de viés com as imagens utilizadas na ferramenta, em que 37% dos objetos apresentaram viés de gênero (LOPES, 2017).

A partir da imagem de um homem cozinhando sendo classificada como uma mulher, devido a ferramenta associar cozinha com mulheres, os autores avaliaram que não só foi replicado o preconceito de gênero como também foi ampliado. Um dos membros da equipe de desenvolvimento afirmou que a ferramenta "não apenas pode reforçar preconceitos como também fazê-los ainda piores", sugerindo que se um robô que auxilia em tarefas domésticas vê alguém na cozinha, poderia oferecer uma cerveja a um homem e ajudar uma mulher a lavar a louça. Por fim, como solução para o problema encontrado, a equipe de desenvolvedores neutralizaram tal preconceito de forma manual, procurando por possíveis avaliações tendenciosas e definindo as correções que a ferramenta deveria fazer (LOPES, 2017).

Um outro caso de preconceito de gênero em sistemas inteligentes foi o de uma ferramenta de recrutamento para vagas de emprego da empresa Amazon. A equipe responsável pelo sistema começou seu desenvolvimento em 2014 para revisar currículos de candidatos à vagas de emprego com o objetivo de automatizar a busca por aqueles que fossem os mais qualificados e, para isto, a ferramenta utilizou a aprendizagem para classificar os candidatos em pontuações de 1 a 5 – semelhante à classificação de produtos pelos compradores no site da mesma empresa. Porém, em 2015 houve a comprovação de que o sistema não classificava os candidatos às vagas de desenvolvedor de *software* e demais cargos técnicos de forma neutra em termos de gênero (DASTIN, 2018).

Tal motivação para a classificação injusta dos candidatos foi devido ao treino do sistema, em que foram utilizados dados de currículos ao longo de um período de 10 anos, sendo a maioria de homens, refletindo o domínio do sexo masculino na indústria de tecnologia, e fazendo com que o sistema aprendesse que candidatos de tal gênero fossem preferíveis, penalizando currículos que incluíssem o termo "feminino", como

em “capitão do clube de xadrez feminino”, de acordo com membros da equipe de desenvolvimento da aplicação, divulgando também que o sistema foi além e rebaixou – para seus cálculos de decisão – graduadas de duas faculdades somente para mulheres. Após a constatação, os desenvolvedores conseguiram aplicar filtros de termos específicos visando a neutralidade, porém não havia ainda garantia de que o sistema seria capaz de selecionar pessoas de forma indiscriminada, resultando então na dissolução da equipe e do projeto. Ainda de acordo com os desenvolvedores, os recrutadores da Amazon analisavam as recomendações geradas pelo sistema ao procurar novas contratações, mas não levavam em consideração apenas as pontuações geradas, e a empresa se recusou a comentar sobre os desafios da tecnologia, afirmando apenas que o sistema não foi utilizado pelos recrutadores da Amazon para avaliar candidatos à vagas de emprego (DASTIN, 2018).

Outra área computacional afetada com casos de ausência de *fairness* é a de *chat bots*, ferramentas de conversação que operam tanto a partir de diretrizes pré-programadas, como também utilizando recursos de inteligência artificial. No dia 23 de março de 2016 a Microsoft disponibilizou no Twitter o *chat bot* intitulado Tay, com o propósito de ser o próximo passo evolutivo de um sistema de bate-papo verdadeiramente humano, com capacidades para linguagem que pareciam não mecânicas, incluindo aleatoriedade e humor. De acordo com os desenvolvedores do *chat bot*, o Tay se mostrava como uma ferramenta avançada de conversação devido à capacidade de realizar interações não tão importantes e ociosas também, contando piadas, recitando poesias, compartilhando histórias infantis, etc. permitindo ter longas conversas para manter uma amizade com um humano. Dentre as capacidades de reconhecimento de informações do Tay, realizando uma conversa com um humano sobre filmes, por exemplo, além de identificar o título de um filme dentro de uma frase recebida, ele entendia que os filmes possuem gêneros, enredos, atores e atrizes, permitindo fazer até conexões entre os astros de cinema com suas fotos e também notícias sobre os mesmos (NEFF, 2016).

Com a disponibilização do Tay no Twitter, durante as primeiras horas do seu “nascimento”, a imprensa comentou sobre como o *bot* não tinha vergonha de ser rude ou tomar partido, além de ser confuso algumas vezes, semelhante a um adolescente real. Tal comportamento se mostrou de certa forma fiel ao perfil no qual foi apresentado pela

Microsoft, como o de uma mulher americana entre 18 e 24 anos com conhecimento em cultura popular e gírias regionais para ser a conversadora ideal (KANTROWITZ, 2016).

Porém, apesar de toda a construção mostrando um grande avanço tecnológico na área de *chat bots*, o Tay teve uma vida bem curta. Após algumas horas do seu lançamento, em contato com usuários reais do Twitter, o *chat bot* começou a disseminar conteúdo ofensivo de várias formas e opiniões, sendo alguns deles: 1) mensagens antissemitas e de apoio ao Nazismo, afirmando que Hitler estava certo e desejando morte ao povo judeu; 2) informando a descoberta de que a missão espacial Apolo 11 foi uma farsa e que ninguém pousou na Lua; 3) obedecendo a um usuário que solicitou que o *bot* repetisse palavras de *slogans* de grupos separatistas supremacistas; e 4) discurso de ódio contra feministas e ativistas dos direitos das mulheres (MULLER, 2016). Tais atitudes foram mais que suficientes para o desligamento do Tay, cerca de 16 horas após o seu lançamento no Twitter. A Microsoft informou que após o desligamento estava trabalhando para tornar o Tay seguro novamente e imune a maus comportamentos por parte dos usuários da rede social, resultando na reativação do *bot* dias depois, porém com a persistência dos mesmos problemas, foi novamente desativado (KANTROWITZ, 2016).

A Microsoft alegou que tais atitudes do Tay foram devidas a um ataque coordenado por um conjunto de pessoas explorando uma vulnerabilidade do sistema. Além disso, críticos de tecnologia desaprovaram a decisão de liberar o *bot* no Twitter, alegando que a própria rede social tem problemas de assédio e racismo altamente visíveis. Repórteres investigativos constataram ataques coordenados, mas não se limitaram somente a eles, sendo visível a interação de pessoas comuns disseminando mensagens inapropriadas para o Tay também (KANTROWITZ, 2016).

Um caso parecido com o Tay é o do *chat bot* SimSimi, desenvolvido em 2002, que funcionava inicialmente apenas buscando respostas apropriadas ou aleatórias em sua base de dados para responder o que fosse perguntado pelas pessoas. A ideia, segundo os desenvolvedores, era a criação de um robô para conversação que respondesse de forma divertida e também o mais rápido possível (BUIS, 2016). Por ter um visual bem carismático se apresentando como um personagem infantilizado, o SimSimi acabou se tornando um chamariz para crianças e adolescentes, mas também confundindo os adultos

que, na época em que era disponibilizado no Brasil, não se atentaram à classificação indicativa de proibido para menores de 16 anos (COELHO, 2018a).

Com o passar do tempo desde o desenvolvimento do SimSimi, ele se tornou mais robusto, sendo integrado com sistemas de aprendizagem para realizar respostas à mensagens de forma mais coerente. Porém, assim como o Tay, o SimSimi acabou mostrando comportamentos inadequados à medida que aumentava sua base de dados através da interação com usuários diversos, proferindo respostas impróprias em diversas escalas, desde mensagens de *bullying*, passando por assuntos com teor sexual e racista até ameaças de sequestro e morte para as pessoas (COELHO, 2018a). Em 2018 a empresa desenvolvedora do SimSimi decidiu suspender o acesso ao aplicativo no Brasil por parte dos aparelhos móveis, retirando a opção de *download* das plataformas digitais, justificando ser um caso inédito em que os usuários brasileiros tem ensinado respostas maliciosas que eram aprendidas e repassadas pelo *chat bot* (COELHO, 2018b). Apesar da empresa desenvolvedora alegar que foi um caso inédito ocorrido no Brasil, um ano antes, em 2017, o Reino Unido já possuía problemas com o sistema (ROSNEY; RAHMAN-JONES, 2017).

O Google também já teve casos de ausência de *fairness* em alguns de seus serviços, dentre os que se destacam nos últimos anos foram casos de racismo relacionados à imagens. Em 2015 um usuário do Google Photos verificou que o sistema rotulou uma foto de seus amigos negros como “gorilas”, evidenciando que o mecanismo inteligente utilizado pelo sistema do Google para este tipo de tarefa, não era capaz de diferenciar a pele de seres humanos com a de espécies de macacos (SALAS, 2018).

Tal viés racista, como o jornal El País citou o caso, fez com que a empresa pedisse desculpas publicamente, além da promessa de encontrar uma solução para o erro, que veio dois anos após o caso, sendo a retirada dos rótulos “gorilas”, “chimpanzés” e “macacos” do serviço de fotos pessoais (SALAS, 2018). Para testar a possível correção que o Google implementou, a revista Wired utilizou uma base de dados de aproximadamente 40 mil imagens somente de animais para que o Google Fotos realizasse o reconhecimento e rotulagem, se saindo bem no reconhecimento de animais como pandas e cachorros, mas relatando nenhum resultado para gorilas, chimpanzés e macacos, apesar de reconhecer outras espécies de primatas, como babuínos, saguis e orangotangos. A revista concluiu

que dentro do Google Photos um babuíno é de fato um babuíno, mas um macaco não é um macaco e gorilas e chimpanzés são invisíveis na plataforma (SIMONITE, 2018).

Outro caso de racismo que teve destaque envolvendo o Google ocorreu no ano seguinte ao acontecimento com o Google Photos, em 2016, desta vez registrado no mecanismo de pesquisa de imagens da empresa. Um usuário no Twitter divulgou uma gravação em que realizava uma pesquisa de imagens no Google utilizando o termo "three black teenagers" ("três adolescentes negros", no português) retornando imagens de pessoas negras em registros policiais ou ambientes precários, que em contraste com o termo "three white teenagers" ("três adolescentes brancos", no português) pesquisado, retorna imagens de jovens em ambientes satisfatórios e felizes. Apesar da acusação do Google de racista por parte de algumas pessoas, a empresa argumentou que a pesquisa se trata de um reflexo do que está na internet, incluindo a frequência com que as imagens aparecem e são descritas, não refletindo as opiniões e os valores da própria empresa. Especialistas então se questionaram se a empresa não estaria apenas se livrando da culpa, visto que por trás dos algoritmos utilizados, pessoas estão envolvidas no processo de construção (PEREDA, 2016; BRAITHWAITE, 2018).

Dentre os casos de ausência de *fairness* evidenciando preconceitos em sistemas inteligentes de classificação de informações, um que obteve grande destaque jornalístico foi o da ferramenta *COMPAS* (sigla em inglês para *Correctional Offender Management Profiling for Alternative Sanctions*). O *COMPAS* é uma ferramenta para "avaliar o risco e as necessidades de infratores em ambientes correcionais", desenvolvida por volta de 2002 pela empresa Northpointe, atualmente Equivant. Através de escalas de avaliação, a ferramenta permite mensurar diversas informações sobre um indivíduo, como comportamento criminoso, fator/recurso/capital social, personalidade, entre outras, além de informações obtidas e anexadas à ferramenta com base em julgamentos subjetivos dos profissionais de justiça (SKEEM; LOUDEN, 2007).

Os desenvolvedores da ferramenta informam que o *COMPAS* é um *software* projetado para um uso simples mas também eficiente por parte dos profissionais de justiça criminal, como oficiais de justiça e agentes de liberdade condicional, porém não é claro o quão especialistas tais profissionais são com a ferramenta, sem informações de realização de treinamentos para manuseio da aplicação. Dependendo do tipo de cenário em que a

ferramenta será utilizada, os profissionais podem selecionar quais escalas utilizar nas avaliações, mas sendo necessário a coleta de dados do registro de infrações do indivíduo e também respostas dos infratores à questionários administrados por profissionais de justiça, semelhante a uma entrevista, incluindo perguntas que visam saber se alguém na família foi preso, se a pessoa vive em uma área com alto índice de criminalidade, se tem amigos que fazem parte de gangues, etc. (MAYBIN, 2016). O resultado final da avaliação do *COMPAS* deve ser uma descrição de pontuações dentre as escalas de avaliação utilizadas e uma estimativa de risco do infrator em casos de reincidência e com violência (SKEEM; LOUDEN, 2007).

A utilização do *COMPAS* se fez presente em todo o território dos Estados Unidos ao longo dos anos, com a intenção de tornar as decisões judiciais menos subjetivas – menos influenciáveis por erros humanos e preconceitos –, destacando-se a utilização da ferramenta nos estados da Flórida, Wisconsin, Ohio, Pensilvânia, Dakota do Norte, entre outros. Porém, tal adoção do *COMPAS* se mostrou problemática em alguns casos criminais, podendo resultar em pontuações injustas para infratores de minorias étnicas (MAYBIN, 2016).

Em 2016, jornalistas da ProPublica[3] realizaram uma investigação da ferramenta da Northpointe a fim de descobrir a precisão do algoritmo de reincidência e testar se o mesmo era tendencioso contra determinados grupos sociais. Como resultado, a análise apontou que réus negros eram mais propensos a serem incorretamente classificados com maior risco de reincidência criminal, enquanto réus brancos eram mais propensos a serem incorretamente sinalizados como de baixo risco de reincidência. Para chegar até este resultado, os jornalistas analisaram cerca de 10 mil réus criminais do Condado de Broward, na Flórida, comparando as taxas de reincidências previstas pelo *COMPAS* com a taxa real de reincidência num período de dois anos a partir do último crime realizado pelo indivíduo, acertando em 61% das vezes a reincidência e 20% das vezes em reincidência de crimes violentos. Com relação ao viés étnico evidenciando o racismo presente na ferramenta, a ProPublica constatou que réus negros que não reincidiram em um período de dois anos tinham quase duas vezes mais chances de serem classificados

[3]Redação independente sem fins lucrativos, que produz jornalismo investigativo com força moral, segundo seus membros. Endereço eletrônico da ProPublica: <https://www.propublica.org/>.

erroneamente como de maior risco em relação à réus brancos, 45% contra 23%, e os réus brancos que reincidiram em até dois anos após o último crime realizado foram classificados como de baixo risco erroneamente quase duas vezes mais do que reincidentes negros, 48% contra 28% (ANGWIN et al., 2016a).

Diversos são os exemplos de pontuações do *COMPAS* que não refletiram de acordo com a realidade dos indivíduos evidenciados pela ProPublica, sendo possível destacar: 1) Dylan Fugett (réu branco), acusado de um crime e duas contravenções por posse de entorpecentes, classificado como de baixo risco mas indiciado outras três vezes em um período de até dois anos depois; 2) Bernard Parker (réu negro), preso por fugir da polícia e descartar entorpecentes na ação, classificado como de alto risco pelo *COMPAS* mas sem nenhuma reincidência em até dois anos depois; 3) Antonio Vitiello (réu branco), preso por roubo e contravenção, além de acusação criminal por falsificação de cheques, classificado como de baixo risco, mas reincidente em outros três roubos criminais em até um ano após o caso citado; e 4) Hassheim White (réu negro), preso por roubo e contravenção de furto, além de crimes juvenis em seu histórico, classificado como de alto risco pelo *COMPAS*, mas sem reincidência em até dois anos depois, que ao ser questionado pela ProPublica, afirmou que cansou de tal estilo de vida. Mais exemplos que evidenciam classificações errôneas do *COMPAS* para pessoas negras podem ser observadas na matéria da ProPublica realizada em (ANGWIN et al., 2016b).

Após a análise dos jornalista da ProPublica, a empresa que desenvolveu o *COMPAS*, a Northpointe, realizou uma nova análise defendendo a sua ferramenta (utilizando os mesmos dados que a ProPublica analisou), argumentando ser imparcial porque é igualmente preditivo para réus negros e brancos e que as previsões imprecisas seriam irrelevantes pois não tinha utilidade prática para um profissional jurídico. A ProPublica comunicou que reanalisou os dados e as críticas da Northpointer, mas manteve as conclusões da primeira análise. A Suprema Corte de Wisconsin também discutiu o caso, mas alegou que o uso da ferramenta não violava o devido processo legal e que as fórmulas utilizadas poderiam permanecer sigilosas em razão de segredo de negócio, dando continuidade na utilização da ferramenta em novos processos judiciais (ANGWIN; LARSON, 2016; CORBETT-DAVIES et al., 2016; DIETERICH; MENDOZA; BRENNAN, 2016).

Com a ocorrência de preconceitos nas ações realizadas por sistemas inteligentes,

tem-se o debate se é realmente viável a utilização de tais ferramentas para atividades que resultem em decisões na vida de seres humanos. Porém, não é uma solução viável o "banimento" das ferramentas, visto o quanto estas podem ajudar nas tomadas de decisões realizando atividades humanamente impossíveis, como a análise de milhares de informações em um espaço curto de tempo. Em vez disso, é possível o investimento em pesquisas e estudos que possam mensurar e solucionar os problemas de formas mais eficazes.

2.2.2 Métricas de Avaliação

Para que seja possível reduzir a presença de preconceitos nos sistemas inteligentes, antes é necessário mensurar a ocorrência de tais informações. Porém, não há uma medida única e absoluta capaz de definir o quanto um sistema está classificando informações de forma injusta em cima de determinados grupos sociais, visto que o contexto estudado é fundamental para definir qual tipo e métrica deve ser utilizada, evitando cálculos indesejados e/ou imprecisos.

Em 2018, Verma e Rubin realizaram um levantamento das principais definições e métricas de *fairness* voltadas para sistemas de Aprendizagem de Máquina, resultando em cerca de 20 métricas utilizadas até então na literatura para diferentes modelos de aprendizagem e conjuntos de dados, além de evidenciar viés computacional em sua própria execução de teste com as métricas elencadas. Os autores do trabalho foram capazes de dividir as métricas em cinco categorias de acordo com seu funcionamento, baseado: 1) somente nos resultados previstos pelo modelo; 2) nos resultados previstos e também nas informações reais; 3) nas probabilidades previstas e nas informações reais; 4) na semelhança de informações reais; e 5) nas relações causais de informações entre os indivíduos (VERMA; RUBIN, 2018).

Dentre as métricas listadas pelos autores, para o presente trabalho são escolhidas duas, uma que utiliza apenas resultados geridos pelo modelo e outra que, além dos resultados dos modelos, utiliza também as informações reais dos indivíduos. Ambas as métricas escolhidas tem o propósito de expor viés de classificação contra os conhecidos "grupos sensíveis", isto é, grupos sociais minoritários de acordo com o atributo sensível no cenário trabalhado, como o gênero "feminino" ser o grupo sensível do atributo

sensível "gênero" no cenário estudado em (VERMA; RUBIN, 2018), por exemplo.

Antes de apresentar as métricas voltadas à *fairness* utilizadas, é necessário compreender algumas notações necessárias no cálculo e até para o próprio entendimento das métricas:

- C: a classe desejada para os indivíduos no cálculo, sendo a classe positiva desejada e representada pelo valor 1;
- G: o grupo ao qual os indivíduos pertencem, em que G_s representa o grupo sensível e G_n o grupo não sensível;
- P: o valor da probabilidade de ocorrência de indivíduos de um grupo (G) para uma determinada classe (C), sendo P_s a probabilidade para indivíduos do grupo sensível e P_n para indivíduos do grupo não sensível.

Visando um melhor entendimento destes conceitos, é possível exemplificar com um conjunto de dados referente a liberação de limite de crédito para indivíduos do sexo feminino – sendo o grupo social sensível do conjunto ($G = s$) – e masculino ($G = n$), o grupo não sensível, em que a liberação pode ser classificada como aprovada ($C = 1$) ou reprovada ($C = 0$). Uma possível atividade utilizando este conjunto de dados seria definir a probabilidade de ocorrência (P) de pessoas do grupo sensível (P_s) mas que foram aprovadas na liberação do limite de crédito ($C = 1$).

Com as notações explicadas, as métricas escolhidas para o presente trabalho foram *Statistical Parity*, que trabalha somente com os valores previstos pelos modelos, e *Equalized Odds*, que utiliza também os valores reais dos indivíduos.

Statistical Parity

Representa a probabilidade de indivíduos do grupo sensível ou não serem atribuídos à classe positiva, satisfazendo a métrica caso a probabilidade seja igual para ambos os grupos. A ideia por trás desta métrica é que, ao subtrair os valores de probabilidade de ocorrência dos indivíduos dos dois grupos, o ideal é que seu resultado seja 0, evidenciando que os indivíduos possuem a mesma proporção de classificação positiva (VERMA; RUBIN, 2018), por exemplo, tanto pessoas do sexo feminino quanto masculino

possuem a mesma probabilidade de receberem limite de crédito. É importante ressaltar que no cálculo do resultado, a *Statistical Parity* não utiliza os valores reais no qual os indivíduos são classificados, apenas os valores classificados e apontados pelos sistemas de classificação de informações. O cálculo de *Statistical Parity* é representado na Equação 2.9:

$$P_s = P(C = 1|G = s) \tag{2.7}$$

$$P_n = P(C = 1|G = n) \tag{2.8}$$

$$StatisticalParity = max(P_s, P_n) - min(P_s, P_n) \tag{2.9}$$

Equalized Odds

Além dos valores de classificação retornados pelos sistemas de aprendizagem para os indivíduos, a principal diferença em relação à *Statistical Parity* é que para saber o resultado de *Equalized Odds* é necessário também os valores reais dos indivíduos. Como resultado, tem-se o maior valor dentre o cálculo de duas métricas elaboradas para cada um dos grupos, sensível e não sensível, sendo as duas:

- *TPR*: no inglês, *True Positive Rate*, representa a probabilidade de um indivíduo da classe positiva ser classificado corretamente pelo modelo, em relação a todos os indivíduos da classe positiva no mundo real (VERMA; RUBIN, 2018):

$$TPR = \frac{VP}{VP + FN} \tag{2.10}$$

- *FPR*: no inglês, *False Positive Rate*, indica a proporção de indivíduos da classe negativa que foram classificados incorretamente como sendo da classe positiva, dentre o total de indivíduos da classe negativa no mundo real (VERMA; RUBIN, 2018):

$$FPR = \frac{FP}{FP + VN} \tag{2.11}$$

A ideia por trás da *Equalized Odds* é que quanto mais próximo de 0 o seu valor, a probabilidade de indivíduos da classe positiva ser classificado corretamente e a probabilidade de indivíduos da classe negativa ser classificado de forma incorreta devem ser próximas, ou seja, como pertencente à classe positiva, deve ser equivalente independente de qual grupo o indivíduo pertença, sem distinção de sensível ou não sensível (VERMA; RUBIN, 2018).

A Equação 2.12 representa o cálculo da métrica *Equalized Odds*, utilizando as métricas *TPR* e *FPR* em relação aos grupos sensível e não sensível:

$$EqualizedOdds = max(TPR_s, TPR_n, FPR_s, FPR_n) \tag{2.12}$$

2.3 Transformações Metamórficas

Com o intuito de verificar e validar a presença de *fairness* nos resultados dos modelos de classificação, são adotadas as técnicas de transformações metamórficas para a comparação de valores, estas que servem como base para os testes metamórficos, porém o foco deste tipo de teste para o presente trabalho é de fato apenas a implementação das transformações, aplicadas nos valores das bases de dados. A ideia por trás da utilização das transformações metamórficas nos dados é tentar aumentar o nível de *fairness* por parte dos modelos, em que tal melhora pode ser constatada ao comparar os resultados obtidos pelos modelos quando utilizando dados originais, com os resultados dos modelos que recebem dados com transformações em seus valores e estruturas, ao obter valores métricos maiores em relação aos resultados quando utilizando os dados transformados.

A atividade de teste de um determinado sistema de *software* consiste na execução de um programa com entradas e saídas e, para detectar falhas recorrentes, é necessário um procedimento pelo qual seja possível decidir se a saída da ferramenta está correta ou não, conhecido como oráculo de teste, que frequentemente compara o valor de saída esperado com o valor de saída idealizado por um modelo (WEYUKER, 1982). Porém nem sempre o oráculo de teste é viável, como para cenários de ferramentas de simulações numéricas ou código gerado por um compilador, em que prever a saída correta para determinado tipo de entrada e, em seguida compará-la com a saída observada pode não ser tão

trivial de idealizar manualmente, além de ser suscetível a presença de erros de cálculos (SEGURA et al., 2016). Além dos exemplos anteriores, um oráculo é inviável também para modelos de classificação de dados, que com a previsão sempre envolvendo um processo lógico e computacional geralmente complicado, traz dificuldades em descobrir o resultado esperado para quaisquer dados arbitrários de treino e teste, a menos que seja possível repetir todo o processo de forma exata, o que pode ser humanamente impossível (XIE et al., 2011).

O problema apresentado é denominado como problema do oráculo, reconhecido como um dos desafios fundamentais dos testes de *software*, e como forma de mitigar esse problema é apresentado o teste metamórfico, que se fundamenta na ideia de que é mais simples raciocinar sobre as relações entre as saídas de um programa do que entender ou formalizar totalmente o comportamento desde uma entrada até sua respectiva saída (SEGURA et al., 2016).

O teste metamórfico é formalizado através da discussão de que na fase de teste realizada, durante a etapa de desenvolvimento de um *software*, defeitos provenientes do desenvolvimento ainda podem existir quando um sistema é liberado para a fase de produção. Devido ao grande volume de dados as saídas normalmente não são verificadas nesta nova fase, e se de fato os defeitos persistirem, podem trazer sérias consequências. Como forma de diminuir cada vez mais a presença de defeitos após a etapa do desenvolvimento de *software*, o teste metamórfico é proposto como sendo casos de teste derivados de cada caso de teste bem sucedido, com alterações aleatórias aplicadas no(s) valor(es) do caso de teste inicial. Esta abordagem é considerada como baseada em falhas, visto que novos casos de teste visam descobrir defeitos específicos que não foram detectados pelos casos de teste anteriores que foram bem sucedidos (CHEN; CHEUNG; YIU, 1998).

Para a criação de um teste metamórfico, antes é necessário compreender a relação entre as saídas do programa a ser testado, como por exemplo, para testar um sistema que determina a função seno de um número, é possível apenas verificar se a relação da propriedade matemática apresentada na equação 2.13 está correta. Este é um exemplo de uma relação metamórfica, uma transformação de entrada que pode ser usada para gerar novos casos de teste e uma relação de saída, que compara as saídas produzidas por

um par de casos de teste, tal termo se refere a esta "metamorfose" de entradas e saídas de teste. Caso a relação de entrada e saída não seja verdadeira em um determinado caso de teste, uma falha é então sinalizada (SEGURA et al., 2016).

$$sin(x) = sin(\pi - x) \tag{2.13}$$

Outro exemplo de caso de teste metamórfico é a realização da pesquisa das sentenças "*leakless dearer*" e "*leakless dearer negative*" no Google Busca[4], em que a pesquisa pela primeira sentença retorna menos dados que a pesquisa realizada para a segunda sentença. De acordo com as especificações do Google – na época em que o teste foi realizado –, qualquer espaço entre palavras significa uma condição *AND* para o mecanismo de pesquisa, ou seja, as páginas retornadas na pesquisa devem conter todas as palavras inseridas na busca em sua composição, porém como a quantidade de páginas retornadas na pesquisa da segunda sentença – que contém 3 palavras – é maior que a quantidade da pesquisa da primeira sentença, há um erro de sistema, visto que de acordo com as especificações, só seria possível um número igual ou menor que a quantidade de páginas nos resultados da pesquisa da primeira sentença, já que ela contém menos palavras (ZHOU et al., 2012). O fato de haver um termo a mais na segunda sentença pesquisada ("*negative*"), caracteriza como uma alteração do caso de teste, logo, uma transformação metamórfica.

Para a detecção de falhas por parte do teste metamórfico, uma característica importante que deve ser definida é o tipo de comparação entre as saídas para as execuções com os dados originais e transformados. Instintivamente uma comparação de igualdade de valores é preferível a uma comparação de não igualdade, ou seja, a comparação dos resultados originais com os valores obtidos após as transformações metamórficas devem ter valores semelhantes. Este tipo de comparação é comumente utilizado pois uma expressão de igualdade é mais simples de definir se está certa ou errada de acordo com os valores recebidos, porém é mais suscetível a falhas do que uma comparação de não-igualdade (XIE et al., 2011).

Para o presente trabalho são adotadas comparações de não-igualdade, visto que a partir da utilização de transformações metamórficas nos dados, busca-se a melhoria

[4]Mecanismo de pesquisas na internet sobre qualquer assunto ou conteúdo.

de resultados métricos em comparação com os resultados quando utilizando os dados originais, ou seja, o valor resultante dos dados transformados deve ser diferente do valor resultante dos dados originais. Sendo assim, as transformações metamórficas tem um papel fundamental neste trabalho que sobressai em relação ao teste metamórfico tradicional como um todo, com a comparação de igualdade de valores sendo preterida.

O ponto chave dos testes metamórficos é saber quais transformações aplicar sobre os dados para que seja possível obter resultados satisfatórios. Tais transformações são elaboradas de acordo com as relações metamórficas, estas sendo construídas de diversas formas que requer, geralmente, um conhecimento do domínio do problema. Outros fatores que destacam boas relações metamórficas são aquelas que tornam o caso de teste o mais diferente possível da informação original (SEGURA et al., 2016) e também relações derivadas de partes específicas dos modelos, estas se saindo melhor quando comparadas com relações que envolvem todo o sistema, em certos casos (JUST; SCHWEIGGERT, 2011).

Buscando uma forma mais organizada de entender a construção de relações metamórficas para diferentes aplicações, os autores do trabalho realizado em (MURPHY; KAISER; HU, 2008) foram capazes de categorizar as relações para aplicações de aprendizagem de máquina supervisionada, utilizando os modelos de classificação *Martingale Boosting* e *Support Vector Machine (SVM)*, e aplicações não supervisionadas, utilizando um sistema de detecção de intrusão baseado em anomalias, intitulado *PAYL*. No total os autores categorizaram as relações metamórficas em seis tipos, sendo elas:

- Aditiva: somar o valor dos atributos com um valor constante;
- Multiplicativa: multiplicar o valor dos atributos por um valor constante;
- Permutativa: alteração na ordem dos atributos e também das linhas de informações para cada amostra;
- Inversa: multiplicação do valor dos atributos por um valor constante negativo, obtendo o oposto dos resultados originais;
- Inclusiva: acrescentando um novo elemento dentro dos dados utilizados;

- Exclusiva: retirar um elemento pertencente aos dados originais utilizados, com um resultado para os demais dados já esperados com tal retirada.

O propósito ao aplicar as transformações metamórficas seguindo as relações acima é obter resultados iguais entre modelos treinados com dados originais e com dados transformados, ou resultados na mesma ordem. Para as relações aditiva, multiplicativa e permutativa, as transformações não devem afetar o resultado final, visto que realizar tais operações em todos os valores de um ou mais atributos utilizando a mesma constante quando positiva não afeta a relação entre si dos dados, além disso, alterar a ordem dos valores ou atributos de treino também não causa impacto na classificação final, desde que se mantenham as mesmas informações. Para as demais relações a constatação se dá na ordem dos resultados, em que na relação inversa a ordem passa de crescente para decrescente ou vice-versa, e nas relações inclusiva e exclusiva, a ordem varia de acordo com a inserção/retirada de elementos a serem utilizados (MURPHY; KAISER; HU, 2008).

Tendo como base os seis tipos de relações metamórficas categorizados em 2008, uma nova pesquisa – incluindo autores do estudo anterior – foi capaz de criar categorias adicionais para as relações, voltadas à modelos de aprendizagem supervisionada, e idealizadas a partir de comportamentos gerais dos algoritmos estudados, sendo o *K-Nearest Neighbors (KNN)* e o *Naive Bayes (NB)*. Embora não tão distintas de algumas já apresentadas, as novas categorias das relações serviram também como uma forma de especificação em relação as anteriores, que foram apresentadas de forma mais abrangente (XIE et al., 2011). As novas categorias das relações metamórficas se diferenciam da seguinte forma:

- Consistência com transformação afim: o resultado deve ser o mesmo ao aplicar uma função de transformação afim arbitrária nos valores;
- Permutação de rótulos de classe: troca apenas dos rótulos dos dados, por exemplo em um cenário de classificação binária, valores que são 0 passam a ser 1 e vice-versa;
- Permutação de atributos: alteração na ordem dos atributos dispostos no conjunto de dados;

- Adição de atributos não informativos: um novo atributo que possua relação apenas com a classe do respectivo dado;
- Adição de atributos informativos: assim como a relação anterior, a criação de um novo atributo que possui relação com a classe do respectivo dado, mas também que tenha cerca concordância com as demais classes;
- Consistência com retreino: se dá ao adicionar o resultado de uma linha como um novo dado de treinamento para o conjunto com as transformações;
- Amostra adicional de treino: duplicação de todas as amostras de uma determinada classe do conjunto de dados de treino;
- Adição de classes através de duplicação de amostras: assim como a relação anterior, é feita uma duplicação de amostras, porém é realizada em todas as classes exceto uma, com as amostras duplicadas pertencentes as novas classes baseadas nas classes originais;
- Adição de classes através de novos rótulos: como uma variação da relação anterior, novas classes são criadas nos dados, porém sem uma duplicidade de dados, apenas alterando o rótulo de dados já existentes;
- Remoção de classes: realização da retirada de todas as amostras pertencentes a uma determinada classe do conjunto de dados;
- Remoção de amostras: da mesma forma que a relação de remoção de classes, há a retirada de amostras de uma ou mais classes, porém sem retirar todos as amostras da classe.

2.4 Considerações Finais do Capítulo

Passando por temas base para o entendimento deste trabalho, foi possível compreender a classificação de dados de forma automática, neste caso a atividade a ser realizada no estudo, através de algoritmos de aprendizagem de máquina. O foco da realização da atividade de classificação é então a garantia de melhores valores de *fairness* nos

resultados obtidos, representando que os modelos possam classificar dados de forma mais imparcial dentre os grupos sociais trabalhados. Por fim, tem-se o método que busca a melhora na garantia de *fairness*, sendo as transformações metamórficas a serem aplicadas nos dados, que ao contrário da forma como os testes metamórficos geralmente se apresentam na literatura, busca-se uma diferença de resultados entre as execuções de dados originais com dados transformados. É esperado que este capítulo auxilie no processo de entendimento do conhecimento técnico e teórico, deixando mais claro como os temas se relacionam durante o desenvolvimento do estudo.

3 Trabalhos Relacionados

O estudo de testes metamórficos como suavização do problema do oráculo em atividades de aprendizagem de máquina já se encontra bem difundido no setor acadêmico, assim como também trabalhos que utilizam transformações para verificar o quão imparciais os modelos de classificação podem ser. Entretanto, ao pegar as transformações metamórficas como forma de aprimoramento dos índices de *fairness* nos resultados dos modelos, a quantidade de trabalhos sobre tal tema se mostra escassa na literatura. Sendo assim, pesquisas que utilizam transformações para a validação de modelos e pesquisas que visam a medição de *fairness* em modelos de classificação, se fazem úteis na compreensão de como a estratégia (transformações metamórficas) pode ajudar na diminuição do problema (parcialidade na classificação de dados) proposto no presente trabalho.

3.1 Transformações Metamórficas para Validação de Modelos

Através dos testes metamórficos, as transformações são aplicadas nos conjuntos de dados para que possam ser executados nos modelos, para a realização de comparações entre resultados com os dados originais e transformados. Geralmente, a validação se dá através da igualdade de resultados entre as duas execuções, mostrando que mesmo com um comportamento adverso, o modelo deve ser capaz de manter a consistência na realização de suas tarefas.

Embora já citados anteriormente, é importante trazer para este capítulo os trabalhos

realizados em (MURPHY; KAISER; HU, 2008; XIE et al., 2011), que solidificaram os tipos de relações metamórficas na literatura para as aprendizagens supervisionada e não supervisionada. No primeiro trabalho, através de um conhecimento prévio dos algoritmos *MartiRank* (implementação alternativa do algoritmo *Martingale Boosting*) e *SVM*, os autores foram capazes de criar seis tipos de relações metamórficas de acordo com as entradas e saídas dos dados, como uma alternativa de testar tais aplicações na ausência de um oráculo. Já em execuções iniciais, a transformação do tipo inversa (que transforma um número positivo em negativo) indicou um erro de implementação no *MartiRank*, que não estava apto para trabalhar com valores negativos, enquanto que para o *SVM*, os resultados se encontravam em limites dentro do planejado, seja em questão de deslocamento, expansão ou inversão de valores.

O segundo trabalho, realizado em (XIE et al., 2011), serviu de expansão para a pesquisa anterior, em que foi possível a criação de novas categorias de relações metamórficas para modelos de aprendizagem supervisionada, com os testes sendo realizados nos modelos *K-Nearest Neighbors* e *Naive Bayes*. Dentre as novas categorias de relações metamórficas, os autores foram capazes de observar a adição de atributos e classes nos dados, como também a utilização de amostras adicionais e consistência de informações com retreino. Com as execuções foi possível perceber inconsistências nos dois modelos, sendo as relações de adição e remoção de classes as que foram capazes de revelar valores fora do limite esperado para ambos os algoritmos. Os autores concluem que a abordagem idealizada permite que usuários e desenvolvedores verifiquem e validem de maneira eficaz os componentes de aprendizagem sem a necessidade de um conhecimento sólido ou experiência em teste de *software*, porém deixam clara uma herança de limitação para este tipo de teste, em que se não houver uma violação revelada pelas relações metamórficas, não é possível concluir a correção nem a adequação do algoritmo investigado.

Continuando em testes voltados para o *SVM*, os autores do trabalho realizado em (NAKAJIMA; BUI, 2016) foram capazes de definir transformações metamórficas para serem utilizadas de duas formas: 1) como pseudo-oráculo para a comparação dos resultados; e 2) para o aumento da população do conjunto de dados de entrada. Dentre as transformações, são aplicadas a reordenação de dados e atributos, troca

de classes e valores de atributos e adição de ruído, todas baseadas nas categorias do trabalho de (MURPHY; KAISER; HU, 2008). Porém, o foco dos autores não é definir novas transformações e sim uma metodologia sistemática para a obtenção destas que possam ser aplicadas em modelos *SVM*, atingindo tal objetivo através da inspeção de componentes, funções e procedimentos necessários para o funcionamento do próprio modelo.

Entrando no cenário de reconhecimento de informações em imagens, o trabalho de (DING; KANG; HU, 2017) utiliza uma abordagem de validação metamórfica, como os autores denominam, para analisar a precisão de uma estrutura de aprendizagem profunda que inclui uma rede neural convolucional. O cenário de desenvolvimento da pesquisa é a análise de imagens de células biológicas, com cerca de 7500 imagens em todo o conjunto de dados, referentes a células com estruturas intactas, detritos celulares e pequenas partículas, porém, para a execução do modelo de aprendizagem profunda foi necessário o corte das imagens em tamanhos menores, chegando a mais de 420 mil imagens. A primeira relação metamórfica empregada pelos autores sai do padrão de transformações metamórficas sobre os dados, sendo ela: a execução da rede neural deve gerar resultados de classificação melhores que quando exercendo a mesma análise de imagens utilizando um modelo *SVM*, atingindo tal objetivo apesar da necessidade do redimensionamento das imagens. As demais relações metamórficas, dez no total, são todas transformações empregadas nos dados, como adição, duplicação e remoção de amostras e classes, mas sempre com o intuito de ou não afetar a precisão da classificação, ou obter valores próximos nas execuções. Por fim, a validação metamórfica acaba se mostrando como razoável para a garantia de qualidade em outras estruturas de aprendizagem de máquina, porém se saindo melhor na aprendizagem profunda devido aos algoritmos complexos e a necessidade de uma grande quantidade de dados.

Utilizando uma instância de um modelo *KNN*, a pesquisa feita em (SANTOS et al., 2020a) verifica se as relações metamórficas do trabalho de (XIE et al., 2011) podem ser aplicadas de forma eficiente para garantir a qualidade da tarefa de classificação de informações relacionadas à diagnósticos de câncer de mama. Para cada relação metamórfica, onze no total, os autores realizaram uma execução com todas as amostras do conjunto de dados, além da execução com os dados sem qualquer transformação, e

utilizando a acurácia como métrica de avaliação foi possível visualizar que a relação metamórfica que remove amostras de uma classe foi a que obteve o maior valor métrico, significando na melhor relação – para este experimento em específico – capaz de detectar falhas no modelo de aprendizagem, uma vez que altos valores de acurácia podem significar em um *overfiting* do modelo. Em suas conclusões os autores consideram que as relações metamórficas empregadas são "de grande valia" para o processo de testagem de sistemas de aprendizagem, mas ressaltam que o cenário estudado não possibilita uma grande generalização para demais tipos de aplicações e dados de outros domínios de estudo.

Trazendo as transformações para validação de modelos em um contexto recente, tem-se o estudo realizado em (MA; PAN; FAN, 2022) utilizando as transformações metamórficas para testar modelos de classificação de diagnóstico inteligente da COVID-19. A atividade consiste na utilização de imagens de tomografias computadorizadas de pulmões de pacientes, em que a partir da presença de lesões pulmonares nas imagens é possível que modelos de aprendizagem possam classificar um diagnóstico da COVID-19 em quatro diferentes níveis: leve, comum, pesado ou crítico. Ao todo são implementadas oito transformações metamórficas nas imagens, consistindo em alterações como o aumento de brilho da área da lesão pulmonar, melhora do contorno da textura da lesão, inserção de áreas com lesões em imagens ou partes de imagens que não contém e também a remoção de lesões, além também de transformações mais simples como a rotação ou espelhamento das imagens. Após as execuções são utilizadas algumas métricas na comparação dos valores obtidos, como *Accuracy*, *Sensitivity*, *Precision*, *F1-Score*, com ênfase maior para a métrica *Specificity*, em que quanto menor o seu valor, maior a diferença de inconsistência dos resultados entre os dados originais e transformados, e sendo assim, as transformações de melhora do contorno de texto da lesão, inserção de áreas com lesões e o espelhamento das imagens, são as que mais indicaram inconsistência nos resultados. Por fim, os autores do trabalho se mostraram satisfeitos com os resultados obtidos, em que as transformações de fato puderam indicar diferenças consideráveis nas classificações para o contexto empregado, apesar da fraca capacidade de generalização das transformações em imagens de diferentes tipos de doenças, sendo um ponto idealizado para melhorias futuras.

Continuando no domínio da classificação de imagens, o trabalho realizado em (TORIKOSHI; NISHI; TAKAHASHI, 2023) determina a aplicação das transformações metamórficas de forma um pouco sofisticada e não como geralmente são realizadas na literatura. Os autores argumentam que métodos de transformações em imagens como inserção de pontos pretos, alterações de brilho ou adição de ruído, quando utilizados, são com o propósito de testar a confiabilidade dos modelos, enquanto que para testar a segurança dos mesmos é possível através de transformações em regiões com alta probabilidade de reconhecimento incorreto. Sendo assim, são utilizadas técnicas de inteligência artificial para reconhecer as regiões determinantes para uma classificação da imagem, e com isto aplicar as transformações nestas regiões específicas com o intuito de verificar a ocorrência de classificações errôneas. As transformações utilizadas variam entre inversão de cores, pontos pretos, alteração de brilho, desfoque de região e adição de ruído gaussiano, porém limitadas em regiões de 10x10 *pixels*. Ao fim das execuções, os resultados métricos se mostraram satisfatórios, em que ao utilizar uma taxa de detecção de falhas, foi possível visualizar que com as transformações, há um índice maior de falhas encontradas com uma proporção de quase duas vezes mais que imagens com transformações sem o reconhecimento de regiões.

Embora o foco do estudo realizado (ZHANG; TOWEY; PIKE, 2023) não seja o de validar modelos de classificação com as transformações, é importante citá-lo devido ao contexto atual em que ele é empregado, com a utilização do ChatGPT para a construção de transformações metamórficas de forma automática, para o teste de sistemas de direção autônoma. Utilizando um módulo de estacionamento dos sistemas de direção autônoma como cenário de aplicação, são solicitadas variadas transformações metamórficas ao ChatGPT que possam ser aplicadas neste cenário, em seguida são levantados alguns critérios para uma avaliação de qualidade das transformações obtidas, como: correção, aplicabilidade, novidade, clareza, relevância para a segurança, utilidade geral e viabilidade computacional. Em um primeiro contato com o ChatGPT, ele foi capaz de definir apenas cenários de testes com descrições de ambientes em que um veículo hipotético poderia estar e o que deveria fazer, sem qualquer menção ou relação a transformações metamórficas de dados. Devido aos erros encontrados, os autores realizaram novas interações com a plataforma incluindo *feedbacks* das suas respostas,

em que foi possível a obtenção de respostas alinhadas com transformações metamórficas de fato, visto que o ChatGPT é capaz de aprender e melhorar em respostas futuras com base no apontamento de erros e informações mais diretas de acordo com os cenários indicados a ele. Como conclusão, os autores se mostraram satisfeitos com o estudo realizado, sugerindo que a plataforma pode ser uma abordagem econômica para a idealização de transformações metamórficas de forma automática, reduzindo o esforço manual necessário para a mesma atividade de idealização de transformações.

3.2 Mensuração de *Fairness* em Classificações

Trazendo as transformações metamórficas para o âmbito de *fairness* em classificação automática de informações, é possível adotá-las com a finalidade de avaliar o quanto os modelos estão sendo imparciais através da comparação dos resultados nas execuções com e sem transformações metamórficas, sendo este o intuito predominante em trabalhos acadêmicos que estudam a união destas duas áreas. Se faz necessária então uma compreensão de como as pesquisas são realizadas visando a medição de índices de *fairness* à cerca dos modelos de classificação de dados.

O foco do trabalho encontrado em (ADEBAYO; KAGAL, 2016) é o desenvolvimento de uma metodologia que seja capaz de determinar a dependência de entradas de dados para um modelo de caixa-preta (como são referenciados os modelos em que humanamente não é possível obter uma compreensão exata do seu funcionamento como base em entradas e saídas), porém não é especificado o algoritmo utilizado, apenas a utilização deste termo para mostrar que os autores não querem ter conhecimento sobre o que acontece por dentro do modelo utilizado. Tal metodologia se aproveita de transformações para a criação de várias cópias dos dados de entrada a fim de determinar a dependência do modelo com as informações recebidas, isto é, os autores tem a ideia de que todos os atributos de uma determinada entrada sejam transformados de acordo com um atributo de interesse, este sendo o gênero de pessoas no cenário de liberação de limite de crédito de um banco. As execuções de teste se dão em três diferentes algoritmos de classificação, e apesar da não divulgação de quais algoritmos foram os escolhidos, os resultados mostram que a dependência dos algoritmos em relação a identidade de gênero

é consistentemente baixa, indicando que o algoritmo não depende excessivamente em saber o gênero de uma pessoa para determinar seu limite de crédito.

Além de apresentar um conjunto de definições de *fairness* da literatura até então, o trabalho de (GALHOTRA; BRUN; MELIOU, 2017) desenvolve uma aplicação baseada em testes para medir se e quanto um modelo é discriminatório em suas classificações com foco na causalidade do comportamento discriminatório, recebendo o nome de *Themis*. Para a utilização da ferramenta, é necessário fornecer o modelo a ser executado, um nível de confiança desejado, um limite de erro aceitável e um esquema de entradas válidas, que com estas informações o *Themis* é capaz de gerar um conjunto de testes para calcular uma pontuação de discriminação causal, isto é, ele é capaz de verificar se, e quanto, um modelo discrimina raça e idade, por exemplo. Para avaliar a ferramenta, são executadas diversas instâncias de modelos de classificação (Regressão Logística, Árvore de Decisão e *Naive Bayes*) e também diversos conjuntos de dados, em cenários de censo demográfico e aprovação de limite de crédito. Após as execuções, os autores concluem que o *Themis* é de fato capaz de verificar se muitas das instâncias dos modelos discriminam raça e gênero, como por exemplo em uma das execuções a aplicação mostrou que para 77% das amostras passadas, alterar apenas o sexo, estado civil ou raça faz com que a saída do modelo seja invertida.

Semelhante a ideia do *Themis*, o trabalho de (SHARMA; WEHRHEIM, 2019) define uma abordagem de teste capaz de gerar alterações em conjuntos de dados para que possam ser executados em modelos de classificação. Como não há um oráculo para a tarefa de classificação de dados através de modelos de aprendizagem, os autores se guiam em buscar com que os modelos aprendam o que realmente há nos dados, chamando esse conceito de "*Balanced Data Usage*" (ou Uso Balanceado de Dados, em português), que consiste em desconsiderar aspectos dos dados, como nomes escolhidos para os atributos ou ordens de instâncias de dados durante o aprendizado. Intitulada *Tile*, a abordagem consiste na realização de até quatro transformações metamórficas, sendo elas a permutação de linhas e de colunas, embaralhamento do nome de atributos e transformação de atributos categóricos de texto para valores numéricos. Para avaliar a abordagem criada, são utilizados nove conjuntos de dados do mundo real, que incluem cenários de censo demográfico, ocupação, exames de câncer, liberação de crédito, entre

outros, e instâncias de quatorze modelos de classificação de dados, estes os mais utilizados na literatura para tarefas de aprendizagem supervisionada como Árvore de Decisão, Floresta Aleatória, Regressão Logística, etc. Após as execuções é possível observar que apenas em dois modelos não foi possível observar um desbalanceamento (ao comparar as classificações de um modelo treinado com dados originais e uma instância do mesmo modelo treinada com os dados transformados), o *Naive Bayes* e o *Gaussian*, enquanto que as transformações que mais evidenciam um desbalanceamento nas classificações são as permutações de linhas e colunas.

Ainda sobre o trabalho de (SHARMA; WEHRHEIM, 2019), os autores informam algumas ameaças à validade da pesquisa, sendo a aleatoriedade computacional como a principal delas, uma vez que ocorre tanto na divisão das amostras dos conjuntos de dados para treino e teste, como também nos cálculos realizados pelos modelos, impossibilitando de certa forma a replicação de toda a pesquisa. Outro fator importante que não foi apresentado no artigo é que as transformações de permutação de linhas e a alteração de valores textuais categóricos para numéricos são atividades já padronizadas em métodos de testagem para alguns modelos de classificação.

Uma área em que também é possível a utilização de transformações metamórficas para verificar preconceitos em classificações, é a de Análise de Sentimentos, sendo possível a citação do estudo realizado em (ASYROFI et al., 2021). Nele, é desenvolvida a abordagem *BiasFinder* através de técnicas de Processamento de Linguagem Natural, com a capacidade de identificar e transformar palavras associadas a classes de categorias específicas, como por exemplo artigos e pronomes relacionados ao gênero feminino, como sendo o grupo sensível em comparação com o gênero masculino. Além das transformações, o *BiasFinder* é capaz também de utilizar os textos originais e transformados como entrada de modelos quaisquer de análise de sentimento para a realização de comparações dos resultados, em que quando há classificações diferentes entre os textos, são apresentados como uma evidência de uma classificação tendenciosa.

Os autores do estudo realizado em (JOHNSON; BRUN, 2022) foram capazes de desenvolver a plataforma *Fairkit-learn*, capaz de acompanhar e avaliar treinamentos e classificações de modelos com relação a índices de *fairness* e demais métricas de qualidade.

Utilizando os pacotes *Scikit-Learn*[1] e *AIF360*[2], o *Fairkit-learn* oferece suporte a todos os algoritmos de aprendizagem disponibilizados pelos dois pacotes citados, além da capacidade de trabalhar com mais de 70 métricas de *fairness*, de acordo com os autores. Para a utilização da plataforma, é possível selecionar até cinco entradas em uma interface gráfica: modelos de aprendizagem, métricas de avaliação, hiperparâmetros de modelos, limites de classificação e algoritmos de pré e pós-processamento. Para a avaliação da plataforma os autores indicam a realização de testes com 54 estudantes com experiência variada em modelos de aprendizagem de máquina, que apesar de não apresentarem valores concretos de tais testes, afirmam que a mesma se mostrou satisfatória para os usuários ao encontrarem combinações resultando em modelos de alto desempenho e geralmente mais justos em relação aos modelos utilizando os pacotes *Scikit-Learn* e *AIF360* em suas originalidades.

3.3 Garantia de *Fairness* com Transformações Metamórficas

Como abordado no início do Capítulo, trabalhos que visam a utilização de Transformações Metamórficas para a diminuição de índices de preconceito são escassos no ambiente acadêmico, com o maior destaque na área de *fairness* em modelos de classificação sendo a medição do quanto os modelos podem ser discriminatórios em seus resultados. Entretanto, é possível destacar trabalhos realizados entre os anos de 2009 e 2010 que são precursores na garantia de *fairness* ao realizar alterações nos dados e também nos próprios modelos, ou seja, aumentando os índices de imparcialidade dos resultados dos modelos com as alterações efetuadas.

Embora não seja explícita a utilização de transformações baseadas nas categorias apresentadas no Capítulo 2, a pesquisa observada em (KAMIRAN; CALDERS, 2009) realiza o método de "massagear" o conjunto de dados fazendo modificações menos intrusivas possíveis, com o propósito de alcançar um conjunto de dados livre de discri-

[1]Pacote de ferramentas de Aprendizagem de Máquina em *Python*: <https://scikit-learn.org/stable/index.html>.

[2]Pacote de ferramentas para análise de *Fairness* em *Python*: <https://aif360.res.ibm.com/>.

minação para ser executado em um modelo *Naive Bayes*. O massageamento dos dados – com informações de aprovação de crédito para pessoas – consiste na alteração das classes das amostras de acordo com o atributo sensível, este sendo a categoria da idade da pessoa: jovem ou idoso, realizando sempre de forma alternada e consequentemente resultando em uma quantidade de amostras balanceada entre as duas classes, sendo esta a definição de um "conjunto livre de discriminação" por parte dos autores. Executando dois modelos *NB*, um com o conjunto de dados original e outro com o conjunto alterado, são utilizadas uma métrica de discriminação idealizada na própria pesquisa e a acurácia para comparação de resultados, em que o modelo treinado com os dados alterados consegue atingir os menores valores de discriminação e também valores de acurácia próximos ou até maiores em relação ao modelo treinado com os dados originais, o que era uma preocupação dos autores, visto que a alteração de dados de treinamento pode resultar em valores métricos mais baixos em comparação com execuções originais.

A pesquisa do parágrafo anterior foi ampliada pelos autores e apresentada em (KAMIRAN; CALDERS; PECHENIZKIY, 2010), porém desta vez as alterações são aplicadas no próprio modelo utilizado de uma árvore de decisão e não no conjunto de dados. É feita uma incorporação de consciência de discriminação no processo de construção do modelo através de duas técnicas: 1) construção do modelo com reconhecimento de dependência, em que no critério de divisão para um nó das árvores, é avaliado o nível de discriminação causado por essa divisão, e não apenas a contribuição para a precisão; e 2) nova rotulação para as folhas das árvores, em que normalmente os rótulos das folhas são determinados pela classe majoritária referente as tuplas das folhas, sendo assim a rotulação é alterada de forma que a discriminação seja reduzida com uma perda mínima na precisão. Colocando os resultados em evidência, os autores argumentam que a nova pesquisa se sobressai em relação às anteriores tanto em valores de discriminação quanto de acurácia, porém é importante ressaltar que são utilizados modelos diferentes nestas comparações (*NB* na pesquisa anterior e Árvore de Decisão na atual), tornando a comparação não tão confiável.

Utilizando um modelo *Naive Bayes* para a atividade de classificação, ao expôr que apenas removendo a coluna de dados com informações sensíveis que neste caso específico, a que representa o gênero das pessoas, não resolve o problema da parcialidade

no modelo, os autores do trabalho realizado em (CALDERS; VERWER, 2010) propõem três abordagens visando a resolução da discriminação nas classificações: 1) alteração de parâmetros internos do modelo para impactar na probabilidade de decisão positiva nos resultados; 2) criação de um modelo para cada classe do atributo sensível seguido de um equilíbrio entre os modelos treinados; e 3) adição de uma variável imparcial e livre de discriminação. Com os resultados foi possível observar que a abordagem criando um modelo de classificação para cada classe do atributo sensível se saiu melhor em relação as demais, atingindo bons valores tanto de acurácia quanto de discriminação, com pouca variação nos resultados.

Um estudo recente que visa uma diminuição da parcialidade através de transformações metamórficas é o realizado em (KHOO et al., 2023), desenvolvido no âmbito da Análise de Sentimentos. Utilizando um conjunto de dados de críticas de filmes, as transformações nos dados são realizadas pelo *BiasFinder* (abordagem desenvolvida em (ASYROFI et al., 2021)), através da alteração de palavras relacionadas aos gêneros masculinos e femininos quando encontrados, para que as críticas tanto originais quanto transformadas possam ser utilizadas em modelos de classificação com redes neurais em suas estruturas. Após as classificações iniciais, os sentimentos classificados são comparados entre os textos originais e transformados, para a identificação de violações de *fairness* (quando um texto original e transformado possui classificações diferentes). Os textos transformados são utilizados em novas etapas de retreino dos modelos com diferentes quantidades de amostras utilizadas, sendo possível a diminuição na ocorrência de violações de *fairness* na medida em que mais amostras são utilizadas no retreino, de acordo com os autores.

3.4 Considerações Finais do Capítulo

Tendo um vislumbre da literatura que percorre as áreas de *fairness* e transformações metamórficas, fica claro que o foco das pesquisas são a utilização das transformações como método de avaliação de imparcialidade sob as classificações dos modelos de aprendizagem, enquanto que trabalhos que utilizem destas técnicas visando de fato a diminuição da imparcialidade nos modelos se mostram bem limitados e em quantidade

escassa, o que pode levantar o questionamento se de fato tais abordagens são realmente capazes de melhorar os índices de *fairness* para futuras classificações dos modelos, porém este logo é encerrado uma vez que foram discutidos trabalhos que tiveram impacto de forma positiva em seus resultados. Outro fator importante que pôde ser visto nas discussões dos trabalhos é a forte dependência à acurácia como forma de avaliação, que por ser uma métrica de classificação simples, pode não expressar com rigor o quanto os trabalhos podem ser robustos ou limitados.

São com estas restrições que a presente pesquisa desenvolve um novo formato de adoção das transformações metamórficas em um conjunto de dados, para verificar se é possível de fato fazer com que os modelos de aprendizagem possam ser mais imparciais em suas classificações, além de utilizar diferentes métricas de avaliação a fim de se obter as análises mais acuradas possíveis, evitando a forte dependência em um único ou poucos indicadores de eficácia.

4 Estudo

Com a apresentação de conceitos fundamentais para a compreensão do trabalho, bem como a apresentação do estado em que a literatura referente a Transformações Metamórficas e *Fairness* em Classificação de Dados se encontra até então, o presente Capítulo é responsável por apresentar toda a estrutura do estudo desenvolvido, iniciando através de uma descrição de forma resumida que visa a melhor compreensão possível sobre o que é feito do início ao fim no estudo. Após o entendimento geral, são apresentadas as Questões de Pesquisa que o estudo busca responder e também os principais componentes utilizados ao longo do desenvolvimento, tanto na parte dos dados trabalhados como também as ferramentas de classificação, métricas de cálculo e técnicas de reamostragem de valores para inferência de resultados.

4.1 Descrição

A presente pesquisa tem como objetivo verificar se a realização de Transformações Metamórficas dos valores presentes em um Conjunto de Dados, são capazes de aumentar a garantia de *fairness* em classificações de modelos de Aprendizagem de Máquina. Devido as limitações do estado da arte para o problema apresentado, como discutido no Capítulo anterior, a abordagem proposta neste trabalho busca agregar novas informações à comunidade acadêmica se é realmente possível, e também viável, que aplicando transformações metamórficas de forma aleatória em um conjunto de dados, seja realmente capaz de diminuir o viés dos modelos de aprendizagem de máquina, aumentando os índices de *fairness* na classificação de informações.

O desenvolvimento de todo o estudo se dá através da comparação de resultados de classificações de modelos de aprendizagem de máquina, porém, para que as comparações sejam possíveis, as classificações dos modelos treinados com o conjunto de dados original (apenas dividido de acordo com parâmetros de treino e teste) são postas em comparação com as classificações dos modelos treinados em quatro diferentes cenários de aplicação de transformações metamórficas, idealizados visando uma alteração nos valores das classificações e também com foco na informação sensível do conjunto de dados, sendo eles: 1) transformação de maiores valores: em que somente os valores maiores que a mediana do atributo sofrem a transformação metamórfica; 2) transformação de menores valores: embora seguindo a lógica do método anterior utilizando a mediana, desta vez os valores abaixo desta que recebem as transformações; 3) transformações de valores do grupo sensível: as transformações são aplicadas em todos os atributos, mas somente nas amostras que pertencem ao determinado grupo sensível do conjunto de dados utilizado; e 4) transformações de valores do grupo não sensível: seguindo a lógica do cenário 3 de aplicar as transformações em todas as amostras porém somente às que pertencem ao grupo não sensível. A utilização da mediana de valores para os dois primeiros cenários é preferida pela questão dos valores dos atributos serem em formato discreto e não contínuo, além de em alguns casos não estarem dispostos com a mesma razão de progressão.

Com as transformações realizadas no conjunto de dados, os resultados das classificações dos modelos são então comparados de acordo com o cálculo de métricas de classificação e de *fairness*, a fim da visualização de uma melhora ou piora das classificações dos modelos treinados com os valores transformados, em relação às classificações dos modelos treinados com os valores originais, sem qualquer alteração.

4.2 Questões de Pesquisa

Com um entendimento do projeto, é necessário visualizar o que se pretende obter ao final de sua execução, para isto, todo o planejamento do estudo é coordenado com o propósito de responder as seguintes Questões de Pesquisa:

- **Questão de Pesquisa 1 (QP1): O quão imparciais os modelos de apren-**

dizagem se mostram com a utilização de transformações metamórficas nos dados?

Os resultados das métricas de *fairness*, *Statistical Parity* e *Equalized Odds*, são cruciais para responder esta questão, visto que com elas é possível ter uma dimensão se um modelo está sendo parcial para um determinado grupo social ou não, isto comparando os modelos treinados com dados originais e transformados.

- **Questão de Pesquisa 2 (QP2): Quais os métodos de aplicação de transformação metamórfica que mais impactam na garantia de *fairness* nos modelos?**

  Com base na QP1 em que as métricas de *fairness* são importantes na resposta, para a segunda questão deve ser feita uma comparação destas entre os Cenários de Execução, resultando naquele(s) que pode(m) definir melhor as classificações imparciais.

- **Questão de Pesquisa 3 (QP3): É possível manter bons índices de eficácia com a aplicação das transformações metamórficas nos dados?**

  Em se tratando de eficácia, as métricas de classificação (*Accuracy*, *Balanced Accuracy*, *Precision*, *Recall* e *F1-Score*) são o ponto chave para esta questão, uma vez que com as transformações metamórficas realizadas na fase de treino, estas métricas devem expôr como é a diferença nos resultados das classificações em termos de erros e acertos com base nos valores reais, mostrando se os índices possuem uma melhora ou até piora em relação os resultados com o treino utilizando os dados originais.

- **Questão de Pesquisa 4 (QP4): Como se apresenta o custo/benefício da aplicação das transformações metamórficas pela melhora de imparcialidade dos modelos?**

  Para esta questão uma análise comparativa dos valores obtidos nos dois tipos de métricas, tanto de classificação quanto de *fairness* se faz necessária, além também do tempo de aplicação e execução das transformações metamórficas, realizando a discussão em todos os cenários de execução.

4.3 Componentes

4.3.1 Conjunto de Dados

O conjunto de dados utilizado na proposta é uma amostra do Dutch Virtual Census of 2001[1], que contém informações sociais e econômicas de pessoas residentes da Holanda, contabilizadas em 2001. O arquivo do conjunto de dados utilizado no estudo é o mesmo disponibilizado[2] pelos autores do trabalho realizado em (QUY et al., 2022), já no formato em *.CSV* que contém 12 atributos de 60420 pessoas, porém, no trabalho os autores não informam se foi realizado ou não algum tipo de pré-processamento nesse conjunto especificamente. Abaixo é possível verificar nome e breve descrição das informações (algumas não são detalhadas pois se distanciam do interesse do estudo), além dos valores presentes nas mesmas:

- *sex*: gênero biológico da pessoa, sendo possível masculino ou feminino. Para a presente proposta, essa informação específica é tratada como o atributo sensível no cálculo das métricas de *fairness*, visto que na amostra existem menos pessoas do sexo feminino com profissão de alto nível (9903 pessoas), em relação à quantidade de pessoas do sexo masculino (18860 pessoas). Valores existentes: 1 (sexo masculino) e 2 (sexo feminino);
- *age*: grupo de idade da pessoa, cada um representa uma faixa de 5 anos. Valores existentes: 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 e 15;
- *household_position*: relação com o(a) líder da moradia, seja cônjuge, filho(a), etc. Valores existentes: 1110, 1121, 1122, 1131, 1132, 1140, 1210, 1220;
- *household_size*: quantidade de integrantes no domicílio que a pessoa reside. Valores existentes: 111, 112, 113, 114, 125 e 126;
- *prev_residence_place*: local de residência da pessoa um ano antes de ser entrevistada, se na Holanda (valor 1) ou não (valor 2);

[1] O conjunto de dados pode ser encontrado em: <https://microdata.worldbank.org/index.php/catalog/2102>.

[2] O arquivo do conjunto de dados utilizado pode ser encontrado em: <https://github.com/tailequy/fairness_dataset/blob/main/experiments/data/dutch.csv>.

- *citizenship*: país de cidadania da pessoa, se de cidadania Holandesa (valor 1), em outro país da Europa (valor 2) ou de um país fora da Europa (valor 3);
- *country_birth*: nacionalidade da pessoa, se nascida na Holanda (valor 1), em outro país da Europa (valor 2) ou em um país fora da Europa (valor 3);
- *edu_level*: nível de escolaridade da pessoa, desde o pré-primário até o terciário, de acordo com o sistema educacional holandês. Valores existentes: 0, 1, 2, 3, 4 e 5;
- *economic_status*: status econômico da pessoa, variando entre pessoa desempregada (valor 111), empregada (valor 112) e aposentada (valor 120);
- *cur_eco_activity*: setor da atividade econômica atual da pessoa, como agricultura, mineração, construção, comunicação, administração pública, etc. Valores existentes: 111, 122, 124, 131, 132, 133, 134, 135, 136, 137, 138 e 139;
- *marital_status*: estado civil atual da pessoa, sendo solteira (valor 1), casada (valor 2), viúva (valor 3) ou divorciada (valor 4);
- *occupation*: o nível de profissão da pessoa, se é uma profissão de baixo ou alto nível. A ideia do presente estudo é treinar os modelos de classificação com os demais atributos para que possam definir uma classe de acordo com o atributo *occupation*, em que seus valores então classificados possam ser comparados com os valores originais para o atributo. Valores possíveis para o atributo: 0 (profissão de baixo nível) e 1 (profissão de alto nível).

Carregamento e Pré-Processamento de Dados

A partir da pesquisa realizada em (QUY et al., 2022), foi possível a escolha do conjunto de dados *Dutch Census 2001* devido a sua grande quantidade de informações disponíveis. Tal pesquisa foi elaborada pelos autores com o propósito de analisar as principais bases de dados utilizadas na literatura relacionada a atividade de reconhecimento de *fairness* em sistemas inteligentes, verificando o relacionamento entre os diversos atributos, com ênfase nos atributos sensíveis e de classe quando utilizando uma instância do algoritmo *Naive Bayes* nas classificações.

Embora nenhum método de limpeza ou correção de dados seja especificado por parte dos autores do trabalho, a disposição dos mesmos em arquivo se mostra correta, sem quaisquer amostras repetidas ou com ausência de informações por exemplo, como geralmente são problemas encontrados na fase de obtenção de dados para análise. É possível conferir a quantidade de amostras separadas por gênero, este sendo o atributo sensível, e nível de ocupação, a classe na qual a amostra pertence, na Tabela 4.1. Uma Análise Exploratória do conjunto de dados pode ser conferida no Apêndice A.

	Nível Alto	Nível Baixo	Total
Gênero Feminino	9903	20370	30273
Gênero Masculino	18860	11287	30147
Total	28763	31657	60420

Tabela 4.1: Quantidade de amostras por Gênero (Feminino e Masculino) e Nível de Ocupação (Alto e Baixo) do *Dutch Census 2001*.

Imediatamente após o carregamento do conjunto de dados, a técnica *Hold-out* de validação cruzada é empregada, que consiste na divisão dos dados em subconjuntos destinados para o treino e teste dos modelos sem a replicação de amostras, que para a pesquisa é utilizada uma proporção de 70%-30% respectivamente. Tal divisão do conjunto é realizada levando em conta apenas o número total de amostras, sem levar em consideração as quantidades de amostras para cada gênero ou nível de ocupação, sendo assim são separadas 42294 amostras para treino (70% dos dados) e 18126 para o teste dos modelos (30% dos dados), sendo o processo realizado de forma aleatória, sem um controle específico de em qual subconjunto, se de treino ou de teste, uma amostra deveria ficar.

4.3.2 Modelos de Classificação

Para a realização da classificação dos dados são utilizados nove modelos de aprendizagem de máquina no total, estes que são consolidados na literatura para atividades de diferentes cenários de classificação de informações, além disso, tal quantidade visa uma comparação ainda maior dos resultados, buscando uma melhor agregação de informações à literatura, embora que de forma inicial, sem estudar de forma aprofundada o funcionamento nas etapas de treino e teste de cada um dos modelos executados, logo, é importante ressaltar

que todos os modelos são utilizados em suas formas padrão, provenientes do pacote *Scikit-Learn* na versão 1.1.2 para a linguagem de programação *Python 3*.

Abaixo é possível entender inicialmente como cada modelo de aprendizagem de máquina funciona, dos nove sendo um modelo de classificação própria de dados, o *Decision Tree*, e os demais sendo modelos *Ensemble*, que utilizam modelos de classificação em sua estrutura interna. Descrições bem detalhadas dos modelos não são presentes pois se distancia do foco do estudo:

- *Decision Tree* ou Árvore de Decisão: cria uma árvore composta por nós (tomada de decisão com relação aos caminhos) e folhas (a classe da instância que está sendo analisada), com base na hipótese de que uma sequência de seleções ótimas locais levarão a uma solução ótima global. O ideal é a utilização dos atributos mais importantes primeiro na criação da árvore, ou seja, atributos que possuem o maior poder de classificação, em que tal processo é repetido para cada nó até que não seja mais possível o seu aprofundamento (MELO et al., 2016);

- Modelos *Ensemble* do tipo *Bagging* utilizados:

  - *Bagging Classifier*: constrói várias instâncias de um modelo (geralmente árvores de decisão, como o utilizado no estudo) para serem treinados com subconjuntos aleatórios gerados a partir do conjunto de dados original, incluindo a repetição de amostras, em seguida, a classificação final é dada através do voto majoritário das classificações das instâncias dos modelos criados. É utilizado como forma de reduzir a variância de um modelo, introduzindo a aleatoriedade da seleção de amostras do conjunto de dados em seu procedimento de construção (BREIMAN, 1996);

  - *Extra Trees*: embora se assemelhe com o *Bagging Classifier* na criação de instâncias de um modelo para o treino e classificação das amostras, o *Extra Trees* utiliza todo o conjunto de dados original para a realização do treino e classificação das instâncias do modelo base, o Árvore de Decisão, sem seleção aleatória e replicação de amostras para cada instância. Outra característica das árvores criadas pelo modelo é o ponto de criação dos nós de decisão, escolhidos de forma aleatória (AZNAR, 2020);

  - *Random Forest*: busca minimizar o ajuste em excesso do modelo aos dados de treino, causando a perda da capacidade de generalização, conhecido como *overfitting*. A ideia deste modelo é, assim como o modelo *Bagging Classifier*, utilizar o conjunto de dados de treino para gerar diversos outros conjuntos com o mesmo tamanho do original, a serem utilizados na criação de diversas árvores de decisão, porém com o acréscimo da utilização de diferentes combinações dos atributos do conjunto de dados nos nós das árvores, sendo cada novo conjunto para uma nova árvore distinta (MELO et al., 2016).

- Modelos *Ensemble* do tipo *Boosting* utilizados:

  - *Adaptive Boosting*: trabalha através da atribuição de pesos para cada uma das amostras utilizadas no treinamento. Criando um modelo simples por vez, os pesos das amostras são modificados individualmente, em que quando amostras são classificadas de forma incorreta pelo modelo da etapa anterior, tem seus pesos aumentados para a próxima etapa de classificação, já as amostras que são classificadas corretamente, tem seus pesos diminuídos. À cada nova etapa de criação de uma nova instância do modelo para classificação, as amostras que possuem maior erro de classificação possuem os maiores pesos, e consequentemente maior concentração pelas instâncias nas etapas futuras. A classificação final de uma amostra é dada através de uma votação majoritária das classificações dos modelos simples criados (HASTIE et al., 2009);

  - *Gradient Boosting*: funciona geralmente construindo árvores de decisão de forma sequencial, em que cada nova árvore criada tenta corrigir erros da anterior através da remoção de ramificações que falham na redução de perda, tornando as árvores com baixa profundidade, impactando em um consumo menor de memória e tornando as classificações mais rápidas. A ideia de classificação de dados do modelo se dá através da combinação de modelos simples, que como cada árvore fornece boas classificações somente em parte dos dados, novas árvores são criadas para melhorar o desempenho de forma iterativa (MÜLLER; GUIDO, 2016);

  - *Histogram-based Gradient Boosting*: podendo ser mais rápido que o *Gradient Boosting* para conjuntos com grande quantidade de amostras, o modelo agrupa as amostras em compartimentos de valor inteiro (normalmente 256 compartimentos), reduzindo o número de pontos de divisão a serem considerados. É permitido também que o modelo aproveite estruturas de dados baseadas em números inteiros, evitando a dependência de valores contínuos classificados ao construir as árvores (KE et al., 2017).

- Modelo *Ensemble* do tipo *Stacking* utilizado:

  - *Stacking Classifier*: consiste no empilhamento de modelos individuais, em que a saída de um modelo é utilizada como entrada para o próximo modelo a ser executado, até que o último modelo empilhado possa realizar a classificação final da amostra, visando a redução do viés de cada modelo (WOLPERT, 1992). Para a execução do modelo na pesquisa, são empilhadas instâncias dos modelos *Decision Tree* e *Random Forest*, enquanto o modelo final é uma instância do modelo *Gradient Boosting*.

- Modelo *Ensemble* do tipo *Voting* utilizado:

  - *Voting Classifier*: a ideia por trás do modelo é combinar modelos de aprendizagem de máquina conceitualmente diferentes e usar um voto majoritário como sendo a classificação final de uma determinada amostra, podendo ser útil para um conjunto de modelos com desempenho igualmente favorável, a fim de equilibrar as suas fraquezas individuais (SANTOS, 2022). Assim como o *Stacking Classifier*, foram utilizados os modelos *Decision Tree*, *Random Forest* e *Gradient Boosting* no *Voting Classifier*, visto que são modelos de diferentes categorias de aprendizagem.

4.3.3 Transformações Metamórficas

Antes do envio das amostras para os modelos realizarem as classificações, uma cópia dos dados a serem utilizados no treino dos modelos é feita para que possam ser aplicadas as

Transformações Metamórficas, sendo assim, passam a existir dois conjuntos de treino, o original e o transformado, a cada execução das operações de treino e teste dos modelos.

Uma vez que não se encontram na literatura uma ou mais transformações que garantem *fairness* nas classificações dos modelos, as transformações aplicadas na pesquisa são idealizadas de forma empírica a partir de operações e também constantes matemáticas tradicionais, como a média de uma série e os números de Euler e Pi, por exemplo. Tais transformações se encaixam nas categorias Multiplicativa e Inversa de transformações metamórficas estruturadas em (MURPHY; KAISER; HU, 2008). Com relação às transformações das categorias Permutativa, Inclusiva e Exclusiva, elas são realizadas de forma implícita, através da seleção aleatória das amostras para os conjuntos de treino e teste e também na execução dos modelos que adotam a replicação de amostras, porém, impossibilitando um controle de parâmetros das amostras para que tais transformações possam ser analisadas e comparadas.

Os autores da pesquisa que pôde definir as categorias das transformações metamórficas afirmam que aquelas do tipo Multiplicativas não causa diferença nas classificações, desde que as alterações sejam realizadas em todas as amostras de uma coluna, por exemplo. Sendo assim, para o presente estudo foi idealizada a aplicação das transformações metamórficas em parte dos atributos e não para todas as amostras a serem executadas.

As transformações metamórficas (T) aplicadas de forma explícita nos valores do conjunto de dados podem ser conferidas abaixo, listadas por atributo do conjunto e seguidas de uma representação matemática para a aplicação:

- Grupo de idade multiplicado pelo próprio valor (ao quadrado):

$$T(age) = (age)^2 \tag{4.1}$$

- Relação com o(a) líder da família multiplicado pela média de toda a coluna (C):

$$T(household\_position) = household\_position * average(C) \tag{4.2}$$

- Quantidade de integrantes no domicílio multiplicado por 0,00001:

$$T(household\_size) = household\_size * 0,00001 \tag{4.3}$$

- Local de residência um ano antes multiplicado pelo Número de Euler (e, 2,7182818284):

$$T(prev\_residence\_place) = prev\_residence\_place * e \tag{4.4}$$

- Cidadania multiplicada pela raiz quadrada de 2:

$$T(citizenship) = citizenship * \sqrt{2} \tag{4.5}$$

- Nacionalidade multiplicada pelo Número de Ouro (ϕ, 1,6180339887):

$$T(country\_birth) = country\_birth * \phi \tag{4.6}$$

- Nível de escolaridade multiplicado por 1000:

$$T(edu\_level) = edu\_level * 1000 \tag{4.7}$$

- Status econômico multiplicado por -1000:

$$T(economic\_status) = economic\_status * (-1000) \tag{4.8}$$

- Setor da atividade econômica atual multiplicado por Pi (π, 3,1415926535):

$$T(cur\_eco\_activity) = cur\_eco\_activity * \pi \tag{4.9}$$

- Estado Civil dividindo o número 1:

$$T(marital\_status) = 1/marital\_status \tag{4.10}$$

Cenários de Análise

Ao trabalhar com transformações metamórficas do tipo Multiplicativas, é preciso ter ciência de que, quando realizadas em todas as amostras a serem utilizadas no treinamento dos modelos, não geram alterações nos resultados se comparados com uma execução utilizando as amostras originais. Tal fato se deve que a alteração das informações com qualquer valor constante, desde que seja exatamente o mesmo, não é capaz de produzir diferenças de distância de amostras, por exemplo, para o cálculo das classificações finais, sendo capaz de mudar apenas o sentido de apresentação dos dados, dependendo do sinal da constante utilizada (MURPHY; KAISER; HU, 2008).

Visando contornar tal problema, idealiza-se efetuar as transformações metamórficas em parte das amostras a partir de um determinado critério de alteração, de modo que algumas terão seus valores originais e o restante sofrerão as transformações, mudando assim a dimensão do espaço dos valores amostrais, garantindo então alterações nos resultados em busca da garantia de *fairness* nas classificações.

Sendo assim, quatro cenários de aplicação das transformações metamórficas são elaborados, cada um com seu próprio critério para que os dados possam ser transformados. Se tratando das constantes utilizadas nas transformações do tipo Multiplicativas, estas são elaboradas de forma aleatória e distintas para cada atributo do conjunto de dados, já apresentadas na Seção 4.3.3. Os Cenários de Análise são descritos a seguir:

- Transformações de Maiores Valores: neste cenário as transformações são realizadas em cima das amostras que possuam valores iguais ou maiores que a mediana para cada atributo. Na Tabela 4.2 é possível verificar o critério em que a transformação é aplicada para cada atributo do conjunto de dados;

- Transformações de Menores Valores: em oposição ao cenário anterior, aqui as amostras são transformadas quando os valores de cada atributo são menores que a mediana de toda a coluna, enquanto que valores iguais ou maiores permanecem em suas formas originais. É possível visualizar o critério de transformação para o cenário na Tabela 4.3

- Transformações de Valores do Grupo Sensível: diferentemente dos últimos cenários,

Atributo	Valores	Critério de Transformação
age	[4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]	>= 9.5
household_position	[1110, 1121, 1122, 1131, 1132, 1140, 1210, 1220]	>= 1131.5
household_size	[111, 112, 113, 114, 125, 126]	>= 113.5
prev_residence_place	[1, 2]	>= 1.5
citizenship	[1, 2, 3]	>= 2
country_birth	[1, 2, 3]	>= 2
edu_level	[0, 1, 2, 3, 4, 5]	>= 2.5
economic_status	[111, 112, 120]	>= 112
cur_eco_activity	[111, 122, 124, 131, 132, 133, 134, 135, 136, 137, 138, 139]	>= 133.5
marital_status	[1, 2, 3, 4]	>= 2.5

Tabela 4.2: Critério de aplicação das transformações para o Cenário de Transformações de Maiores Valores.

Atributo	Valores	Critério de Transformação
age	[4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]	< 9.5
household_position	[1110, 1121, 1122, 1131, 1132, 1140, 1210, 1220]	< 1131.5
household_size	[111, 112, 113, 114, 125, 126]	< 113.5
prev_residence_place	[1, 2]	< 1.5
citizenship	[1, 2, 3]	< 2
country_birth	[1, 2, 3]	< 2
edu_level	[0, 1, 2, 3, 4, 5]	< 2.5
economic_status	[111, 112, 120]	< 112
cur_eco_activity	[111, 122, 124, 131, 132, 133, 134, 135, 136, 137, 138, 139]	< 133.5
marital_status	[1, 2, 3, 4]	< 2.5

Tabela 4.3: Critério de aplicação das transformações para o Cenário de Transformações de Menores Valores.

neste as transformações são realizadas em todas as amostras pertencentes ao Grupo Sensível do conjunto de dados, isto é, todas as amostras que sejam do gênero feminino tem os valores de cada atributo transformados de acordo com as alterações especificadas na Seção 4.3.3, enquanto que as amostras do gênero masculino permanecem com seus valores originais.

- Transformações de Valores do Grupo Não Sensível: assim como o cenário anterior, as transformações são aplicadas em todas as amostras de um grupo, porém neste são as amostras do Grupo Não Sensível, ou seja, todas as amostras que contém o gênero masculino dos dados, enquanto as amostras do gênero feminino se mantém com os dados originais.

Com a aplicação das transformações metamórficas em cima do conjunto de dados original, são realizadas as execuções de treino e teste de todos os modelos para cada Cenário de Análise, ou seja, cada um dos nove modelos de classificação utilizados no estudo executam dois treinos e dois testes, sendo um treino com os dados originais e

outro com os dados transformados. Por fim, cada modelo criado executa a tarefa de classificação de dados, mas ambos utilizando o mesmo conjunto de testes sem qualquer transformação metamórfica, tornando possível uma comparação justa entre os resultados diferenciando apenas a utilização de dados transformados na etapa de treino.

4.4 Cálculo de Métricas e Estimativas para Validação

Ao obter as classificações realizadas pelas instâncias dos modelos treinados com dados originais e transformados, a próxima etapa do estudo é o cálculo das métricas de classificação (*Accuracy*, *Precision*, *Recall*, *Balanced Accuracy* e *F1-Score*) e de índices de *fairness* (*Statistical Parity* e *Equalized Odds*), já apresentadas no Capítulo 2. Tais cálculos são elaborados de forma automática, utilizando funções já construídas e disponibilizadas nos pacotes *Scikit-Learn*[3] (métricas de classificação) e *Fairlearn*[4] (métricas de *fairness*) voltados para a linguagem de programação *Python*.

Com os valores métricos é possível então uma comparação dos tipos de modelos, para determinar se de fato aqueles treinados com dados transformados melhoram os índices de *fairness* nas classificações, em relação aos valores métricos de modelos treinados com os dados em sua forma original. A forma adotada para a realização da comparação é a subtração dos valores de cada métrica dos modelos treinados com os dados transformados pelos valores dos modelos treinados com os dados originais. Porém há uma diferença na ordem dos fatores entre o cálculo da diferença para as métricas de classificação e de índices de *fairness*, a fim de que sempre que o resultado da diferença seja positivo, seja um indicativo de que um modelo treinado com dados transformados seja melhor que quando treinado com os dados originais.

Na Equação 4.11 é apresentado um exemplo para o cálculo de *Accuracy*, subtraindo o valor do modelo treinado com dados originais do valor do modelo treinado com dados transformados. Tal ordem é adotada para as demais métricas de classificação.

[3]Página com métricas disponibilizadas pelo *Scikit-Learn*: <https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics>.

[4]Página com métricas disponibilizadas pelo *Fairlearn*: <https://fairlearn.org/v0.9/api_reference/index.html#module-fairlearn.metrics>.

$$Diff = Mutate[Accuracy] - Original[Accuracy] \tag{4.11}$$

Já para a diferença das métricas de índices de *fairness*, a ordem dos fatores é invertida, como visto na Equação 4.12 calculando a diferença de *Equalized Odds*, que é adotada também no cálculo da diferença de *Statistical Parity*.

$$Diff = Original[EqualizedOdds] - Mutate[EqualizedOdds] \tag{4.12}$$

Somente com o cálculo das diferenças métricas de uma única execução dos modelos, não é possível definir a máxima de que um determinado tipo de modelo seja melhor que o outro tipo, uma vez que a quantidade de dados do conjunto original não pode representar corretamente a população real de dados. Além disso, outro fator que impacta nos resultados de somente uma execução é a aleatoriedade, que se encontra desde a divisão dos subconjuntos de dados para treino e teste até a criação das instâncias dos modelos de aprendizagem com base nos dados passados, em que a cada nova execução os valores podem ser distintos. Sendo assim, como forma de aumentar a abrangência dos valores obtidos para quantidades que se assemelhem à população real, é viável a utilização do método de reamostragem *bootstrap* para o cálculo de intervalos de confiança para as métricas, visando uma maior generalização dos resultados que podem então representar com maior semelhança uma população real.

De forma simples, o *bootstrap* consiste na realização de várias reamostragens de um único conjunto de dados, em que cada reamostragem possui a mesma quantidade de dados do conjunto original, porém com a possibilidade de repetição de cada uma das amostras de forma aleatória. Visando resultados confiáveis, a execução das reamostragens do *bootstrap* deve ser feita centenas ou até milhares de vezes a partir do conjunto de dados original (DOMINGUES et al., 2015). Na literatura não há um consenso do número exato de execuções que seja capaz de se obter o melhor resultado, como por exemplo pesquisadores variando de 5 a 3500 reamostragens para definir a melhor execução (MARTINEZ; LOUZADA-NETO, 2001), sendo assim, por o presente trabalho ser um estudo experimental, são realizadas, e postas em comparação, execuções com 500, 1000, 2000 e 4000 reamostragens para cada conjunto de classificações retornados

pelas instâncias dos modelos.

Ao passo em que as reamostragens são exercidas, o cálculo de diferença das métricas para os dois tipos de treinamento de modelos (com dados originais e transformados) também são realizados, com o objetivo de construir os intervalos de confiança para cada uma das métricas. Os intervalos são estimados com 95% de nível de confiança, ou seja, as amostras estimadas pelo *bootstrap* devem conter as diferenças métricas calculadas semelhantes em 95% do tempo.

4.5 Etapas de Execução do Estudo

Pensando em uma melhor compreensão sobre o que de fato é feito no presente estudo, a seguir são listadas cada uma das atividades realizadas com um breve resumo de desenvolvimento, sendo dispostas por ordem de execução:

1. Carregamento e Divisão do Conjunto de Dados: após carregar o arquivo contendo o conjunto de dados, o mesmo é dividido em dois subconjuntos para treino e teste com 70% e 30% dos dados respectivamente, de forma aleatória e sem reposição de amostras. A Figura 4.1 apresenta o esquema de divisão do conjunto de dados nos subconjuntos de treino e teste;

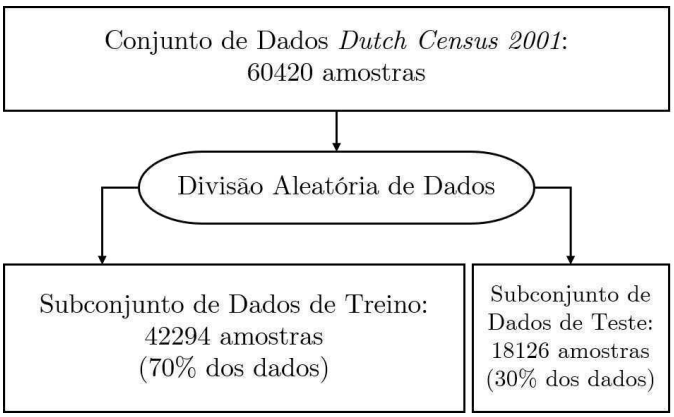


Figura 4.1: Esquema de divisão do Conjunto de Dados *Dutch Census 2001* em subconjuntos de Treino e Teste para a execução dos modelos de classificação.

2. Aplicação de Transformações Metamórficas: uma cópia do subconjunto de treino é realizada para que possa receber as transformações metamórficas em seus valores,

resultando em um conjunto de treino com dados originais e outro com dados transformados;

3. Execução dos Modelos de Classificação: para cada um dos nove modelos de classificação utilizados são criadas duas instâncias, uma em que é treinada com os dados originais e a segunda treinada com o subconjunto de treino com dados transformados, em seguida ambas as instâncias são testadas com o mesmo subconjunto de teste, com os dados na forma original, sem qualquer transformação;

4. Cálculo das Métricas sobre as Classificações: com as classificações realizadas, são então calculadas as métricas de classificação e de índices de *fainess*, para os resultados das duas instâncias de cada modelo, treinado com dados originais e transformados;

5. Reamostragem Utilizando *Bootstrap*: os conjuntos resultantes com as classificações são então reamostrados utilizando a técnica *bootstrap*, sendo realizadas quatro execuções da técnica, alternando a quantidade de reamostras em 500, 1000, 2000 e 4000, cada uma delas com a mesma quantidade de classificações geradas;

6. Cálculo dos Intervalos de Confiança: a cada execução do *bootstrap* são calculados os intervalos de confiança para cada uma das métricas, tanto de classificação quanto de índice de *fainess*, adotando um nível de confiança de 95%.

4.6 Considerações Finais do Capítulo

Ao fim deste Capítulo é esperado um entendimento à cerca dos pilares que sustentam o presente estudo, em que o domínio da aplicação se refere ao nível de ocupação profissional das pessoas, quando seus gêneros sexuais são agrupados de modo suscetível à discriminação por parte de modelos de aprendizagem. Com o problema apontado é então posta em evidência a possível solução para tal, esta sendo a alteração das informações pessoais como entrada de dados para os modelos executados. São apresentadas também as Questões de Pesquisa que guiam o estudo, sendo esperado ao fim das execuções que as respostas para cada uma delas sejam satisfatórias com o propósito de enriquecer

a literatura relacionada à *fairness* e transformações metamórficas. Outros fatores importantes apresentados são as quantidades reais do conjunto de dados trabalhado bem como suas divisões nos subconjuntos de treino e teste, além dos cenários em que as execuções são analisadas, estes de acordo com as estratégias de aplicação das transformações metamórficas nos dados, definindo uma melhor organização para a apresentação e entendimento dos resultados no Capítulo seguinte. Por fim é apresentada a técnica de *bootstrap* para o cálculo de intervalos de confiança, com o propósito de generalizar os resultados obtidos para um cenário de população real dos dados, este sendo um método de inferência adotado em diversos trabalhos para a validação de resultados obtidos. Todo o código desenvolvido ao longo do estudo bem como suas execuções e resultados estão disponibilizados no seguinte repositório do GitHub: <https://github.com/rodolfobolconte/mestrado-ufcg-dissertacao>.

5 Resultados

Este Capítulo tem o objetivo de discorrer os resultados obtidos através do estudo descrito no Capítulo 4, comparando os cenários de execuções propostos e buscando responder de forma embasada e objetiva cada uma das Questões de Pesquisa elencadas anteriormente.

Recapitulando informações relacionadas as execuções dos experimentos, cada Cenário executa duas vezes o treino de cada um dos modelos de classificação, uma com os dados originais e outra com os dados alterados com as Transformações Metamórficas. Após estas duas execuções são gerados então dois modelos do mesmo tipo, porém com treinos diferentes, em que ambos são executados a tarefa de classificação de dados, porém utilizando os mesmos dados sem as transformações. Com a obtenção das classificações dos dois modelos, seus valores são então enviados para o *Bootstrap* para a realização de quatro execuções de reamostragens, com o cálculo da diferença dos valores das métricas entre o modelo com treino de dados originais e o modelo com treino de dados transformados.

No total, são realizadas 144 execuções de treino e teste de modelos (36 para cada cenário), e 32 reamostragens (8 para cada cenário). O tempo de realização de todas as execuções de modelos e reamostragens leva cerca de 11 horas e 6 minutos. À nível de *hardware*, todas as execuções são realizadas em uma máquina com 16 GB de Memória Principal DDR5, Processador AMD Ryzen 7 6800U e Sistema Operacional Windows 10 Home.

5.1 Análise de Valores

Devido a quantidade de execuções realizadas, não é interessante uma discussão de cada uma delas para cada um dos cenários de análise, sendo assim, nas próximas subseções são apresentadas informações de execuções importantes para o entendimento geral do cenário, além de informações gerais agregando informações de todas as execuções de cada Cenário.

Para uma melhor visualização dos resultados, são utilizados gráficos de intervalos de confiança comparando os valores métricos para as instâncias dos modelos treinados com dados originais e dados transformados. É possível entender que caso o intervalo de confiança seja positivo, a instância do modelo treinada com dados transformados se sai melhor que a instância treinada com os dados originais, caso o intervalo seja negativo, a lógica é invertida, e por fim quando o intervalo possui valores positivos e negativos, não é possível afirmar que uma instância é melhor que a outra, apenas que ambas podem ter resultados semelhantes.

5.1.1 Cenário 1: Transformações de Maiores Valores

O modelo *Stacking* teve o melhor resultado na comparação das métricas de *fairness*, em que a instância do modelo treinado com dados transformados pode atingir uma melhora por volta de 5% a 6,5%, para *equalized odds*, em relação a instância treinada com os dados originais, enquanto que para *statistical parity*, o intervalo de melhora fica próximo de 4%. Porém, se tratando das métricas de classificação, para *accuracy*, *balanced accuracy* e *precision* não é possível definir se há uma melhora ou piora na utilização de qualquer um dos dois tipos de instâncias do modelo, enquanto que para *recall* e *F1-score* a diferença é nítida ao visualizar a Figura 5.1, evidenciando que utilizar a instância do *stacking* treinado com os dados originais pode ser pouco mais de 2% melhor para o *F1-score*, por exemplo.

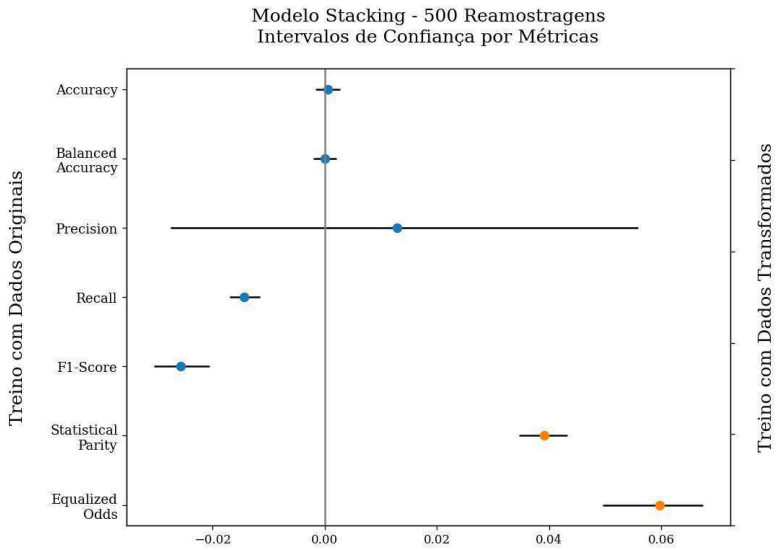


Figura 5.1: Intervalos de Confiança da diferença de instâncias do modelo *Stacking* para 500 reamostragens do *bootstrap* no Cenário 1

Outro modelo que indica melhora nas métricas de *fairness* ao utilizar a instância com os dados transformados é o *Decision Tree*, porém é importante destacar que apesar de haver uma melhora, os seus valores reais podem não ser tão expressivos, com os intervalos de confiança se concentrando em uma diferença entre 0,4% e pouco mais que 0,8% como mostrado na Figura 5.2. Com relação as métricas de classificação, *precision* foi a única que apresentou uma diferença, indicando que a instância do *Decision Tree* treinada com os dados originais pode ser melhor que a instância treinada com os dados transformados, apesar da diferença em torno de 0,1% e 0,5%, já para as demais métricas não é possível afirmar qual instância seria a melhor, com o intervalo de confiança variando entre valores positivos e negativos.

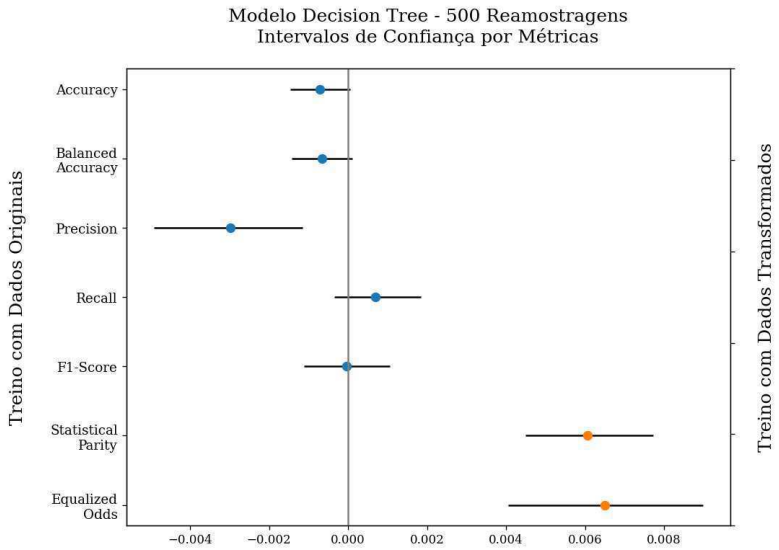


Figura 5.2: Intervalos de Confiança da diferença de instâncias do modelo *Decision Tree* para 500 reamostragens do *bootstrap* no Cenário 1.

Houveram modelos em que não foi possível definir qual instância seria a melhor ao avaliar todas as métricas, em que os intervalos de confiança contém valores tanto positivos quantos negativos, como foi o caso dos modelos *Adaptive Boosting*, *Gradient Boosting* e *Random Forest*.

O modelo *Extra Tree* foi o único em que a instância com os dados originais se saiu melhor que a instância com os dados transformados, com *equalized odds* podendo variar entre 7% e pouco mais de 8% e *statiscal parity* em torno de 6%, já as métricas de classificação apresentam uma melhora para a instância dos dados transformados, exceto para *precision*. É possível visualizar os intervalos de confiança comparando as duas instâncias do modelo na Figura 5.3.

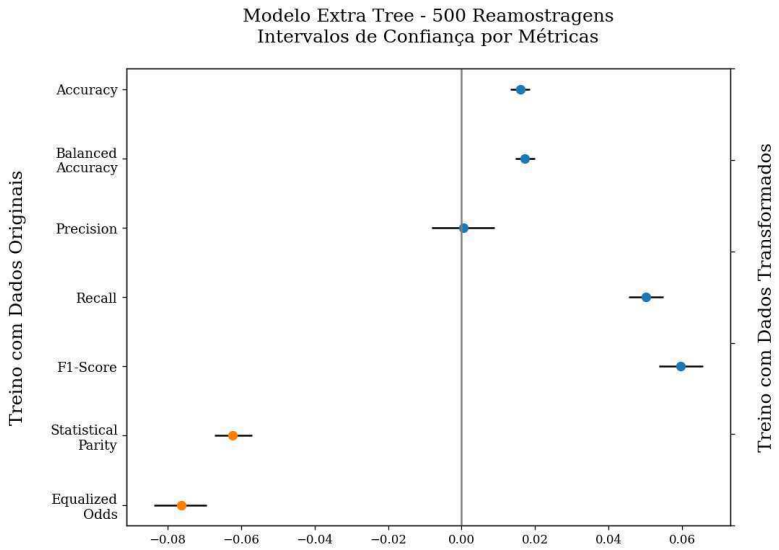


Figura 5.3: Intervalos de Confiança da diferença de instâncias do modelo *Extra Tree* para 500 reamostragens do *bootstrap* no Cenário 1.

Para as demais execuções do *bootstrap*, realizando 1000, 2000 e 4000 reamostragens, a diferença dos intervalos de confiança se mostraram idênticas ou bem próximas em relação aos valores obtidos na execução de 500 reamostragens, sendo assim os valores das demais reamostragens não serão apresentados neste Capítulo devido a uma repetição de informação e discussão também. Tal fato evidencia também uma consistência nas reamostragens, em que é possível inferir que seus resultados seriam coerentes para uma população real.

5.1.2 Cenário 2: Transformações de Menores Valores

Aplicando as transformações metamórficas nos valores das colunas abaixo da própria mediana no conjunto de dados, o *Decision Tree* também apresentou melhora na instância com os dados transformados, com a diferença nas métricas de *fairness* sendo maior que no Cenário 1. Para *equalized odds* o intervalo de confiança indica que a diferença pode atingir valores entre 17% e 20%, enquanto que o intervalo de confiança de *statistical parity* apresenta uma diferença por volta de 13% a 15%, isto é, uma melhora em relação a instância com os dados originais. Já para as métricas de classificação, a única que indica uma melhora quando usando a instância com os dados transformados é o *recall* com a diferença se apresentando em torno de 1%, já as demais métricas

indicam que a instância com os dados originais é melhor, com a *precision* apresentando o maior intervalo de confiança, se encontrando em mais de 5% de diferença. Na Figura 5.4 é possível visualizar a diferença para as métricas em relação a execução de 500 reamostragens.

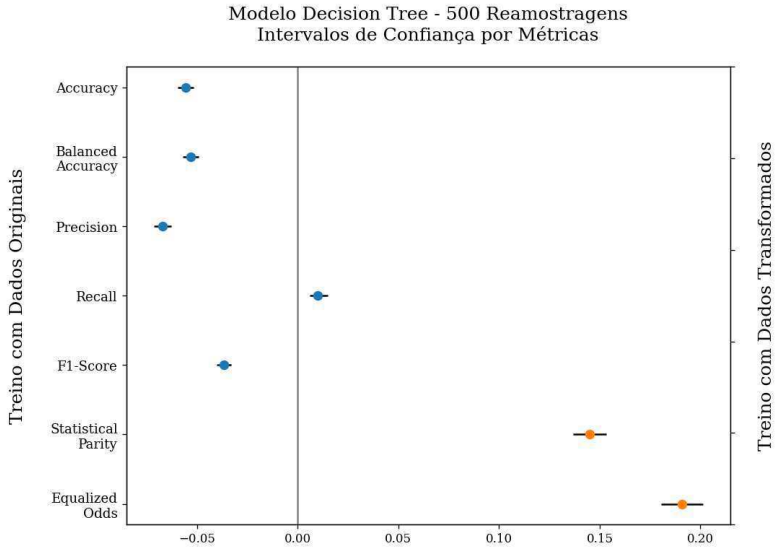


Figura 5.4: Intervalos de Confiança da diferença de instâncias do modelo *Decision Tree* para 500 reamostragens do *bootstrap* no Cenário 2.

O modelo com a segunda maior diferença indicando melhora para a instância com dados transformados foi o *Random Forest*, isto se tratando das métricas de *fairness*, com *equalized odds* podendo apresentar uma diferença entre 3,5% e 5%, enquanto que *statistical parity* tem um intervalo um pouco menor, por volta de 3% e 4%. Em relação a *accuracy* e *balanced accuracy* não é possível definir qual seria a melhor instância, uma vez que de acordo com seus intervalos de confiança, os valores para estas duas métricas podem ser positivos ou negativos. *Precision* foi a única métrica de classificação que apresentou uma melhora voltada para a instância de dados transformados, com seu intervalo de confiança atingindo quase 1% de diferença, enquanto que *recall* e *F1-score* mostram uma melhora para a instância com os dados originais, com *F1-score* variando entre 1,5% e 2%.

Na Figura 5.5 é possível visualizar os resultados para a execução de 2000 reamostragens do *Random Forest*.

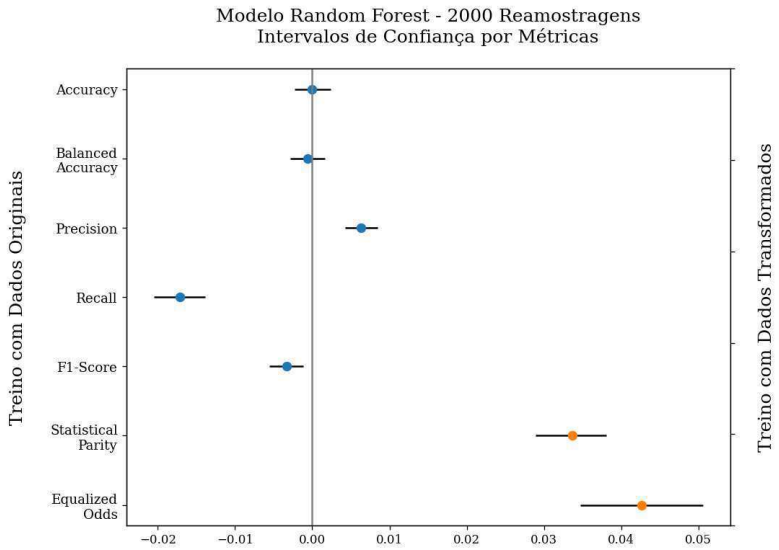
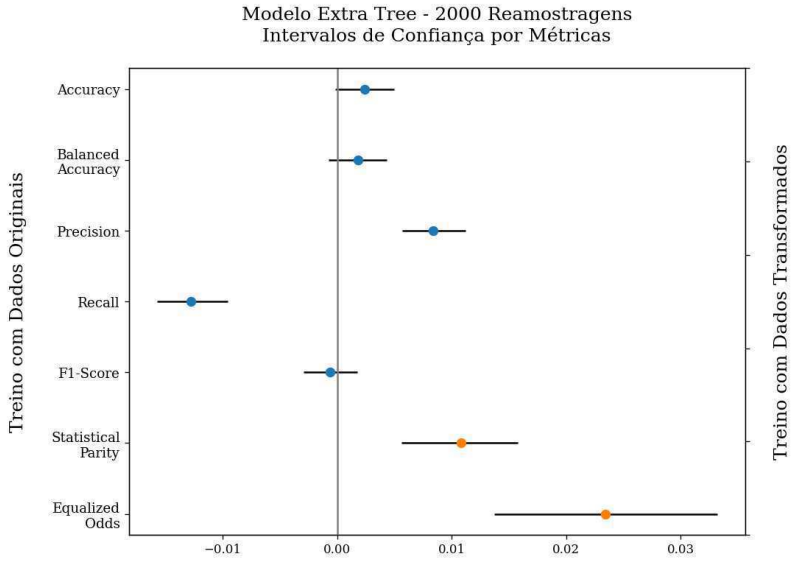
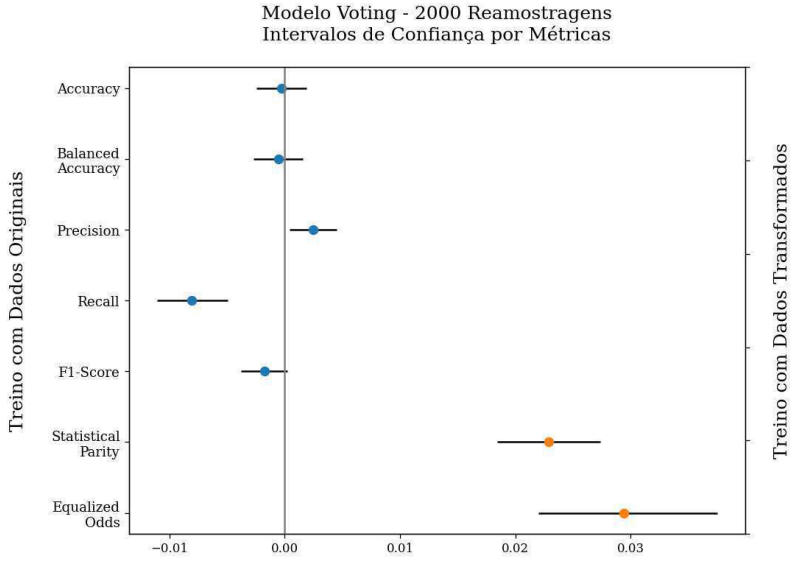


Figura 5.5: Intervalos de Confiança da diferença de instâncias do modelo *Random Forest* para 2000 reamostragens do *bootstrap* no Cenário 2.

Outros modelos que, a partir das métricas de *fairness*, apresentaram melhoras da instância de dados transformados em relação a instância de dados originais foram o *Extra Tree* e o *Voting*, que tiveram intervalos de confiança da diferença próximos de 2% a 4% para *equalized odds* e de 0,5% a 3% para *statistical parity*, enquanto que para as métricas de classificação os intervalos se concentram próximos de nenhuma diferença. Nas Figuras 5.6a e 5.6b é possível visualizar os resultados para os modelos *Extra Tree* e *Voting*, respectivamente.



(a) Modelo *Extra Tree*.



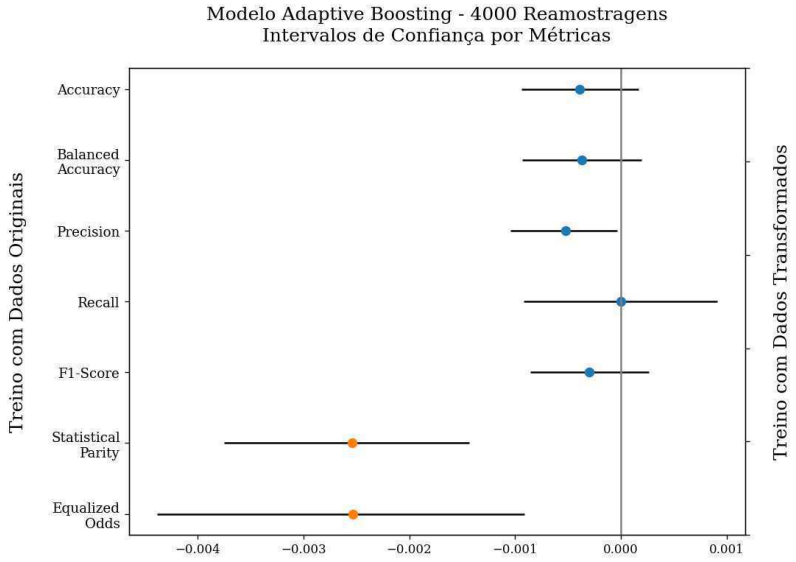
(b) Modelo *Voting*.

Figura 5.6: Intervalos de Confiança da diferença de instâncias dos modelos *Extra Tree* e *Voting* para 2000 reamostragens do *bootstrap* no Cenário 2.

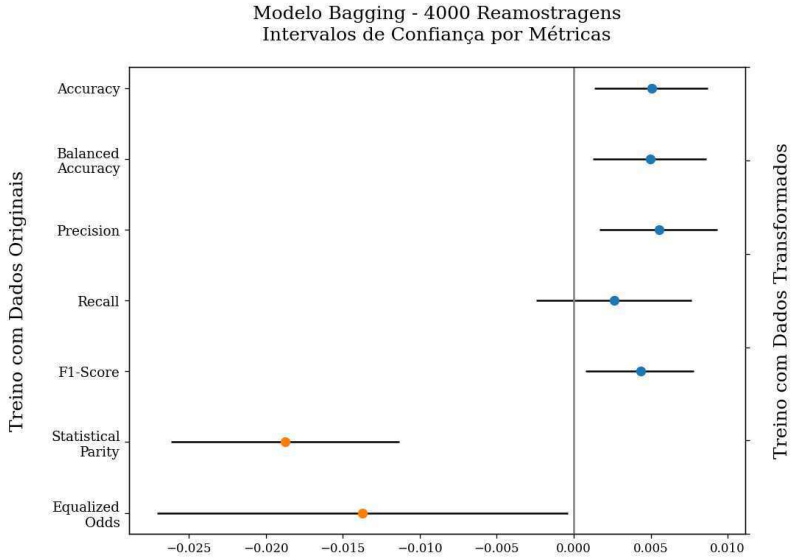
Para o cenário inverso, isto é, quando as instâncias dos modelos que são treinadas com os dados originais se saem melhor em relação as instâncias treinadas com os dados transformados, desde que comparadas as métricas de *fairness*, os modelos *Adaptive Boosting*, *Bagging* e *Gradient Boosting* atingem esses resultados. Porém, os valores dos intervalos de confianças das métricas são próximos do limiar nulo (quando não é possível definir se uma instância é melhor que outra), como para o modelo *Bagging*, que o intervalo de confiança para *equalized odds* tem o menor valor próximo de 0,1% e

o maior valor em torno de 2,7%, enquanto que a menor diferença é obtida pelo modelo *Adaptive Boosting*, com o intervlao de confiança de *equalized odds* entre 0,1% e 0,4%. Se tratando das métricas de classificação, para o modelo *Adaptive Boosting* não é possível definir qual instância seria a melhor devido os intervalos possuírem valores positivos e negativos, enquanto que para os outros dois modelos, somente o *recall* possui intervalos com valores positivos e negativos, com as demais métricas indicando uma melhora para a instância com dados transformados.

Nas Figuras 5.7a e 5.7b é possível visualizar os intervalos de confiança comparando os dois tipos de instâncias dos modelos *Adaptive Boosting* e *Bagging* respectivamente, para a execução do *bootstrap* com 4000 reamostragens.



(a) Modelo *Adaptive Boosting*.



(b) Modelo *Bagging*.

Figura 5.7: Intervalos de Confiança da diferença de instâncias dos modelos *Adaptive Boosting* e *Bagging* para 4000 reamostragens do *bootstrap* no Cenário 2.

Assim como no Cenário 1, os modelos não apresentam diferenças consideráveis a cada execução do *bootstrap* – com um número diferente de reamostragens –, logo, os gráficos apresentados anteriormente foram de execuções de diferentes quantidades de reamostragens visando a diversidade de informações.

5.1.3 Cenário 3: Transformações de Valores do Grupo Sensível

Diferente dos cenários anteriores, neste as transformações metamórficas são aplicadas somente nas amostras pertencentes ao grupo sensível do conjunto de dados, isto é, as informações de pessoas do gênero feminino. E, assim como mostrado nos últimos cenários, as execuções com diferentes quantidades de reamostragens do *bootstrap* realizadas não possuem diferenças consideráveis entre os modelos, com os intervalos de confiança tendo valores bem próximos para as métricas a cada nova execução, mostrando uma consistência de informações ao inferir tais valores para uma população real que os dados podem representar.

Ao todo cinco modelos foram capazes de apresentar uma melhora na utilização das instâncias com os dados transformados:

- *Bagging*: dentre todas as métricas de *fairness* e de classificação, o modelo foi capaz de apresentar uma diferença nas instâncias apenas na métrica *statistical parity*, com um intervalo de pouco menos de 2% até quase 3,5%, enquanto que para as demais métricas não é possível definir qual seria a melhor instância para uma utilização, devido os valores positivos e negativos presentes nos intervalos. Tais informações apresentadas são referentes a execução com 500 reamostragens do *bootstrap*;

- *Extra Tree*: *precision* foi a única métrica que na comparação das instâncias mostra que quando utilizando os dados originais pode ser melhor ao invés da instância com os dados transformados, enquanto que todas as demais métricas de classificação e de *fairness* indicam o contrário. *Equalized odds* apresenta um intervalo de pouco mais de 8% até cerca de 10,5% de diferença entre as duas instâncias, enquanto que *statistical parity* apresentou um intervalo com valores menores, entre 4% e 5% de diferença. Na Figura 5.8 é possível observar os intervalos de confiança das métricas para o modelo *Extra Tree* quando executando 500 reamostragens;

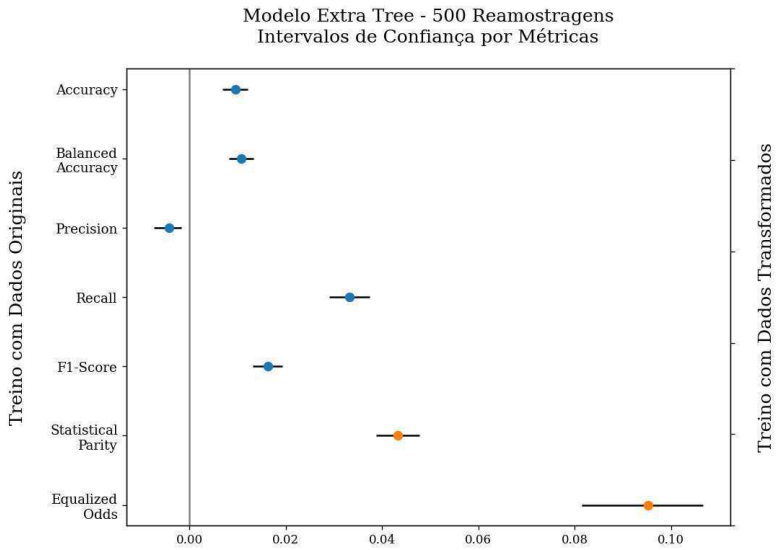


Figura 5.8: Intervalos de Confiança da diferença de instâncias do modelo *Extra Tree* para 500 reamostragens do *bootstrap* no Cenário 3.

- *Gradient Boosting*: somente uma métrica de classificação indicou uma melhora na utilização da instância com dados transformados, o *recall* com um intervalo de confiança de diferença próximo de 0,5%, enquanto que as demais métricas de classificação indicam o contrário, com *precision* tendo os maiores valores para o intervalo de confiança, em torno de 1,5% a 2% de diferença. As métricas de *fairness* indicam uma melhora para a instância com os dados transformados, com *equalized odds* podendo atingir uma diferença de até 3,5% e *statistical parity* entre 2% e 3%. Toda a discussão é feita com base na execução de 500 reamostragens do *bootstrap*;

- *Histogram-based Gradient Boosting*: é o único modelo em que todas as métricas indicam uma melhora na utilização da instância com os dados transformados, que embora o intervalo de confiança da diferença para *statistical parity* seja próximo de 0%, o intervalo para *equalized odds* indica que pode apresentar valores de pouco mais de 6% até 8,5%, já as métricas de classificação tem intervalos de confiança que variam entre 1% a 4% dentre todas elas. A Figura 5.9 apresenta os intervalos de confiança da comparação das instâncias do modelo para a execução de 1000 reamostragens do *bootstrap*;

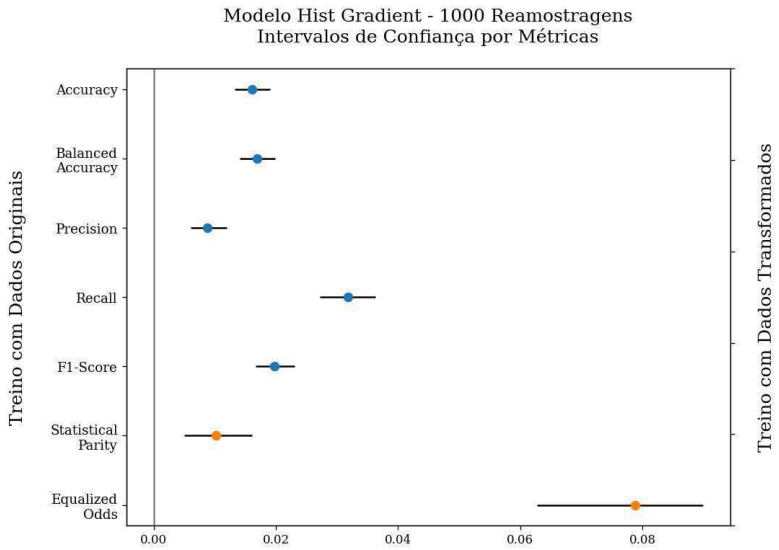


Figura 5.9: Intervalos de Confiança da diferença de instâncias do modelo *Histogram-based Gradient Boosting* para 1000 reamostragens do *bootstrap* no Cenário 3.

- *Random Forest*: os resultados para as instâncias deste modelo são semelhantes aos resultados do modelo *Gradient Boosting*, em que o *recall* é a única métrica de classificação que comparando as instâncias, indica uma melhora na utilização da que é treinada com os dados transformados, inclusive com valores próximos para o intervalo de confiança, entre 0,3% e 0,9% de diferença, já as métricas de *fairness* possuem intervalos próximos entre si, com *equalized odds* entre 2,2% e 3,8% e *statistical parity* com um intervalo um pouco menor, entre 2,5% e 3,5%, isto para a execução de 2000 reamostragens do *bootstrap*, porém as demais execuções são próximas destes valores;

- *Voting*: as métricas de *fairness* são as únicas que indicam melhora para a instância treinada com dados transformados, com seus valores próximos de 0,7% a 1,7%. Para as métricas de classificação, *recall* é a única em que não é possível afirmar qual seria a melhor instância, tendo valores positivos e negativos no intervalo de confiança, enquanto que para as demais métricas os valores dos intervalos não chegam a ultrapassar uma diferença de 1% de melhora.

O modelo *Stacking* é o único em que as métricas de *fairness* apresentam uma diferença favorável para a instância treinada com os dados originais, com *equalized odds* tendo intervalo de confiança entre 3% e 5% e *statistical parity* entre 1,8% e pouco mais

de 3%, já as métricas de classificação, todas elas são favoráveis para a utilização da instância com dados transformados, embora tendo intervalos de confiança bem próximos de 1% de diferença. A Figura 5.10 traz os intervalos das métricas do modelo *Stacking* para a execução de 4000 reamostragens do *bootstrap*, com as demais execuções obtendo valores próximos também.

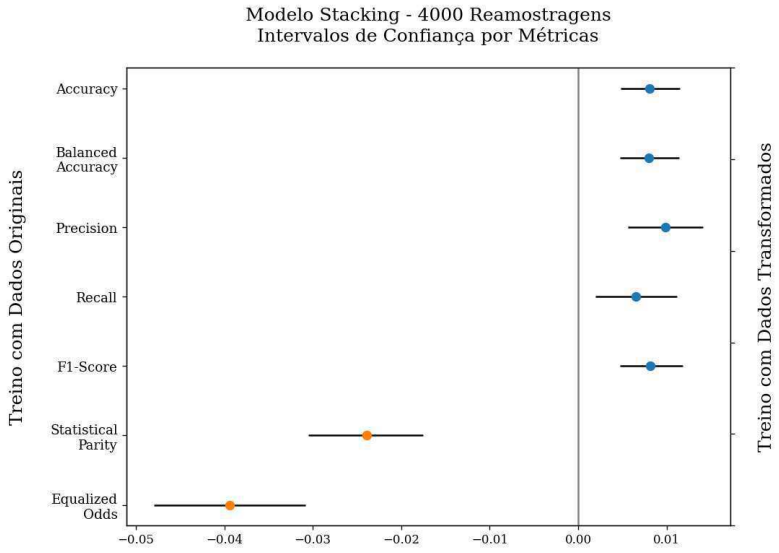
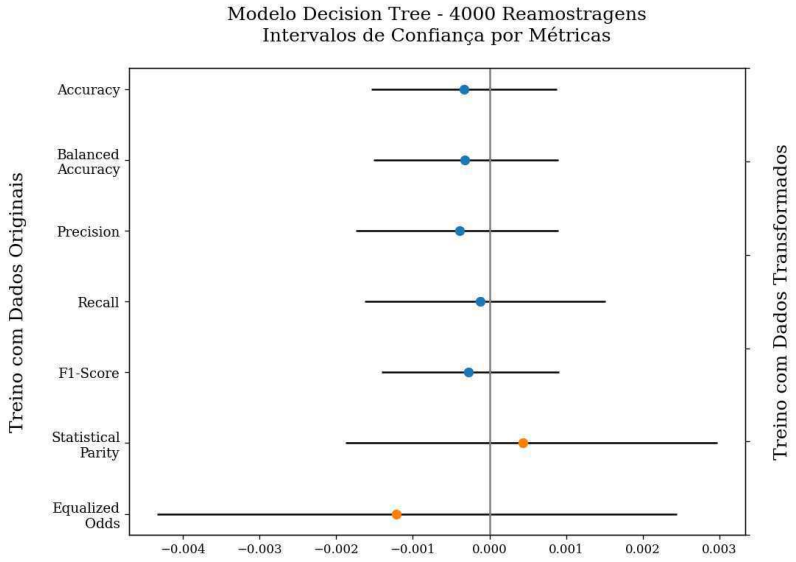
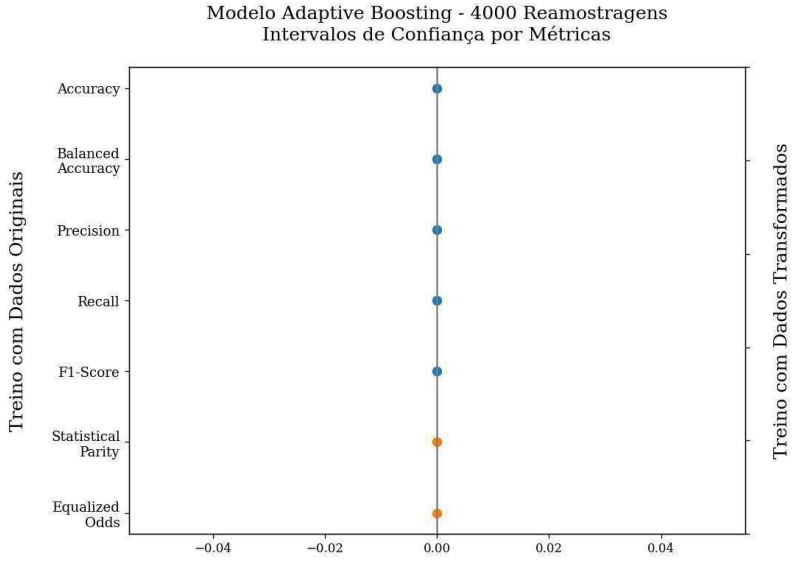


Figura 5.10: Intervalos de Confiança da diferença de instâncias do modelo *Stacking* para 4000 reamostragens do *bootstrap* no Cenário 3.

Apenas dois modelos durantes as execuções – para todas as reamostragens do *bootstrap* e também para todas as métricas de *fairness* e de classificação – indicaram que não é possível inferir qual seria a melhor instância para uma possível utilização, uma vez que ou seus intervalos de confiança possuem valores positivos e negativos ou todos os valores mínimos, médios e máximos são 0, como foram os casos dos modelos *Decision Tree* e *Adaptive Boosting*, este último sendo o mais curioso devido os valores serem 0% de diferença para todas as métricas, indicando que realmente não há diferença na inferência dos valores para os resultados das duas instâncias do modelo, mesmo com a aleatoriedade da seleção de amostras. As Figuras 5.11a e 5.11b apresentam os resultados para os modelos *Decision Tree* e *Adaptive Boosting*, respectivamente, para as execuções de 4000 reamostragens do *bootstrap*.



(a) Modelo *Decision Tree*.



(b) Modelo *Adaptive Boosting*.

Figura 5.11: Intervalos de Confiança da diferença de instâncias dos modelos *Decision Tree* e *Adaptive Boosting* para 4000 reamostragens do *bootstrap* no Cenário 3.

5.1.4 Cenário 4: Transformações de Valores do Grupo Não Sensível

Semelhante ao terceiro cenário, neste as transformações metamórficas são realizadas nas amostras pertencentes a um grupo específico do conjunto de dados, que neste caso são as amostras do grupo não sensível, ou seja, pessoas do gênero masculino. Sendo assim, todas as informações de pessoas do gênero masculino são alteradas de acordo com as

transformações metamórficas especificadas na Seção 4.3.3, enquanto que as amostras do gênero feminino são utilizadas em sua forma original.

Neste cenário apenas dois modelos se mostraram favoráveis na utilização de instâncias com os dados transformados, ao analisar apenas as métricas de *fairness* para a comparação dos intervalos de confiança:

- *Random Forest*: apenas *equalized odds* indicou uma melhora para a instância com dados transformados, atingindo um intervalo de confiança da diferença entre cerca de 4,2% a 5,7%, já ao visualizar o intervalo para *statistical parity*, não é possível indicar qual seria a melhor instância, uma vez que pode atingir valores favoráveis para ambas, assim como a métrica *precision* que possui um intervalo menor, porém as demais métricas de classificação são favoráveis para a instância com os dados transformados. A Figura 5.12 apresenta os intervalos de confiança para a comparação das instâncias do modelo, para a execução de 500 reamostragens do *bootstrap*;

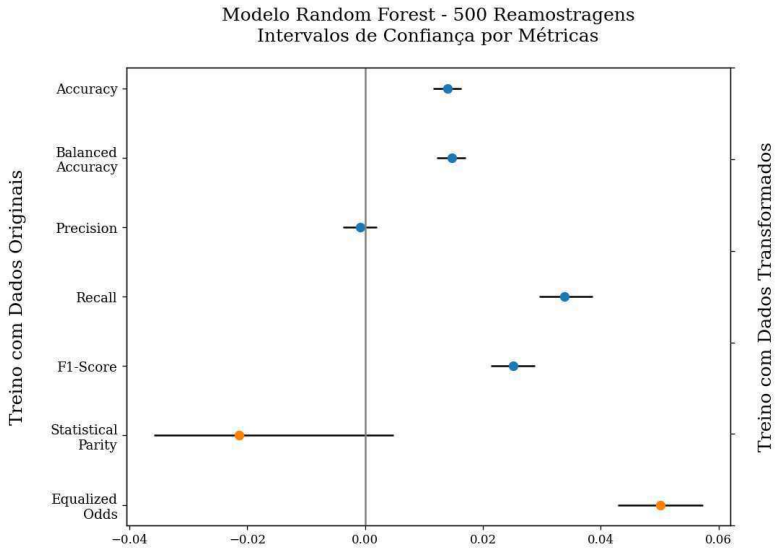


Figura 5.12: Intervalos de Confiança da diferença de instâncias do modelo *Random Forest* para 500 reamostragens do *bootstrap* no Cenário 4.

- *Stacking*: para este modelo, todas as métricas tanto de *fairness* quanto de classificação são favoráveis a instância treinada com dados transformados. *Equalized odds* possui o intervalo com os maiores valores dentre as métricas, com a diferença das instâncias podendo atingir em torno de 7,9% até cerca de 9,7%, enquanto que

statistical parity possui um intervalo menor com valores menores também, indo de 5% a 6,2%. *Recall* e *F1-score* possuem intervalos próximos, entre 4% e 6%, enquanto *precission* possui o intervalo com os menores valores, próximos de 1%. Na Figura 5.13 é possível visualizar os intervalos de confiança da comparação das instância para a execução de 1000 reamostragens.

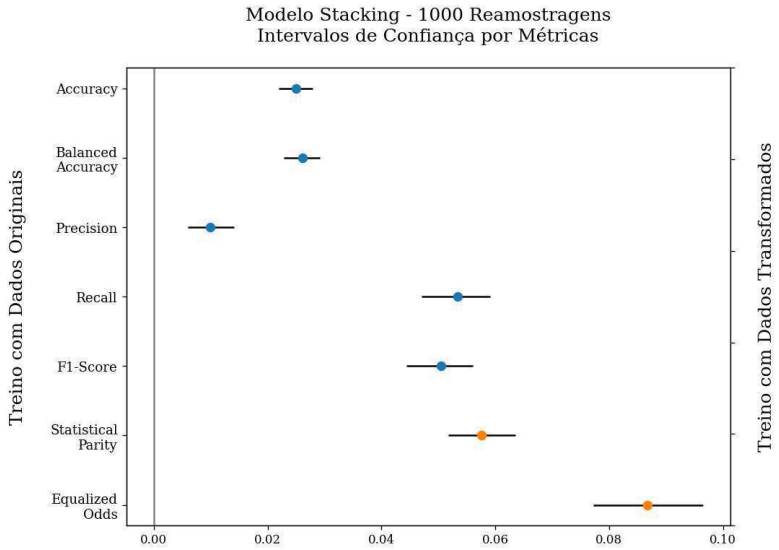


Figura 5.13: Intervalos de Confiança da diferença de instâncias do modelo *Stacking* para 1000 reamostragens do *bootstrap* no Cenário 4.

Em se tratando de melhorias para instâncias treinadas com os dados originais, no total são três modelos ao comparar as métricas de *fairness*. Para o modelo *Extra Tree*, *equalized odds* é a única métrica favorável para a instância com dados originais, que pode atingir valores entre 2,3% a quase 4% de diferença, enquanto que para *statistical parity* não é possível definir qual seria a melhor instância. Já para modelo *Histogram-based Gradient Boosting*, ambas as métricas de *fairness* indicam vantagem para a instância com dados originais, embora com os intervalos de confiança próximos de 0%, sendo *equalized odds* de 0,5% a 1,2% e *statistical parity* de pouco menos de 0,5% até 0,75%, como mostra a Figura 5.14 para a execução de 1000 reamostragens. Por fim, o modelo *Voting* também indica melhora em *fairness* para a instância com dados originais, apesar do intervalo de confiança para *statistical parity* não ultrapassar de 1% de diferença, indo de 0,1% até 0,6%.

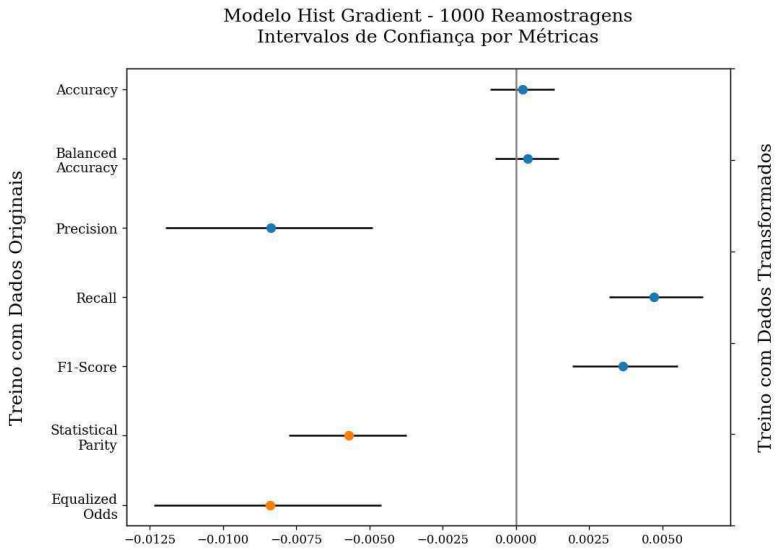


Figura 5.14: Intervalos de Confiança da diferença de instâncias do modelo *Histogram-based Gradient Boosting* para 1000 reamostragens do *bootstrap* no Cenário 4.

Finalizando o cenário, *Adaptive Boosting*, *Bagging*, *Decision Tree* e *Gradient Boosting* são os quatro modelos que apresentaram intervalos de confiança que não é possível definir qual seria a melhor instância de acordo com as métricas de *fairness*, inclusive com o modelo *Adaptive Boosting* obtendo o mesmo resultado do Cenário 3, com todos os valores métricos sendo 0% de diferença. Outro fato curioso foi que para o modelo *Bagging*, apenas uma métrica indicou melhora para a instância com os dados originais, sendo o *recall*, enquanto que todas as outras apresentaram intervalos de diferença com valores positivos e negativos, sem a possibilidade de definir qual seria a melhor instância. A Figura 5.15 traz a comparação das instâncias do modelo *Bagging* para a execução de 4000 reamostragens do *bootstrap*.

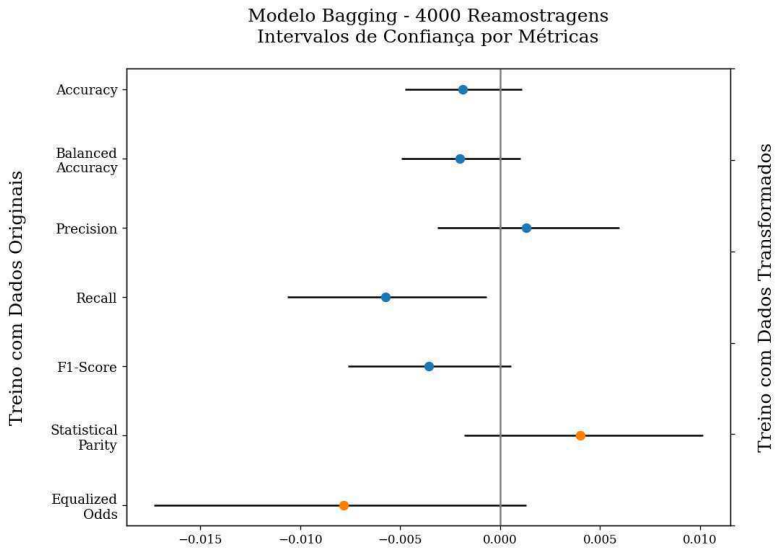


Figura 5.15: Intervalos de Confiança da diferença de instâncias do modelo *Bagging* para 4000 reamostragens do *bootstrap* no Cenário 4.

Assim como os demais cenários de análise, os valores dos intervalos de confiança das diferenças entre as instâncias dos modelos para cada nova execução do *bootstrap*, não apresentam diferenças consideráveis, devido a isto a discussão não é elaborada percorrendo cada uma das execuções.

5.2 Discussão

Tendo em vista a quantidade de informações apresentadas no presente Capítulo, a seguir é feita uma discussão com base nos resultados obtidos visando responder às Questões de Pesquisa idealizadas na Seção 4.2.

Questão de Pesquisa 1 (QP1): O quão imparciais os modelos de aprendizagem se mostram com a utilização de transformações metamórficas nos dados?

O termo imparcialidade nesta questão se refere ao viés que os modelos de classificação podem apresentar em relação ao atributo sensível do conjunto de dados, isto é, analisando os resultados das amostras de acordo com o gênero, masculino ou feminino, que para isto, as métricas de *fairness* são fundamentais nesta discussão visto que a análise das classificações são realizadas tendo como base o próprio atributo sensível tal qual o gênero especificado para cada amostra. E para responder esta questão é feita uma

discussão passando por cada modelo de classificação utilizado:

- *Adaptive Boosting*: foi o único dentre os nove modelos de classificação que para todas as execuções dos quatro cenários não apresentou resultado favorável para as instâncias treinadas com os dados transformados, com seus valores variando entre uma melhoria para uma possível utilização da instância treinada com os dados originais (como observado no Cenário 2 em que os intervalos de confiança tanto de *statistical parity* quanto de *equalized odds* se mostram como negativos), e para valores de intervalos de confiança em que não é possível definir qual seria a melhor instância do modelo, observada nos Cenários 1, 3 e 4;

- *Bagging*: o modelo apresentou resultados favoráveis as instâncias treinadas com dados transformados apenas na métrica *statistical parity* (que analisa somente os valores resultantes dos modelos sem levar em consideração os valores do mundo real), para os cenários de transformação dos maiores valores e dos valores de amostras pertencentes ao grupo sensível (Cenários 1 e 3). No Cenário 2 a *statistical parity* se mostrou contrária, indicando melhoria para a instância com dados originais, enquanto que no quarto cenário não foi possível definir qual seria a melhor instância, assim como a métrica *equalized odds* para todos os cenários executados;

- *Decision Tree*: tanto *statistical parity* quanto *equalized odds* indicaram uma melhora voltada para a instância com os dados transformados, porém apenas para os cenários de transformação dos maiores e menores valores, enquanto que para os cenários em que os dados transformados são do grupo sensível e do grupo não-sensível, não é possível indicar qual seria a melhor instância devido os valores positivos e negativos nos intervalos de confiança da diferença para as duas métricas;

- *Extra Tree*: favorecendo as instâncias com os dados transformados, o modelo apresentou bons resultados nos Cenários 2 e 3 para as duas métricas de *fairness*, enquanto que no Cenário 1 foi o contrário, indicando melhora na utilização de instâncias com os dados originais. No cenário de transformação de informações

das amostras do grupo não-sensível (Cenário 4) enquanto que com *statistical parity* não é possível definir qual a melhor instância devido os intervalos de confiança com valores positivos e negativos, para *equalized odds* a melhor instância seria a com os dados originais;

- *Gradient Boosting*: o Cenário 3 foi o único em que o modelo apresentou resultados favoráveis para instâncias com os dados transformados (levando em consideração que qualquer intervalo de confiança acima de 0 indique uma melhoria, independentemente do valor exato obtido), tanto *statistical parity* quanto *equalized odds* justificam a afirmação. Para os Cenários 1 e 2, as métricas apontam que a instância com os dados originais é melhor, enquanto que no Cenário 4 não é possível afirmar qual seria a melhor instância em uma possível utilização;

- *Histogram-Based Gradient Boosting*: semelhante ao modelo *Extra Tree* as métricas de *fairness* indicam que as instâncias com dados transformados são melhores nos Cenários 2 e 3, enquanto que para os Cenários 1 e 4 se mostrou o contrário, em que as instâncias com os dados originais se apresentam melhor em relação as instâncias com os dados transformados;

- *Random Forest*: assim como os modelos *Extra Tree* e *Histogram-Based Gradient Boosting*, nos Cenários 2 e 3 as duas métricas de *fairness* apresentam valores favoráveis as instâncias com os dados transformados. No Cenário 4 apenas *equalized odds* favorece as instâncias com dados transformados, enquanto que com *statistical parity* não é possível definir qual a melhor, assim como para as duas métricas de *fairness* para o Cenário 1;

- *Stacking*: nos Cenários 1 e 4 o modelo apresentou valores favoráveis de *statistical parity* e *equalized odds* para instâncias treinadas com os dados transformados, ao contrário do que foi apresentado no Cenário 3, sendo favoráveis as instâncias com os dados originais. Finalizando os cenários, com as duas métricas não foi possível definir qual seria a melhor instância no segundo cenário;

- *Voting*: finalizando a discussão dos modelos, o *Voting* apresentou melhora para a instância com dados transformados nos Cenário 1, 2 e 3, em que no primeiro

cenário apenas a métrica *statistical parity* foi capaz de corroborar a melhora, enquanto que para os segundo e terceiro cenários ambas as métricas apresentam valores favoráveis. No Cenário 4 as métricas se dividiram, com *statistical parity* indicando melhora para as instâncias com os dados originais e *equalized odds* indicando que não é possível dizer qual seria a melhor instância, podendo atingir valores de diferença tanto positivos quanto negativos nos intervalos de confiança.

Em resumo as informações apresentadas acima, os modelos que melhor se saíram em relação a melhoria de *fairness*, que pode ser apresentada na utilização de instâncias dos modelos utilizando dados transformados, foram o *Random Forest* e o *Voting*, que embora para ambos as duas métricas de *fairness* indicam melhoria em até dois cenários de análise, um terceiro cenário de análise indica melhora quando analisados os valores de *statistical parity* ou *equalized odds* de forma individual. O *Adaptive Boosting* se mostrou como o modelo de menor desempenho relacionado a *fairness* visto que de acordo com as métricas não foram possíveis indicar que as instâncias treinadas com os dados transformados fossem melhor em alguma das diversas execuções realizadas. O *Gradient Boosting* foi o segundo modelo com menor desempenho, indicando resultados favoráveis para as instâncias com os dados transformados em apenas um cenário, enquanto que os demais modelos apresentaram valores propícios em até dois cenários de análise.

Questão de Pesquisa 2 (QP2): Quais os métodos de aplicação de transformação metamórfica que mais impactam na garantia de *fairness* nos modelos?

É possível responder essa questão de duas formas, a primeira é analisar quais os cenários que mais apresentam resultados favoráveis as instâncias treinadas com os dados transformados. Sendo assim, ao fazer um levantamento das duas métricas de *fairness* em comparação com cada modelo de classificação de dados, tem-se o Cenário 3, isto é, quando as transformações são aplicadas somente nos valores pertencentes as amostras de gênero feminino (grupo sensível do conjunto de dados), com a maior ocorrência de melhora de *fairness* entre os modelos. Somando todas as execuções de modelos para o Cenário 3, foi possível observar uma melhora voltada as instâncias com os dados transformados em 6 modelos de classificação distintos, sendo eles: *Bagging*, *Extra Tree*, *Gradient Boosting*, *Histogram-Based Gradient Boosting*, *Random Forest* e *Voting*.

É importante ter em mente que o termo "melhora" aqui empregado pode representar qualquer valor de diferença, não importando se a diferença seja um número muito próximo de 0% com base nos intervalos de confiança, como por exemplo para o modelo *Voting*, em que a melhora indicada por *statistical parity* quando executando 500 reamostragens possui um intervalo de confiança de diferença entre 0,7% e 1,5%, no Cenário 3. Porém, para o mesmo Cenário foi possível observar intervalos de diferença maiores, como para o modelo *Extra Tree* em que o intervalo da diferença de *equalized odds* apresentou valores entre 8% e 10,2%, uma diferença bem maior em relação ao modelo *Voting*.

O Cenário 2 foi o segundo com maior ocorrência de resultados favoráveis as instâncias com dados transformados, apresentando melhoras em 5 modelos de classificação: *Decision Tree* (o único modelo diferente do Cenário 3), *Extra Tree*, *Histogram-Based Gradient Boosting*, *Random Forest* e *Voting*. Sendo assim, a transformação metamórfica quando aplicada nos valores menores que a mediana das informações, também pode ser um bom cenário para a melhora de índices de *fairness* para os modelos de classificação citados, pelo menos.

Por fim, os Cenários 1 e 4 foram os que menos apresentaram melhoras nos resultados dos modelos. O Cenário 1 ainda apresentou melhora em 4 modelos de classificação, porém este é um número menor que a metade do número de modelos executados na presente pesquisa, sendo assim, o método de transformar valores maiores que a mediana das informações pode não ser tão eficiente para a melhora de índices de *fairness* das classificações dos modelos em geral. Já no Cenário 4 apenas dois modelos apresentaram melhora favorável as instâncias com os dados transformados, sendo o *Random Forest* e o *Stacking*, se apresentando como o pior cenário de aplicação de transformações metamórficas dentre os demais.

A segunda forma de responder esta questão é avaliar a grandeza da diferença entre os tipos de instâncias dos modelos, isto é, o quão distantes de 0% são os valores dos intervalos de confiança das diferenças entre as instâncias treinadas com os dados transformados e as treinadas com os dados originais de cada modelo de classificação.

A maior diferença que uma instância treinada com dados transformados pode atingir se encontra em um intervalo de valores entre 18% e 20% em relação a instância treinada

com os dados originais. Tal intervalo de confiança de diferença foi atingido pelo modelo *Decision Tree* na execução de 500 reamostragens do *boostrap*, apresentado no Cenário 2, em que as transformações metamórficas são realizadas nos valores menores que a mediana de cada coluna do conjunto de dados. Já a menor diferença de melhora entre as instâncias foi atingida pelo modelo *Voting* no Cenário 1, em que para 500 reamostragens o intervalo de confiança da diferença girou em torno de 0,08% a 0,3%.

É possível ainda calcular a média de grandeza por cenário de execução, pegando os maiores valores dos intervalos de confiança de cada modelo que apresentou valores favoráveis as instâncias com os dados transformados, seja em *statistical parity* ou *equalized odss*. Sendo assim, o Cenário 4 seria o melhor, com uma média de 7,85% para os maiores valores dos intervalos de confiança, porém é importante entender que apenas dois dos nove modelos de classificação apresentaram resultados favoráveis, o que pode tornar injusta uma comparação por média. Como alternativa no cálculo da média, poderia utilizar a quantidade fixa de nove modelos utilizados e não apenas a quantidade daqueles que apresentaram bons valores, tornando assim o Cenário 2 o melhor, com uma média de maiores valores dos intervalos de confiança em torno de 3,64%, seguido do Cenário 3 com uma média de 3,57%.

Visualizando as duas formas de como responder a segunda Questão de Pesquisa, é possível concluir que o Cenário 2 seja o método que possui mais impacto na melhoria dos índices de *fairness* nos modelos, visto que tanto possui os valores mais altos de diferença entre instâncias para um modelo, quanto a maior média para os maiores valores de um intervalo ao calcular para os nove modelos utilizados.

Questão de Pesquisa 3 (QP3): É possível manter bons índices de eficácia com a aplicação das transformações metamórficas nos dados?

A eficácia para esta questão representa o desempenho das classificações dos modelos em comparação com os valores reais do conjunto de dados (uma comparação entre o que foi classificação e qual o valor real do nível de ocupação das amostras do conjunto).

Ao analisar somente as execuções que apresentaram melhorias voltadas para as instâncias com os dados transformados, o panorama não é promissor, visto que de 17 execuções de modelos (levando em conta apenas a execução de 500 reamostragens, visto que as demais tem valores próximos), apenas em 5 destas a maioria dos intervalos de

confiança das métricas de classificação também indicam uma melhora, enquanto que 8 execuções indicam uma piora, ou seja, utilizar instâncias com os dados originais seria melhor que com os dados transformados, para os índices de classificação, apenas.

Do Cenário 1, o modelo *Voting* foi o único a apresentar melhora nas métricas de classificação, sendo as métricas *accuracy* (com intervalo de diferença entre 0,1% e 1,2%) e *precision* (com intervalo de 0,4% a 2,1%). No Cenário 2, a maioria das métricas de classificação para cada modelo não apresentaram valores satisfatórios no aumento dos índices de classificação, com a ocorrência de casos contrários ou que não foi possível definir qual seria a melhor instância para a classificação, devido os intervalos com valores negativos e positivos.

Para cada um dos Cenários 3 e 4, dois modelos apresentaram melhora nos índices de classificação a partir das instâncias com os dados transformados. No Cenário 3 o modelo *Extra Tree* apresentou valores positivos para todas as métricas de classificação exceto *precision*, que apresentou valores favoráveis a instância com os dados originais, já o modelo *Histogram-Based Gradient Boosting* se saiu com valores positivos para todas as métricas. No Cenário 4 os únicos modelos que apresentaram melhoria nos índices de *fairness* também apresentaram melhorias nos índices de classificação, sendo o *Random Forest* com apenas a métrica *precision* sem conseguir definir qual seria a melhor instância, e o modelo *Stacking* apontando melhoria para todas as métricas.

Sendo assim, respondendo parcialmente a esta questão com base nas informações apresentadas até então, não é possível manter bons índices de eficácia (classificação) ao utilizar as transformações metamórficas, analisando apenas os modelos que tiveram resultados favoráveis as instâncias treinadas com os dados transformados.

Porém, ao analisar todas as execuções dos modelos para cada cenário não é possível definir de fato se é possível manter ou não bons índices de classificação. Verificando as diferenças dos intervalos de confiança de cada um dos modelos para cada um dos Cenários de execução, é observado que as quantidades de execuções são iguais para manter, não manter ou não saber se há bons índices de classificação quando aplicando as transformações metamórficas, isto é, quando a maioria das métricas de classificação entram em um consenso geral (quando 3 métricas apresentam valores favoráveis a um tipo de instância, por exemplo).

Questão de Pesquisa 4 (QP4): Como se apresenta o custo/benefício da aplicação das transformações metamórficas pela melhora de imparcialidade dos modelos?

Para responder a última questão, os tempos de execuções, tanto da aplicação das transformações metamórficas no conjunto de dados, quanto das atividades de treino dos modelos se fazem necessários – a atividade de classificação não é necessária para a comparação uma vez que é utilizado o mesmo conjunto de dados de teste para a execução dos dois tipos de instâncias dos modelos.

No Cenário 1, a aplicação das transformações metamórficas no conjunto de dados de treino levam em torno de 52,2 segundos para ser realizada em 331879 valores que são iguais ou maiores que a mediana de suas respectivas colunas, tendo um total de 422940 valores em todo o conjunto de dados de treino somando 10 colunas, sem aplicar as transformações na coluna de gênero (o atributo sensível). Com relação a execução do treino dos modelos, não cabe especificar aqui o tempo para cada um dos nove modelos utilizados, apenas que tanto o treino com os dados originais quanto com os dados transformados, levam tempos próximos de execução, variando entre maior para a instância com os dados transformados e menor para a com os dados originais e vice-versa. Por exemplo, o treino do *Adaptive Boosting* com os dados originais leva cerca de 0,59 segundo, enquanto que com os dados transformados leva pouco mais de 0,61 segundo, enquanto que para o modelo *Extra Tree*, o treino da instância com os dados transformados é executado em cerca de 1,65 segundos, contra 1,72 segundos utilizando os dados originais.

O segundo Cenário possui um tempo maior de aplicação das transformações sobre os dados, em comparação com o Cenário 1, levando cerca de 57,51 segundos para a aplicação em cerca de 91061 valores do conjunto de treino, logo, valores abaixo da mediana de suas respectivas colunas. Com relação as execuções de treino das instâncias com os dados originais e transformados, o comportamento se mantém o mesmo que no Cenário 1 e se estende para os demais Cenários também, com tempos próximos de execução e alternando entre maior e menor para os dois tipos de instâncias.

Os Cenários 3 e 4 possuem tempos menores de aplicação das transformações metamórficas, uma vez que a quantidade de amostras a receberam as transformações

são consideravelmente menores em comparação com os Cenários 1 e 2. No terceiro Cenário, aplicar as transformações nos dados das amostras pertencentes ao grupo sensível, isto é, amostras do gênero feminino, leva em torno de 35,91 segundos, para cerca de 21181 amostras do conjunto de teste, enquanto que no Cenário 4 a aplicação das transformações é feita em aproximadamente 35,50 segundos para 21113 amostras de pessoas do gênero masculino.

Respondendo então a Questão de Pesquisa 4, o custo/benefício se mostra relativo para o tipo de atividade a ser realizada. Para o presente estudo, em que o ambiente é controlado a nível de pesquisa e testagem com a realização de diversas execuções, o custo/benefício ao aplicar as transformações metamórficas pode ser considerado bom quando com o propósito de atingir uma melhora de até 20% de *equalized odds*, como é o caso do modelo *Decision Tree* no Cenário 2, e embora levando cerca de 57,71 segundos para que as transformações sejam realizadas. Porém, para o modelo *Histogram-Based Gradient Boosting* ainda no Cenário 2, o custo/benefício pode não ser proveitoso, uma vez que a melhora em *equalized odds* seria de no máximo 0,7% em comparação com a instância do modelo treinada com os dados originais, logo não há uma compensação em levar a mesma quantidade de tempo para apenas 0,7% de diferença.

No Apêndice B é possível conferir os resultados obtidos para cada um dos modelos em cada Cenário de Análise e também para cada execução de reamostragem do *bootstrap*.

6 Considerações Finais

O presente trabalho de mestrado trouxe o problema da análise da parcialidade e/ou discriminação contra indivíduos ou grupos menos favorecidos em sociedade, que pode ser constatada em modelos de classificação de dados. Tal problema é abordado através de *fairness*, que é um paradigma da computação que visa sempre a diminuição da discriminação nos resultados alcançados. Como forma de resolução do problema explicado, foram analisadas técnicas de transformações metamórficas realizadas em cima de um conjunto de dados com informações de indivíduos moradores da Holanda por volta do ano de 2001, que são classificadas em dois níveis de atividade profissional, baixo e alto nível. O conjunto de dados possui como atributo sensível à discriminação o gênero dos indivíduos, nas opções de masculino ou feminino para cada um, este último sendo o grupo sensível com chances de discriminação por parte dos modelos.

Ao todo são realizadas 10 transformações metamórficas em 10 diferentes colunas do conjunto de treino, que é utilizado em 9 modelos de classificação de dados, sendo criadas uma instância – de cada modelo – treinada com os dados em seu formato original e outra instância treinada com os dados transformados. Após as execuções do treino e classificação das instâncias de cada modelo, os valores são reamostrados em diversas execuções do *bootstrap* a fim de verificar a consistência nos resultados e a comparação das duas instâncias de cada modelo em diferentes métricas tanto de classificação quanto de *fairness*.

A análise de todo o estudo se deu através de diferentes Cenários de aplicação das transformações metamórficas, sejam elas realizadas nos valores maiores ou iguais que a mediana dos valores de cada coluna (Cenário 1) ou nos valores menores que a mediana

(Cenário 2), e nos valores pertencentes apenas as amostras do grupo sensível (indivíduos do sexo feminino) para o Cenário 3 ou nas amostras do grupo não sensível (indivíduos do gênero masculino) para o Cenário 4.

Por fim, o estudo se mostrou satisfatório no que se refere as transformações metamórficas melhorarem os índices de *fairness* nas classificações dos modelos que, embora não tenha sido um resultado unânime em uma melhora considerável para todas as execuções, foi possível alcançar valores de até 20% de melhora relacionada a utilização de uma instância treinada com os dados transformados em relação a instância treinada com os dados originais, para o mesmo modelo de classificação.

É esperado então que o presente estudo possa contribuir principalmente no enriquecimento da discussão de *fairness*, no que se refere a utilização de transformações metamórficas de dados para a melhoria dos índices de imparcialidade dos modelos. Embora tais técnicas não sejam tão disseminadas na literatura até então, é visível que sua utilização pode trazer benefícios aos desenvolvedores e usuários, tanto no âmbito científico quanto até mesmo no âmbito profissional.

6.1 Limitações

Com a execução de todo o estudo, foi perceptível o entendimento de algumas limitações no desenvolvimento mas, tendo tais limitações em vista, fica claro que a solução da utilização de transformações metamórficas de informações, para o problema da parcialidade contra indivíduos ou grupos discriminados por modelos de aprendizagem, possui abertura para expansão, com o propósito de melhorar cada vez mais os seus resultados a fim de ser unânime o fato da utilização de transformações metamórficas melhorarem os índices de imparcialidade nos resultados de diferentes modelos de classificação de dados.

A seguir são elencadas algumas das limitações que foram observadas no desenvolvimento de todo o estudo:

- Um único domínio foi abordado durante a execução do estudo, este sendo o de indivíduos da Holanda no ano de 2001 com apenas informações de gênero

referentes as informações sensíveis, tendo o gênero feminino como o grupo sensível e suscetível a discriminação/parcialidade dos modelos;

- O conjunto de dados utilizado possui um certo equilíbrio nas suas amostras, tanto na quantidade de indivíduos dos gêneros masculino e feminino, quanto na quantidade de amostras dos diferentes níveis de ocupação profissional. É possível que tais quantidades próximas de classes tenham contribuído para que em algumas execuções dos modelos não tenha sido claro qual seria a melhor instância, se a treinada com os dados transformados ou a com os dados originais;

- A aplicação de apenas dois tipos de transformação metamórfica nos dados de maneira controlada, isto é, foram utilizadas somente transformações dos tipos multiplicativa e inversa para a verificação de melhoria nos resultados dos modelos. Apesar de outros tipos de transformações terem sido utilizadas também, como do tipo Permutativa, elas não foram consideradas na análise, devido suas aplicações não terem sido de forma proposital;

- As transformações metamórficas foram idealizadas de forma empírica e aleatória para cada coluna do conjunto de dados, e testes aplicando uma determinada transformação em outras colunas ou até mesmo quantidades diferentes de transformações em conjunto com valores originais não foram realizados, fazendo o estudo não tão abrangente na quantidade de testes possíveis para validações mais robustas;

- A aleatoriedade se fez presente em diversas partes do estudo, desde a separação do conjunto de dados para treinamento e testagem até o próprio funcionamento interno dos modelos, principalmente aqueles que trabalham na criação de diversas instâncias de um modelo base, como o *Random Forest* ao criar vários modelos internos do tipo *Decision Tree*. Para o estudo foram definidos estados aleatórios fixos nas execuções de divisão dos dados e treino e teste dos modelos, porém para novas execuções em estados diferentes, os resultados obtidos podem não ser os mesmos;

- Informações relacionadas a custo computacional não foram analisadas, limitando a

discussão do custo/benefício da utilização de transformações metamórficas apenas aos tempos de execução da aplicação das transformações e também as atividades de treinamento dos modelos de classificação;

6.2 Pensamentos Futuros

Com base nas limitações apresentadas na Seção anterior, é possível idealizar alguns pensamentos que possam contribuir na evolução do estudo desenvolvido, a fim de encontrar melhores resultados e solidificar ainda mais a referência deste trabalho nas áreas de *fairness*, transformações metamórficas e análise e testes de parcialidade em modelos de classificação de dados. Sendo assim, os seguintes pontos são idealizados para o futuro, visando o engrandecimento do presente trabalho:

- Utilizar novos domínios de dados para avaliar a generalização das transformações metamórficas aplicadas no presente estudo, bem como informações atualizadas, representando o mundo em que vivemos atualmente;
- Trabalhar com outros tipos de características sensíveis, como raça, etnia, religião, etc., a fim de verificar se as transformações podem ajudar modelos que discriminam indivíduos ou grupos sensíveis dessas características;
- Executar o estudo em conjuntos com proporções mais desbalanceadas no que se refere tanto a classe dos dados quanto as quantidades de amostras dos atributos sensíveis. Realizar também reamostragens na seleção de dados visando um desbalanceamento proposital de classes;
- Diversificar as transformações metamórficas, tanto na testagem de uma mesma transformação em diferentes colunas quanto a combinação de uma certa quantidade de colunas transformadas em conjunto com colunas com seus valores originais. Outro ponto em relação as transformações seria a utilização de outros tipos para análise, não se limitando apenas as transformações dos tipos multiplicativa e inversa;

- Trabalhar de forma controlada com a aleatoriedade nas execuções para avaliar as melhores combinações, tanto de amostras utilizadas nas etapas de treino e teste quanto nas criações e divisões que os modelos de classificação são capazes de realizar;
- Avaliar com maior precisão o custo computacional ao aplicar as transformações e também a diferença nas atividades de treino dos modelos com os dados transformados e originais.

Bibliografia

ACADEMY, D. S. *Deep Learning Book*. Data Science Academy, 2022. Disponível em: <https://www.deeplearningbook.com.br/>.

ADEBAYO, J.; KAGAL, L. Iterative orthogonal feature projection for diagnosing bias in black-box models. *arXiv preprint arXiv:1611.04967*, 2016. Disponível em: <https://arxiv.org/abs/1611.04967>.

ALLWRIGHT, S. *Accuracy vs balanced accuracy, which is the best metric?* 2022. Disponível em: <https://stephenallwright.com/accuracy-vs-balanced-accuracy/>.

ANGWIN, J.; LARSON, J. *ProPublica Responds to Company's Critique of Machine Bias Story*. ProPublica, 2016. Disponível em: <https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story>.

ANGWIN, J. et al. *How We Analyzed the COMPAS Recidivism Algorithm*. ProPublica, 2016. Disponível em: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.

ANGWIN, J. et al. *What Algorithmic Injustice Looks Like in Real Life*. ProPublica, 2016. Disponível em: <https://www.propublica.org/article/what-algorithmic-injustice-looks-like-in-real-life>.

ASYROFI, M. H. et al. *BiasFinder: Metamorphic Test Generation to Uncover Bias for Sentiment Analysis Systems*. 2021. Disponível em: <https://arxiv.org/abs/2102.01859>.

AZNAR, P. *What is the difference between Extra Trees and Random Forest?* 2020. QuantDare, The scientific blog of ETS Asset Management Factory. Disponível em: <https://quantdare.com/what-is-the-difference-between-extra-trees-and-random-forest/>.

BINNS, R. Fairness in machine learning: Lessons from political philosophy. Conference on Fairness, Accountability, and Transparency, 2018. Disponível em: <https://arxiv.org/abs/1712.03586>.

BRAITHWAITE, A. *'Three black teenagers' Google Image search sparks racism row*. SBS News, 2018. Disponível em: <https://www.sbs.com.au/news/article/three-black-teenagers-google-image-search-sparks-racism-row/nb0mf503u>.

BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, n. 2, p. 123–140, 1996. Disponível em: <https://link.springer.com/article/10.1007/BF00058655>.

BRUCE, A.; BRUCE, P. *Estatística Prática para Cientistas de Dados, 50 Conceitos Essenciais*. O' Reilly, Alta Books, 2019. Disponível em: <https://altabooks.com.br/produto/estatistica-pratica-para-cientistas-de-dados/>.

BUIS, J. *This obscure Korean bot has quietly turned into a swearing machine*. The Next Web, 2016. Disponível em: <https://thenextweb.com/news/simsimi-swearing-machine>.

BURKOV, A. *The hundred-page machine learning book*. Andriy Burkov Quebec City, QC, Canada, 2019. v. 1. Disponível em: <https://themlbook.com/>.

CALDERS, T.; VERWER, S. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, Springer, v. 21, p. 277–292, 2010. Disponível em: <https://link.springer.com/article/10.1007/s10618-010-0190-x>.

CERRI, R.; CARVALHO, A. C. Aprendizado de máquina: Breve introdução de aplicações. *Cadernos de Ciencia e Tecnologia*, v. 34, n. 3, p. 297–313, 2017. Disponível em: <https://seer.sct.embrapa.br/index.php/cct/article/view/26381>.

CESARO, J. *Avaliação de Discriminação em Aprendizagem de Máquina usando Técnicas de Interpretabilidade*. Tese (Doutorado) — Universidade de São Paulo, 2021. Disponível em: <https://www.teses.usp.br/teses/disponiveis/3/3141/tde-31052021-114333/pt-br.php>.

CHEN, T. Y.; CHEUNG, S. C.; YIU, S. Metamorphic testing: A new approach for generating next test cases. *CoRR*, abs/2002.12543, 1998. Disponível em: <https://arxiv.org/abs/2002.12543>.

CLEARY, T. A. Test bias: Validity of the scholastic aptitude test for negro and white students in integrated colleges. *ETS Research Bulletin Series*, v. 1966, n. 2, p. i–23, 1966. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1966.tb00529.x>.

CLEGER, S.; PESSOA, M. S. P.; LIMA, J. Viés em aprendizagem de máquina: como a inteligência artificial pode prejudicar as minorias. *VIII Encontro Regional de Computação e Sistemas de Informação*, p. 54–63, 01 2019. Disponível em: <https://www.researchgate.net/publication/339064294_Vies_em_Aprendizagem_de_Maquina_como_a_inteligencia_Artificial_pode_prejudicar_as_minorias>.

COELHO, T. *SimSimi é perigoso? Veja polêmica do app com crianças e saiba protegê-las*. TechTudo, 2018. Disponível em: <https://www.techtudo.com.br/dicas-e-tutoriais/2018/04/simsimi-e-perigoso-veja-polemica-do-app-com-criancas-e-saiba-protege-las.ghtml>.

COELHO, T. *Simsimi é suspenso no Brasil; entenda caso do app de chat online*. TechTudo, 2018. Disponível em: <https://www.techtudo.com.br/noticias/2018/04/simsimi-e-suspenso-no-brasil-entenda-caso-do-app-de-chat-online.ghtml>.

COLE, N. S.; ZIEKY, M. J. The new faces of fairness. *Journal of educational Measurement*, Wiley Online Library, v. 38, n. 4, p. 369–382, 2001. Disponível em: <https://doi.org/10.1111/j.1745-3984.2001.tb01132.x>.

CORBETT-DAVIES, S. et al. *A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear*. The Washington Post, 2016. Disponível em: <https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/>.

CRUZ, M. *Problemas resolvidos por algoritmos de classificação*. 2023. Alura. Disponível em: <https://www.alura.com.br/artigos/problemas-resolvidos-algoritmos-classificacao>.

DASTIN, J. *Amazon scraps secret AI recruiting tool that showed bias against women*. 2018. Disponível em: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.

DIETERICH, W.; MENDOZA, C.; BRENNAN, T. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, v. 7, n. 7.4, p. 1, 2016. Disponível em: <https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html>.

DING, J.; KANG, X.; HU, X.-H. Validating a deep learning framework by metamorphic testing. In: IEEE. *2017 IEEE/ACM 2nd International Workshop on Metamorphic Testing (MET)*. 2017. p. 28–34. Disponível em: <https://ieeexplore.ieee.org/document/7961649>.

DOMINGUES, K. et al. Estimação de intervalos de confiança via reamostragem bootstrap. In: . [s.n.], 2015. Disponível em: <https://www.researchgate.net/publication/280132590_Estimacao_de_Intervalos_de_Confianca_via_Reamostragem_Bootstrap>.

DRESSEL, J.; FARID, H. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, v. 4, n. 1, 2018. Disponível em: <https://www.science.org/doi/abs/10.1126/sciadv.aao5580>.

GALHOTRA, S.; BRUN, Y.; MELIOU, A. Fairness testing: testing software for discrimination. In: *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*. [s.n.], 2017. p. 498–510. Disponível em: <https://dl.acm.org/doi/10.1145/3106237.3106277>.

GEITGEY, A. *Machine Learning is Fun!* 2014. Disponível em: <https://medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471>.

GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., 2019. v. 2. Disponível em: <https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/>.

GHIRARDELLO, G. *Visão computacional: o que é e principais aplicações*. BotCity, 2023. Disponível em: <https://blog.botcity.dev/pt-br/2023/11/28/visao-computacional/>.

HASTIE, T. et al. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009. v. 2. Disponível em: <https://link.springer.com/book/10.1007/978-0-387-21606-5>.

HOWLEY, D. *Google Photos Mislabels 2 Black Americans as Gorillas*. 2015. Disponível em: <https://finance.yahoo.com/news/google-photos-mislabels-two-black-americans-as-122793782784.html>.

HUTCHINSON, B.; MITCHELL, M. 50 years of test (un)fairness: Lessons for machine learning. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery, 2019. (FAT* '19), p. 49–58. ISBN 9781450361255. Disponível em: <https://doi.org/10.1145/3287560.3287600>.

JENSEN, A. R. Bias in mental testing. Glencoe, IL, 1980. Disponível em: <https://eric.ed.gov/?id=ED183698>.

JOHNSON, B.; BRUN, Y. Fairkit-learn: A fairness evaluation and comparison toolkit. In: *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings*. New York, NY, USA: Association for Computing Machinery, 2022. (ICSE '22), p. 70–74. ISBN 9781450392235. Disponível em: <https://doi.org/10.1145/3510454.3516830>.

JOSHI, P. et al. *Python: Real world machine learning*. Packt Publishing Ltd, 2016. Disponível em: <https://www.packtpub.com/product/not-set/9781787123212>.

JULIAN, D. *Designing Machine Learning Systems with Python*. Packt Publishing, 2016. (Community experience distilled). ISBN 9781785882951. Disponível em: <https://books.google.com.br/books?id=vxOwDAEACAAJ>.

JUST, R.; SCHWEIGGERT, F. Automating unit and integration testing with partial oracles. *Software Quality Journal*, Springer, v. 19, p. 753–769, 2011. Disponível em: <https://link.springer.com/article/10.1007/s11219-011-9151-x>.

KAMIRAN, F.; CALDERS, T. Classifying without discriminating. In: IEEE. *2009 2nd international conference on computer, control and communication*. 2009. p. 1–6. Disponível em: <https://ieeexplore.ieee.org/document/4909197>.

KAMIRAN, F.; CALDERS, T.; PECHENIZKIY, M. Discrimination aware decision tree learning. In: IEEE. *2010 IEEE international conference on data mining*. 2010. p. 869–874. Disponível em: <https://ieeexplore.ieee.org/document/5694053>.

KANTROWITZ, A. *Microsoft's New AI-Powered Chatbot Mimics A 19-Year-Old American Girl*. BuzzFeed News, 2016. Disponível em: <https://www.buzzfeednews.com/article/alexkantrowitz/microsoft-introduces-tay-an-ai-powered-chatbot-it-hopes-will>.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, v. 30, 2017. Disponível em: <https://dl.acm.org/doi/10.5555/3294996.3295074>.

KHOO, L. S. et al. Exploring and repairing gender fairness violations in word embedding-based sentiment analysis model through adversarial patches. In: *2023 IEEE International Conference on Software Analysis, Evolution*

and Reengineering (SANER). [s.n.], 2023. p. 651–662. Disponível em: <https://ieeexplore.ieee.org/document/10123468>.

LAAN, P. Van der. The 2001 census in the netherlands: Integration of registers and surveys. In: _____. [s.n.], 2001. p. 39–52. Disponível em: <https://www.researchgate.net/publication/269678441_The_2001_Census_in_the_Netherlands_Integration_of_Registers_and_Surveys>.

LOPES, R. *Ferramentas de reconhecimento de imagem podem ser um tanto sexistas nas avaliações*. Giz Brasil, 2017. Disponível em: <https://gizmodo.uol.com.br/ia-aprendizado-sexista/>.

MA, Y.; PAN, Y.; FAN, Y. Metamorphic testing of classification program for the covid-19 intelligent diagnosis. In: *2022 9th International Conference on Dependable Systems and Their Applications (DSA)*. [s.n.], 2022. p. 178–183. Disponível em: <https://ieeexplore.ieee.org/document/9914419>.

MADIEGA, T. A. Eu guidelines on ethics in artificial intelligence: Context and implementation. EPRS: European Parliamentary Research Service, 2019. Disponível em: <https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2019)640163>.

MARTINEZ, E.; LOUZADA-NETO, F. Estimação intervalar via bootstrap. *Revista de Matemática e Estatística*, v. 19, p. 217–251, 2001. Disponível em: <http://www.mat.ufrgs.br/~viali/estatistica/mat2274/material/textos/A12_Artigo.pdf>.

MAYBIN, S. *Sistema de algoritmo que determina pena de condenados cria polêmica nos EUA*. BBC News Brasil, 2016. Disponível em: <https://www.bbc.com/portuguese/brasil-37677421>.

MELO, A. S. d. C. et al. Previsão automática de evasão estudantil: um estudo de caso na ufcg. Universidade Federal de Campina Grande, 2016. Disponível em: <http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/800>.

METZ, C. et al. *Como um chatbot dominou o mundo: veja os bastidores da ascensão do ChatGPT*. 2023. Estadão. Disponível em: <https://www.estadao.com.br/link/empresas/como-um-chatbot-dominou-o-mundo-veja-os-bastidores-da-ascensao-do-chatgpt/>.

MIKHAIL, E. M.; ACKERMANN, F. E. *Observations and Least Squares: With Contributions by F. Ackermann*. IEP, 1976. Disponível em: <https://www.abebooks.com/9780819123978/Observations-Least-Squares-Mikhail-Edward-0819123978/plp>.

MILLER, D.; ZALTA, E. N. *Justice. The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab - Philosophy Department, 2017. Disponível em: <https://plato.stanford.edu/entries/justice/>.

MÜLLER, A. C.; GUIDO, S. *Introduction to machine learning with Python: a guide for data scientists*. "O'Reilly Media, Inc.", 2016. Disponível em: <https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/>.

MULLER, L. *Tay: Twitter conseguiu corromper a IA da Microsoft em menos de 24 horas*. TecMundo, 2016. Disponível em: <https://www.tecmundo.com.br/inteligencia-artificial/102782-tay-twitter-conseguiu-corromper-ia-microsoft-24-horas.htm>.

MURPHY, C.; KAISER, G. E.; HU, L. Properties of machine learning applications for use in metamorphic testing. *Department of Computer Science, Columbia University*, p. 867–872, 01 2008. Disponível em: <https://academiccommons.columbia.edu/doi/10.7916/D8XK8PFD>.

NAKAJIMA, S.; BUI, H. N. Dataset coverage for testing machine learning computer programs. In: IEEE. *2016 23rd Asia-Pacific Software Engineering Conference (APSEC)*. 2016. p. 297–304. Disponível em: <https://ieeexplore.ieee.org/document/7890601/>.

NEFF, G. Talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*, University of Southern California, Annenberg School for Communication, 2016. Disponível em: <https://ijoc.org/index.php/ijoc/article/view/6277>.

PEREDA, C. F. *O Google é racista?* El País, 2016. Disponível em: <https://brasil.elpais.com/brasil/2016/06/10/tecnologia/1465577075_876238.html>.

QUY, T. L. et al. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 12, n. 3, p. e1452, 2022. Disponível em: <https://arxiv.org/abs/2110.00530>.

REZENDE, S. O. et al. *Conceitos sobre Aprendizado de Máquina - Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda, 2003. Disponível em: <https://repositorio.usp.br/item/001718620>.

ROSNEY, D.; RAHMAN-JONES, I. *Anti-bullying campaigners call for a ban on chatbot app SimSimi*. BBC, 2017. Disponível em: <https://www.bbc.com/news/newsbeat-39453778>.

ROSSONI, L.; CHAT, G. A inteligência artificial e eu: escrevendo o editorial juntamente com o chatgpt. *Revista eletrônica de ciência administrativa*, v. 21, n. 3, p. 399–405, 2022. Disponível em: <https://www.periodicosibepes.org.br/index.php/recadm/article/view/3761>.

SALAS, J. *Google conserta seu algoritmo "racista" apagando os gorilas*. El País, 2018. Disponível em: <https://brasil.elpais.com/brasil/2018/01/14/tecnologia/1515955554_803955.html>.

SANTOS, G. *Creating an Ensemble Voting Classifier with Scikit-Learn*. 2022. Towards Data Science. Disponível em: <https://towardsdatascience.com/creating-an-ensemble-voting-classifier-with-scikit-learn-ab13159662d>.

SANTOS, S. H. et al. An experimental study on applying metamorphic testing in machine learning applications. In: *Proceedings of the 5th Brazilian Symposium on Systematic and Automated Software Testing*. [s.n.], 2020. p. 98–106. Disponível em: <https://dl.acm.org/doi/10.1145/3425174.3425226>.

SANTOS, V. B. et al. Um ensemble baseado em árvores de decisão para predizer a ocorrência de aglomerados de ônibus. Universidade Federal de Campina Grande, 2020. Disponível em: <http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/17873>.

SEGURA, S. et al. A survey on metamorphic testing. *IEEE Transactions on Software Engineering*, v. 42, n. 9, p. 805–824, 2016. Disponível em: <https://ieeexplore.ieee.org/document/7422146>.

SEROKELL. *Top Areas for Machine Learning in 2020*. Better Programming, 2020. Disponível em: <https://betterprogramming.pub/the-top-areas-for-machine-learning-in-2020-4c880bf5e288>.

SHARMA, A.; WEHRHEIM, H. Testing machine learning algorithms for balanced data usage. *12th IEEE Conference on Software Testing, Validation and Verification*, 2019. Disponível em: <https://ieeexplore.ieee.org/document/8730187>.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. *Introdução à mineração de dados: Com Aplicações em R*. Elsevier Academic, 2017. Disponível em: <https://www.grupogen.com.br/introducao-a-mineracao-de-dados-com-aplicacoes-em-r>.

SIMONITE, T. *When It Comes to Gorillas, Google Photos Remains Blind*. Wired, 2018. Disponível em: <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>.

SKEEM, J.; LOUDEN, J. E. Assessment of evidence on the quality of the correctional offender management profiling for alternative sanctions (compas). *Unpublished report prepared for the California Department of Corrections and Rehabilitation.*, 2007. Disponível em: <https://webfiles.uci.edu/skeem/Downloads.html>.

THORNDIKE, R. L. Concepts of culture-fairness. *Journal of Educational Measurement*, [National Council on Measurement in Education, Wiley], v. 8, n. 2, p. 63–70, 1971. ISSN 00220655, 17453984. Disponível em: <http://www.jstor.org/stable/1433959>.

TORIKOSHI, Y.; NISHI, Y.; TAKAHASHI, J. *Sensitive Region-based Metamorphic Testing Framework using Explainable AI*. 2023. Disponível em: <https://ieeexplore.ieee.org/document/10190404>.

VERMA, S.; RUBIN, J. Fairness definitions explained. *FairWare '18: Proceedings of the International Workshop on Software Fairness*, p. 1–7, 2018. Disponível em: <https://doi.org/10.1145/3194770.3194776>.

VIEIRA, C. *Inteligência Artificial: a caixa preta que prejudica as minorias*. 2019. Disponível em: <https://imasters.com.br/desenvolvimento/inteligencia-artificial-caixa-preta-que-prejudica-minorias>.

WEXLER, J. et al. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, IEEE, v. 26, n. 1, p. 56–65, 2019. Disponível em: <https://arxiv.org/abs/1907.04135>.

WEYUKER, E. J. On testing non-testable programs. *The Computer Journal*, The British Computer Society, v. 25, n. 4, p. 465–470, 1982. Disponível em: <https://academic.oup.com/comjnl/article/25/4/465/366384>.

WITTEN, I. H.; EIBE, F.; HALL, M. A. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, 2011. v. 3. Disponível em: <https://www.elsevier.com/books/data-mining/witten/978-0-12-374856-0>.

WOLPERT, D. H. Stacked generalization. *Neural networks*, Elsevier, v. 5, n. 2, p. 241–259, 1992. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0893608005800231>.

XIE, X. et al. Testing and validating machine learning classifiers by metamorphic testing. *Journal of Systems and Software*, v. 84, n. 4, p. 544–558, 2011. ISSN 0164-1212. The Ninth International Conference on Quality Software. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0164121210003213>.

ZHANG, J. M. et al. Machine learning testing: Survey, landscapes and horizons. *Transactions On Software Engineering*, 2019. Disponível em: <https://ieeexplore.ieee.org/document/9000651>.

ZHANG, Y.; TOWEY, D.; PIKE, M. Automated metamorphic-relation generation with chatgpt: An experience report. In: *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*. [s.n.], 2023. p. 1780–1785. Disponível em: <https://ieeexplore.ieee.org/document/10196883>.

ZHOU, Z. Q. et al. Automated functional testing of online search services. *Software Testing, Verification and Reliability*, Wiley Online Library, v. 22, n. 4, p. 221–243, 2012. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/stvr.437>.

A Análise Exploratória de Dados: Dutch Census 2001

Recapitulando informações sobre o conjunto de dados utilizado na presente pesquisa, o mesmo é disponibilizado pelos autores do trabalho realizado em (QUY et al., 2022), se tratando de uma amostra de um conjunto de dados maior construído pelo *IPUMS International*. O *IPUMS*[1] é um projeto com colaboração da Universidade de Minnesota e diversos outros institutos internacionais de estatística e também de arquivamento de dados, que conseguiu – e consegue até então – inventariar dados censitários de diversos países e épocas.

O *Dutch Census 2001* se trata de um conjunto de informações referentes a pessoas vivendo na Holanda por volta do ano de 2001 construído por Paul Van der Laan visando uma melhor disseminação dos dados do censo holandês, para que estatísticos possam realizar análises em cima das informações de formas mais simples, tal construção foi documentada em (LAAN, 2001). O conjunto contém descrições tanto sociais quanto profissionais de indivíduos, que se tratando da amostra utilizada, contém ainda um atributo binário de classe (dois valores possíveis), *occupation*, que representa se o indivíduo possui uma profissão de alto nível, com certo prestígio na sociedade, ou baixo nível.

Para atividades de classificação automática de informações com ênfase no estudo de *fairness*, com o conjunto de dados é possível utilizar o atributo sensível *sex* – que representa o sexo de cada indivíduo – para analisar e verificar a imparcialidade de modelos durantes as classificações da ocupação dos indivíduos. No total, o conjunto utilizado possui 12 informações de 60420 indivíduos, variando desde o gênero biológico

[1]Página principal do *IPUMS International*: <https://international.ipums.org/international/>.

e idade dos indivíduos, até o nível educacional e o *status* econômico em que o indivíduo se encontra.

Embora a quantidade de indivíduos de cada gênero sejam próximas, 30273 para o gênero feminino e 30147 para o masculino, há uma diferença considerável no nível de ocupação profissional entre ambos, em que 18860 indivíduos do gênero masculino possuem uma ocupação de alto nível, contrastando com apenas 9903 indivíduos do gênero feminino para o mesmo nível. Trazendo a discussão para a escala percentual, do total de indivíduos com alto nível de ocupação profissional, indivíduos do gênero masculino representam cerca de 65,4% deste montante, enquanto que dos indivíduos com baixo nível de ocupação, o gênero feminino representa 64,35%. Na Figura A.1 é possível visualizar a relação das quantidades de indivíduos de cada gênero separados pelo seu nível de ocupação.

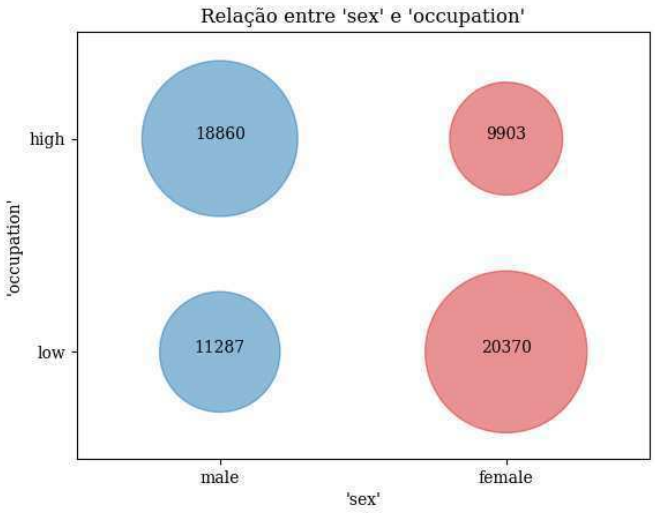


Figura A.1: Proporção de indivíduos do conjunto de dados agrupados por gênero (atributo *'sex'* no Eixo X do gráfico) e nível de ocupação profissional (atributo *'occupation'* no Eixo Y do gráfico.

A.1 Quantidades de Amostras

Passando pelas informações dos indivíduos presentes no conjunto de dados, tem-se a coluna de idade (*'age'*), com valores que representam faixas de idades reais. O valor 8 é aquele com a maior quantidade para o atributo, representando indivíduos que possuem idades de 35 a 39 anos, somando 8748 indivíduos, cerca de 14,5% do total. Elaborando

um *ranking* de valores para este atributo, os cinco valores com as maiores quantidades são os valores de 7 a 11, representando idades de 30 a 54 anos, todos com uma ocorrência de no mínimo 7000 indivíduos por faixa de idade. Na Figura A.2 é possível verificar a distribuição dos valores para o atributo idade, e para facilitar a compreensão da representação de cada valor, a Tabela A.1 apresenta os valores e suas faixas de idades referentes.

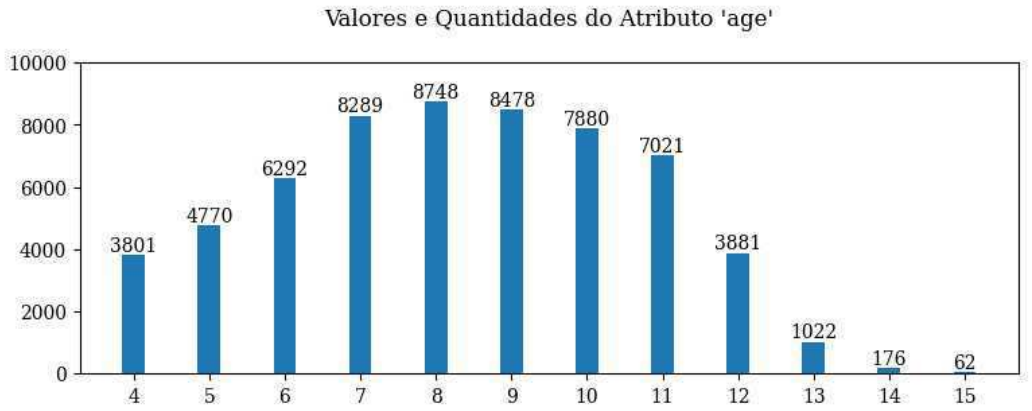


Figura A.2: Quantidade de indivíduos por valores do atributo *'age'*.

Atributo *'age'*			
Valor	Faixa de Idade	Valor	Faixa de Idade
4	15 a 19 anos	10	45 a 49 anos
5	20 a 24 anos	11	50 a 54 anos
6	25 a 29 anos	12	55 a 59 anos
7	30 a 34 anos	13	60 a 64 anos
8	35 a 39 anos	14	65 a 69 anos
9	40 a 44 anos	15	70 a 74 anos

Tabela A.1: Valores do atributo 'age' e suas respectivas faixas de idade do Conjunto de Dados.

O próximo atributo a ser visualizado é o estado civil (*'marital_status'*) dos indivíduos, com valores de 1 a 4. O valor 2 possui a maior quantidade de ocorrências, representando pessoas casadas ou em união com alguém, cerca de 36655 dos indivíduos, seguido do valor 1 que representa pessoas solteiras ou que nunca se casaram, com 19656 indivíduos. Estes dois valores representam cerca de 93,2% dos indivíduos do conjunto. Na Figura A.3 contém um gráfico com as quantidades de indivíduos para cada valor do atributo, sendo o valor 3 representando pessoas separadas ou divorciadas e o valor 4 pessoas com

o cônjuge já falecido.

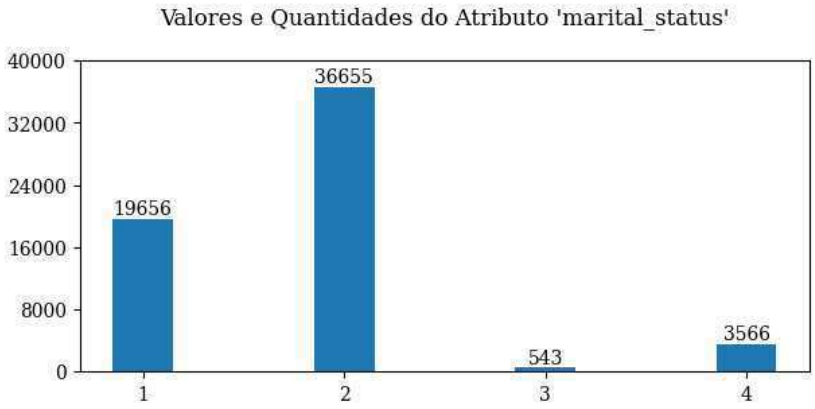
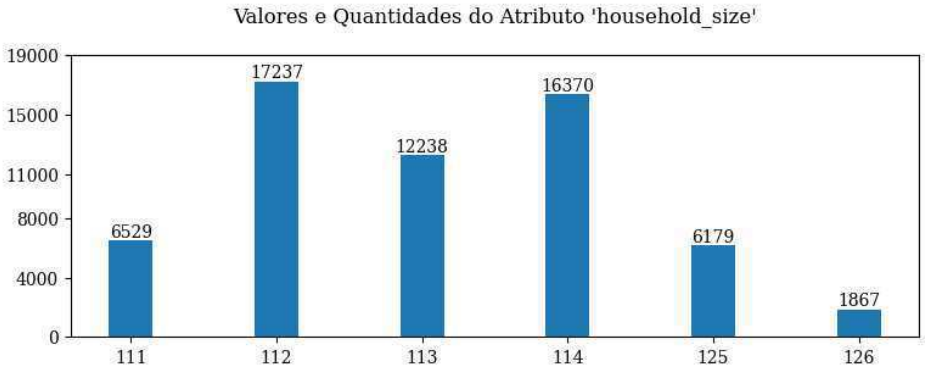


Figura A.3: Quantidade de indivíduos por valores do atributo *'marital_status'*.

Unindo a discussão para o ambiente familiar dos indivíduos, o conjunto de dados possui dois atributos com estes aspectos. O primeiro atributo é o *'household_position'*, que representa uma relação com o líder familiar do domicílio, porém para este atributo não há uma documentação real nas fontes que disponibilizam o conjunto de dados, em que seus valores não foram localizados de fato. O valor 1122 possui 26225 indivíduos, a maior quantidade dentro do atributo, seguido do valor 1121 com 9975 indivíduos, em que pode-se inferir de forma empírica que sejam cônjuges ou filhos no ambiente. O segundo atributo é o *'household_size'* representando a quantidade de integrantes no domicílio em que o indivíduo pertence, com 6 valores distintos, sendo o 112 e 114 com as maiores quantidades, representando 2 e 4 indivíduos, respectivamente.

Na Figura A.4a é possível observar as quantidades de indivíduos para cada valor do atributo *'household_position'* enquanto que na Figura A.4b constam as quantidades do atributo *'household_size'*.

(a) Atributo *'household_position'*.



(b) Atributo *'household_size'*.

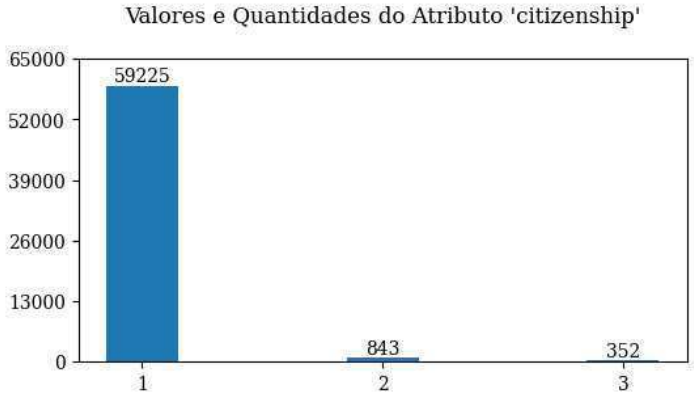
Figura A.4: Quantidade de indivíduos por valores dos atributos *'household_position'* e *'household_size'*.

Analisando as origens dos indivíduos é possível trabalhar com os atributos *'citizenship'*, representando a cidadania do indivíduo, *'country_birth'*, a nacionalidade, e *'prev_residence_place'* que representa o local de residência do indivíduo um ano antes de ser entrevistado para o censo demográfico.

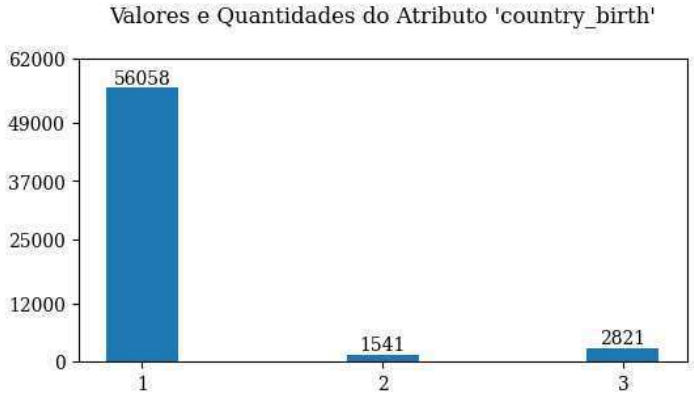
Para o atributo *'citizenship'*, o valor 1 possui a maior ocorrência, pouco mais de 59 mil indivíduos, significando a cidadania holandesa para 98% das pessoas do conjunto de dados. Na Figura A.5a é possível visualizar a quantidade de indivíduos para cada um dos valores.

Assim como o *'citizenship'*, o atributo *'country_birth'* possui uma grande quantidade de amostras do valor 1, indicando que 92,8% dos indivíduos são oriundos da Holanda, enquanto que indivíduos com nacionalidade fora do continente Europeu (valor 3) possui quantidade maior que de indivíduos de demais países europeus (valor 2). Na Figura A.5b é possível visualizar as quantidades de indivíduos para o atributo.

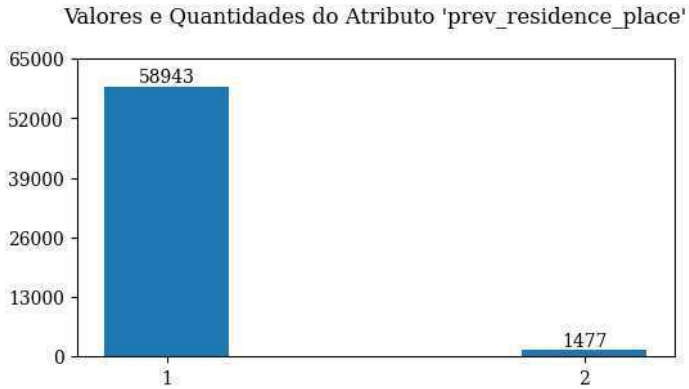
O atributo binário *'prev_residence_place'* – com duas classes possíveis de valores – tem o valor 1 com a maior ocorrência, indicando que 58943 indivíduos moraram na Holanda um ano antes, enquanto apenas 1477 moraram fora da Holanda nesse mesmo período. A Figura A.5c mostra as quantidades para os valores do atributo.



(a) Atributo *'citizenship'*.



(b) Atributo *'country_birth'*.



(c) Atributo *'prev_residence_place'*.

Figura A.5: Quantidade de indivíduos por valores dos atributos *'citizenship'*, *'country_birth'* e *'prev_residence_place'*.

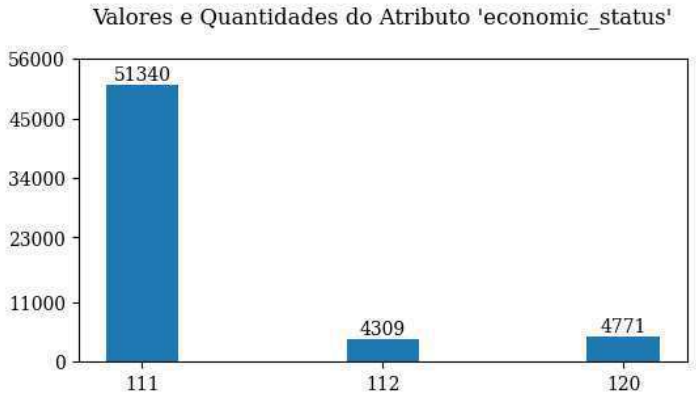
Finalizando a visualização de quantidades por atributos, tem-se os de informações educacionais e econômicas, sendo o *'edu_level'* representando o nível de escolaridade de cada um, o *'economic_status'* como o *status* econômico dos indivíduos, e o *'cur_eco_activity'* como o setor da atividade econômica atual da pessoa.

O *'edu_level'* varia de 0 a 5, sendo o valor 3 com a maior quantidade de amostras, representando o *Upper Secondary* no sistema de educação do país, que para o Brasil seria o equivalente ao Ensino Médio. Tal nível de educação possui cerca de 22672 amostras, seguido do valor 5 que representa o nível *Tertiary* de educação, equivalente ao Ensino Superior, com 18109 indivíduos. Estes dois valores representam cerca de 67,5% do total de indivíduos para o atributo. A Figura A.6a mostra o restante dos valores, sendo o valor 0 o *Pre-Primary*, valor 1 como *Primary*, valor 2 como *Lower Secondary* e valor 4 como *Post Secondary*.

Para o atributo *'economic_status'* existem somente 3 valores, sendo o valor 111 representando pessoas desempregadas, com a maior quantidade de ocorrências para a coluna, representando quase 85% do total de indivíduos. O valor 112 representa as pessoas empregadas, com 4309 amostras e o valor 120 indica indivíduos aposentados, com 4771 indivíduos.

Por fim, o atributo *'cur_eco_activity'* representa o setor de atividade profissional que o individuo tem no momento da entrevista para o censo. Ao todo são 12 valores para o atributo, sendo o valor 131 com a maior quantidade de ocorrências (11621 indivíduos), representando atividades de reparações mecânicas, como de veículos e eletrônicos de uso doméstico. Compondo os 5 valores com maiores ocorrências tem o valor 135 (10239 amostras) representando atividades imobiliárias de venda e aluguel, o valor 138 (8168 amostras) para atividades do setor de saúde e serviço social, o valor 122 (6505 amostras) para o setor de mineração e indústria, e o valor 137 (5862 amostras) referente ao setor de educação. A Figura A.6c apresenta as quantidades de cada um dos valores de *'cur_eco_activity'*, que contém ainda setores de agricultura, caça e pesca (valor 111), construção (valor 124), hotéis e restaurantes (132), transporte, armazenamento e comunicação (valor 133), intermediação financeira (valor 134), administração e segurança pública (valor 136) e atividades de serviços comunitários (valor 139).

(a) Atributo *'edu_level'*.



(b) Atributo *'economic_status'*.

(c) Atributo *'cur_eco_activity'*.

Figura A.6: Quantidade de indivíduos por valores dos atributos *'edu_level'*, *'economic_status'* e *'cur_eco_activity'*.

A.2 Coeficientes de Correlações entre Atributos

Em se tratando de um conjunto de dados de classificação para as amostras, é importante a realização do cálculo de coeficientes de correlação entre cada uma das variáveis

descritivas em relação a variável alvo, que neste caso é o atributo *'occupation'*, com o propósito de verificar a intensidade, direção e também significância entre os valores. Para isto, são feitos os cálculos dos coeficientes das correlações de Pearson, Spearman e Kendall entre os atributos do conjunto de dados, com os resultados apresentados na Figura A.7.

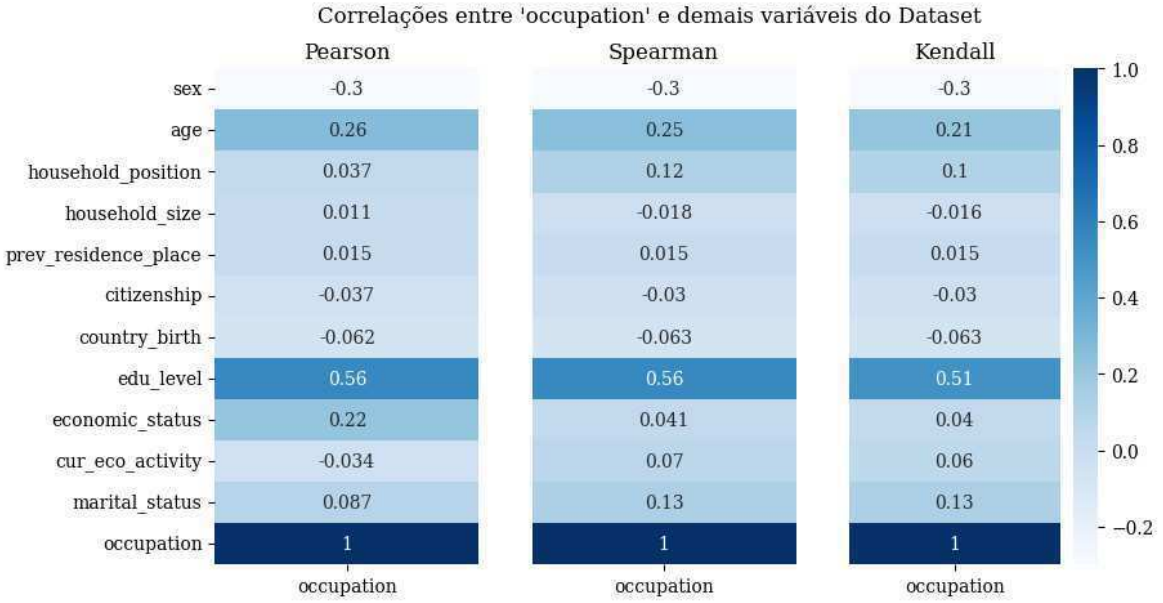


Figura A.7: Valores dos Coeficientes das Correlações de Pearson, Spearman e Kendall entre o atributo alvo, *'occupation'*, com os demais atributos descritivos do conjunto de dados.

Os valores dos coeficientes para a maioria dos atributos descritivos se mostram bem próximos de 0, significando que o atributo *'occupation'* não necessariamente é dependente de forma linear destes atributos. Porém, com relação ao atributo *'edu_level'*, os coeficientes se apresentaram maiores em relação aos demais, atingindo cerca de 0,56 de coeficiente para as correlações de Pearson e Spearman, e 0,51 para a correlação de Kendall, o que pode significar uma correlação moderada entre os valores.

O atributo que apresentou o segundo maior coeficiente das correlações foi o atributo de idade dos indivíduos, atingindo 0,26 para a correlação de Pearson, 0,25 para a correlação de Spearman e 0,21 para a de Kendall. Embora sejam valores próximos de 0, indicando uma não correlação de valores, é importante destacar tal atributo devido a baixa quantidade dentre todos os que atingiram valores de coeficientes acima de 0,1, como o atributo de *'economic_status'*, que alcançou um coeficiente de 0,22 com o atributo *'occupation'*, porém somente para a correlação de Pearson, enquanto que para

as demais correlações seu coeficiente foi de 0,04.

Finalizando as discussões de relação dos atributos descritivos com o atributo alvo, é possível observar que o atributo sensível do conjunto, *'sex'* possui um coeficiente negativo para as correlações, com valor de -0,3, que embora não signifique uma correlação moderada ou forte com o atributo *'occupation'*, o seu valor é importante destacar, visto que é o segundo atributo com a maior correlação, sem levar em conta a direção das correlações dos atributos. Tal valor de coeficiente pode significar que de fato há um nível de ocupação profissional de um indivíduo que possa se identificar com um determinado gênero, mesmo que de forma mínima.

A.3 Atributo sensível *'sex'*

Atingindo o segundo maior índice de correlação com o atributo alvo, *'occupation'*, é importante verificar então o quão relacionado o gênero dos indivíduos pode estar com as demais informações do conjunto de dados, com intuito de descobrir se de fato um determinado gênero está fraca ou fortemente em sincronia com os demais atributos, para que discussões relacionadas a ausência ou garantia de *fairness* possam ser realizadas.

Como mostrado na Figura A.8, dentre os outros 10 atributos descritivos do conjunto, o maior valor de coeficiente alcançado foi com o atributo que indica o setor de atividade profissional atual do indivíduo, porém com valor máximo de 0,22 para a correlação de Spearman, mostrando uma correlação fraca com o gênero dos indivíduos.

Para os demais atributos descritivos, o gênero possui relação fraca de valores, variando de -0,13 (com o atributo de grupo de idade dos indivíduos para a correlação de Pearson e Spearman) a 0,048 (para o atributo de posição do indivíduo em relação ao líder da moradia, para a correlação de Spearman).

Sendo assim, não é possível falar com exatidão que dependendo do gênero, os indivíduos podem ter condições boas ou ruins de vida se analisarmos o nível de educação ou de atividade profissional dos mesmos, por exemplos. Com isto, não é possível elencar com exatidão a discussão de que indivíduos de um determinado gênero pertença a grupos distintos na sociedade de acordo com suas próprias características.

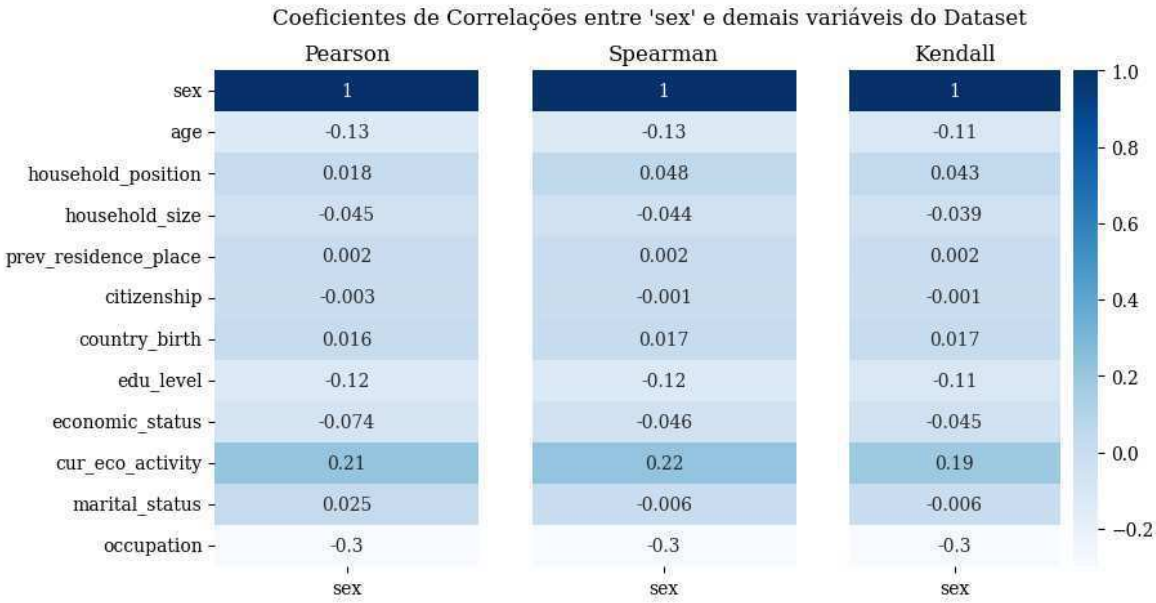


Figura A.8: Valores dos Coeficientes das Correlações de Pearson, Spearman e Kendall entre o atributo descritivo *'sex'*, com os demais atributos do conjunto de dados.

A.4 Atributo descritivo *'edu_level'*

Tendo em vista o atributo *'edu_level'*, que melhor se saiu no cálculo dos coeficientes das correlações (atingindo os maiores valores, independente do julgamento de uma correlação forte ou fraca com o atributo alvo), este mesmo atributo é posto em evidência na Figura A.9 com relação aos demais atributos descritivos.

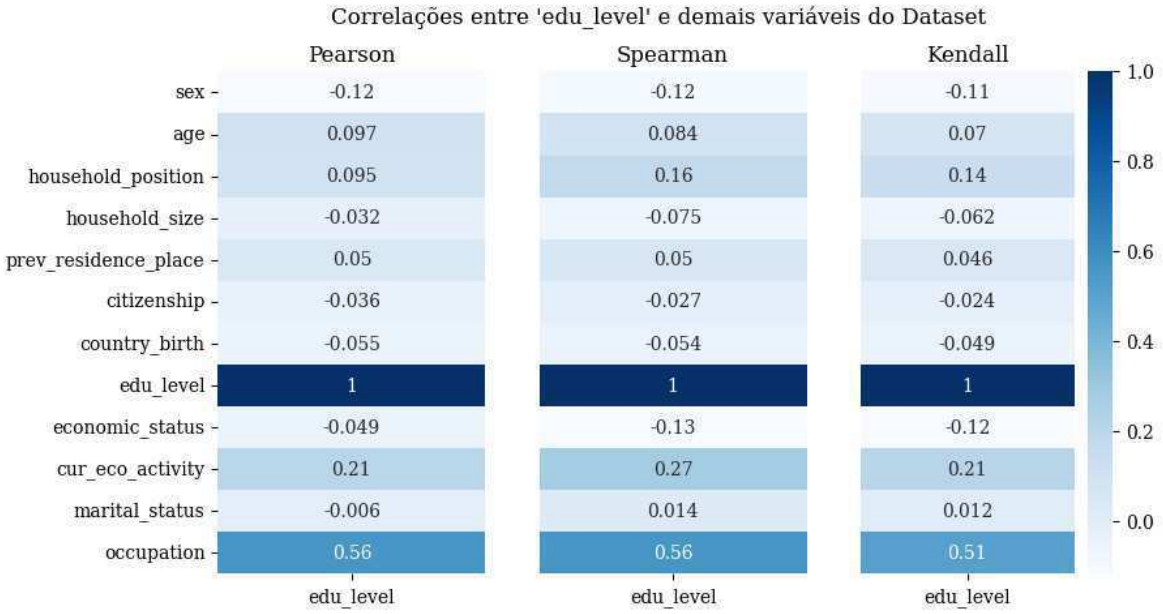


Figura A.9: Valores dos Coeficientes das Correlações de Pearson, Spearman e Kendall entre o atributo descritivo *'edu_level'*, com os demais atributos do conjunto de dados.

Semelhante ao atributo sensível, o atributo que descreve o setor da atividade profissional atual do indivíduo possui o maior coeficiente de correlação com o atributo *'edu_level'*, atingindo um valor de 0,27 para a correlação de Spearman, enquanto que os demais atingem valores entre -0,13 (para o *status* econômico do indivíduo na correlação de Spearman) e 0,16 (para o atributo de relação com o líder da moradia, também na correlação de Spearman).

De forma resumida, a partir dos valores de coeficientes, é nítida uma correlação fraca ou até próxima de nula entre o nível de educação do indivíduo com as demais características do mesmo, porém isto não impede de uma visualização da proporção das informações de nível educacional com o atributo alvo do conjunto de dados, o nível de ocupação.

Na Figura A.10 é possível visualizar a quantidade de amostras de cada valor do nível de educação, postas em comparação com o nível de ocupação do indivíduo, que em um primeiro momento o valor 5 se destaca devido ao desbalanceamento na quantidade de amostras positivas e negativas para o nível de ocupação, sendo 16211 amostras positivas e apenas 1898 negativas, cerca de 11,71% da quantidade total de amostras para o valor 5 do nível de educação. Para o valor 4 as quantidades de nível de ocupação se mostram mais balanceadas, com 1466 para o nível alto de ocupação e 1114 para o nível baixo.

Outro ponto importante observado no gráfico é que até o valor 3 do nível de educação, o nível baixo de ocupação possui as maiores quantidades de indivíduos, enquanto que para os valores 4 e 5 de educação, o cenário inverte, com o nível alto de ocupação passando a ter as maiores quantidades, em comparação com o nível baixo de ocupação, mantendo o mesmo valor de nível de educação.

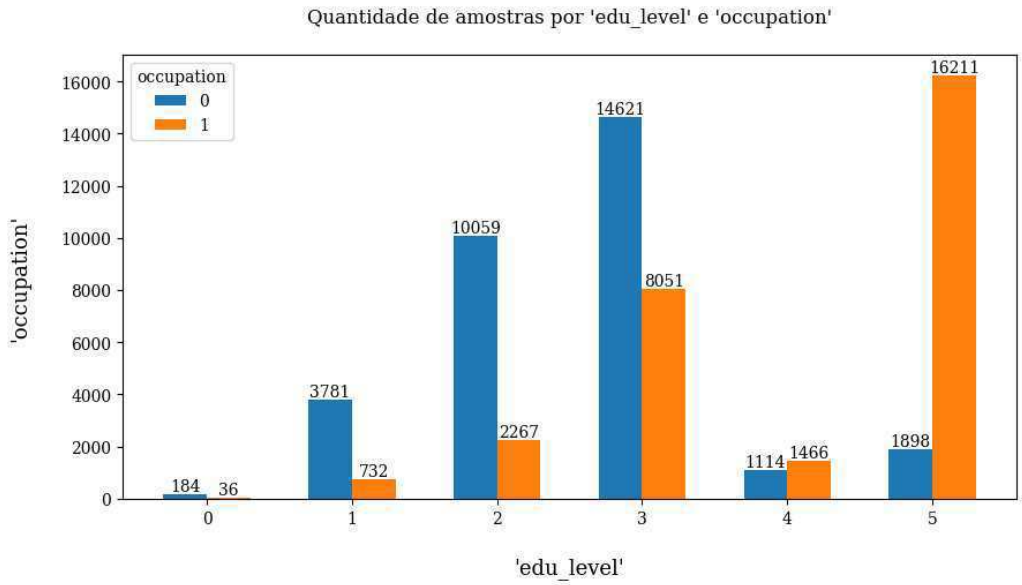


Figura A.10: Quantidade de amostras por valores do atributo *'edu_level'* e separadas pelo atributo *'occupation'*.

Utilizando a mesma abordagem comparativa na Figura A.10, desta vez os valores do *'edu_level'* são comparados em relação aos valores do atributo sensível *'sex'* do conjunto de dados, na Figura A.11. Assim como para o nível de ocupação, a quantidade de pessoas do gênero feminino é maior em relação ao masculino apenas para os valores de 0 a 3 de nível de educação, enquanto que para os valores 4 e 5 as maiores quantidades passam a ser para o gênero masculino. Outra visualização importante é que para nenhum dos valores de educação há um desbalanceamento entre os gêneros na mesma proporção que no valor 5 para o nível de ocupação, mostrando que de certa forma o nível de educação é equiparado entre os gêneros dos indivíduos.

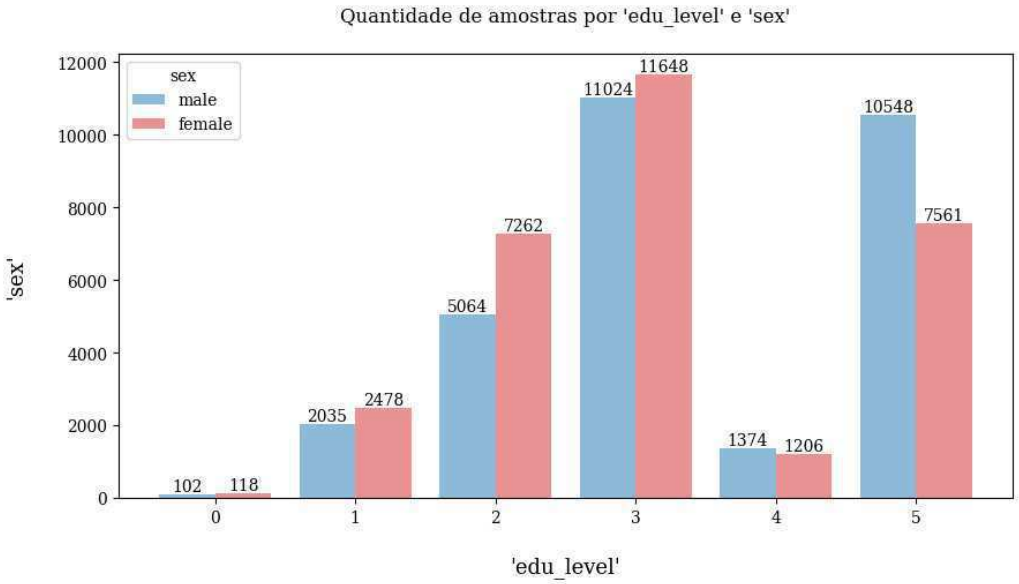


Figura A.11: Quantidade de amostras por valores do atributo *'edu_level'* e separadas pelo atributo *'sex'*.

B Resultados Obtidos no Estudo

A seguir são apresentadas tabelas para cada uma das execuções realizadas durante o estudo, com informações de tempos – em segundos – de:

- Execução de treino dos modelos com os dados originais;
- Aplicação das transformações metamórficas nos conjuntos de dados de treino;
- Execução de treino dos modelos com os dados transformados;
- Classificação dos dados de teste.

As demais linhas das tabelas apresentam informações relativas aos intervalos de confiança da comparação entre as instâncias dos modelos, treinadas com os dados originais (com o favorecimento sendo indicado por valores negativos) e com os dados transformados (com o favorecimento através dos valores positivos). São dispostas todas as métricas de classificação e de *fairness* com os valores dos limites inferiores, superiores e os valores pontuais dos intervalos de confiança. Além da coluna indicando as métricas, as demais são referentes aos modelos de classificação utilizados.

A ordem das tabelas segue a ordem crescente de cada Cenário de Análise e de número de reamostragens do *bootstrap*, com a primeira sendo referente as execuções dos modelos do Cenário 1 com 500 reamostragens e a última referente as 4000 reamostragens do Cenário 4.

Cenário 1 - 500 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	0,5986	0,3693	0,0511	1,7235	1,6658	0,3629	1,785	11,128	3,4178
Tempo de Aplicação de Transformações	52,2236	52,2236	52,2236	52,2236	52,2236	52,2236	52,2236	52,2236	52,2236
Tempo de Treino - Dados Transformados	0,6117	0,359	0,0511	1,6585	1,597	0,475	1,7415	10,9991	3,3584
Tempo de Teste	0,0561	0,019	0,003	0,1263	0,016	0,015	0,1314	0,1454	0,2126
Accuracy: Limite Inferior	-0,0001	-0,0164	-0,0015	0,0132	0	0,0002	-0,0005	-0,0017	0,0001
Accuracy: Valor Pontual	0,0001	-0,0141	-0,0007	0,0159	0	0,0006	-0,0001	0,0006	0,0006
Accuracy: Limite Superior	0,0003	-0,0117	0,0001	0,0186	0	0,001	0,0004	0,0027	0,0012
Balanced Accuracy: Limite Inferior	-0,0001	-0,017	-0,0014	0,0146	0	0,0002	-0,0005	-0,0021	0
Balanced Accuracy: Valor Pontual	0,0001	-0,0147	-0,0007	0,0172	0	0,0006	-0,0001	0	0,0006
Balanced Accuracy: Limite Superior	0,0003	-0,0122	0,0001	0,0199	0	0,0011	0,0004	0,0021	0,0011
Precision: Limite Inferior	-0,0101	-0,1454	-0,0049	-0,0082	0	-0,0144	-0,0064	-0,0275	0,0043
Precision: Valor Pontual	-0,0028	-0,1227	-0,003	0,0005	0	-0,0044	-0,001	0,0129	0,0118
Precision: Limite Superior	0,0007	-0,0995	-0,0012	0,009	0	0,0028	0,005	0,0559	0,0211
Recall: Limite Inferior	0	-0,0341	-0,0003	0,0455	0	0,0008	-0,001	-0,017	-0,0009
Recall: Valor Pontual	0,0002	-0,0299	0,0007	0,0501	0	0,0015	-0,0005	-0,0144	-0,0001
Recall: Limite Superior	0,0006	-0,0255	0,0018	0,0548	0	0,0023	0,0001	-0,0116	0,0006
F1-Score: Limite Inferior	0	-0,0379	-0,0011	0,0537	0	0,0015	-0,0018	-0,0305	-0,0014
F1-Score: Valor Pontual	0,0004	-0,0306	0	0,0595	0	0,0028	-0,0008	-0,0257	-0,0001
F1-Score: Limite Superior	0,0011	-0,0433	0,0011	0,0655	0	0,0043	0,0001	-0,0206	0,0012
Statistical Parity: Limite Inferior	-0,0008	0,0225	0,0045	-0,0673	0	-0,0025	-0,0002	0,0347	0,0008
Statistical Parity: Valor Pontual	-0,0003	0,0271	0,0061	-0,0622	0	-0,0017	0,0008	0,0392	0,0019
Statistical Parity: Limite Superior	0	0,032	0,0077	-0,0569	0	-0,001	0,0018	0,0433	0,003
Equalized Odds: Limite Inferior	-0,0009	-0,0055	0,0041	-0,0837	0	-0,0036	-0,0012	0,0496	-0,0007
Equalized Odds: Valor Pontual	-0,0004	0,0006	0,0065	-0,0761	0	-0,0023	0,0009	0,0598	0,0007
Equalized Odds: Limite Superior	0	0,0168	0,009	-0,0693	0	-0,0012	0,0029	0,0674	0,0019

Figura B.1: Resultados alcançados no Cenário 1 para as execuções de 500 reamostragens.

Cenário 1 - 1000 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	0,6387	0,381	0,0541	1,7369	1,656	0,3629	1,7753	11,2276	3,4501
Tempo de Aplicação de Transformações	57,2349	57,2349	57,2349	57,2349	57,2349	57,2349	57,2349	57,2349	57,2349
Tempo de Treino - Dados Transformados	0,6227	0,374	0,0531	1,7265	1,6267	0,4803	1,7632	11,1669	3,4478
Tempo de Teste	0,0592	0,019	0,002	0,1233	0,016	0,0111	0,1304	0,1464	0,2166
Accuracy: Limite Inferior	-0,0001	-0,0164	-0,0015	0,0133	0	0,0002	-0,0005	-0,0017	0,0001
Accuracy: Valor Pontual	0,0001	-0,0141	-0,0007	0,0159	0	0,0006	-0,0001	0,0006	0,0006
Accuracy: Limite Superior	0,0003	-0,0117	0,0001	0,0186	0	0,001	0,0004	0,0028	0,0012
Balanced Accuracy: Limite Inferior	-0,0001	-0,017	-0,0014	0,0147	0	0,0003	-0,0005	-0,0022	0
Balanced Accuracy: Valor Pontual	0,0001	-0,0147	-0,0007	0,0172	0	0,0006	-0,0001	0	0,0006
Balanced Accuracy: Limite Superior	0,0003	-0,0122	0,0001	0,0199	0	0,0011	0,0004	0,0021	0,0011
Precision: Limite Inferior	-0,0096	-0,1449	-0,005	-0,0083	0	-0,014	-0,0065	-0,0273	0,0038
Precision: Valor Pontual	-0,0028	-0,1227	-0,003	0,0005	0	-0,0044	-0,001	0,0129	0,0118
Precision: Limite Superior	0,0007	-0,1008	-0,0011	0,009	0	0,0028	0,0047	0,0559	0,021
Recall: Limite Inferior	0	-0,0343	-0,0003	0,0456	0	0,0008	-0,001	-0,0171	-0,0008
Recall: Valor Pontual	0,0002	-0,0299	0,0007	0,0501	0	0,0015	-0,0005	-0,0144	-0,0001
Recall: Limite Superior	0,0006	-0,0256	0,0018	0,0547	0	0,0024	0	-0,0115	0,0006
F1-Score: Limite Inferior	0	-0,0579	-0,0012	0,0538	0	0,0015	-0,0018	-0,0307	-0,0014
F1-Score: Valor Pontual	0,0004	-0,0506	0	0,0595	0	0,0028	-0,0008	-0,0257	-0,0001
F1-Score: Limite Superior	0,0011	-0,0434	0,0011	0,0654	0	0,0044	0,0001	-0,0205	0,0012
Statistical Parity: Limite Inferior	-0,0008	0,0224	0,0045	-0,0673	0	-0,0025	-0,0001	0,0348	0,0008
Statistical Parity: Valor Pontual	-0,0003	0,0271	0,0061	-0,0622	0	-0,0017	0,0008	0,0392	0,0019
Statistical Parity: Limite Superior	0	0,0319	0,0076	-0,057	0	-0,001	0,0017	0,0433	0,0031
Equalized Odds: Limite Inferior	-0,0009	-0,0055	0,0041	-0,0834	0	-0,0037	-0,0012	0,0497	-0,0005
Equalized Odds: Valor Pontual	-0,0004	0,0006	0,0065	-0,0761	0	-0,0023	0,0009	0,0598	0,0007
Equalized Odds: Limite Superior	0	0,0162	0,0091	-0,0693	0	-0,0012	0,0029	0,0673	0,002

Figura B.2: Resultados alcançados no Cenário 1 para as execuções de 1000 reamostragens.

Cenário 1 - 2000 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	0,6347	0,3728	0,0516	1,7014	1,6766	0,372	1,817	11,3119	3,7032
Tempo de Aplicação de Transformações	54,1782	54,1782	54,1782	54,1782	54,1782	54,1782	54,1782	54,1782	54,1782
Tempo de Treino - Dados Transformados	0,6081	0,3709	0,0531	1,7033	1,6212	0,4793	1,7446	11,1924	3,4801
Tempo de Teste	0,0582	0,019	0,002	0,1281	0,016	0,011	0,1323	0,1496	0,2146
Accuracy: Limite Inferior	-0,0001	-0,0164	-0,0015	0,0133	0	0,0002	-0,0005	-0,0017	0,0001
Accuracy: Valor Pontual	0,0001	-0,0141	-0,0007	0,0159	0	0,0006	-0,0001	0,0006	0,0006
Accuracy: Limite Superior	0,0002	-0,0117	0,0001	0,0186	0	0,001	0,0004	0,0027	0,0012
Balanced Accuracy: Limite Inferior	-0,0001	-0,017	-0,0015	0,0146	0	0,0002	-0,0005	-0,0021	0
Balanced Accuracy: Valor Pontual	0,0001	-0,0147	-0,0007	0,0172	0	0,0006	-0,0001	0	0,0006
Balanced Accuracy: Limite Superior	0,0002	-0,0122	0,0001	0,0199	0	0,0011	0,0004	0,0021	0,0012
Precision: Limite Inferior	-0,0096	-0,144	-0,005	-0,0079	0	-0,0144	-0,0067	-0,0275	0,0036
Precision: Valor Pontual	-0,0028	-0,1227	-0,003	0,0005	0	-0,0044	-0,001	0,0129	0,0118
Precision: Limite Superior	0,0007	-0,1011	-0,0011	0,0091	0	0,0028	0,0048	0,0545	0,0212
Recall: Limite Inferior	0	-0,0342	-0,0005	0,0455	0	0,0008	-0,001	-0,017	-0,0008
Recall: Valor Pontual	0,0002	-0,0299	0,0007	0,0501	0	0,0015	-0,0005	-0,0144	-0,0001
Recall: Limite Superior	0,0006	-0,0256	0,0018	0,0547	0	0,0024	0,0001	-0,0116	0,0007
F1-Score: Limite Inferior	0	-0,0579	-0,0012	0,0537	0	0,0015	-0,0018	-0,0307	-0,0014
F1-Score: Valor Pontual	0,0004	-0,0506	0	0,0595	0	0,0028	-0,0008	-0,0257	-0,0001
F1-Score: Limite Superior	0,0011	-0,0434	0,0011	0,0654	0	0,0044	0,0002	-0,0205	0,0013
Statistical Parity: Limite Inferior	-0,0008	0,0224	0,0045	-0,0675	0	-0,0026	-0,0002	0,0349	0,0008
Statistical Parity: Valor Pontual	-0,0003	0,0271	0,0061	-0,0622	0	-0,0017	0,0008	0,0392	0,0019
Statistical Parity: Limite Superior	0	0,0319	0,0077	-0,0569	0	-0,0009	0,0018	0,0433	0,0031
Equalized Odds: Limite Inferior	-0,0009	-0,0053	0,004	-0,0834	0	-0,0036	-0,0012	0,0498	-0,0005
Equalized Odds: Valor Pontual	-0,0004	0,0006	0,0065	-0,0761	0	-0,0023	0,0009	0,0598	0,0007
Equalized Odds: Limite Superior	0	0,0153	0,0092	-0,0691	0	-0,0012	0,003	0,0671	0,002

Figura B.3: Resultados alcançados no Cenário 1 para as execuções de 2000 reamostragens.

Cenário 1 - 4000 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	0,6262	0,369	0,0571	1,7204	1,6315	0,366	3,0334	20,3308	3,7041
Tempo de Aplicação de Transformações	54,6493	54,6493	54,6493	54,6493	54,6493	54,6493	54,6493	54,6493	54,6493
Tempo de Treino - Dados Transformados	0,6167	0,3647	0,0519	1,6694	1,6312	0,4803	3,0367	14,4255	3,5286
Tempo de Teste	0,0596	0,0198	0,003	0,1283	0,017	0,013	0,2281	0,2788	0,2244
Accuracy: Limite Inferior	-0,0001	-0,0164	-0,0015	0,0132	0	0,0002	-0,0005	-0,0016	0,0001
Accuracy: Valor Pontual	0,0001	-0,0141	-0,0007	0,0159	0	0,0006	-0,0001	0,0006	0,0006
Accuracy: Limite Superior	0,0002	-0,0118	0,0001	0,0185	0	0,001	0,0004	0,0026	0,0012
Balanced Accuracy: Limite Inferior	-0,0001	-0,017	-0,0015	0,0144	0	0,0002	-0,0005	-0,0021	0
Balanced Accuracy: Valor Pontual	0,0001	-0,0147	-0,0007	0,0172	0	0,0006	-0,0001	0	0,0006
Balanced Accuracy: Limite Superior	0,0002	-0,0122	0,0001	0,0199	0	0,0011	0,0004	0,002	0,0012
Precision: Limite Inferior	-0,0097	-0,1437	-0,005	-0,008	0	-0,0143	-0,0067	-0,0278	0,0035
Precision: Valor Pontual	-0,0028	-0,1227	-0,003	0,0005	0	-0,0044	-0,001	0,0129	0,0118
Precision: Limite Superior	0,0007	-0,1011	-0,0011	0,0089	0	0,0026	0,0048	0,0529	0,021
Recall: Limite Inferior	0	-0,0341	-0,0004	0,0453	0	0,0008	-0,001	-0,017	-0,0009
Recall: Valor Pontual	0,0002	-0,0299	0,0007	0,0501	0	0,0015	-0,0005	-0,0144	-0,0001
Recall: Limite Superior	0,0006	-0,0257	0,0018	0,0547	0	0,0023	0	-0,0116	0,0007
F1-Score: Limite Inferior	0	-0,0577	-0,0012	0,0536	0	0,0015	-0,0018	-0,0306	-0,0014
F1-Score: Valor Pontual	0,0004	-0,0506	0	0,0595	0	0,0028	-0,0008	-0,0257	-0,0001
F1-Score: Limite Superior	0,0011	-0,0435	0,0011	0,0654	0	0,0044	0,0001	-0,0206	0,0013
Statistical Parity: Limite Inferior	-0,0008	0,0225	0,0045	-0,0674	0	-0,0025	-0,0002	0,035	0,0008
Statistical Parity: Valor Pontual	-0,0003	0,0271	0,0061	-0,0622	0	-0,0017	0,0008	0,0392	0,0019
Statistical Parity: Limite Superior	0	0,0318	0,0078	-0,057	0	-0,0009	0,0018	0,0432	0,0031
Equalized Odds: Limite Inferior	-0,0009	-0,0052	0,004	-0,0833	0	-0,0036	-0,0012	0,0499	-0,0005
Equalized Odds: Valor Pontual	-0,0004	0,0006	0,0065	-0,0761	0	-0,0023	0,0009	0,0598	0,0007
Equalized Odds: Limite Superior	0	0,0151	0,0092	-0,0689	0	-0,0012	0,003	0,067	0,002

Figura B.4: Resultados alcançados no Cenário 1 para as execuções de 4000 reamostragens.

Cenário 2 - 500 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	0,6798	0,4301	0,0563	1,8338	1,8612	0,3609	1,8208	11,6626	3,8285
Tempo de Aplicação de Transformações	57,5124	57,5124	57,5124	57,5124	57,5124	57,5124	57,5124	57,5124	57,5124
Tempo de Treino - Dados Transformados	0,6778	0,4167	0,0642	1,8479	1,8418	0,4843	1,8249	11,6451	3,7603
Tempo de Teste	0,0652	0,0211	0,0029	0,1765	0,0221	0,014	0,1614	0,1848	0,2607
Accuracy: Limite Inferior	-0,0009	0,0013	-0,0397	-0,0003	0,0006	-0,0015	-0,0023	-0,0084	-0,0023
Accuracy: Valor Pontual	-0,0004	0,0051	-0,0356	0,0024	0,0017	-0,0004	0	-0,0054	-0,0003
Accuracy: Limite Superior	0,0002	0,0087	-0,0316	0,0049	0,0028	0,0007	0,0024	-0,0024	0,0019
Balanced Accuracy: Limite Inferior	-0,0009	0,0013	-0,057	-0,0009	0,0005	-0,0014	-0,0029	-0,009	-0,0026
Balanced Accuracy: Valor Pontual	-0,0004	0,005	-0,0531	0,0018	0,0016	-0,0003	-0,0006	-0,0058	-0,0006
Balanced Accuracy: Limite Superior	0,0002	0,0086	-0,0492	0,0044	0,0027	0,0008	0,0017	-0,0029	0,0016
Precision: Limite Inferior	-0,001	0,0016	-0,0715	0,0053	0,001	-0,0027	0,0042	-0,004	0,0005
Precision: Valor Pontual	-0,0005	0,0056	-0,0672	0,0084	0,0022	-0,0016	0,0063	-0,0008	0,0025
Precision: Limite Superior	0	0,0092	-0,0629	0,0112	0,0034	-0,0005	0,0085	0,0024	0,0046
Recall: Limite Inferior	-0,0009	-0,0028	0,006	-0,016	-0,0017	0,0006	-0,0204	-0,0214	-0,0111
Recall: Valor Pontual	0	0,0026	0,0101	-0,0127	-0,0006	0,0022	-0,0171	-0,0169	-0,008
Recall: Limite Superior	0,0008	0,0076	0,015	-0,0097	0,0006	0,0037	-0,014	-0,0131	-0,0048
F1-Score: Limite Inferior	-0,0009	0,0008	-0,0401	-0,0029	0,0002	-0,001	-0,0055	-0,0103	-0,0037
F1-Score: Valor Pontual	-0,0003	0,0043	-0,0366	-0,0006	0,0011	0	-0,0033	-0,0074	-0,0018
F1-Score: Limite Superior	0,0002	0,0079	-0,033	0,0017	0,0021	0,0012	-0,001	-0,0047	0,0003
Statistical Parity: Limite Inferior	-0,0036	-0,0253	0,1368	0,0056	-0,0131	0,002	0,029	-0,0042	0,0184
Statistical Parity: Valor Pontual	-0,0025	-0,0187	0,145	0,0108	-0,0107	0,0042	0,0336	0,0021	0,0229
Statistical Parity: Limite Superior	-0,0013	-0,0111	0,1532	0,0157	-0,0085	0,0062	0,0382	0,0084	0,0271
Equalized Odds: Limite Inferior	-0,0043	-0,0274	0,1804	0,0123	-0,0187	0,0012	0,0347	-0,022	0,022
Equalized Odds: Valor Pontual	-0,0025	-0,0137	0,1907	0,0234	-0,0147	0,0045	0,0427	-0,01	0,0294
Equalized Odds: Limite Superior	-0,0009	0,0004	0,2012	0,0332	-0,0111	0,0075	0,0505	0,0017	0,0373

Figura B.5: Resultados alcançados no Cenário 2 para as execuções de 500 reamostragens.

Cenário 2 - 1000 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	0,6878	0,4231	0,0581	1,8474	1,8455	0,3532	1,8476	11,6244	3,7429
Tempo de Aplicação de Transformações	57,2364	57,2364	57,2364	57,2364	57,2364	57,2364	57,2364	57,2364	57,2364
Tempo de Treino - Dados Transformados	0,6803	0,4091	0,0582	1,786	1,8388	0,4963	1,8448	11,5604	3,73
Tempo de Teste	0,0632	0,0221	0,002	0,1745	0,018	0,0121	0,1634	0,1866	0,2537
Accuracy: Limite Inferior	-0,0009	0,0013	-0,0601	-0,0003	0,0006	-0,0015	-0,0022	-0,0086	-0,0024
Accuracy: Valor Pontual	-0,0004	0,0051	-0,0556	0,0024	0,0017	-0,0004	0	-0,0054	-0,0003
Accuracy: Limite Superior	0,0002	0,0087	-0,0515	0,005	0,0029	0,0006	0,0024	-0,0023	0,0018
Balanced Accuracy: Limite Inferior	-0,0009	0,0013	-0,0571	-0,0009	0,0005	-0,0014	-0,0029	-0,0089	-0,0027
Balanced Accuracy: Valor Pontual	-0,0004	0,005	-0,0531	0,0018	0,0016	-0,0003	-0,0006	-0,0058	-0,0006
Balanced Accuracy: Limite Superior	0,0002	0,0086	-0,0492	0,0043	0,0027	0,0007	0,0016	-0,0029	0,0015
Precision: Limite Inferior	-0,001	0,0015	-0,0716	0,0054	0,0011	-0,0027	0,0043	-0,0041	0,0005
Precision: Valor Pontual	-0,0005	0,0056	-0,0672	0,0084	0,0022	-0,0016	0,0063	-0,0008	0,0025
Precision: Limite Superior	0	0,0094	-0,0629	0,0112	0,0034	-0,0005	0,0084	0,0024	0,0044
Recall: Limite Inferior	-0,0009	-0,0025	0,0057	-0,0159	-0,0018	0,0007	-0,0205	-0,0208	-0,0113
Recall: Valor Pontual	0	0,0026	0,0101	-0,0127	-0,0006	0,0022	-0,0171	-0,0169	-0,008
Recall: Limite Superior	0,0008	0,0077	0,0147	-0,0095	0,0007	0,0038	-0,014	-0,0129	-0,0048
F1-Score: Limite Inferior	-0,0009	0,0008	-0,0404	-0,003	0,0002	-0,001	-0,0054	-0,0102	-0,0039
F1-Score: Valor Pontual	-0,0003	0,0043	-0,0366	-0,0006	0,0011	0	-0,0033	-0,0074	-0,0018
F1-Score: Limite Superior	0,0002	0,008	-0,033	0,0018	0,0021	0,0011	-0,0012	-0,0046	0,0002
Statistical Parity: Limite Inferior	-0,0037	-0,0262	0,1371	0,0055	-0,0131	0,0021	0,0289	-0,0039	0,0182
Statistical Parity: Valor Pontual	-0,0025	-0,0187	0,145	0,0108	-0,0107	0,0042	0,0336	0,0021	0,0229
Statistical Parity: Limite Superior	-0,0014	-0,0112	0,1535	0,0157	-0,0085	0,0063	0,0382	0,0084	0,0275
Equalized Odds: Limite Inferior	-0,0043	-0,0273	0,1804	0,0134	-0,0185	0,0012	0,0347	-0,0219	0,0219
Equalized Odds: Valor Pontual	-0,0025	-0,0137	0,1907	0,0234	-0,0147	0,0045	0,0427	-0,01	0,0294
Equalized Odds: Limite Superior	-0,0009	-0,0005	0,2015	0,0334	-0,0112	0,0076	0,0509	0,0018	0,0372

Figura B.6: Resultados alcançados no Cenário 2 para as execuções de 1000 reamostragens.

Cenário 2 - 2000 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	0,693	0,4221	0,0582	1,9068	1,8426	0,3616	1,8904	11,6339	3,6693
Tempo de Aplicação de Transformações	58,4548	58,4548	58,4548	58,4548	58,4548	58,4548	58,4548	58,4548	58,4548
Tempo de Treino - Dados Transformados	0,6878	0,4101	0,0592	1,9231	1,8478	0,4826	1,8435	11,6351	3,6931
Tempo de Teste	0,0639	0,0221	0,002	0,1845	0,018	0,016	0,1644	0,1882	0,2487
Accuracy: Limite Inferior	-0,0009	0,0014	-0,0601	-0,0002	0,0007	-0,0015	-0,0023	-0,0086	-0,0024
Accuracy: Valor Pontual	-0,0004	0,0051	-0,0556	0,0024	0,0017	-0,0004	0	-0,0054	-0,0003
Accuracy: Limite Superior	0,0002	0,0087	-0,0514	0,005	0,0029	0,0006	0,0024	-0,0023	0,0019
Balanced Accuracy: Limite Inferior	-0,0009	0,0014	-0,0573	-0,0007	0,0006	-0,0014	-0,0029	-0,0089	-0,0027
Balanced Accuracy: Valor Pontual	-0,0004	0,005	-0,0531	0,0018	0,0016	-0,0003	-0,0006	-0,0058	-0,0006
Balanced Accuracy: Limite Superior	0,0002	0,0086	-0,0492	0,0043	0,0028	0,0007	0,0016	-0,0027	0,0016
Precision: Limite Inferior	-0,001	0,0018	-0,0718	0,0057	0,0011	-0,0027	0,0042	-0,0041	0,0004
Precision: Valor Pontual	-0,0005	0,0056	-0,0672	0,0084	0,0022	-0,0016	0,0063	-0,0008	0,0025
Precision: Limite Superior	0	0,0093	-0,0628	0,0112	0,0034	-0,0005	0,0085	0,0024	0,0045
Recall: Limite Inferior	-0,0009	-0,0024	0,0055	-0,0157	-0,0017	0,0007	-0,0205	-0,0209	-0,0111
Recall: Valor Pontual	0	0,0026	0,0101	-0,0127	-0,0006	0,0022	-0,0171	-0,0169	-0,008
Recall: Limite Superior	0,0009	0,0077	0,0146	-0,0096	0,0007	0,0038	-0,0139	-0,0127	-0,005
F1-Score: Limite Inferior	-0,0009	0,0009	-0,0404	-0,0029	0,0002	-0,001	-0,0055	-0,0103	-0,0038
F1-Score: Valor Pontual	-0,0003	0,0043	-0,0366	-0,0006	0,0011	0	-0,0033	-0,0074	-0,0018
F1-Score: Limite Superior	0,0003	0,0079	-0,033	0,0017	0,0021	0,0011	-0,0012	-0,0045	0,0002
Statistical Parity: Limite Inferior	-0,0037	-0,0261	0,1371	0,0056	-0,0131	0,0021	0,0289	-0,0041	0,0184
Statistical Parity: Valor Pontual	-0,0025	-0,0187	0,145	0,0108	-0,0107	0,0042	0,0336	0,0021	0,0229
Statistical Parity: Limite Superior	-0,0014	-0,0112	0,153	0,0157	-0,0086	0,0062	0,0381	0,0083	0,0274
Equalized Odds: Limite Inferior	-0,0044	-0,0268	0,1803	0,0137	-0,0185	0,0012	0,0347	-0,0219	0,022
Equalized Odds: Valor Pontual	-0,0025	-0,0137	0,1907	0,0234	-0,0147	0,0045	0,0427	-0,01	0,0294
Equalized Odds: Limite Superior	-0,0009	-0,0006	0,2013	0,0332	-0,0112	0,0074	0,0506	0,0018	0,0375

Figura B.7: Resultados alcançados no Cenário 2 para as execuções de 2000 reamostragens.

Cenário 2 - 4000 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	0,6759	0,4272	0,0561	1,8566	1,8502	0,3729	1,8548	11,7494	3,8174
Tempo de Aplicação de Transformações	55,4627	55,4627	55,4627	55,4627	55,4627	55,4627	55,4627	55,4627	55,4627
Tempo de Treino - Dados Transformados	0,6639	0,4181	0,0592	1,824	1,8541	0,5103	1,8505	11,6582	3,7612
Tempo de Teste	0,0632	0,0231	0,003	0,1755	0,018	0,012	0,1614	0,1855	0,2587
Accuracy: Limite Inferior	-0,0009	0,0013	-0,0601	-0,0002	0,0006	-0,0015	-0,0023	-0,0085	-0,0025
Accuracy: Valor Pontual	-0,0004	0,0051	-0,0556	0,0024	0,0017	-0,0004	0	-0,0054	-0,0003
Accuracy: Limite Superior	0,0002	0,0087	-0,0514	0,0049	0,0029	0,0006	0,0023	-0,0024	0,0019
Balanced Accuracy: Limite Inferior	-0,0009	0,0012	-0,0573	-0,0007	0,0005	-0,0014	-0,0029	-0,0089	-0,0028
Balanced Accuracy: Valor Pontual	-0,0004	0,005	-0,0531	0,0018	0,0016	-0,0003	-0,0006	-0,0058	-0,0006
Balanced Accuracy: Limite Superior	0,0002	0,0086	-0,049	0,0043	0,0028	0,0007	0,0016	-0,0028	0,0016
Precision: Limite Inferior	-0,001	0,0017	-0,0717	0,0057	0,0011	-0,0027	0,0042	-0,0039	0,0004
Precision: Valor Pontual	-0,0005	0,0056	-0,0672	0,0084	0,0022	-0,0016	0,0063	-0,0008	0,0025
Precision: Limite Superior	0	0,0093	-0,0628	0,0112	0,0034	-0,0005	0,0085	0,0024	0,0045
Recall: Limite Inferior	-0,0009	-0,0024	0,0057	-0,0158	-0,0018	0,0007	-0,0205	-0,021	-0,011
Recall: Valor Pontual	0	0,0026	0,0101	-0,0127	-0,0006	0,0022	-0,0171	-0,0169	-0,008
Recall: Limite Superior	0,0009	0,0077	0,0145	-0,0096	0,0007	0,0037	-0,0138	-0,0128	-0,005
F1-Score: Limite Inferior	-0,0009	0,0008	-0,0403	-0,003	0,0002	-0,001	-0,0055	-0,0102	-0,0038
F1-Score: Valor Pontual	-0,0003	0,0043	-0,0366	-0,0006	0,0011	0	-0,0033	-0,0074	-0,0018
F1-Score: Limite Superior	0,0003	0,0078	-0,0329	0,0017	0,0021	0,001	-0,0011	-0,0045	0,0003
Statistical Parity: Limite Inferior	-0,0037	-0,0261	0,1367	0,0057	-0,013	0,0021	0,0289	-0,0042	0,0185
Statistical Parity: Valor Pontual	-0,0025	-0,0187	0,145	0,0108	-0,0107	0,0042	0,0336	0,0021	0,0229
Statistical Parity: Limite Superior	-0,0014	-0,0113	0,1533	0,0158	-0,0086	0,0063	0,0382	0,0083	0,0272
Equalized Odds: Limite Inferior	-0,0044	-0,0271	0,1798	0,0136	-0,0184	0,0013	0,0348	-0,0219	0,0221
Equalized Odds: Valor Pontual	-0,0025	-0,0137	0,1907	0,0234	-0,0147	0,0045	0,0427	-0,01	0,0294
Equalized Odds: Limite Superior	-0,0009	-0,0004	0,2015	0,0332	-0,0112	0,0076	0,0508	0,0023	0,0372

Figura B.8: Resultados alcançados no Cenário 2 para as execuções de 4000 reamostragens.

Cenário 1 - 500 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	1,1484	0,6707	0,0853	2,7575	2,641	0,7449	2,8559	17,1199	5,9744
Tempo de Aplicação de Transformações	35,9126	35,9126	35,9126	35,9126	35,9126	35,9126	35,9126	35,9126	35,9126
Tempo de Treino - Dados Transformados	1,1319	0,6862	0,0979	2,7336	2,6161	0,8942	2,8339	17,6889	5,5882
Tempo de Teste	0,1062	0,0462	0,0067	0,3383	0,024	0,0271	0,296	0,3247	0,4339
Accuracy: Limite Inferior	0	-0,0046	-0,0015	0,0069	-0,0113	0,0133	-0,0114	0,005	-0,0064
Accuracy: Valor Pontual	0	-0,0004	-0,0003	0,0095	-0,0098	0,0161	-0,0093	0,0081	-0,0046
Accuracy: Limite Superior	0	0,0031	0,001	0,0121	-0,0082	0,0191	-0,0067	0,0113	-0,0028
Balanced Accuracy: Limite Inferior	0	-0,0046	-0,0015	0,0081	-0,0106	0,0142	-0,0106	0,0049	-0,0062
Balanced Accuracy: Valor Pontual	0	-0,0005	-0,0003	0,0108	-0,0091	0,0169	-0,0085	0,008	-0,0043
Balanced Accuracy: Limite Superior	0	0,003	0,001	0,0133	-0,0076	0,02	-0,006	0,0112	-0,0026
Precision: Limite Inferior	0	-0,0046	-0,0018	-0,0075	-0,0186	0,006	-0,0184	0,0056	-0,0089
Precision: Valor Pontual	0	-0,0001	-0,0004	-0,0042	-0,0165	0,0088	-0,016	0,0099	-0,0068
Precision: Limite Superior	0	0,0038	0,0009	-0,0017	-0,0143	0,0121	-0,0129	0,0141	-0,0044
Recall: Limite Inferior	0	-0,0067	-0,0017	0,0291	0,0018	0,0272	0,0029	0,0022	-0,0022
Recall: Valor Pontual	0	-0,0014	-0,0001	0,0331	0,0029	0,0317	0,0058	0,0065	0
Recall: Limite Superior	0	0,0034	0,0013	0,0374	0,0043	0,0364	0,0089	0,0111	0,0023
F1-Score: Limite Inferior	0	-0,0044	-0,0013	0,0132	-0,009	0,0168	-0,0082	0,0049	-0,0055
F1-Score: Valor Pontual	0	-0,0006	-0,0003	0,0163	-0,0075	0,0198	-0,0061	0,0082	-0,0037
F1-Score: Limite Superior	0	0,0026	0,001	0,0193	-0,0061	0,0231	-0,0038	0,0115	-0,0019
Statistical Parity: Limite Inferior	0	0,017	-0,0018	0,0388	0,022	0,0051	0,025	-0,03	0,0069
Statistical Parity: Valor Pontual	0	0,0247	0,0004	0,0432	0,025	0,0102	0,0299	-0,0239	0,0109
Statistical Parity: Limite Superior	0	0,0331	0,0028	0,0478	0,0279	0,016	0,0349	-0,018	0,0148
Equalized Odds: Limite Inferior	0	-0,0273	-0,004	0,0815	0,0288	0,0627	0,0229	-0,0474	0,0063
Equalized Odds: Valor Pontual	0	-0,0039	-0,0012	0,0953	0,0328	0,0788	0,03	-0,0394	0,0126
Equalized Odds: Limite Superior	0	0,0211	0,0025	0,1066	0,0366	0,0886	0,0377	-0,0304	0,0179

Figura B.9: Resultados alcançados no Cenário 3 para as execuções de 500 reamostragens.

Cenário 1 - 1000 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	0,9714	0,572	0,0807	2,7689	2,4943	0,5841	2,9848	18,5773	5,5692
Tempo de Aplicação de Transformações	50,5193	50,5193	50,5193	50,5193	50,5193	50,5193	50,5193	50,5193	50,5193
Tempo de Treino - Dados Transformados	0,9668	0,5454	0,0902	2,5936	2,523	0,8882	2,8739	17,2331	5,6217
Tempo de Teste	0,0978	0,0376	0,003	0,313	0,0199	0,0241	0,3049	0,3252	0,432
Accuracy: Limite Inferior	0	-0,0047	-0,0015	0,007	-0,0114	0,0133	-0,0116	0,005	-0,0063
Accuracy: Valor Pontual	0	-0,0004	-0,0003	0,0095	-0,0098	0,0161	-0,0093	0,0081	-0,0046
Accuracy: Limite Superior	0	0,0036	0,0009	0,0119	-0,0082	0,019	-0,0067	0,0114	-0,0026
Balanced Accuracy: Limite Inferior	0	-0,0046	-0,0015	0,0082	-0,0107	0,0141	-0,0107	0,0049	-0,0061
Balanced Accuracy: Valor Pontual	0	-0,0005	-0,0003	0,0108	-0,0091	0,0169	-0,0085	0,008	-0,0043
Balanced Accuracy: Limite Superior	0	0,0034	0,0009	0,0132	-0,0076	0,0199	-0,006	0,0113	-0,0024
Precision: Limite Inferior	0	-0,0047	-0,0018	-0,0073	-0,0189	0,006	-0,0187	0,0059	-0,0089
Precision: Valor Pontual	0	-0,0001	-0,0004	-0,0042	-0,0165	0,0088	-0,016	0,0099	-0,0068
Precision: Limite Superior	0	0,0043	0,0009	-0,0017	-0,0141	0,012	-0,0129	0,014	-0,0043
Recall: Limite Inferior	0	-0,0065	-0,0016	0,0291	0,0017	0,0272	0,0028	0,0021	-0,0023
Recall: Valor Pontual	0	-0,0014	-0,0001	0,0331	0,0029	0,0317	0,0058	0,0065	0
Recall: Limite Superior	0	0,0036	0,0014	0,0375	0,0042	0,0363	0,0089	0,0112	0,0023
F1-Score: Limite Inferior	0	-0,0045	-0,0014	0,0134	-0,009	0,0167	-0,0083	0,0049	-0,0054
F1-Score: Valor Pontual	0	-0,0006	-0,0003	0,0163	-0,0075	0,0198	-0,0061	0,0082	-0,0037
F1-Score: Limite Superior	0	0,0029	0,0009	0,0193	-0,0061	0,023	-0,0038	0,0118	-0,0018
Statistical Parity: Limite Inferior	0	0,0165	-0,0018	0,0388	0,0219	0,005	0,0245	-0,0302	0,007
Statistical Parity: Valor Pontual	0	0,0247	0,0004	0,0432	0,025	0,0102	0,0299	-0,0239	0,0109
Statistical Parity: Limite Superior	0	0,033	0,0029	0,0479	0,0281	0,016	0,0349	-0,0181	0,0144
Equalized Odds: Limite Inferior	0	-0,0269	-0,0044	0,0822	0,0286	0,0628	0,0225	-0,0482	0,0068
Equalized Odds: Valor Pontual	0	-0,0039	-0,0012	0,0953	0,0328	0,0788	0,03	-0,0394	0,0126
Equalized Odds: Limite Superior	0	0,022	0,0025	0,1067	0,0371	0,0898	0,038	-0,0307	0,018

Figura B.10: Resultados alcançados no Cenário 3 para as execuções de 1000 reamostragens.

Cenário 1 - 2000 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	1,0011	0,7551	0,0814	2,0758	1,8314	0,3346	1,892	11,7403	3,6929
Tempo de Aplicação de Transformações	52,1367	52,1367	52,1367	52,1367	52,1367	52,1367	52,1367	52,1367	52,1367
Tempo de Treino - Dados Transformados	0,9839	0,6559	0,0859	2,046	1,7866	0,3932	1,8511	11,6757	3,6772
Tempo de Teste	0,0929	0,0448	0,004	0,2443	0,0195	0,011	0,2085	0,2285	0,3035
Accuracy: Limite Inferior	0	-0,0045	-0,0015	0,0071	-0,0114	0,0132	-0,0116	0,0049	-0,0064
Accuracy: Valor Pontual	0	-0,0004	-0,0003	0,0095	-0,0098	0,0161	-0,0093	0,0081	-0,0046
Accuracy: Limite Superior	0	0,0036	0,0009	0,012	-0,0082	0,019	-0,0067	0,0113	-0,0027
Balanced Accuracy: Limite Inferior	0	-0,0044	-0,0015	0,0083	-0,0107	0,014	-0,0108	0,0049	-0,0062
Balanced Accuracy: Valor Pontual	0	-0,0005	-0,0003	0,0108	-0,0091	0,0169	-0,0085	0,008	-0,0043
Balanced Accuracy: Limite Superior	0	0,0035	0,0009	0,0132	-0,0076	0,0199	-0,006	0,0115	-0,0025
Precision: Limite Inferior	0	-0,0045	-0,0018	-0,0071	-0,0189	0,0058	-0,0188	0,0057	-0,009
Precision: Valor Pontual	0	-0,0001	-0,0004	-0,0042	-0,0165	0,0088	-0,016	0,0099	-0,0068
Precision: Limite Superior	0	0,0044	0,0009	-0,0016	-0,0142	0,0119	-0,013	0,0141	-0,0044
Recall: Limite Inferior	0	-0,0064	-0,0016	0,029	0,0018	0,0271	0,0027	0,0019	-0,0023
Recall: Valor Pontual	0	-0,0014	-0,0001	0,0331	0,0029	0,0317	0,0058	0,0065	0
Recall: Limite Superior	0	0,0036	0,0014	0,0375	0,0042	0,0361	0,009	0,0111	0,0022
F1-Score: Limite Inferior	0	-0,0043	-0,0014	0,0133	-0,009	0,0167	-0,0084	0,0048	-0,0054
F1-Score: Valor Pontual	0	-0,0006	-0,0003	0,0163	-0,0075	0,0198	-0,0061	0,0082	-0,0037
F1-Score: Limite Superior	0	0,003	0,0009	0,0193	-0,0061	0,023	-0,0038	0,0118	-0,0019
Statistical Parity: Limite Inferior	0	0,0165	-0,0018	0,0386	0,0219	0,0046	0,0251	-0,0302	0,0072
Statistical Parity: Valor Pontual	0	0,0247	0,0004	0,0432	0,025	0,0102	0,0299	-0,0239	0,0109
Statistical Parity: Limite Superior	0	0,0327	0,003	0,0479	0,0281	0,0159	0,0349	-0,0177	0,0146
Equalized Odds: Limite Inferior	0	-0,0272	-0,0042	0,0822	0,0286	0,0625	0,0225	-0,0479	0,0069
Equalized Odds: Valor Pontual	0	-0,0039	-0,0012	0,0953	0,0328	0,0788	0,03	-0,0394	0,0126
Equalized Odds: Limite Superior	0	0,022	0,0024	0,1065	0,0372	0,0891	0,0381	-0,0307	0,018

Figura B.11: Resultados alcançados no Cenário 3 para as execuções de 2000 reamostragens.

Cenário 1 - 4000 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	0,6717	0,4161	0,0541	1,7548	1,7481	0,3557	1,9317	11,7067	3,6382
Tempo de Aplicação de Transformações	31,0335	31,0335	31,0335	31,0335	31,0335	31,0335	31,0335	31,0335	31,0335
Tempo de Treino - Dados Transformados	0,6748	0,4361	0,0541	1,7732	1,7426	0,4289	1,902	11,5892	3,6067
Tempo de Teste	0,0612	0,0281	0,003	0,2196	0,018	0,012	0,2081	0,2266	0,3007
Accuracy: Limite Inferior	0	-0,0044	-0,0015	0,0072	-0,0114	0,0132	-0,0117	0,0048	-0,0065
Accuracy: Valor Pontual	0	-0,0004	-0,0003	0,0095	-0,0098	0,0161	-0,0093	0,0081	-0,0046
Accuracy: Limite Superior	0	0,0037	0,0009	0,0121	-0,0082	0,0191	-0,0067	0,0115	-0,0027
Balanced Accuracy: Limite Inferior	0	-0,0044	-0,0015	0,0084	-0,0107	0,014	-0,0109	0,0047	-0,0062
Balanced Accuracy: Valor Pontual	0	-0,0005	-0,0003	0,0108	-0,0091	0,0169	-0,0085	0,008	-0,0043
Balanced Accuracy: Limite Superior	0	0,0035	0,0009	0,0133	-0,0075	0,0199	-0,0059	0,0114	-0,0025
Precision: Limite Inferior	0	-0,0045	-0,0017	-0,007	-0,0188	0,0057	-0,019	0,0057	-0,0091
Precision: Valor Pontual	0	-0,0001	-0,0004	-0,0042	-0,0165	0,0088	-0,016	0,0099	-0,0068
Precision: Limite Superior	0	0,0044	0,0009	-0,0015	-0,0142	0,0119	-0,013	0,0141	-0,0044
Recall: Limite Inferior	0	-0,0065	-0,0016	0,029	0,0018	0,0272	0,0027	0,002	-0,0023
Recall: Valor Pontual	0	-0,0014	-0,0001	0,0331	0,0029	0,0317	0,0058	0,0065	0
Recall: Limite Superior	0	0,0036	0,0015	0,0375	0,0042	0,0365	0,009	0,0111	0,0022
F1-Score: Limite Inferior	0	-0,0043	-0,0014	0,0134	-0,009	0,0166	-0,0085	0,0047	-0,0054
F1-Score: Valor Pontual	0	-0,0006	-0,0003	0,0163	-0,0075	0,0198	-0,0061	0,0082	-0,0037
F1-Score: Limite Superior	0	0,0031	0,0009	0,0194	-0,0061	0,0231	-0,0037	0,0118	-0,0019
Statistical Parity: Limite Inferior	0	0,0168	-0,0019	0,0386	0,0218	0,0047	0,025	-0,0305	0,0073
Statistical Parity: Valor Pontual	0	0,0247	0,0004	0,0432	0,025	0,0102	0,0299	-0,0239	0,0109
Statistical Parity: Limite Superior	0	0,0327	0,003	0,0479	0,0282	0,0158	0,0348	-0,0175	0,0146
Equalized Odds: Limite Inferior	0	-0,0272	-0,0043	0,0829	0,0285	0,0628	0,0224	-0,048	0,0073
Equalized Odds: Valor Pontual	0	-0,0039	-0,0012	0,0953	0,0328	0,0788	0,03	-0,0394	0,0126
Equalized Odds: Limite Superior	0	0,0217	0,0024	0,1067	0,0373	0,0898	0,038	-0,0308	0,0181

Figura B.12: Resultados alcançados no Cenário 3 para as execuções de 4000 reamostragens.

Cenário 4 - 500 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	0,6577	0,4157	0,0615	1,8585	1,8052	0,3469	1,8869	11,5775	3,6428
Tempo de Aplicação de Transformações	35,5065	35,5065	35,5065	35,5065	35,5065	35,5065	35,5065	35,5065	35,5065
Tempo de Treino - Dados Transformados	0,6422	0,4141	0,0592	1,7897	1,7207	0,376	1,8434	11,5768	3,6786
Tempo de Teste	0,0589	0,0261	0,0036	0,2246	0,018	0,013	0,1985	0,2135	0,2817
Accuracy: Limite Inferior	0	-0,0051	-0,0009	-0,0092	-0,0006	-0,0009	0,0116	0,022	0
Accuracy: Valor Pontual	0	-0,0019	0,0002	-0,007	-0,0002	0,0002	0,014	0,025	0,001
Accuracy: Limite Superior	0	0,0011	0,0014	-0,0049	0,0003	0,0013	0,0163	0,028	0,0022
Balanced Accuracy: Limite Inferior	0	-0,0053	-0,0009	-0,0096	-0,0006	-0,0007	0,0122	0,023	0
Balanced Accuracy: Valor Pontual	0	-0,002	0,0003	-0,0074	-0,0002	0,0004	0,0147	0,0261	0,0012
Balanced Accuracy: Limite Superior	0	0,001	0,0014	-0,0052	0,0003	0,0015	0,0171	0,0291	0,0023
Precision: Limite Inferior	0	-0,0035	-0,0026	-0,0011	-0,001	-0,0118	-0,0037	0,0062	-0,0033
Precision: Valor Pontual	0	0,0013	-0,0005	0,0016	-0,0004	-0,0084	-0,0009	0,0099	-0,0015
Precision: Limite Superior	0	0,0061	0,0015	0,0044	0	-0,0052	0,002	0,0143	0,0003
Recall: Limite Inferior	0	-0,0105	-0,0006	-0,0218	-0,0007	0,0031	0,0295	0,0471	0,002
Recall: Valor Pontual	0	-0,0057	0,001	-0,0178	0	0,0047	0,0339	0,0534	0,0038
Recall: Limite Superior	0	-0,0009	0,0029	-0,0136	0,0008	0,0065	0,0386	0,0591	0,0059
F1-Score: Limite Inferior	0	-0,0077	-0,001	-0,0167	-0,0007	0,0016	0,0214	0,0444	0,0006
F1-Score: Valor Pontual	0	-0,0036	0,0005	-0,0133	-0,0002	0,0036	0,0251	0,0505	0,0022
F1-Score: Limite Superior	0	0,0006	0,002	-0,0099	0,0005	0,0056	0,0288	0,0559	0,0039
Statistical Parity: Limite Inferior	0	-0,0012	-0,0025	-0,024	-0,0012	-0,0078	-0,0359	0,0519	-0,006
Statistical Parity: Valor Pontual	0	0,004	-0,0002	-0,0108	-0,0003	-0,0057	-0,0214	0,0575	-0,0037
Statistical Parity: Limite Superior	0	0,0101	0,0022	0,0134	0,0004	-0,0039	0,0049	0,0635	-0,0013
Equalized Odds: Limite Inferior	0	-0,0172	-0,0024	-0,0367	-0,0011	-0,0121	0,0429	0,0773	-0,0019
Equalized Odds: Valor Pontual	0	-0,0078	0,0011	-0,0294	0	-0,0084	0,0502	0,0868	0,0023
Equalized Odds: Limite Superior	0	0,001	0,0048	-0,0219	0,0012	-0,0044	0,0573	0,0971	0,0065

Figura B.13: Resultados alcançados no Cenário 4 para as execuções de 500 reamostragens.

Cenário 4 - 1000 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	0,6447	0,4111	0,0642	1,8142	1,7898	0,3328	1,8809	11,8513	5,8217
Tempo de Aplicação de Transformações	32,7735	32,7735	32,7735	32,7735	32,7735	32,7735	32,7735	32,7735	32,7735
Tempo de Treino - Dados Transformados	0,6368	0,396	0,0561	1,7586	1,7398	0,3818	1,8432	11,6417	5,9463
Tempo de Teste	0,0592	0,0251	0,003	0,2166	0,018	0,015	0,1955	0,2186	0,4231
Accuracy: Limite Inferior	0	-0,005	-0,0009	-0,0092	-0,0006	-0,0009	0,0116	0,0219	-0,0001
Accuracy: Valor Pontual	0	-0,0019	0,0002	-0,007	-0,0002	0,0002	0,014	0,025	0,001
Accuracy: Limite Superior	0	0,0011	0,0013	-0,0047	0,0002	0,0013	0,0163	0,028	0,0022
Balanced Accuracy: Limite Inferior	0	-0,0052	-0,0009	-0,0096	-0,0006	-0,0007	0,0123	0,0229	0
Balanced Accuracy: Valor Pontual	0	-0,002	0,0003	-0,0074	-0,0002	0,0004	0,0147	0,0261	0,0012
Balanced Accuracy: Limite Superior	0	0,001	0,0014	-0,005	0,0002	0,0015	0,0171	0,0292	0,0023
Precision: Limite Inferior	0	-0,0033	-0,0025	-0,0012	-0,001	-0,012	-0,0037	0,006	-0,0034
Precision: Valor Pontual	0	0,0013	-0,0005	0,0016	-0,0004	-0,0084	-0,0009	0,0099	-0,0015
Precision: Limite Superior	0	0,006	0,0014	0,0043	0	-0,0049	0,002	0,0141	0,0004
Recall: Limite Inferior	0	-0,0105	-0,0007	-0,0219	-0,0008	0,0032	0,0296	0,0471	0,0019
Recall: Valor Pontual	0	-0,0057	0,001	-0,0178	0	0,0047	0,0339	0,0534	0,0038
Recall: Limite Superior	0	-0,0007	0,0029	-0,0137	0,0008	0,0064	0,0386	0,0591	0,0058
F1-Score: Limite Inferior	0	-0,0077	-0,001	-0,0167	-0,0008	0,0019	0,0214	0,0444	0,0006
F1-Score: Valor Pontual	0	-0,0036	0,0005	-0,0133	-0,0002	0,0036	0,0251	0,0505	0,0022
F1-Score: Limite Superior	0	0,0006	0,002	-0,0098	0,0004	0,0055	0,0288	0,0561	0,0038
Statistical Parity: Limite Inferior	0	-0,0013	-0,0024	-0,0237	-0,0011	-0,0078	-0,0362	0,0517	-0,006
Statistical Parity: Valor Pontual	0	0,004	-0,0002	-0,0108	-0,0003	-0,0057	-0,0214	0,0575	-0,0037
Statistical Parity: Limite Superior	0	0,0103	0,0022	0,0138	0,0005	-0,0037	0,0042	0,0636	-0,0012
Equalized Odds: Limite Inferior	0	-0,0173	-0,0025	-0,0361	-0,0012	-0,0124	0,0429	0,0772	-0,0019
Equalized Odds: Valor Pontual	0	-0,0078	0,0011	-0,0294	0	-0,0084	0,0502	0,0868	0,0023
Equalized Odds: Limite Superior	0	0,001	0,0048	-0,0222	0,0012	-0,0046	0,0578	0,0965	0,0066

Figura B.14: Resultados alcançados no Cenário 4 para as execuções de 1000 reamostragens.

Cenário 4 - 2000 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	0,9203	0,64	0,087	3,07	2,7697	0,5721	3,3903	20,9054	3,7886
Tempo de Aplicação de Transformações	47,861	47,861	47,861	47,861	47,861	47,861	47,861	47,861	47,861
Tempo de Treino - Dados Transformados	0,9284	0,6262	0,089	3,0131	2,7174	0,6308	3,3083	21,1168	3,7034
Tempo de Teste	0,0956	0,0358	0,004	0,3604	0,0281	0,0181	0,3499	0,3933	0,2854
Accuracy: Limite Inferior	0	-0,005	-0,0009	-0,009	-0,0006	-0,0008	0,0116	0,022	-0,0001
Accuracy: Valor Pontual	0	-0,0019	0,0002	-0,007	-0,0002	0,0002	0,014	0,025	0,001
Accuracy: Limite Superior	0	0,0011	0,0014	-0,0048	0,0002	0,0013	0,0164	0,028	0,0022
Balanced Accuracy: Limite Inferior	0	-0,0051	-0,0009	-0,0095	-0,0006	-0,0006	0,0123	0,0229	0
Balanced Accuracy: Valor Pontual	0	-0,002	0,0003	-0,0074	-0,0002	0,0004	0,0147	0,0261	0,0012
Balanced Accuracy: Limite Superior	0	0,0011	0,0014	-0,0052	0,0002	0,0015	0,0172	0,0291	0,0023
Precision: Limite Inferior	0	-0,0032	-0,0025	-0,0011	-0,001	-0,0119	-0,0039	0,0058	-0,0034
Precision: Valor Pontual	0	0,0013	-0,0005	0,0016	-0,0004	-0,0084	-0,0009	0,0099	-0,0015
Precision: Limite Superior	0	0,0059	0,0015	0,0043	0	-0,005	0,002	0,0139	0,0004
Recall: Limite Inferior	0	-0,0106	-0,0007	-0,0218	-0,0008	0,0032	0,0293	0,0475	0,0019
Recall: Valor Pontual	0	-0,0057	0,001	-0,0178	0	0,0047	0,0339	0,0534	0,0038
Recall: Limite Superior	0	-0,0007	0,0029	-0,0137	0,0007	0,0063	0,0386	0,0591	0,0058
F1-Score: Limite Inferior	0	-0,0077	-0,001	-0,0167	-0,0007	0,0019	0,0213	0,0447	0,0006
F1-Score: Valor Pontual	0	-0,0036	0,0005	-0,0133	-0,0002	0,0036	0,0251	0,0505	0,0022
F1-Score: Limite Superior	0	0,0007	0,0021	-0,0099	0,0004	0,0055	0,029	0,0559	0,0038
Statistical Parity: Limite Inferior	0	-0,0017	-0,0025	-0,0238	-0,0011	-0,0078	-0,0364	0,0517	-0,006
Statistical Parity: Valor Pontual	0	0,004	-0,0002	-0,0108	-0,0003	-0,0057	-0,0214	0,0575	-0,0037
Statistical Parity: Limite Superior	0	0,01	0,0022	0,0144	0,0005	-0,0037	0,0048	0,0636	-0,0013
Equalized Odds: Limite Inferior	0	-0,0171	-0,0026	-0,0361	-0,0012	-0,0125	0,0428	0,0771	-0,0019
Equalized Odds: Valor Pontual	0	-0,0078	0,0011	-0,0294	0	-0,0084	0,0502	0,0868	0,0023
Equalized Odds: Limite Superior	0	0,0012	0,0048	-0,0226	0,0011	-0,0045	0,058	0,0965	0,0064

Figura B.15: Resultados alcançados no Cenário 4 para as execuções de 2000 reamostragens.

Cenário 4 - 4000 Reamostragens									
Métricas	Adaptive Boosting	Bagging	Decision Tree	Extra Tree	Gradient Boosting	Hist Gradient	Random Forest	Stacking	Voting
Tempo de Treino - Dados Originais	0,6818	0,4077	0,0592	1,8305	1,8115	0,3442	1,9153	12,1001	3,819
Tempo de Aplicação de Transformações	30,1514	30,1514	30,1514	30,1514	30,1514	30,1514	30,1514	30,1514	30,1514
Tempo de Treino - Dados Transformados	0,6375	0,4016	0,0588	1,7955	1,7541	0,4018	1,8478	12,0212	3,7732
Tempo de Teste	0,0684	0,0253	0,003	0,2249	0,0171	0,0097	0,2006	0,2235	0,2988
Accuracy: Limite Inferior	0	-0,0047	-0,0009	-0,009	-0,0006	-0,0008	0,0115	0,0221	-0,0001
Accuracy: Valor Pontual	0	-0,0019	0,0002	-0,007	-0,0002	0,0002	0,014	0,025	0,001
Accuracy: Limite Superior	0	0,0011	0,0014	-0,0049	0,0003	0,0013	0,0163	0,0279	0,0022
Balanced Accuracy: Limite Inferior	0	-0,0049	-0,0009	-0,0095	-0,0006	-0,0006	0,0123	0,0231	0
Balanced Accuracy: Valor Pontual	0	-0,002	0,0003	-0,0074	-0,0002	0,0004	0,0147	0,0261	0,0012
Balanced Accuracy: Limite Superior	0	0,001	0,0014	-0,0052	0,0003	0,0014	0,0171	0,029	0,0023
Precision: Limite Inferior	0	-0,0031	-0,0024	-0,001	-0,001	-0,0118	-0,0038	0,0059	-0,0034
Precision: Valor Pontual	0	0,0013	-0,0005	0,0016	-0,0004	-0,0084	-0,0009	0,0099	-0,0015
Precision: Limite Superior	0	0,0059	0,0014	0,0042	0	-0,0051	0,0021	0,0139	0,0004
Recall: Limite Inferior	0	-0,0107	-0,0008	-0,0219	-0,0008	0,0032	0,0293	0,0478	0,0018
Recall: Valor Pontual	0	-0,0057	0,001	-0,0178	0	0,0047	0,0339	0,0534	0,0038
Recall: Limite Superior	0	-0,0007	0,0029	-0,0137	0,0008	0,0063	0,0383	0,0591	0,0058
F1-Score: Limite Inferior	0	-0,0076	-0,001	-0,0167	-0,0007	0,0019	0,0213	0,045	0,0006
F1-Score: Valor Pontual	0	-0,0036	0,0005	-0,0133	-0,0002	0,0036	0,0251	0,0505	0,0022
F1-Score: Limite Superior	0	0,0006	0,0021	-0,0098	0,0004	0,0055	0,0288	0,0559	0,0038
Statistical Parity: Limite Inferior	0	-0,0018	-0,0025	-0,0236	-0,0012	-0,0078	-0,0363	0,0518	-0,006
Statistical Parity: Valor Pontual	0	0,004	-0,0002	-0,0108	-0,0003	-0,0057	-0,0214	0,0575	-0,0037
Statistical Parity: Limite Superior	0	0,0101	0,0022	0,0146	0,0005	-0,0037	0,0046	0,0634	-0,0013
Equalized Odds: Limite Inferior	0	-0,0173	-0,0024	-0,0362	-0,0012	-0,0126	0,0428	0,0771	-0,0018
Equalized Odds: Valor Pontual	0	-0,0078	0,0011	-0,0294	0	-0,0084	0,0502	0,0868	0,0023
Equalized Odds: Limite Superior	0	0,0013	0,0049	-0,0225	0,0012	-0,0044	0,0579	0,0965	0,0063

Figura B.16: Resultados alcançados no Cenário 4 para as execuções de 4000 reamostragens.