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

ÓRION DARSHAN WINTER DE LIMA

JUSTIÇA EM APRENDIZAGEM DE MÁQUINA NA ESTIMATIVA DE RISCO DE CONTRATOS PÚBLICOS

CAMPINA GRANDE - PB

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

Justiça em aprendizagem de máquina na estimativa de risco de contratos públicos

Órion Darshan Winter de Lima

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

Nazareno Ferreira de Andrade
(Orientador)

Campina Grande, Paraíba, Brasil

JUSTIÇA EM APRENDIZAGEM DE MÁQUINA NA ESTIMATIVA DE RISCO DE CONTRATOS PÚBLICOS

ÓRION DARSHAN WINTER DE LIMA

DISSERTAÇÃO APROVADA EM 21/02/2020

NAZARENO FERREIRA DE ANDRADE, Dr., UFCG
Orientador(a)

FÁBIO JORGE ALMEIDA MORAIS, Dr., UFCG
Examinador(a)

FLAVIO VINICIUS DINIZ DE FIGUEIREDO, Dr., UFMG
Examinador(a)

CAMPINA GRANDE - PB

L732j Lima, Órion Darshan Winter de.
Justiça em aprendizagem de máquina na estimativa de risco de contratos públicos/Órion Darshan Winter de Lima. - Campina Grande, 2020.
70 f. : il. Color.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2020.
"Orientação: Prof. Dr. Nazareno Ferreira de Andrade".
Referências.

1. Aprendizagem de Máquina. 2. Justiça. 3. Estimativa de Risco. 4. Gastos Públicos. I. Andrade, Nazareno Ferreira de. I. Título.

CDU 004.8(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA Itapuana Soares Dias CRB-15/93

Resumo

O governo brasileiro firma contratos com empresas para a aquisição de produtos e prestação de serviços. Porém, devido à alta demanda para a fiscalização desses contratos, os órgãos de controle necessitam realizar uma priorização dos mesmos, normalmente através de uma estimativa de risco de contratos ou empresas. Com seu sucesso em outros contexto, técnicas de aprendizagem de máquina vêm sendo empregadas na estimativa de risco por esses órgãos. Ao mesmo tempo, trabalhos recentes mostraram que sistemas de apoio a decisão semelhantes à estimativa de risco com aprendizagem de máquina podem ser injustos. Essa observação aponta para o risco de que modelos criados com aprendizagem de máquina por órgãos de controle possam ser injustos. Esta dissertação apresenta uma avaliação da justiça nos modelos de estimativa de risco semelhantes aos utilizados por órgãos de controle federais e estaduais brasileiros, utilizando bases de dados disponíveis para esses órgãos. Os modelos de estimativa de risco estudados incluem tanto métodos ad-hoc disponíveis nos órgãos quanto de aprendizagem de máquina, utilizando uma metodologia de estimativa de risco análoga à publicada em um artigo de um órgão de controle federal. Além disso, foram empregados três métodos do estado da arte que objetivam mitigar as injustiças evidenciadas pelos modelos de estimativa de risco. Nossos resultados apontam que empresas jovens são mais falsamente acusadas quando comparadas com empresas consolidadas em todos os cenários, além de outras classes sensíveis também serem em contextos específicos. Além disso, não houve um consenso de qual método de mitigação de injustiças tem melhor desempenho. Apesar disso, em todos os cenários estudados existe um modelo que melhora a justiça pelo menos uma classe sensível. Em metade dos cenários houve uma melhora da eficácia seguida da justiça, enquanto na outra metade houve um *trade-off* entre justiça e eficácia. Desta forma, espera-se que os órgãos de controle possam atentar às injustiças presentes nos modelos de estimativa de risco e proporcionar um método de mitigá-las.

Abstract

Brazilian government signs contracts with companies for products acquirement and services provision. However, due to high demand for contracts audition, control units need to prioritize these contracts, which are usually done through risk estimation of contracts or companies. With its success in other contexts, machine learning techniques have been used in risk estimation by this agencies. Meanwhile, recent works have shown that decision making systems similar to machine learning risk estimation can be unfair. This observation points out the risk that machine learning models created by control units may be unfair. This Master's thesis presents an assessment of justice in risk estimation models similar to those used by Brazilian federal and state units, using databases available for these units. The risk estimation studied models include both ad-hoc methods available in agencies and machine learning method, using a risk estimation methodology analogous to one published in an article by a federal control unit. Furthermore, three state-of-the-art methods were applied with the objective of mitigate injustices by the risk estimation models. Our results show that young companies are more falsely accused compared with the consolidated ones in all scenarios, besides others sensible classes in specific contexts. Furthermore, there wasn't a consensus in which mitigation method had the best performance. Despite that, in all studied scenarios there was a model that improves justice at least in one sensible class. Half of the scenarios had an improvement of efficiency and justice, but the other half had a trade-off between justice and efficiency. In this way, it is expected that control units can pay attention to injustice present in risk estimation model and provide a method to mitigate them.

Agradecimentos

Agradeço primeiramente à minha namorada Nayra e à minha família, por todo apoio que me deram, por todos momentos de descontração e felicidade que me proporcionaram. A presença de vocês foi imprescindível para minha persistência neste mestrado.

Meus agradecimentos também ao meu orientador Nazareno, por todos os ensinamentos passados, por toda presença e assistência, por ser meu mentor nas reuniões semanais e na correria das escritas de artigo e dissertação.

Agradeço também à todos do Laboratório Analytics e do Laboratório de Sistemas Distribuídos por terem contribuído para meu desenvolvimento pessoal e humano. Agradeço por todas as discussões conceituais da nossa área, políticas e filosóficas, como também por momentos de descontração e piadas dignas de stand up.

Agradeço à todos professores da graduação e pós-graduação que contribuíram para minha formação profissional, assim como aos que tiraram dúvidas durante o mestrado, sendo esses membros internos da UFCG ou externos.

Agradeço também aos meus amigos, tanto da época de escola, quanto da universidade, pelos momentos de alegria e confraternização.

Agradeço por fim aos funcionários da COPIN e à CAPES por todos procedimentos e recursos necessários para a realização de uma pós-graduação.

Conteúdo

1 Introdução **1**

1.1 Motivação 2

1.1.1 Histórico 4

1.2 Objetivos e Contribuições 4

1.3 Estrutura do Documento 5

2 Fundamentação teórica **6**

2.1 Licitações 6

2.2 Estimativa de Risco 7

2.2.1 Estimativa de risco de contratos públicos e empresas 7

2.2.2 Aprendizagem de Máquina para Estimativa de Riscos 8

2.3 Justiça em Aprendizagem de Máquina 9

2.4 Técnicas de Mitigação de Injustiça 10

2.4.1 Disparate Impact Remover 10

2.4.2 Calibrated Equalized Odds 11

2.4.3 Adversarial Debiasing 11

3 Trabalhos Relacionados **13**

3.1 Investigação de gastos públicos 13

3.1.1 Estimativa de risco 13

3.1.2 Investigações de fraudes 14

3.2 Justiça em Aprendizagem de Máquina 15

3.2.1 Análise de Justiça 15

3.2.2 Mitigação de Injustiças 15

4 Materiais e Métodos **17**

4.1 Bases de Dados 18

4.1.1 Classes sensíveis 20

4.2 Tratamento dos Dados 21

4.3 Treinamento dos Modelos 23

4.4 Infraestrutura 24

5 Análise de Justiça **32**

5.1 Metodologia 32

5.2 Justiça no Estado da Prática 34

5.2.1 Empresas na esfera municipal com abordagem ad-hoc dos especialistas 35

5.2.2 Empresas na esfera municipal com aprendizagem de máquina . . . 36

5.2.3 Empresas na esfera federal com aprendizagem de máquina 37

5.2.4 Contratos na esfera municipal com aprendizagem de máquina . . . 38

5.3 Discussão 38

6 Mitigação de Injustiça **41**

6.1 Metodologia 41

6.1.1 Melhora na justiça 42

6.1.2 Parâmetros e critério 42

6.2 Resultados da Mitigação de Injustiça 45

6.2.1 Empresas na esfera municipal com abordagem ad-hoc dos especialistas 45

6.2.2 Empresas na esfera municipal com aprendizagem de máquina . . . 47

6.2.3 Empresas na esfera federal com aprendizagem de máquina 52

6.2.4 Contratos na esfera municipal com aprendizagem de máquina . . . 55

6.3 Discussão 58

7 Conclusões **63**

7.1 Discussão 63

7.2 Limitações 65

7.3 Trabalhos Futuros 66

Lista de Símbolos

AIF360 - *Artificial Intelligence Fairness 360 Open Source Toolkit*

AM - *Aprendizagem de Máquina*

BF - *Programa Bolsa Família*

CGU - *Controladoria Geral da União*

COMPAS - *Correctional Offender Management Profiling for Alternative Sanctions*

KNN - *K-Nearest Neighbors*

MPPB - *Ministério Público da Paraíba*

SMOTE - *Synthetic Minority Over-sampling Technique*

SVM - *Support Vector Machine*

TFP - *Taxa de Falso Positivo*

TRAMITA - *Sistema de Tramitação de Processos e Documentos do TCE-PB*

Lista de Figuras

4.1 Fluxo de atividades realizadas. 17

5.1 Disparidade das métricas das empresas da esfera municipal (abordagem ad-hoc). 35
5.2 Disparidade das métricas das empresas da esfera municipal (abordagem AM). 36
5.3 Disparidade das métricas das empresas da esfera federal. 37
5.4 Disparidade das métricas dos contratos municipais na Paraíba. 38

6.1 Disparidade das métricas após mitigação de injustiças das empresas da esfera municipal (abordagem ad-hoc) através do calibrated equalized odds. 45
6.2 Eficácia após mitigação de injustiças das empresas da esfera municipal (abordagem ad-hoc) através do calibrated equalized odds. 46
6.3 Disparidade das métricas após mitigação de injustiças das empresas da esfera municipal (abordagem AM) através do adversarial debiasing. 47
6.4 Eficácia após mitigação de injustiças das empresas da esfera municipal (abordagem AM) através do adversarial debiasing. 48
6.5 Disparidade das métricas após mitigação de injustiças das empresas da esfera municipal (abordagem AM) através do calibrated equalized odds. 49
6.6 Eficácia após mitigação de injustiças das empresas da esfera municipal (abordagem AM) através do calibrated equalized odds. 50
6.7 Disparidade das métricas após mitigação de injustiças das empresas da esfera municipal (abordagem AM) através do disparate impact remover. 50
6.8 Eficácia após mitigação de injustiças das empresas da esfera municipal (abordagem AM) através do disparate impact remover. 51

6.9 Disparidade das métricas após mitigação de injustiças das empresas da esfera federal através do adversarial debiasing. 51
6.10 Eficácia após mitigação de injustiças das empresas da esfera federal através do adversarial debiasing. 52
6.11 Disparidade das métricas após mitigação de injustiças das empresas da esfera federal através do calibrated equalized odds. 53
6.12 Eficácia após mitigação de injustiças das empresas da esfera federal através do calibrated equalized odds. 54
6.13 Disparidade das métricas após mitigação de injustiças das empresas da esfera federal através do disparate impact remover. 54
6.14 Eficácia após mitigação de injustiças das empresas da esfera federal através do disparate impact remover. 55
6.15 Disparidade das métricas após mitigação de injustiças dos contratos da esfera municipal através do adversarial debiasing. 56
6.16 Eficácia após mitigação de injustiças dos contratos da esfera municipal através do adversarial debiasing. 56
6.17 Disparidade das métricas após mitigação de injustiças dos contratos da esfera municipal através do calibrated equalized odds. 57
6.18 Eficácia após mitigação de injustiças dos contratos da esfera municipal através do calibrated equalized odds. 58
6.19 Disparidade das métricas após mitigação de injustiças dos contratos da esfera municipal através do disparate impact remover. 59
6.20 Eficácia após mitigação de injustiças dos contratos da esfera municipal através do disparate impact remover. 59

Lista de Tabelas

4.1 Características sensíveis dos conjuntos de dados. 21

4.2 Melhores hiperparâmetros para os contratos municipais na Paraíba 25

4.3 Melhores hiperparâmetros para as empresas da esfera municipal 26

4.4 Melhores hiperparâmetros para as empresas da esfera federal 27

4.5 Eficácia dos modelos de estimativa de risco das empresas da esfera federal . 28

4.6 Eficácia dos modelos de estimativa de risco das empresas da esfera municipal 29

4.7 Eficácia dos modelos de estimativa de risco dos contratos da esfera municipal 30

6.1 Melhores algoritmos de mitigação para cada cenário 60

Capítulo 1

Introdução

O governo brasileiro utiliza licitações para realizar obras, serviços, inclusive de publicidade, compras, alienações e locações, firmados através de um contrato regido pela Lei 8.666/93 [1]. No nível federal, a CGU é responsável por auditar os gastos feitos por todos os ministérios e 2.947 unidades descentralizadas, assinando aproximadamente 25 mil novos contratos cada ano, além de 61 mil contratos em progresso, com um valor total de cerca de 232 bilhões de dólares [22]. Esta fiscalização é de suma importância para verificar se os contratos estão sendo cumpridos de forma satisfatória, bem como descobrir eventuais esquemas de corrupção, os quais desviam a verba destinada aos gastos públicos. Este desvio afeta diretamente a sociedade, como atestam Sun e Sales [22]:

> Escândalos de corrupção de larga escala que ocorreram recentemente no Brasil, como a Operação Lava Jato, a qual envolveu 16 companhias (incluindo a companhia estatal de petróleo Petrobras) e aproximadamente 9,5 bilhões de dólares americanos, mostram o dano potencial a sociedade. A corrupção não somente resulta em perdas financeiras, mas também representa serviço incompleto ou de qualidade inferior.

Um dos passos das investigações dos gastos públicos do governo brasileiro é a estimativa de risco de empresas ou contratos, a qual identifica o risco de uma empresa rescindir contratos. Baseada nessa estimativa de risco é feita uma priorização das empresas ou contratos mais suspeitos de terem cometido irregularidades na execução do contrato, seja essa

irregularidade em decorrência do descumprimento contratual ou através de um esquema de corrupção. Essa estimativa de risco e priorização são necessárias por conta da alta demanda de contratos para serem auditados em comparação à quantidade de servidores disponíveis. Com a priorização é possível investigar primeiramente contratos mais suspeitos, os quais têm maior chance de serem problemáticos, tendo em vista que nem sempre é possível investigar todos.

Frequentemente a estimativa de risco é realizada manualmente ou por meio de regras especificadas por especialistas. Contudo, mais recentemente esse processo vem sendo automatizado através de modelos de aprendizagem de máquina. A existência de um histórico de investigações de contratos e empresas proporciona um cenário favorável para a utilização de aprendizagem de máquina supervisionada. Desta forma, a rotulagem indicando se um contrato ou uma empresa é arriscada pode ser utilizada de modo que um algoritmo aprenda os padrões da base de dados. Essa estimativa de risco é caracterizada pela predição do risco do governo celebrar um certo contrato com uma empresa, ou seja, o risco da execução do contrato por uma determinada empresa.

A seguir é descrita a motivação deste trabalho, bem como os seus objetivos e contribuições. Por último, será detalhada a estrutura deste documento.

1.1 Motivação

Sistemas de apoio a tomada de decisão baseados em aprendizagem de máquina são hoje amplamente utilizados em vários cenários que afetam a vida de pessoas [21], tal como a análise de risco de crédito [24] e estimativa de risco criminal [13]. Sistemas como esses podem ajudar humanos a decidir se um aumento do limite do cartão de crédito ou um empréstimo solicitado deve ser aprovado ou não, ou até mesmo se o juiz em uma audiência deve conceder a liberdade condicional a um preso. Para o primeiro caso, uma pessoa que está passando por dificuldades financeiras ou precisando de dinheiro para montar o próprio negócio pode ter um empréstimo negado apenas por morar em uma determinada região da cidade ou tem um salário baixo, mesmo sendo uma boa pagadora. Para o segundo caso ainda é mais grave, pois uma pessoa pode ter sua liberdade condicional negada apenas pela cor de sua pele! Sistemas injustos têm disparidades nas taxas de classificação quando comparados diferentes grupos,

ou seja, eles tendem a errar mais em grupos com características sensíveis. Com a prevalência desse tipo de sistema em áreas com potencial de alto impacto na vida das pessoas, e de potenciais injustiças também, trabalhos recentes têm levantado preocupações sobre possíveis vieses não intencionais de sistemas de aprendizagem de máquina [18].

Assim como na análise de risco de crédito e predição de risco criminal em geral, injustiça na estimativa de risco para contratos públicos pode também causar grande prejuízo à vida de pessoas. Por exemplo, se sistemas baseados em aprendizagem de máquina fizerem com que um grupo vulnerável seja mais investigado — além do que seria necessário por seu risco real —, as investigações podem gerar prejuízo para o grupo. Este prejuízo pode se dar, por exemplo, através de um ônus para provar a inocência frente a uma denúncia. O viés de confirmação do investigador, onde as pessoas tendem a procurar, perceber, interpretar e criar novas evidências de modo a verificar suas crenças preexistentes pode levar investigadores a levantar suspeitas indevidas [19], afetando os indivíduos do grupo sensível. Caso uma empresa tenha um contrato rescindido, por exemplo, são afetados não só os sócios da empresa, mas todos os empregados da empresa, assim como suas famílias. Empresas de pequeno porte, bem como características relacionadas, são o foco deste trabalho. Como essas empresas ainda estão se estabelecendo no mercado entendemos que são mais vulneráveis à sanções decorrentes de investigações, pois um prejuízo financeiro inesperado para provar a inocência frente a uma acusação pode acarretar, em um caso mais extremo, no fechamento da empresa.

Como visto no início deste capítulo, a estimativa de risco para investigações de contratos com a administração pública é imprescindível, tendo em vista a limitação de recurso por parte dos órgãos de controle. Atualmente existem soluções computacionais que realizam a priorização de contratos ou empresas automaticamente com boa precisão, como a proposta por Sun e Sales [22], indicando que esta abordagem funciona em certos contextos. Todavia, apesar deste cenário ter potencial impacto em muitas empresas, pessoas e na gestão pública, não foram encontrados estudos com objetivo de avaliar e mitigar injustiças no uso de aprendizagem de máquina para a estimativa de risco em contratos públicos e empresas brasileiras.

1.1.1 Histórico

A ideia deste trabalho de dissertação surgiu através de conversas com servidores dos órgãos de controle, como a Controladoria Geral da União, Ministério Público da Paraíba e Tribunal de Contas do Estado da Paraíba. Ao entender o processo utilizado para estimativa de risco utilizado na investigação de contratos públicos e empresas, percebeu-se a similaridade com outros cenários de estimativa de risco. Como esse também é um cenário de forte impacto na vida das pessoas, como citado na Seção 1.1, despertou-se o interesse de estudo nesta área de justiça em aprendizagem de máquina. Através de parcerias firmadas com esses órgãos, as bases de dados utilizadas nessas investigações foram cedidas para fins de pesquisa. Com isso, o interesse de estudo da justiça na estimativa de risco de contratos públicos e empresas pôde ser materializado neste trabalho.

1.2 Objetivos e Contribuições

O objetivo geral deste trabalho é avaliar se os modelos atualmente utilizados para estimativas de risco de contratos públicos e empresas, utilizados em auditorias de órgãos de controles, são justos. Além disso, também objetivamos quantificar a eficácia de técnicas do estado da arte para mitigar injustiças propiciadas pelos modelos de aprendizagem de máquina. Para realizar esse objetivo foram feitas as seguintes atividades:

- Avaliação experimental da eficácia de cinco modelos de aprendizagem de máquina utilizados para estimativa de risco de contratos públicos e empresas, combinados com 3 técnicas da literatura utilizadas para mitigar o desbalanceamento das classes da variável alvo.

- Avaliação experimental da justiça dos modelos de estimativa de risco de contratos públicos e empresas mais eficazes. Comparação entre as diferentes classes sensíveis frente a sua eficácia e justiça.

- Avaliação experimental de três técnicas do estado da arte para mitigação de injustiças aplicados sobre a estimativa de risco de contratos públicos e empresas. Comparação entre as diferentes classes sensíveis frente a sua eficácia e justiça.

A partir dessas atividades é possível avaliar e mitigar injustiças presentes na estimativa de risco utilizada pelos órgãos de controle para estimar risco de contratos públicos e empresas. Este trabalho contribui ao proporcionar, de forma pioneira, uma visão sobre a ética no processo de estimativa de risco para auditoria de contratos públicos e empresas. Dessa maneira, contribuímos para que seja possível tornar a priorização de contratos a serem auditados um processo mais justo, fazendo com que empresas pertencentes à grupos vulneráveis não sejam investigadas mais do que seria necessário pelo seu risco real.

No decorrer do mestrado do autor, o artigo intitulado *Fairness in Risk Estimation of Brazilian Public Contracts* [26] foi publicado nos Anais do VII Symposium on Knowledge Discovery, Mining and Learning (KDMiLe). Desta forma, trechos deste artigo fazem parte deste documento, tendo em vista que ele é parte deste trabalho de dissertação.

1.3 Estrutura do Documento

Os capítulos que seguem estão estruturados da seguinte forma:

No Capítulo 2 serão explicados os aspectos necessários para o entendimento deste trabalho, desde o contexto de investigações de contratos feitos com a administração pública, até a área de pesquisa de justiça em aprendizagem de máquina, como também a de estimativa de risco. No Capítulo 3 serão apresentados trabalhos relacionados tanto na linha de estimativa de risco, no contexto investigativo de contratos na administração pública e em outros contextos, como na justiça em aprendizagem de máquina. No Capítulo 4 serão explicados as bases de dados utilizadas e o tratamento das mesmas, assim como os métodos utilizados para estimativa de risco, bem como análise e mitigação da injustiça na estimativa de risco. Os resultados obtidos através dos experimentos realizados serão expostos nos Capítulos 5.2 e 6.2. Por último, no Capítulo 7 serão apontadas as conclusões deste trabalho de dissertação, bem como as limitações encontradas e oportunidades de trabalhos futuros na área.

Capítulo 2

Fundamentação teórica

Neste capítulo são detalhados os principais conceitos utilizados neste trabalho de dissertação. Na Seção 2.1 são abordados conceitos de licitações. Na Seção 2.2 são abordados conceitos de estimativa de risco, realizados tanto de maneira ad-hoc quanto via aprendizagem de máquina. Na Seção 2.3 são detalhados conceitos sobre justiça em aprendizagem de máquina. Por último, na Seção 2.4 são descritos as técnicas de mitigação de injustiça utilizados no decorrer desse trabalho.

2.1 Licitações

O Artigo 37, inciso XXI, da Constituição da República Federativa do Brasil de 1988, regulamentado através da Lei 8.666/93, determina que os contratos administrativos sejam precedidos de licitação pública, bem como o Art. 175 da Carta Magna, ao tratar das outorgas de Concessões e Permissões, também faz referência à obrigatoriedade de licitar, imposta ao ente estatal [4]. Licitação é um procedimento administrativo, no qual a administração pública convoca interessados na apresentação de propostas, observada a igualdade entre os participantes, a fim de selecionar a que se revele mais adequada, uma vez preenchidos os requisitos mínimos necessários ao bom cumprimento das obrigações a que eles se propõem, os quais devem ser divulgados previamente [7; 8]. Além disso, de acordo com a Lei 8.666/93 [1], são delimitados tanto os objetos quanto os entes estatais envolvidos na licitação:

> Esta Lei estabelece normas gerais sobre licitações e contratos ad-

> ministrativos pertinentes a obras, serviços, inclusive de publicidade, compras, alienações e locações no âmbito dos Poderes da União, dos Estados, do Distrito Federal e dos Municípios.

A exigência de um procedimento licitatório busca contornar o risco de existirem escolhas impróprias e escusas, tendo em vista que várias pessoas podem concorrer em igualdade de condições e a Administração Pública pode escolher a proposta mais vantajosa, além de atuar na busca do Desenvolvimento Nacional [4].

Uma vez que um contrato é celebrado, após da realização do processo licitatório, existe uma obrigação contratual estabelecida pelo estado para com uma pessoa física ou jurídica. O Artigo 58, inciso III, da Lei no 8.666/93, confere à Administração a prerrogativa (poder-dever) de fiscalizar a execução dos contratos administrativos [7]. Essa fiscalização é realizada pelos órgãos de controle, tais como Ministérios Públicos, Tribunais de Contas e Controladoria Geral da União. Devido a grande quantidade de contratos administrativos celebrados, os órgãos de controle necessitam priorizar os contratos mais suscetíveis à quebras contratuais. Essa priorização muitas vezes é realizada através de estimativas de risco, tema o qual é tratado na Seção 2.2.

2.2 Estimativa de Risco

A estimativa de risco, por vezes chamada de avaliação ou análise de risco, mede o risco de um certo evento acontecer em um determinado cenário. Normalmente são utilizados scores para mensurar o risco, porém esse valor pode ser dividido em categorias ou dicotomizado apenas em alto ou baixo risco. A estimativa de risco é comumente utilizada para a avaliação de crédito [10; 24]. Quando uma pessoa ou empresa solicita crédito a um banco, é necessário primeiramente verificar o risco do indivíduo ser inadimplente. Desta forma, essa estimativa de risco pode ser utilizada para apoio à tomada de decisão do crédito.

2.2.1 Estimativa de risco de contratos públicos e empresas

No contexto de gastos públicos, mais especificamente gastos realizados através de licitações, existe uma grande quantidade de contratos que devem ser investigados frente a quantidade

de servidores envolvidos na auditoria. Dessa forma é necessário realizar uma estimativa de risco de uma certa empresa não cumprir suas obrigações contratuais para que os contratos mais arriscados sejam investigados primeiro.

A estimativa de risco pode ser feita de maneira ad-hoc, através de uma série de critérios especificados por especialistas. De forma adicional, uma ponderação pode ser aplicada sobre esses critérios de acordo com a relevância de determinado critério. Esses critérios, ponderados ou não, são somados de maneira a formar um score de risco. Mais recentemente, os órgãos de controle têm utilizado aprendizagem de máquina para realizar a estimativa de risco.

2.2.2 Aprendizagem de Máquina para Estimativa de Riscos

Com o avanço da tecnologia, o uso da aprendizagem de máquina para estimativa de risco vem se tornando cada vez mais comum. Seja através de análise de risco de crédito [10; 24], predição de risco de doenças [23], avaliação de risco de enchentes [14], estimativa de risco de contratos públicos e empresas [22] ou de avaliações de risco em sentenças criminais [13], a aprendizagem de máquina trouxe resultados promissores para a estimativa de risco.

A estimativa de risco realizada através de aprendizagem de máquina pode ser modelada em diversos contextos, como visto anteriormente. Porém, independente de contexto objetiva-se utilizar aprendizagem de máquina para estimar quais observações têm mais alto risco de forma eficiente. A eficiência é verificada tanto pela acuracidade dos modelos, como também pela velocidade com que a estimativa de risco é realizada, uma vez que o treinamento foi realizado.

Diferentemente de uma abordagem ad-hoc, onde uma série de critérios é pré-definida para que a estimativa de risco seja realizada, a estimativa de risco via aprendizagem de máquina consegue aprender padrões da base de dados de treinamento. A abordagem mais comum é utilizar a aprendizagem de máquina supervisionada para estimar o risco, utilizando um conjunto de características para prever a variável alvo, o risco. Assim como a estimativa de risco ad-hoc, a estimativa de risco via aprendizagem de máquina pode ser codificada tanto como um score, através de classes, como também dicotomizada em apenas alto ou baixo risco.

Essa abordagem a uma primeira vista parece ser imparcial, porém a aprendizagem de

máquina para estimativa de risco pode aprender vieses não intencionais, como detalhado a seguir.

2.3 Justiça em Aprendizagem de Máquina

O conceito de justiça é bastante vasto, porém definições tipicamente tangem princípios morais e filosóficos. No contexto de aprendizagem de máquina, várias definições foram propostas recentemente na literatura sobre o que significa justiça e discriminação e como pode ser definido matematicamente [3]. A maioria desses conceitos utilizam uma comparação entre as taxas de classificação entre grupos [3], comparando as classes sensíveis com o grupo de referência. Desta forma, tomamos esse conceito de justiça para este trabalho utilizando como métrica de taxa de classificação a taxa de falsos positivos (TFP).

As **classes sensíveis**, também chamado de grupo protegido, são classes do conjunto de dados que podem ser impactadas negativamente caso uma decisão equivocada for tomada contra essa classe em particular. O **grupo de referência**, por sua vez, são as demais classes da mesma característica dessa classe sensível que não são tão prejudicados caso uma decisão errada seja tomada. Uma **característica sensível**, também chamado de característica protegida, é uma característica categórica dos dados composta por classes sensíveis e o grupo de referência. Assim como Saleiro et al. [18], utilizamos neste trabalho o grupo de referência composto apenas por uma classe, enquanto as demais são consideradas classes sensíveis. Em um contexto de estimativa de risco de reincidência de detentos, uma característica sensível poderia ser a raça do detento, uma classe sensível poderia ser a raça negra e o grupo de referência poderia ser a raça branca.

No contexto de investigações de gastos públicos, assim como em outros contextos, as pessoas que rotulam ou que geram os dados rotulados através de suas ações normalmente carregam um viés e transferem este viés para os dados. Como consequência, a máquina aprende esse viés para gerar seus modelos, tornando o processo menos justo e mais tendencioso. Esse viés em questão é chamado de viés de confirmação. Nele pessoas tendem a procurar, perceber, interpretar e criar novas evidências de forma a verificar suas crenças preexistentes [19].

A avaliação da justiça na aprendizagem de máquina pode ser realizada de diversas ma-

neiras. Speicher et al. [21], por exemplo, propõe uma métrica de avaliação de justiça que engloba tanto justiça entre uma classe sensível e um grupo de referência, como também dentro das observações da classe sensível. Neste trabalho focamos apenas na avaliação entre a classe sensível e o grupo de referência, utilizando o *toolkit* Aequitas, proposto por Saleiro et al. [18]. A principal métrica que utilizamos aqui para medir a justiça é a disparidade entre a classe sensível e o grupo de referência, mais especificamente a disparidade da taxa de falsos positivos, explicada em mais detalhes na Seção 5.1.

2.4 Técnicas de Mitigação de Injustiça

Com a crescente discussão sobre as implicações éticas da aprendizagem de máquina em contextos sensíveis, pesquisadores da área passaram a estudar técnicas de mitigação de injustiça (eg. [9; 12; 16; 25]). Uma vez identificado que existem vieses contra uma ou mais classes sensíveis do conjunto de dados, o objetivo dessa classe de algoritmos é mitigar as injustiças ao mesmo tempo que preserva as métricas de eficácia.

2.4.1 Disparate Impact Remover

Um dos algoritmos de mitigação utilizados neste trabalho é o *disparate impact remover* [9]. Ele é classificado como um algoritmo de pré-processamento, pois modifica os dados originais objetivando aumentar a justiça nos grupos ao passo que preserva o ranking de estimativa de risco dentro dos grupos. Esse algoritmo objetiva mitigar as injustiças baseada em uma métrica particular de justiça, o *disparate impact*. Essa métrica é a razão entre a probabilidade da variável alvo ser positiva dado que é da classe sensível e a probabilidade da variável sensível ser positiva dado que é do grupo de referência. Por exemplo, seja C a variável alvo de contratar uma pessoa, sendo positivo se a pessoa for contratada, e X se a pessoa pertence a uma classe sensível, então a razão das probabilidades é a seguinte: $\frac{P(C=SIM|X=NÃO)}{P(C=SIM|X=SIM)}$. Caso esta razão esteja abaixo de um limiar, por exemplo 80%, então é dito que existe um impacto desigual (disparate impact) entre a classe sensível e o grupo de referência. O método de remover o impacto desigual envolve manipular os valores tanto das observações da classe sensível, quanto do grupo de referência, de maneira que fiquem mais próximos da distribuição mediana. Apesar de alterar um pouco a distribuição, esse método preserva o ranking

das observações, de modo que os dados, além de se tornarem mais justos, ainda consigam predizer a variável resposta.

2.4.2 Calibrated Equalized Odds

Outro algoritmo de mitigação de utilizado neste trabalho é o *calibrated equalized odds* [16]. Esse é um algoritmo de pós-processamento, pois otimiza a calibração do estimador de risco mudando a variável de saída com objetivo de melhorar as chances equalizadas (Equalized Odds) [2]. A noção da não discriminação através da chance equalizada é baseada na taxa de falso positivo e falso negativo para cada grupo, desta forma garante que nenhum tipo de erro afete desproporcionalmente um grupo [16]. Além da chance equalizada, esta abordagem também satisfaz restrições de calibrações. A calibração diz que as probabilidades devem carregar significado semântico, por exemplo, dado um classificador $h1$ e um conjunto de características x, se existem 100 pessoas em um grupo $G1$ e a probabilidade de classificar na classe positiva seja $h1(x) = 0,6$, então é esperado que 60 deles pertençam à classe positiva [16].

2.4.3 Adversarial Debiasing

Por último, o *adversarial debiasing* [25] utiliza redes neurais adversariais para maximizar a acurácia e simultaneamente reduzir a habilidade de detectar o atributo sensível baseado na predição. Esta redução da detecção do atributo sensível pode se dar através dos conceitos de *Demographic Parity*, *Equality of Odds* e *Equality of Opportunity*. O conceito de Demographic Parity garante que um preditor $\hat{Y}$ seja independente da variável sensível Z. Ou seja, a $P(\hat{Y} = \hat{y})$ é igual para todos os valores das variáveis sensíveis Z: $P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y}|Z = z)$. O conceito de Equality of Odds garante que um preditor $\hat{Y}$ e uma variável sensível Z sejam condicionalmente independentes dado o rótulo real Y. Ou seja, para todos os possíveis valores do rótulo real Y, $P(\hat{Y} = \hat{y})$ é a mesma para todos os valores da variável sensível Z: $P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|Z = z, Y = y)$. Já o conceito de Equality of Opportunity, se a variável alvo Y for discreta, garante que um preditor $\hat{Y}$ a variável sensível Z sejam condicionalmente independentes de $Y = y$, dada uma classe y. Ou seja, para um valor particular do rótulo real Y, $P(\hat{Y} = \hat{y})$ é a mesma para todos

os valores da variável sensível Z: $P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|Z = z, Y = y)$. Esse algoritmo leva a uma classificação mais justa, tendo em vista que ele não carrega nenhuma informação de discriminação de grupo que o algoritmo possa explorar [2]. Esse algoritmo é considerado da classe *in-processing* em virtude de mudar o método de aprendizagem.

Capítulo 3

Trabalhos Relacionados

Neste capítulo serão descritos os principais trabalhos relacionados com este trabalho de dissertação. Na Seção 3.1 serão detalhados os trabalhos de investigação de gastos públicos, tanto no contexto de estimativa de risco de contratos, quanto em identificação de fraudes. Na Seção 3.2 serão descritos os trabalhos relacionados no contexto de justiça em aprendizagem de máquina. Apesar de existirem trabalhos relacionados tanto no contexto de investigação de gastos públicos como no de justiça em aprendizagem de máquina, não foram encontrados estudos que abordassem ambos contextos.

3.1 Investigação de gastos públicos

A investigação dos gastos públicos é um processo complexo e custoso. Com a introdução de técnicas de aprendizagem de máquina, órgãos de controle podem otimizar o esforço investido nesse processo através da priorização da investigação dos gastos baseado nas predições dos modelos [20]. Trabalhos relacionados à investigação dos gastos públicos estão descritos a seguir.

3.1.1 Estimativa de risco

Um dos principais trabalhos relacionados com este trabalho de dissertação é o realizado por Sun e Sales [22]. Esse trabalho objetiva identificar empresas suspeitas de cometer alguma irregularidade baseado no histórico de outras empresas que cometeram irregularidades ou

não. A base de dados utilizada nesse trabalho é uma das que utilizamos em nossa pesquisa, detalhada na Seção 4.1, a qual chamamos de *empresas da esfera federal*. Essa base nos foi gentilmente cedida pelos autores. Além da mesma base, replicamos também a metodologia para o treinamento dos modelos de Sun e Sales, tendo em vista que temos objetivo de analisar a justiça no cenário atual da estimativa de risco utilizada na investigação de contratos públicos e empresas em órgãos de controle.

Outro estudo relacionado no contexto de estimativa de risco é o desenvolvido por Domingos et al [20]. Eles abordam o uso de aprendizagem de máquina, mais especificamente *deep auto-encoders*, para identificar anomalias nas compras de tecnologia da informação para priorização das investigações relacionadas às compras. Gomes et al. [11] também utilizam *deep auto-encoders* para priorização das investigações, mas com o objetivo de identificar anomalias nos gastos dos deputados, provindos da Cota para Exercício da Atividade Parlamentar (CEAP).

Apesar de todos os artigos citados terem o objetivo de estimar o risco de um contrato ou empresa para priorização da investigação, eles diferem em suas abordagens. Sun e Sales [22] utilizam aprendizagem supervisionada para realizar a estimativa de risco, onde a base de dados utilizada continha um rótulo indicando se a empresa cometeu uma irregularidade. Já Domingos et al. [20] e Gomes et al. [11] utilizaram uma abordagem não supervisionada para detectar anomalias e investigaram mais a fundo as mais anômalas.

3.1.2 Investigações de fraudes

Ralha e Silva [17] abordam a detecção de um tipo específico de fraude: os cartéis de licitação. Para realizar essa detecção, eles abordam o uso de sistemas multi-agente, clusterização e regras de associação para identificar a formação de cartéis em processos de licitação e contratação pública.

Carvalho et al. [5] por sua vez aborda outro tipo específico de fraude: o fracionamento de despesas, um método utilizado para fraudar a modalidade licitatória onde o administrador público fraciona as despesas de modo a realizar dispensas licitatórias, eliminando assim a concorrência. Para atacar o problema, eles abordam o uso de redes bayesianas para identificar e prevenir o fracionamento de despesas.

Apesar desses trabalhos também fazerem parte do contexto de investigação de gastos

públicos, eles focam na identificação de casos específicos de fraude. Como se trata de investigações de gastos públicos através de aprendizagem de máquina, esses trabalhos também compartilham grande parte de sua metodologia com este trabalho de dissertação, no tocante ao treinamento de modelos.

3.2 Justiça em Aprendizagem de Máquina

Com o avanço das técnicas de aprendizagem de máquina, passou-se a questionar as implicações éticas dos modelos. Em cenários em que a decisão dos modelos possa impactar negativamente a vida de pessoas, a atenção para questões éticas deve ser redobrada. Trabalhos relacionados tanto na análise de justiça quanto na mitigação de injustiças são descritos a seguir.

3.2.1 Análise de Justiça

No contexto de justiça em aprendizagem de máquina Angwin et al. [13] causaram amplo impacto ao mostrar que um sistema de estimativa de risco de detentos, o COMPAS, utilizado como sistema de apoio a decisão nas audiências que determinavam se um detento teria sua liberdade condicional aprovada, era enviesada contra negros. A partir da publicação desse trabalho, as implicações éticas que a aprendizagem de máquina traz passaram a ser mais debatidas.

Posteriormente, Saleiro et al. [18] desenvolveram o *toolkit* Aequitas. Como até então não havia um consenso de quais métricas deveriam ser utilizadas para medir justiça na aprendizagem de máquina, esse *toolkit* proporcionou um arcabouço das principais métricas para a análise da justiça em aprendizagem de máquina. As métricas usadas neste trabalho são advindas do Aequitas. Como cada contexto traz particularidades, escolhemos um subconjunto das métricas para utilizar no nosso trabalho, detalhado na Seção 5.1.

3.2.2 Mitigação de Injustiças

Tendo em vista que análises em alguns cenários apresentaram injustiças contra classes sensíveis, novas abordagens de mitigações de injustiças vêm sendo criadas. Seu objetivo é não

só mitigar as injustiças presentes em cenários sensíveis, mas também fazer com que essa mitigação não venha acompanhada de uma piora da eficácia dos modelos. Como diferentes abordagens de mitigação foram criadas, Bellamy et al. [2] desenvolveram o *toolkit AIF360* com objetivo de reunir algoritmos de mitigação de injustiças em uma ferramenta. Assim, tanto pesquisadores quanto a indústria podem utilizar a evoluir algoritmos do estado da arte para mitigar injustiças.

Dentre os algoritmos de mitigação que o *toolkit AIF360* proporciona, escolhemos três algoritmos de mitigação para experimentação nesse trabalho. O primeiro foi desenvolvido por Zhang et al. [25] e trata-se de um algoritmo de redes neurais adversariais que realiza o treinamento via aprendizagem de máquina ao passo que mitiga as injustiças. Já Feldman et al. [9] desenvolveram um algoritmo de mitigação que modifica as variáveis de entrada, de modo que as variáveis sensíveis não sejam tão afetadas na predição. Por último, Pleiss et al. [16] desenvolveram um algoritmo que mitiga as injustiças ao fazer pequenos ajustes na predição do modelo. Esses algoritmos são mais detalhados na Seção 2.4.

Capítulo 4

Materiais e Métodos

Este trabalho de dissertação é composto de dois experimentos, os quais exploram justiça na estimativa de risco de contratos e empresas, descrito na figura 4. O primeiro experimento é uma avaliação da justiça dos modelos de estimativa de risco de contratos públicos e empresas atualmente utilizados por órgãos de controle, que por brevidade será chamado de *análise de justiça*. O segundo experimento é uma avaliação experimental de modelos de mitigação de injustiças sobre o mesmo cenário, o qual chamamos de *mitigação de injustiças*. Para a realização desses dois experimentos foi necessário também realizar o treinamento de modelos de estimativa de risco dos contratos e empresas, bem como a avaliação de eficácia destes modelos.

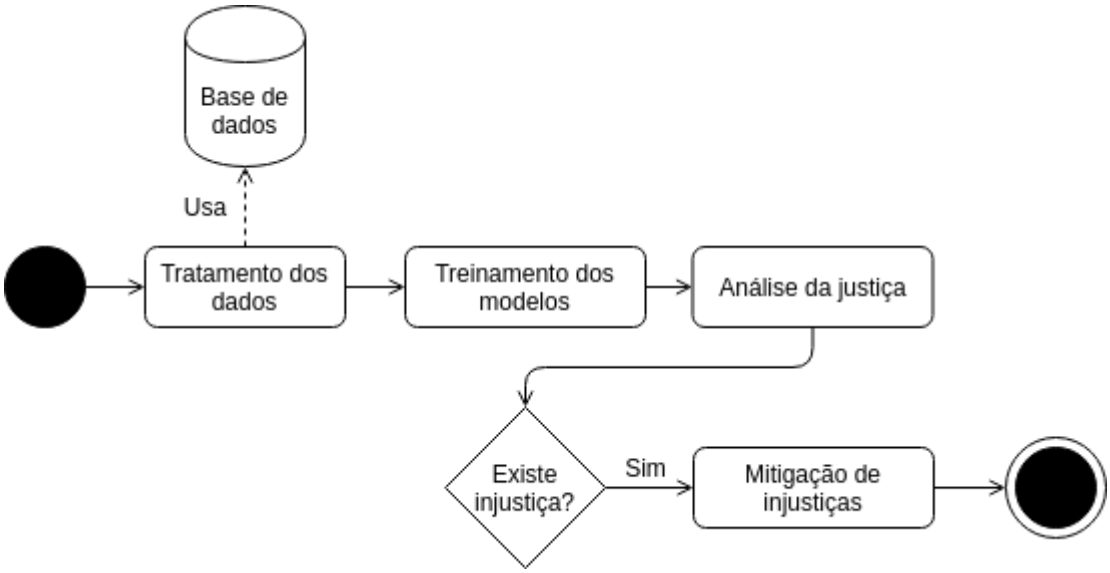

Figura 4.1: Fluxo de atividades realizadas.

As estimativas de risco são baseadas em informações históricas sobre contratos e empresas que estão materializados em três bases de dados detalhados na Seção 4.1. Os tratamentos nos dados necessários para realizar os experimentos são especificados na Seção 4.2. A metodologia do treinamento dos modelos é explicada na Seção 4.3. Por último, a infraestrutura usada, no tocante à linguagem de programação e bibliotecas, bem como suas versões, é descrita na Seção 4.4. Como os experimentos de análise da justiça e mitigação de injustiças presente na estimativa de risco são a parte principal deste trabalho, dedicamos dois capítulos para esses. Dessa forma, a metodologia empregada nesses experimentos é detalhada nos Capítulos 5 e 6.

4.1 Bases de Dados

Para que o trabalho pudesse ser realizado, foi necessário primeiramente ter acesso a dados referentes a contratos e empresas, os quais os órgãos de controle utilizam para estimativa de risco. Além disso, é preciso além de possuir características de empresas e contratos, termos também uma variável resposta que indica se a empresa cumpriu ou não com suas obrigações contratuais, ou seja, se houve a prestação adequada do serviço ou do bem fornecido a um ente público. Os dados utilizados nesta pesquisa vieram de três bases de dados, as quais descrevem essas entidades em diferentes contextos: empresas em nível federal, empresas em nível municipal e contratos em nível municipal. Os dados usados nos experimentos foram modificados a partir de dados utilizados por órgãos de controle externo estaduais e federais fornecidos aos autores e que descrevemos a seguir. Assim como em outros trabalhos da área [22; 20; 11], não descrevemos em detalhes todas as características usadas pelos órgãos de controle para fiscalizar empresas e contratos. Os órgãos entendem que o sigilo da definição das características é estratégico para a investigação.

A primeira das três bases foi criada pela CGU e é baseada em dados usados para fiscalização de contratos com o Governo Federal. Essa base é resultado de um cruzamento de 7 bancos de dados diferentes que medem 183 características de pouco mais de 10 mil empresas referentes à sua capacidade operacional, perfil de participação em licitações, histórico de punições e descobertas, conflitos de interesses e ligações políticas. Esta base foi criada e utilizada por Sales et. al [22] para avaliar o uso de redes neurais na estimativa de risco

de empresas que têm contratos de R$1 milhão ou mais com o Governo Federal. Para cada empresa, existe um rótulo informando se a empresa tem ao menos um contrato que não foi executado de acordo com as suas especificações nos dois anos seguintes à criação dos dados. As características da empresa são de 2011 à 2014, enquanto o rótulo são de 2015 à 2016. Por brevidade, nos referimos a essa base como *empresas da esfera federal*.

A segunda base de dados é proveniente do MPPB e complementada por nós. Esta base possui características operacionais de cerca de 40 mil empresas do estado da Paraíba e 31 características criadas por especialistas para aferir o risco do governo firmar um contrato com uma empresa. Essas características das empresas são datadas de 2014 à abril de 2017. O MPPB usa uma soma ponderada dessas características para auxiliar na análise ad-hoc de risco das empresas listadas. Em nosso experimento, o rótulo que define se uma empresa é arriscada no gabarito vem do TRAMITA. Definimos que uma empresa é arriscada se ela teve um ou mais contratos rescindidos em uma data posterior à criação dos dados do MPPB, ou seja, após abril de 2017. A limiarização da soma ponderada utilizada na estimativa de risco ad-hoc está descrita na Seção 4.2. Chamamos essa base de *empresas da esfera municipal*.

A última base de dados foi criada a partir de dados fornecidos pelo Tribunal de Contas do Estado da Paraíba (TCE-PB), e possui 30 características de cerca de 13 mil contratos celebrados por entes públicos na Paraíba e informações das empresas contratadas no momento da celebração dos contratos. Essa base de dados possui informações sobre contratos celebrados entre 2014 e 2017. Como forma de rotular contratos arriscados, novamente utilizamos a base de dados TRAMITA, porém agora identificando se o contrato foi rescindido em um momento posterior à celebração do mesmo. Diferente dos casos descritos até aqui, essa base nos permite estudar modelos que estimem risco de *contratos*, e não de empresas – as quais podem ter múltiplos contratos em um período –, um cenário ao mesmo tempo mais esparso e de mais relevância prática para os órgãos de controle. Essa base de dados foi criada pelo grupo de pesquisa do autor e será chamada de *contratos municipais na Paraíba*.

Todos os conjuntos de dados utilizados têm um grande desbalanceamento entre observações com alto e baixo risco, onde as observações de baixo risco são maioria. As bases das empresas da esfera federal, empresas da esfera municipal e contratos municipais na Paraíba possuem 92,7%, 99,6% e 98,9% das observações da classe negativa (baixo risco), respectivamente.

4.1.1 Classes sensíveis

Para que a avaliação da justiça, bem como a mitigação, pudessem ser realizadas, é necessário identificar as características sensíveis nas bases de dados. Uma característica é dita sensível caso o julgamento diferente das classes da característica possa impactar negativamente os indivíduos de certas classes, tocando assim em questões morais. Para identificar se existe injustiça na estimativa de risco é necessário comparar as classes sensíveis com suas respectivas classes pertencentes ao grupo de referência. Por exemplo, no caso do sistema de estimativa de risco de reincidência criminal COMPAS [13], uma característica sensível é a raça das pessoas classificadas. Dentro de uma característica sensível, *classes sensíveis* são definidas pelo conjunto das diferentes classes as quais não pertencem ao grupo de referência. O grupo de referência é a classe de uma característica sensível onde não se espera que haja efeito negativo caso haja um julgamento enviesado. Por exemplo, no caso da característica raça usada no sistema de estimativa de risco de reincidência criminal COMPAS [13], as pessoas da raça negra são uma classe sensível, enquanto brancos são o grupo de referência.

As características sensíveis em nossos dados foram identificadas inspecionando-se as variáveis já usadas pelos órgãos de controle ou disponíveis e aparentemente relacionadas ao risco a partir de nosso experiência com esse órgão, e buscando variáveis que possuam classes relacionadas a grupos mais vulneráveis aos custos implicados em responder uma investigação.

Ao final desse processo, as características sensíveis identificadas são as seguintes: porte da empresa, idade da empresa, se a empresa está sediada em cidade de interior e se a empresa tem ou teve um sócio integrante ou ex-integrante do programa bolsa família, esta última será referida como *sócio relacionado ao BF* por brevidade. Entendemos que as empresas que possuem essas características são mais frágeis, de forma que elas podem ser mais prejudicadas em comparação às que não possuem essas características, caso um julgamento indevido seja feito. Assim, ao proporcionar um ambiente mais justo nas investigações, queremos incentivar que elas participem de processos licitatórios. Algumas dessas características não estão disponíveis em todas as bases de dados, e a definição das classes sensíveis também é contingente do formato dos dados. O conjunto de características e classes sensíveis e de referências das diferentes bases é detalhado na Tabela 4.1.

Tabela 4.1: Características sensíveis dos conjuntos de dados.

Conjunto de dados	**Características sensíveis**	**Classes sensíveis**	**Grupo de referência**
Contratos municipais na Paraíba	Empresa de pequeno porte? e idade da empresa	Empresa de pequeno porte: sim; idade da empresa: jovem e nova	Empresa de pequeno porte: não; idade da empresa: consolidada
Empresas da esfera federal	Empresa de pequeno porte?, empresa do interior? e idade da empresa	Empresa de pequeno porte: sim; empresa do interior: sim; idade da empresa: jovem e nova	Empresa de pequeno porte: não; empresa do interior: não; idade da empresa: consolidada
Empresas da esfera municipal	Empresa do interior?, idade da empresa, porte da empresa e sócio relacionado ao BF?	Empresa do interior: sim; idade da empresa: jovem e nova; porte da empresa: empresa pequeno porte e microempresa; sócio relacionado ao BF: sim	Empresa do interior: não; idade da empresa: consolidada; porte da empresa: demais; sócio relacionado ao BF: não

4.2 Tratamento dos Dados

Para realização da análise de justiça e mitigação de injustiças foi necessário primeiramente limiarizar algumas características sensíveis numéricas para características sensíveis categóricas, visto que a análise de justiça objetiva identificar disparidade entre classes de uma característica sensível, descrito em mais detalhes no Capítulo 5. Vale ressaltar que cada base de dados possui diferentes conjuntos de características sensíveis, tendo em vista a diferença estrutural das bases. A seguir são citadas as limiarizações feitas:

- A cidade na qual a empresa é sediada foi limiarizada em se era a capital ou uma cidade do interior
- A idade da empresa foi limiarizada em nova (menos de 3 anos), jovem (3 a 10 anos) ou consolidada (mais de 10 anos).
- O porte da empresa foi limiarizado em pequeno porte ou não, mas em uma das bases foi possível categorizar em microempresa, pequeno porte e demais.

- Se a empresa teve sócio relacionado ao bolsa família foi derivado de quatro variáveis sensíveis: se a empresa possui sócio integrante do bolsa família, se a empresa possuiu sócio integrante do bolsa família, se a empresa possui sócio ex-integrante do programa bolsa família e se a empresa possuiu sócio ex-integrante do bolsa família.

A base de dados *empresas da esfera municipal* possui uma característica particular, na qual existe uma soma ponderada utilizada na estimativa de risco ad-hoc do órgão, como citado na Seção 4.1. Para transformar essa soma ponderada, a qual possuía valores inteiros, em binária foi utilizado o limiar de 160 pontos. Este limiar proporcionou que fossem classificadas como arriscadas cerca de 4,3% das empresas com maior estimativa de risco. Este limiar próximo de 5% foi definido utilizando um critério similar ao de Gomes et al. [11], onde eles limiarizaram observações acima do 95 percentil. Porém, eles utilizaram esse valor para limiarizar observações como anômalas através de *deep auto-encoders*, enquanto aqui utilizamos esse valor para limiarizar observações como de alto risco em uma abordagem ad-hoc.

A base de dados de empresas da esfera federal não possui valores faltantes, porém a de contratos da esfera municipal possui cerca de 0,75% dos dados faltantes, enquanto os dados de empresas da esfera municipal possui cerca de 57% de dados faltantes. Apesar do grande número de dados faltantes para empresas da esfera municipal, percebeu-se que o risco total atribuído a uma empresa era a soma das características ponderadas não nulas, ou seja, os valores nulos das características ponderadas deveriam ser substituídos por zero. Desta forma, apenas 1,8% dos dados realmente eram faltantes para a base de empresas da esfera municipal.

Para o treinamento dos modelos de estimativa de risco foi necessário tanto remover algumas observações, quanto imputar dados faltantes e codificar variáveis categóricas em variáveis *dummy* numéricas. As observações que possuíam valores faltantes nas características sensíveis foram removidas, tendo em vista que não era possível realizar a análise ou mitigação das injustiças para as mesmas. Para a imputação de dados, nas variáveis categóricas que possuíam valores nulos foi atribuído a classe "desconhecido", enquanto para as variáveis numéricas foi atribuída a média dessa variável. Para a transformação de variáveis categóricas em numéricas, as variáveis categóricas foram transformadas em variáveis dummy, ou seja, cada categoria cria uma nova coluna, onde esta possui 1 se uma dada observação pertence àquela classe e 0 caso contrário. Algumas outras poucas variáveis tiveram o tratamento dos

valores faltantes feito de forma diferente, específico para a semântica de cada variável. Porém, como o sigilo dessas informações é estratégico, não serão explanados esses tratamentos.

4.3 Treinamento dos Modelos

Para definir os modelos de aprendizagem de máquina a serem utilizados na estimativa de risco, primeiramente testamos cinco classes de modelos em cada base de dados e buscamos o modelo que teve maior eficácia em cada uma das bases, para em seguida utilizá-lo no estudo da justiça. Ou seja, o objetivo é encontrar o modelo com maior eficácia que estime o risco de uma empresa tenha pelo menos um contrato rescindido, ou que um determinado seja rescindido, dependendo do contexto. Diferentes famílias de modelos foram utilizado para o treinamento: florestas aleatórias, regressões logísticas, redes neurais, máquinas de vetor de suporte e *k-nearest neighbors*.

Como mencionado na Seção 4.1, as bases de dados tem variável alvo bastante desbalanceada. Desta forma fez-se necessário a utilização de técnicas de balanceamento para melhoria dos modelos. As seguintes técnicas foram experimentadas para cada modelo e cada base de dados de treinamento: sem balanceamento, *undersampling* aleatório e SMOTE.

Para o treinamento dos modelos, nosso processo é o mesmo descrito por Sun e Sales [22]: utilizamos validação cruzada com 5 *folds*, balanceando os dados de treino, citado anteriormente, e particionamos o conjunto em treino e teste em 80% e 20%. Os melhores hiperparâmetros, assim como os melhores modelos, foram buscados considerando o f1 score (4.1), tendo em vista o desbalanceamento das classes de risco no conjunto de teste.

$$f_1 score = 2 \cdot \frac{precisao \cdot revocacao}{precisao + revocacao} \tag{4.1}$$

A busca pelos melhores hiperparâmetros de cada modelo foi realizada através de um *grid search* com validação cruzada, o qual faz uma combinação de cada parâmetro buscando pela combinação mais eficaz. A seguir estão descritas os parâmetros utilizados para cada modelo:

- A regressão logística utilizou o algoritmo de otimização *liblinear* e o número máximo de iterações sendo 1000. A técnica de regularização variou entre *l1* e *l2* e a inversa da força de regularização variou entre 1, 0,1 e 0,01.

- A floresta aleatória variou entre *gini* e *entropy* para o ganho de informação em cada split. O número de árvores na floresta variou entre 10, 20, 40, 80 e 100, enquanto a profundidade máxima da árvore variou entre 1, 2, 4 e 8.

- A rede neural utilizou a otimização de pesos *lbfgs* em decorrência do baixo número de observações nas bases de dados. Além disso, variou a função de ativação entre sigmóide logística, tangente hiperbólica (tanh) e *rectified linear unit function* (relu). O termo de regularização L2 variou entre 0,01, 0,001 e 0.0001, enquanto as camadas escondidas da rede neural variou entre (150, 90, 45), (90, 45) e (90, 45, 22).

- A máquina de vetor de suporte teve 10000 como número máximo de iterações e 1e-2 para a tolerância do critério de parada. Além disso, variou seu kernel entre polinomial, sigmóide e *radial basis function* (RBF).

- O *k-nearest neighbors* teve o número de vizinhos variado entre 2, 3, 5, 8 e 13.

Os melhores hiperparâmetros para cada base de dados, técnica de balanceamento e modelo estão descritos nas Tabelas 4.2, 4.3, 4.4.

O resultado da eficácia dos modelos está representado nas Tabelas 4.5, 4.6 e 4.7. Como pode ser visto, os melhores modelos obtidos foram com random forest como algoritmo e o SMOTE como técnica de balanceamento, considerando o f1 score. Apenas para o conjunto de dados de contratos da esfera municipal que o melhor f1 score foi obtido através do KNN e sem técnica de balanceamento. Apesar do f1 score ter sido ligeiramente maior, preferimos utilizar o random forest em conjunto do SMOTE, pois a revocação foi consideravelmente maior. Quando temos uma baixa revocação o algoritmo considera muitas observações de alto risco como sendo de baixo risco.

4.4 Infraestrutura

Para a realização dos experimentos foi utilizada a linguagem de programação Python (versão 3.6.8). Para tratamento dos dados foi utilizado o Pandas [1] em conjunto com o NumPy [2]. Para

[1] `https://pandas.pydata.org/pandas-docs/version/0.25.1/`
[2] `https://pypi.org/project/numpy/1.17.2/`

Tabela 4.2: Melhores hiperparâmetros para os contratos municipais na Paraíba

Técnica de balanceamento	**Modelo**	**Melhores hiperparâmetros**
Sem balanceamento	regressão logística	'C': 1.0, 'penalty': 'l1'
Sem balanceamento	random forest	'criterion': 'gini', 'max_depth': 1, 'n_estimators': 10
Sem balanceamento	redes neurais	'activation': 'tanh', 'alpha': 0.01, 'hidden_layer_sizes': (150, 90, 45)
Sem balanceamento	SVM	'kernel': 'rbf'
Sem balanceamento	KNN	'n_neighbors': 3
SMOTE	regressão logística	'C': 1.0, 'penalty': 'l1'
SMOTE	random forest	'criterion': 'gini', 'max_depth': 8, 'n_estimators': 100
SMOTE	redes neurais	'activation': 'tanh', 'alpha': 0.01, 'hidden_layer_sizes': (150, 90, 45)
SMOTE	SVM	'kernel': 'rbf'
SMOTE	KNN	'n_neighbors': 2
Undersampling	regressão logística	'C': 0.01, 'penalty': 'l1'
Undersampling	random forest	'criterion': 'entropy', 'max_depth': 2, 'n_estimators': 20
Undersampling	redes neurais	'activation': 'logistic', 'alpha': 0.0001, 'hidden_layer_sizes': (150, 90, 45)
Undersampling	SVM	'kernel': 'poly'
Undersampling	KNN	'n_neighbors': 3

Tabela 4.3: Melhores hiperparâmetros para as empresas da esfera municipal

Técnica de balanceamento	**Modelo**	**Melhores hiperparâmetros**
Sem balanceamento	regressão logística	'C': 1.0, 'penalty': 'l1'
Sem balanceamento	random forest	'criterion': 'gini', 'max_depth': 1, 'n_estimators': 10
Sem balanceamento	redes neurais	'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (90, 45)
Sem balanceamento	SVM	'kernel': 'poly'
Sem balanceamento	KNN	'n_neighbors': 2
SMOTE	regressão logística	'C': 1.0, 'penalty': 'l1'
SMOTE	random forest	'criterion': 'gini', 'max_depth': 8, 'n_estimators': 100
SMOTE	redes neurais	'activation': 'logistic', 'alpha': 0.0001, 'hidden_layer_sizes': (150, 90, 45)
SMOTE	SVM	'kernel': 'rbf'
SMOTE	KNN	'n_neighbors': 2
Undersampling	regressão logística	'C': 0.1, 'penalty': 'l2'
Undersampling	random forest	'criterion': 'entropy', 'max_depth': 4, 'n_estimators': 10
Undersampling	redes neurais	'activation': 'logistic', 'alpha': 0.0001, 'hidden_layer_sizes': (90, 45)
Undersampling	SVM	'kernel': 'rbf'
Undersampling	KNN	'n_neighbors': 13

Tabela 4.4: Melhores hiperparâmetros para as empresas da esfera federal

Técnica de balanceamento	**Modelo**	**Melhores hiperparâmetros**
Sem balanceamento	regressão logística	'C': 1.0, 'penalty': 'l1'
Sem balanceamento	random forest	'criterion': 'gini', 'max_depth': 8, 'n_estimators': 10
Sem balanceamento	redes neurais	'activation': 'relu', 'alpha': 0.001, 'hidden_layer_sizes': (150, 90, 45)
Sem balanceamento	SVM	'kernel': 'poly'
Sem balanceamento	KNN	'n_neighbors': 3
SMOTE	regressão logística	'C': 1.0, 'penalty': 'l1'
SMOTE	random forest	'criterion': 'entropy', 'max_depth': 8, 'n_estimators': 100
SMOTE	redes neurais	'activation': 'logistic', 'alpha': 0.0001, 'hidden_layer_sizes': (90, 45, 22)
SMOTE	SVM	'kernel': 'poly'
SMOTE	KNN	'n_neighbors': 2
Undersampling	regressão logística	'C': 1.0, 'penalty': 'l1'
Undersampling	random forest	'criterion': 'gini', 'max_depth': 8, 'n_estimators': 100
Undersampling	redes neurais	'activation': 'logistic', 'alpha': 0.0001, 'hidden_layer_sizes': (90, 45, 22)
Undersampling	SVM	'kernel': 'rbf'
Undersampling	KNN	'n_neighbors': 13

Tabela 4.5: Eficácia dos modelos de estimativa de risco das empresas da esfera federal

Técnica de balanceamento	**Modelo**	**F1 score**	**Revocação**	**Precisão**	**Acurácia**
SMOTE	Regressão Logística	0.346972	0.679487	0.232967	0.804220
SMOTE	Random Forest	0.420253	0.532051	0.347280	0.887635
SMOTE	Redes Neurais	0.209859	0.955128	0.117880	0.449460
SMOTE	SVM	0.142206	1	0.076546	0.076546
SMOTE	KNN	0.172662	0.230769	0.137931	0.830716
Undersampling	Regressão Logística	0.347059	0.756410	0.225191	0.782139
Undersampling	Random Forest	0.352480	0.865385	0.221311	0.756624
Undersampling	Redes Neurais	0.219325	0.916667	0.124564	0.500491
Undersampling	SVM	0.142466	1.000000	0.076696	0.078508
Undersampling	KNN	0.212069	0.788462	0.122510	0.551521
Sem balanceamento	Regressão Logística	0.204878	0.134615	0.428571	0.920020
Sem balanceamento	Random Forest	0.127660	0.076923	0.375000	0.919529
Sem balanceamento	Redes Neurais	0.142206	1.000000	0.076546	0.076546
Sem balanceamento	SVM	0	0	0	0.923454
Sem balanceamento	KNN	0.029412	0.019231	0.062500	0.902846

Tabela 4.6: Eficácia dos modelos de estimativa de risco das empresas da esfera municipal

Técnica de balanceamento	**Modelo**	**F1 score**	**Revocação**	**Precisão**	**Acurácia**
SMOTE	Regressão Logística	0.024331	0.652174	0.012397	0.849399
SMOTE	Random Forest	0.073801	0.434783	0.040323	0.968578
SMOTE	Redes Neurais	0.023661	0.826087	0.012003	0.803706
SMOTE	SVM	0.006376	1	0.003198	0.102654
SMOTE	KNN	0.009174	0.043478	0.005128	0.972959
Undersampling	Regressão Logística	0.049550	0.478261	0.026128	0.947171
Undersampling	Random Forest	0.029186	0.826087	0.014855	0.841763
Undersampling	Redes Neurais	0.019846	0.782609	0.010050	0.777416
Undersampling	SVM	0.005759	1	0.002888	0.005759
Undersampling	KNN	0.026798	0.826087	0.013620	0.827241
Sem balanceamento	Regressão Logística	0	0	0	0.996995
Sem balanceamento	Random Forest	0	0	0	0.997121
Sem balanceamento	Redes Neurais	0.004630	0.043478	0.002445	0.946169
Sem balanceamento	SVM	0.005742	1	0.002879	0.002879
Sem balanceamento	KNN	0	0	0	0.996745

Tabela 4.7: Eficácia dos modelos de estimativa de risco dos contratos da esfera municipal

Técnica de balanceamento	**Modelo**	**F1 score**	**Revocação**	**Precisão**	**Acurácia**
SMOTE	Regressão Logística	0.040619	0.656250	0.020958	0.686572
SMOTE	Random Forest	0.140000	0.437500	0.083333	0.945656
SMOTE	Redes Neurais	0.030769	0.750000	0.015707	0.522275
SMOTE	SVM	0.018983	0.875000	0.009596	0.085624
SMOTE	KNN	0.099010	0.312500	0.058824	0.942496
Undersampling	Regressão Logística	0.024057	0.687500	0.012243	0.436019
Undersampling	Random Forest	0.025466	0.875000	0.012921	0.322907
Undersampling	Redes Neurais	0.024176	0.687500	0.012304	0.438863
Undersampling	SVM	0.020019	1	0.010111	0.010111
Undersampling	KNN	0.026810	0.625000	0.013699	0.541232
Sem balanceamento	Regressão Logística	0	0	0	0.989889
Sem balanceamento	Random Forest	0	0	0	0.989889
Sem balanceamento	Redes Neurais	0	0	0	0.989889
Sem balanceamento	SVM	0.114286	0.062500	0.666667	0.990205
Sem balanceamento	KNN	0.181818	0.125000	0.333333	0.988626

o treinamento dos modelos foi utilizado o scikit-learn [3], enquanto para o balanceamento das classes foi utilizado o imbalanced-learn [4]. Para a análise da justiça foi utilizada a aequitas [5]. Por último, para a mitigação das injustiças foi utilizada a biblioteca AIF360 [6].

[3] https://scikit-learn.org/stable/whats_new/v0.21.html#version-0-21-2
[4] https://pypi.org/project/imbalanced-learn/0.4.3/
[5] https://pypi.org/project/aequitas/0.36.0/
[6] https://pypi.org/project/aif360/0.2.2/

Capítulo 5

Análise de Justiça

Este capítulo contém o detalhamento do primeiro experimento, no qual é realizado uma análise da justiça da estimativa de risco em contratos públicos e empresas. Todos os modelos de estimativa de risco, independente da base de dados, obtiveram melhores resultados com florestas aleatórias como algoritmo de treinamento e *SMOTE* como técnica de balanceamento. Na Seção 5.1 é descrita a metodologia experimental utilizada para avaliar a justiça. Na Seção 5.2 são descritos os resultados obtidos através do experimento. Por último, na Seção 5.3 são discutidos os resultados.

Apesar deste experimento ter sido publicado anteriormente no artigo *Fairness in Risk Estimation of Brazilian Public Contracts* [26], os resultados aqui apresentados são ligeiramente diferentes. Essa diferença se deu em decorrência de uma nova execução do treinamento dos modelos de estimativa de risco, onde foram adicionadas características sensíveis aos dados de treinamento, acarretando em mudanças na análise da justiça. Apesar das pequenas mudanças nos resultados, as conclusões permanecem as mesmas.

5.1 Metodologia

Para realizar a análise da justiça utilizamos o *toolkit Aequitas*, proposto por Saleiro et al. [18]. Esse *toolkit* define métricas de justiça e viés para modelos de aprendizagem de máquina (AM) que permitem verificar se existe disparidade no comportamento do modelo para classes sensíveis do conjunto de dados. Sejam VP e VN as quantidades de verdadeiros positivos e verdadeiros negativos na classificação de um conjunto de entidades, e FP e FN as quanti-

dades de falsos positivos e falsos negativos na mesma classificação. As principais métricas utilizadas na análise de justiça neste trabalho são: a taxa de falsos positivos (5.1), a revocação (5.2) e a precisão (5.3).

$$TFP = \frac{FP}{FP + VN} \tag{5.1}$$

$$revocacao = \frac{VP}{VP + FN} \tag{5.2}$$

$$precisao = \frac{VP}{VP + FP} \tag{5.3}$$

A TFP mede a proporção que o modelo estima erroneamente como de alto risco, quando na verdade era de baixo risco. A revocação mede a proporção que o modelo acerta dentre os casos de alto risco. A precisão mede a proporção que o modelo acerta entre os casos estimados como de alto risco. No contexto de justiça na estimativa de risco de gastos públicos é mais grave acusar falsamente uma empresa de alto risco do que acusar falsamente de baixo risco, pois ao acusar falsamente de alto risco pode propiciar um viés de confirmação ao auditor. Desta forma, a TFP foi escolhida como métrica de justiça. A precisão e revocação foram escolhidas pois são medidas de eficácia bastante utlizadas no contexto de aprendizagem de máquina.

A *disparidade* diz respeito à métrica avaliada para diferentes classes de uma mesma característica sensível. A disparidade $disp$ de uma métrica M entre uma classe do grupo de referência GR e uma classe sensível s é calculada, para uma categoria sensível c, através da seguinte fórmula:

$$disp_{M,c} = \frac{M_s}{M_{GR}} \tag{5.4}$$

Por exemplo, para a disparidade na TFP (5.1), temos $disp_{TFP,c} = \frac{TFP_s}{TFP_{GR}}$. Existe disparidade segundo uma métrica em uma característica sensível dos dados quando o valor da disparidade se distancia de 1, entre pelo menos uma classe sensível e a classe do grupo de referência.

Como os conjunto de dados analisados são uma amostra das estimativas de interesse, reportamos nossos resultados estimando intervalos de confiança para as métricas de interesse

na população das empresas ou contratos. Ou seja, para cada cenário e métrica, calculamos os intervalos de confiança da disparidade, os quais foram calculados para as diferentes classe sensível, de forma a comparar com o grupo de referência. Todos os intervalos são estimados com 95% de confiança através de 10.000 bootstraps. Dado o desbalanceamento das classes sensíveis e da classe positiva na variável de resposta, aplicamos bootstrap estratificado em função da conjunção das características sensíveis com a variável de resposta. A escolha da utilização do bootstrap estratificado, que utiliza amostragens estratificadas, se deve ao fato de que ele é um método mais representativo que bootstrap que utiliza amostragens aleatórias, onde a representatividade se refere a uma característica da população [15]. Por fim, ao comentar os resultados, optamos por interpretar os intervalos estimados, em lugar de dicotomizar resultado em significativos ou não [6].

Vale ressaltar também que, apesar de possuírmos três bases de dados, serão apresentados quatro resultados na próxima seção. Além das três estimativas de risco realizada via aprendizagem de máquina, a base de dados *empresas da esfera municipal* já era composta previamente por uma estimativa de risco de empresas realizada através de uma soma ponderada de características da empresa criada por especialistas. Como essa estimativa de risco não foi realizada via aprendizagem de máquina, chamaremos aqui de estimativa de risco com uma abordagem ad-hoc.

5.2 Justiça no Estado da Prática

Nesta seção são apresentados os resultados da análise da justiça. É relevante comentar previamente que alguns dos intervalos de confiança contém zero. Isso acontece em virtude da distribuição dos dados, pois as classes sensíveis são uma menor parte deles. Mesmo utilizando o bootstrap estratificado, por vezes a reamostragem não selecionou nenhuma observação que fosse falso positivo, para a taxa de falsos positivos, nem verdadeiros positivos, para a precisão e revocação. Como as métricas são razões e têm como numerador as observações destacadas, então os valores são zerados. Associada as métricas das classes sensíveis zeradas, a fórmula da disparidade, descrita na Seção 5.1, tem no numerador a métrica da classe sensível, então a disparidade também obtém valor zero.

5.2.1 Empresas na esfera municipal com abordagem ad-hoc dos especialistas

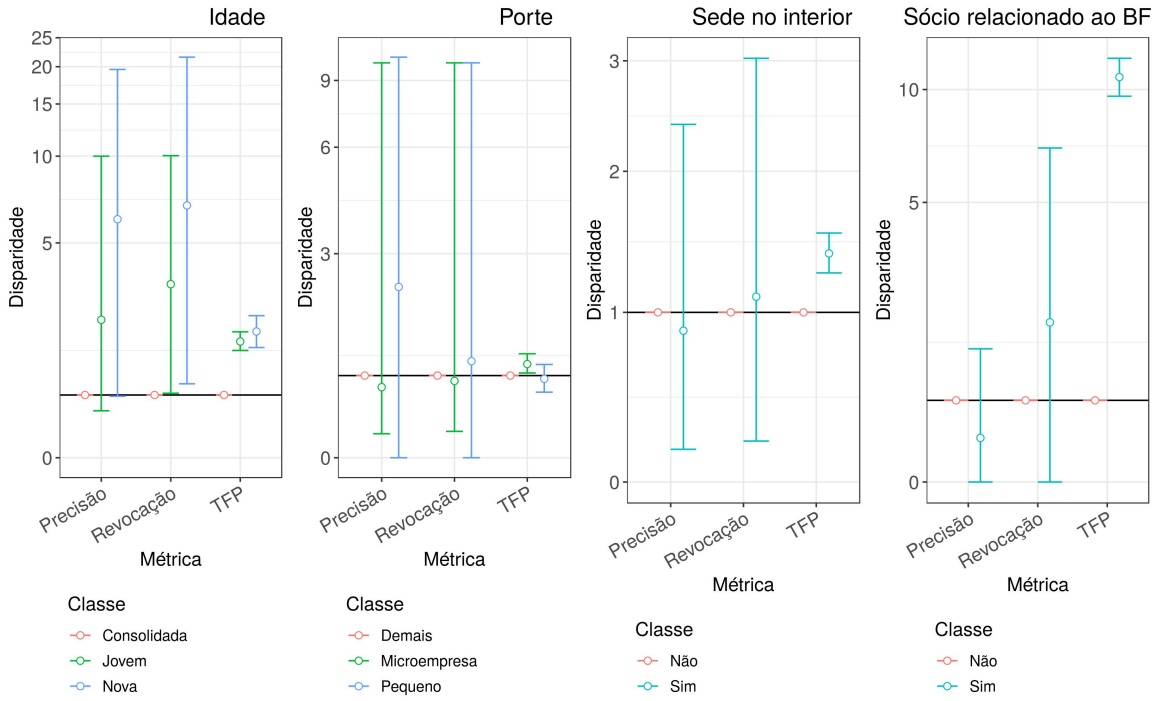

Figura 5.1: Disparidade das métricas das empresas da esfera municipal (abordagem ad-hoc).

A Figura 5.1 mostra as disparidades na estimativa de risco ad-hoc das empresas da esfera municipal. Há uma grande diferença na TFP comparando as empresas que têm ou tiveram sócios vinculados ao bolsa família com as que não tiveram sócios com vínculo; estimamos que a TFP da classe sensível é entre 9,6 e 12 vezes maior que a do grupo de referência. Além disso, empresas novas e jovens têm TFP entre 1,9 e 2,6, e entre 1,8 e 2,2 vezes maior comparadas às empresas consolidadas, respectivamente. A revocação também foi maior para empresas novas e jovens em relação à empresas consolidadas, mas é difícil quantificar o quão maior, visto que estimamos que é plausível que a disparidade esteja entre 1,2 e 21,5 para as novas e entre 1,03 e 10 para as jovens. A disparidade na precisão das empresas jovens é entre 0,74 e 10 vezes maior e as novas e é entre 0,98 e 19,6 vezes maior comparada com as consolidadas. Ou seja, a precisão para as empresas jovens e novas pode ser desprezivelmente menor, igual ou muito maior que a da consolidadas. A TFP também é maior para empresas em cidades do interior entre 1,3 e 1,5 vezes, uma disparidade relativamente pequena, comparada as demais categorias sensíveis. Por último, microempresas também apontaram uma

diferença na TFP quando comparada com empresas maiores, disparidade essa estimada entre 1,03 e 1,3.

Juntos, os resultados apontam que as empresas que têm ou tiveram sócios vinculados ao bolsa família, empresas novas e jovens, assim como empresas do interior e microempresas são estimadas com frequência desproporcionalmente alta como sendo empresas de alto risco em relação ao grupo de referência.

5.2.2 Empresas na esfera municipal com aprendizagem de máquina

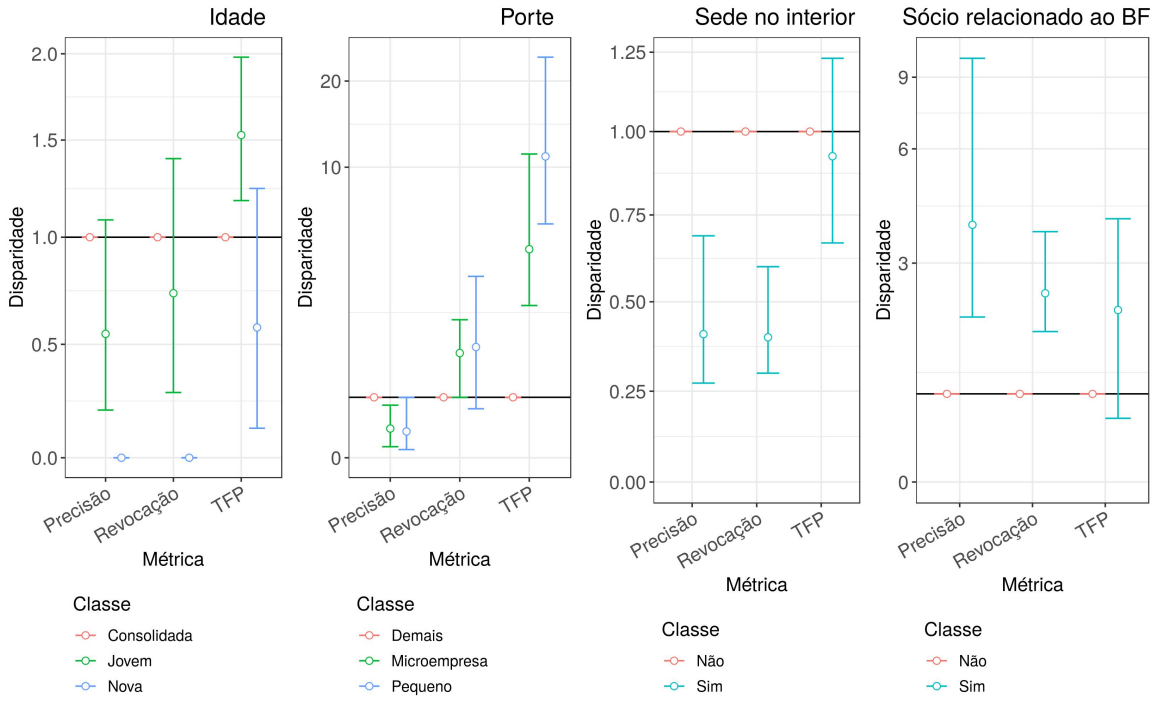

Figura 5.2: Disparidade das métricas das empresas da esfera municipal (abordagem AM).

Ao analisar a estimativa de risco com aprendizagem de máquina das empresas da esfera municipal (Figura 5.2), percebemos que a TFP é maior para empresas de pequeno porte e microempresas em relação às demais. Estimamos que, comparadas com grandes empresas, empresas de pequeno porte e microempresas são classificadas erroneamente como arriscadas mais que 6,3 e 3,1 vezes mais, respectivamente. Sobre a idade da empresa, é possível notar que empresas jovens têm maior TFP que empresas consolidadas, entre 1,2 e 2 vezes. Em contrapartida, a precisão de empresas jovens se for maior que empresas consolidadas é uma diferença pequena (até 1,1 vezes), mas pode ser muito menor (0,2 vezes).

Juntos, os resultados apontam que empresas jovens, empresas de pequeno porte e microempresas são estimadas com frequência desproporcionalmente alta como sendo empresas de alto risco em relação ao grupo de referência. Além disso, ao comparar a abordagem de aprendizagem de máquina com a ad-hoc, é possível perceber uma melhora da injustiça tanto com empresas com sócios relacionados ao bolsa família, quanto empresas do interior e empresas jovens. Porém, ao utilizar a aprendizagem de máquina houve uma piora na injustiça para com empresas de pequeno porte e microempresas.

5.2.3 Empresas na esfera federal com aprendizagem de máquina

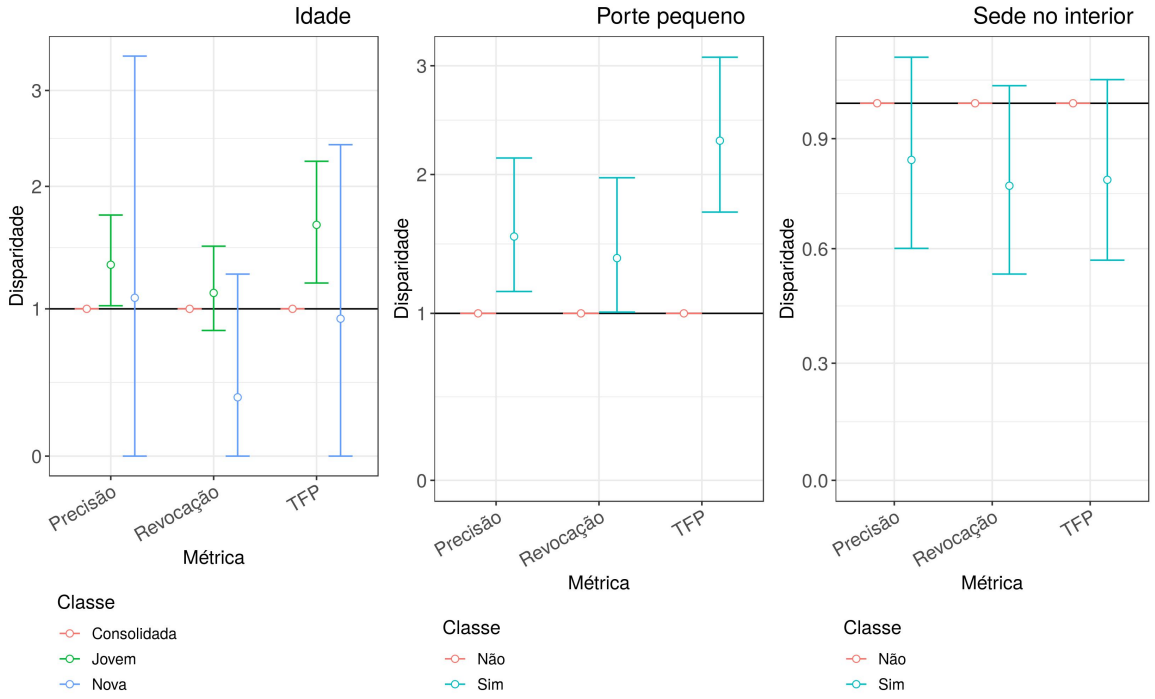

Figura 5.3: Disparidade das métricas das empresas da esfera federal.

No caso do modelo de aprendizagem de máquina para empresas da esfera federal, a Figura 5.3 mostra que ele gera uma disparidade na TFP, precisão e revocação das empresas de pequeno porte. Os valor são, respectivamente [1,7;3,1] [1,1 ;2,1] e [1,01;2], a partir dos quais vê-se que estimamos que o efeito na TFP é claramente alto, enquanto nos outros dois casos pode ser alto ou baixo.

Empresas jovens tiveram tanto TFP quanto precisão maiores em comparação a empresas consolidadas, entre 1,2 e 2,2 vezes e 1,02 e 1,7 vezes respectivamente, enquanto sua

revocação ou foi um pouco menor (0,84 vezes) ou foi razoavelmente maior (1,5 vezes).

Desta forma, os resultados apontam que empresas de pequeno porte e jovens são estimadas erroneamente como sendo de alto risco mais frequentemente que o grupo de referência.

5.2.4 Contratos na esfera municipal com aprendizagem de máquina

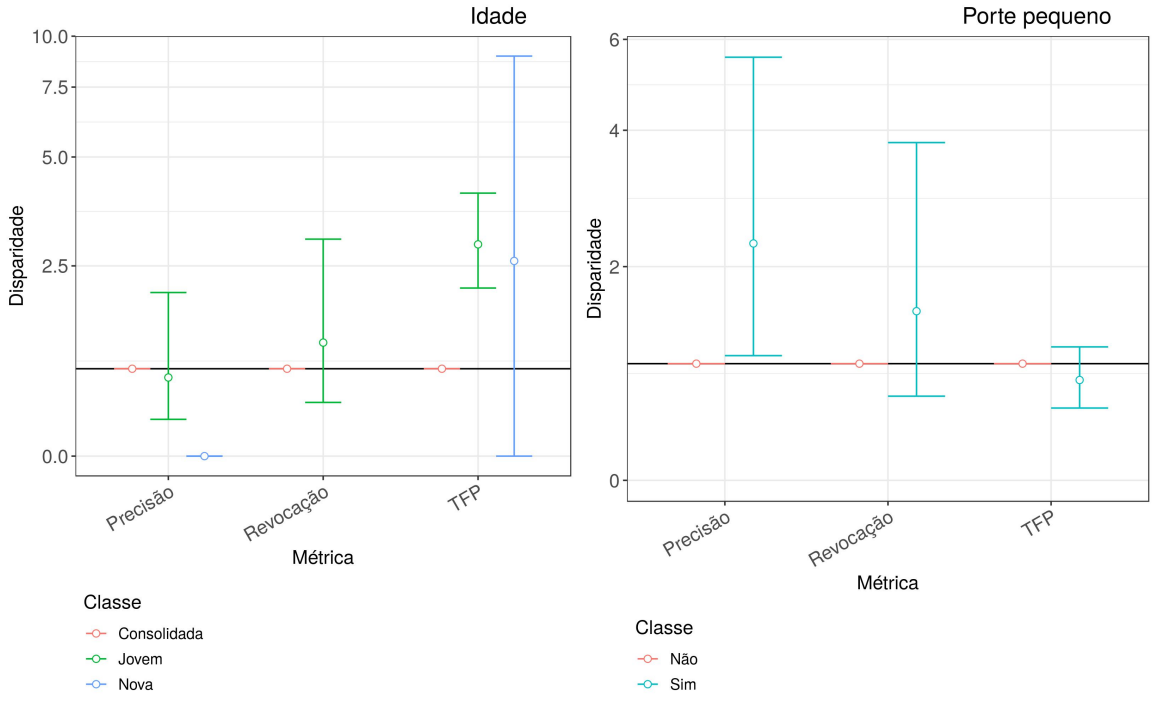

Figura 5.4: Disparidade das métricas dos contratos municipais na Paraíba.

Por último, como pode ser visto na Figura 5.4, o modelo de aprendizagem de máquina baseado na base de dados de contratos municipais na Paraíba apontou uma maior TFP para empresas jovens em relação à empresas consolidadas, entre 2.1 e 4 vezes.

Da mesma forma que nos outros cenários, os resultados apontam que empresas jovens foram mais frequentemente acusadas como de alto risco de forma errada, quando comparada com empresas consolidadas.

5.3 Discussão

Nossa análise aponta que existe uma diferença relevante da TFP em pelo menos uma classe sensível para todos modelos analisados. Além disso, em todos os modelos analisados, em-

presas jovens, com idade entre 3 e 10 anos, tiveram disparidade na TFP comparado à empresas consolidadas. Essas diferenças apontam que empresas de classes sensíveis são estimadas como mais arriscadas de forma injusta em comparação ao grupo de referência.

A disparidade existe tanto na abordagem ad-hoc quanto na abordagem de AM, observando a disparidade das empresas da esfera municipal. Existe tanto no âmbito federal quanto municipal, e tanto no nível da empresa quanto do contrato. Em particular, o fato de que modelos treinados tanto com características criadas por especialistas em controle quanto por pesquisadores de AM levaram a vieses semelhantes é marcante. Isso aponta a relevância de considerarmos injustiça na estimativa de risco para controle de contratos públicos. Discriminar empresas jovens, por exemplo, pode desencorajar a participação de startups em licitações, bem como empresas pequenas mais tradicionais.

Ao observar a abordagem ad-hoc, fica claro que existe uma grande disparidade com empresas com sócios relacionados ao bolsa família. Muitas dessas são pequenas ou microempresas em que os sócios têm ou tiveram baixa renda. Apesar das dificuldades financeiras que eles passaram, os métodos desconfiam sobremaneira desses sócios, estimando injustamente alto risco.

Ao comparar a abordagem ad-hoc com a de AM, é possível perceber que existem mais injustiças na abordagem ad-hoc. Ao utilizar a abordagem de AM reduziu a injustiça dos sócios relacionados ao bolsa família e empresas do interior. Para a idade da empresa, a injustiça notada para empresas novas também desapareceu, mas o mesmo não aconteceu para empresas jovens. Em contrapartida, a utilização de AM propiciou um grande aumento de injustiça para o porte da empresa, seja para empresas de pequeno porte ou micro-empresas.

Ao observar a revocação e precisão juntamente com a TFP, é possível perceber que existem casos em que a TFP é maior para a classe sensível ao mesmo tempo que a precisão ou a revocação é maior. Em outros casos, a diferença na TFP não é seguida de uma diferença da precisão ou revocação. Quando a disparidade não está associada a uma maior precisão ou revocação, temos uma situação de injustiça sem que o modelo tenha mais eficácia classificando a classe sensível. Na situação onde a injustiça está associada com uma maior eficácia, existe um dilema moral: o modelo classifica com maior eficácia a classe na qual ele também produz mais falsos positivos. Esse dilema, por sua vez, aumenta o risco de que um modelo injusto seja posto em prática visando a sua eficácia. Porém nesses casos, como também

em outros cenários, atuar para evitar a injustiça é ainda mais importante. Descobrir muitas irregularidades a troco de prejudicar muitas empresas é imoral.

Capítulo 6

Mitigação de Injustiça

Este capítulo detalha o segundo experimento, no qual é realizada uma avaliação experimental de técnicas de mitigação de injustiça sobre estimativa de risco de contratos públicos e empresas. A estimativa de risco utilizada neste experimento foi a mesma utilizada no experimento de análise de justiça. Os melhores modelos de estimativa de risco foram obtidos através do algoritmo de florestas aleatórias e a melhor técnica de balanceamento foi o *SMOTE*, como descrito na Seção 4.3. Na Seção 6.1 é descrita a metodologia deste experimento de mitigação de injustiças. Na Seção 6.2 são descritos os resultados obtidos através do experimento. Por último, na Seção 6.3 serão discutidos os resultados obtidos.

6.1 Metodologia

Para realizar este experimento utilizamos o *AI Fairness 360 Open Source Toolkit* (AIF360) [2]. Este *toolkit* proporciona alguns algoritmos de mitigação de viés em aprendizagem de máquina do estado da arte. Dentre eles, escolhemos três algoritmos de mitigação: *adversarial debiasing* [25], *calibrated equalized odds* [16] e *disparate impact remover* [9]. A escolha desses três algoritmos se deu por conta de suas diferentes características, de forma que pudessem ser explorados diferentes famílias de algoritmos. O primeiro algoritmo é classificado como in-processing, o segundo de pré-processamento e o último de pós-processamento. O primeiro tem um arcabouço para treinar ao passo que mitiga injustiças, o segundo modifica as variáveis de entrada para que o modelo seja treinado novamente, já o último faz pequenos ajustes na variável de saída, todos com objetivo de mitigar injustiças.

6.1.1 Melhora na justiça

Para avaliar se houve uma melhora da justiça ao aplicar um algoritmo de mitigação, definimos a seguinte fórmula:

$$\mathcal{N}(x) = \begin{cases} x^{-1}, & \text{se } x < 1 \\ x, & \text{caso contrário} \end{cases} \tag{6.1}$$

$$\mathcal{MJ}(disp_i^{TFP}, disp_j^{TFP}) = \frac{\mathcal{N}(disp_i^{TFP}) - \mathcal{N}(disp_j^{TFP})}{\mathcal{N}(disp_i^{TFP}) - 1} \tag{6.2}$$

$\mathcal{MJ}$ mede a melhora de justiça, comparando a disparidade da TFP antes de aplicar o algoritmo de mitigação $disp_i^{TFP}$ e a disparidade depois de aplicar o algoritmo $disp_j^{TFP}$, disparidades essas definidas na Equação 5.4. Como as disparidades são *odds ratio*, utilizamos $\mathcal{N}$ para normalizar a razão. Ou seja, ao aplicar $\mathcal{N}$ na disparidade, ela sempre será maior ou igual à um. Tendo em vista que a melhor disparidade possível é um, onde não existem diferenças entre a TFP de uma classe sensível e a TFP do grupo de referência, $\mathcal{MJ}$ computa quantas vezes diminuiu a disparidade ao aplicar o algoritmo de mitigação.

Quando existe uma disparidade menor que um, isso significa que a disparidade é contrária à esperada. Por exemplo, um valor menor que um significa que o grupo de referência tem TFP maior que uma determinada classe sensível. Assim, ao aplicar $\mathcal{N}$ à disparidade da TFP com valor entre zero e um, tem o mesmo efeito de utilizar o odds ratio do grupo de referência em relação à classe sensível, onde sem aplicar $\mathcal{N}$ seria o odds ratio da classe sensível em relação ao grupo de referência.

6.1.2 Parâmetros e critério

Nenhum dos algoritmos de mitigação permite utilizar todas as características sensíveis conjuntamente, mas apenas uma por vez. Então para cada conjunto de dados e algoritmo de mitigação, variou-se a característica sensível e parâmetros do algoritmo de mitigação, quando esse dispunha de parâmetros. O método *disparate impact remover* é o único que possui um parâmetro, chamado de nível de remoção de injustiça. Desta forma, variou-se esse nível entre 0 e 1 com intervalos de 0,1. Além disso, como o algoritmo *calibrated equalized odds* estima uma probabilidade, colocamos um limiar de 0,5. Desta forma, os valores com pro-

babilidade menor que 0,5 foram estimados como de baixo risco e os demais como de alto risco.

Para definir os melhores hiperparâmetros e características sensíveis considerando tanto eficácia quanto viés, comparamos os resultados dos modelos i e j considerando a média aritmética da diferença na eficácia e da melhora na justiça. Para medir a diferença na eficácia de, usamos a diferença proporcional do f1-score:

$$de(f1_i, f1_j) = \frac{f1_j - f1_i}{f1_i}, \tag{6.3}$$

sendo $f1_i$ o f1-score do modelo i e $f1_j$ o f1-score do modelo j.

Para medir a diferença na justiça considerando as diferentes classes sensíveis, usamos a mediana da melhora na justiça, a qual utiliza a disparidade segundo a TFP. Foram consideradas apenas as classes sensíveis, sem o grupo de referência, pois a melhora na justiça considerando o grupo de referência é sempre igual a zero. Seja $disp_i$ a disparidade na TFP entre uma classe sensível e a de referência no modelo i, e $disp_j$ a disparidade na TFP entre a mesma classe sensível e a de referência no modelo j, a melhora na justiça $\mathcal{MJ}$ é aplicada sobre as disparidades. A mediana das diferenças na justiça $\mathcal{M}$, dado dois modelos i e j e as classes sensíveis $\{c_i, ..., c_n\}$, é calculada da seguinte forma:

$$\mathcal{M}(i,j) = mediana(\mathcal{MJ}(disp_{TFP}^{i,c_1}, disp_{TFP}^{j,c_1}), ..., \mathcal{MJ}(disp_{TFP}^{i,c_n}, disp_{TFP}^{j,c_n}) \tag{6.4}$$

A média comparativa é definida da seguinte forma:

$$mc(result_i, result_j) = \frac{de(f1_i, f1_j) + \mathcal{M}(i,j)}{2} \tag{6.5}$$

Tendo em vista que a $\mathcal{M}(i,j)$ mede a tendência central das melhoras de justiça para as classes sensíveis, considerando taxa de falsos positivos, e que essa medida é uma proporção, foi necessário também colocar a diferença de f1-score em forma de proporção. Assim, essa foi a opção mais viável para encontrar o melhor *trade off* entre eficácia e justiça. Ao comparar o algoritmo de mitigação aplicado a uma característica sensível e com um determinado parâmetro, caso exista, queremos comparar se a medida de justiça melhorou no geral. Como podem existir classes sensíveis que apresentam valores extremos, tanto positivos quanto negativos, optamos por utilizar a mediana de forma a tornar menos sensível a esses valores

extremos. Apesar de querermos um modelo que seja mais justo, também é desejado que ele tenha boa eficácia. Desta forma, utilizamos o f1-score de cada resultado, medindo assim a diferença proporcional. Tendo em vista que tanto a medida de melhora na justiça, quanto a diferença proporcional da eficácia podem assumir valores negativos, não foi possível utilizar a média harmônica. Desta forma, utilizamos a média aritmética entre os dois valores, colocando assim igual importância tanto para a eficácia quanto para a redução de injustiças no processo de mitigação.

Assim como para o primeiro experimento, devido aos conjuntos de dados serem apenas uma amostra das estimativas de risco existentes de interesse, calculamos intervalos de confiança. De forma a medir o efeito sobre a justiça, para cada cenário e algoritmo de mitigação calculamos intervalos de confiança para a melhora da justiça $\mathcal{MJ}$ (6.2) nas diferentes classes sensíveis. Para medir o efeito dos algoritmos sobre a eficácia, calculamos também os intervalos de confiança da diferença proporcional da eficácia de (6.3), mas sem detalhar por classe sensível. Além disso, a mesma ideia foi utilizada para medir a diferença proporcional na precisão e revocação. A diferença proporcional da precisão dp é descrita da seguinte maneira:

$$dp(p_i, p_j) = \frac{p_j - p_i}{p_i}, \tag{6.6}$$

sendo p_i a precisão do modelo i e p_j a precisão do modelo j.

Da mesma forma, a diferença proporcional da revocação dr é descrita da seguinte forma:

$$dr(r_i, r_j) = \frac{r_j - r_i}{r_i}, \tag{6.7}$$

sendo r_i a precisão do modelo i e r_j a precisão do modelo j.

Os intervalos de confiança são estimados com 95% de confiança através de 10.000 bootstraps. Tendo em vista o desbalanceamento das classes sensíveis e da variável resposta, aplicamos bootstrap estratificado em função da conjunção de características sensíveis com a variável resposta. Por fim, interpretamos os resultados através de intervalos de confiança ao invés de testes de hipótese [6].

6.2 Resultados da Mitigação de Injustiça

Nesta seção são apresentados os resultados da mitigação das injustiças. É relevante comentar previamente que alguns dos intervalos de confiança são demasiadamente grande. Isso acontece em virtude da distribuição dos dados, pois as classes sensíveis são uma menor parte deles. Mesmo utilizando o bootstrap estratificado, por vezes a reamostragem não selecionou nenhuma observação que fosse falso positivo, numerador da TFP. Como as métricas de melhora da justiça considera a diferença na disparidade da TFP entre antes de aplicar o algoritmo e depois de aplicá-lo, o intervalo de confiança tem seus resultados bastante afetados.

6.2.1 Empresas na esfera municipal com abordagem ad-hoc dos especialistas

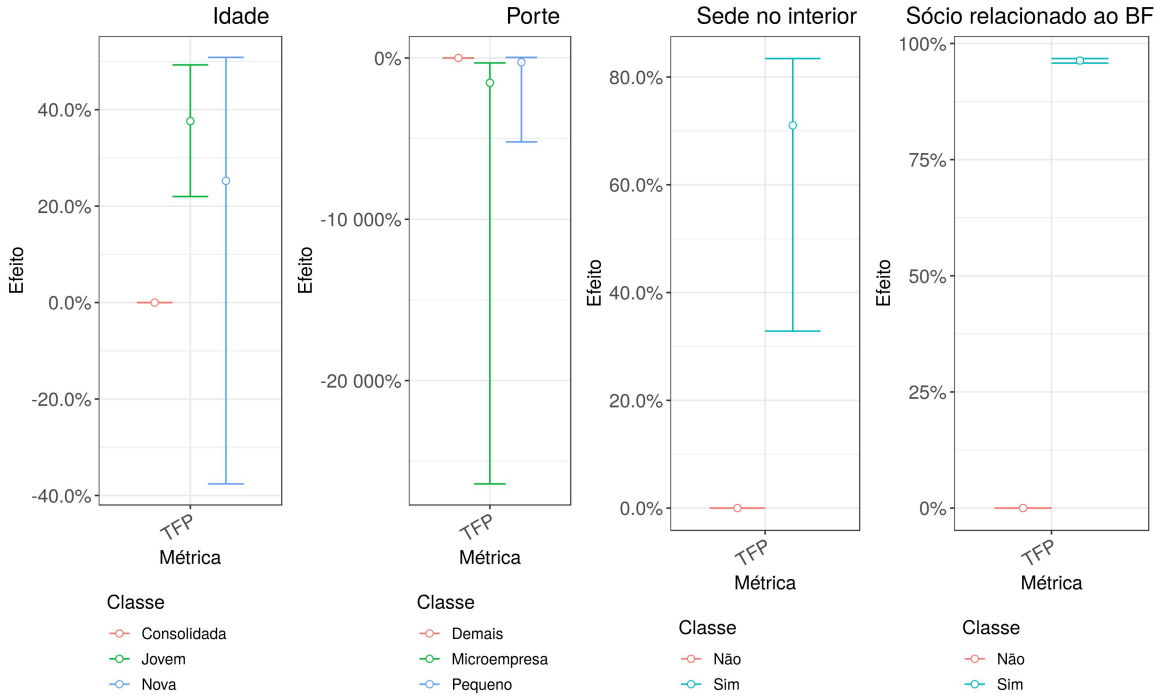

Figura 6.1: Disparidade das métricas após mitigação de injustiças das empresas da esfera municipal (abordagem ad-hoc) através do calibrated equalized odds.

Mitigação através do pós-processamento Calibrated Equalized Odds

Para o conjunto de dados de empresas na esfera municipal, só foi possível utilizar o algoritmo de mitigação *Calibrated Equalized Odds*, pois é um algoritmo de pós-processamento que utiliza apenas os resultados da estimativa de risco e as variáveis sensíveis para mitigar as injustiças. A Figura 6.1 mostra que ao aplicar o algoritmo, empresas do interior tiveram uma melhora na justiça entre 32,9% e 83,4%. Empresas jovens tiveram uma melhora na justiça entre 22% e 49,3%. Já sócios relacionados ao bolsa família tiveram uma melhora na justiça entre 95,8% e 96,7%. Ao contrário do esperado, microempresas tiveram uma piora na justiça entre 307% e 26404%.

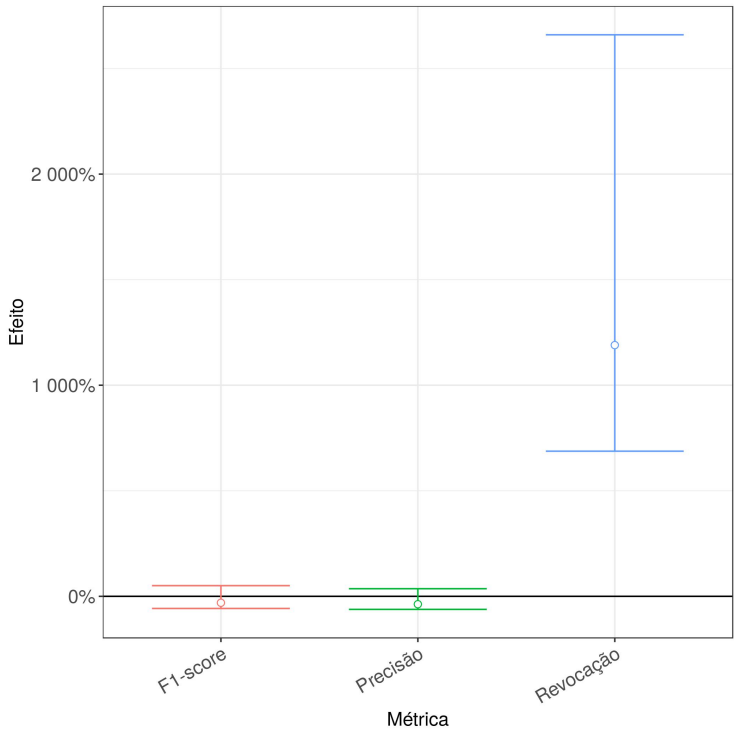

Figura 6.2: Eficácia após mitigação de injustiças das empresas da esfera municipal (abordagem ad-hoc) através do calibrated equalized odds.

Em decorrência dessa melhora evidente para três das quatro características sensíveis, em pelo menos uma das classes sensíveis, a Figura 6.2 mostra que houve uma melhora na eficácia da revocação entre 687,5% e 2660%.

Desta forma, os resultados do algoritmo de mitigação *Calibrated Equalized Odds* para as empresas na esfera municipal com a abordagem ad-hoc dos especialistas teve uma melhora

na justiça em três das quatro características sensíveis, ao passo que melhorou a eficácia do modelo substancialmente. Ou seja, caso este algoritmo de mitigação seja aplicado no cenário de empresas na esfera municipal com abordagem ad-hoc, isso tornará o processo de estimativa de risco mais justo e mais eficaz. A justiça melhorou para empresas sediadas no interior, empresas jovens e empresas relacionadas ao BF. A eficácia melhorou de modo que mais empresas de alto risco fossem detectadas, ao passo que o modelo acerta a mesma proporção de empresas ditas como de alto risco.

6.2.2 Empresas na esfera municipal com aprendizagem de máquina

Mitigação através do in-processing Adversarial Debiasing

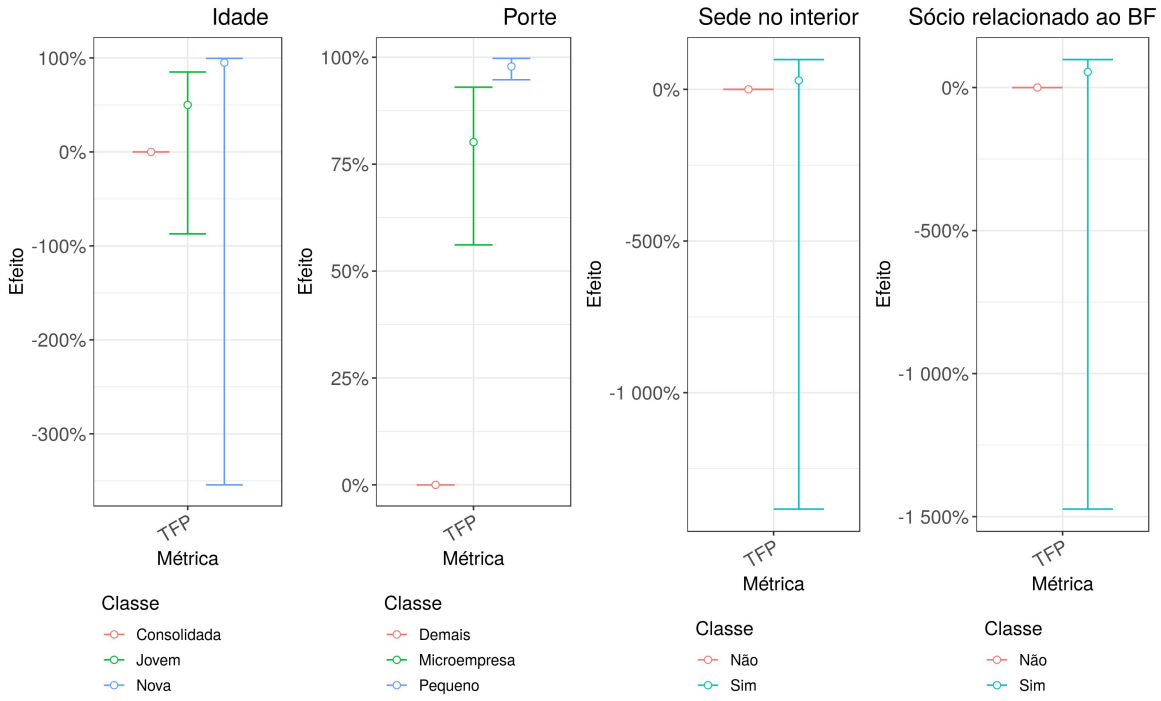

Figura 6.3: Disparidade das métricas após mitigação de injustiças das empresas da esfera municipal (abordagem AM) através do adversarial debiasing.

O algoritmo *Adversarial Debiasing* obteve uma melhora na justiça no porte da empresa, tanto em microempresas, com uma melhora na justiça entre 56,1% e 93%, quanto em empresas pequenas, com uma melhora entre 94,7% e 99,7%, como pode ser visto na Figura 6.3.

Ao observar o efeito na eficácia, observamos na Figura 6.4 que houve uma melhora na

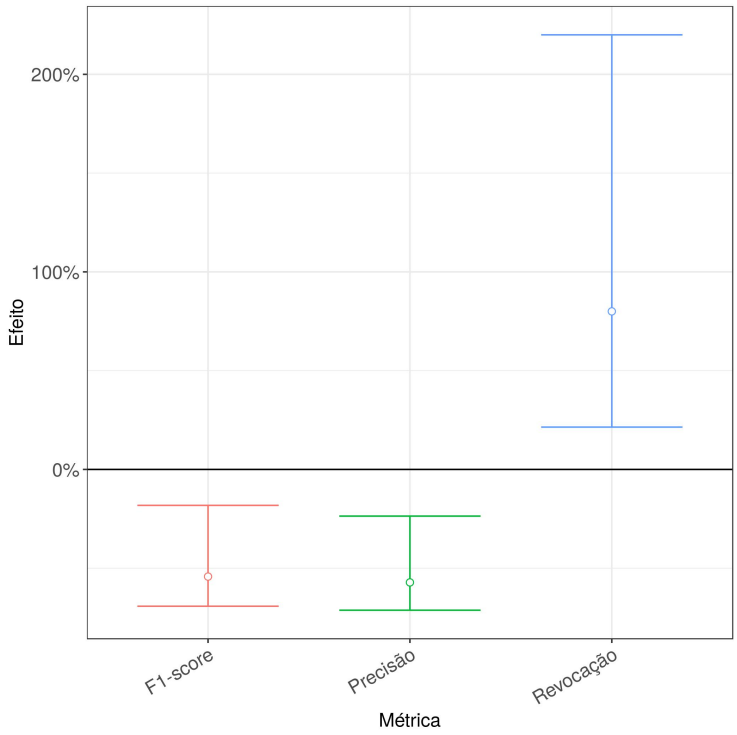

Figura 6.4: Eficácia após mitigação de injustiças das empresas da esfera municipal (abordagem AM) através do adversarial debiasing.

revocação entre 21,4% e 220%, porém houve uma piora na precisão entre 23,6% e 71,3%, que acarretou em uma piora no f1-score entre 18,2% e 69,2%.

Juntos, os resultados mostram que existe um grande *trade-off* no uso do algoritmo de mitigação *Adversarial Debiasing*. Apesar de obter uma melhora na justiça para todas as classes sensíveis de uma característica sensível e uma melhora na eficácia da revocação, o algoritmo acarretou em uma piora tanto da precisão, quanto do f1-score. Isso significa que ao aplicar esse algoritmo de mitigação no cenário de empresas na esfera municipal com aprendizagem de máquina, isso tornará o processo mais justo para microempresas e pequenas empresas. Porém, apesar do modelo detectar mais empresas que são de fato alto risco, ele erra mais das que estima como de alto risco.

Mitigação através do pós-processamento Calibrated Equalized Odds

O algoritmo *Calibrated Equalized Odds* aplicado ao conjunto de dados de empresas da esfera municipal, em uma abordagem de aprendizagem de máquina não proporcionou nenhuma modificação. Com isso, não houve melhoras ou pioras tanto da justiça, quanto da eficácia,

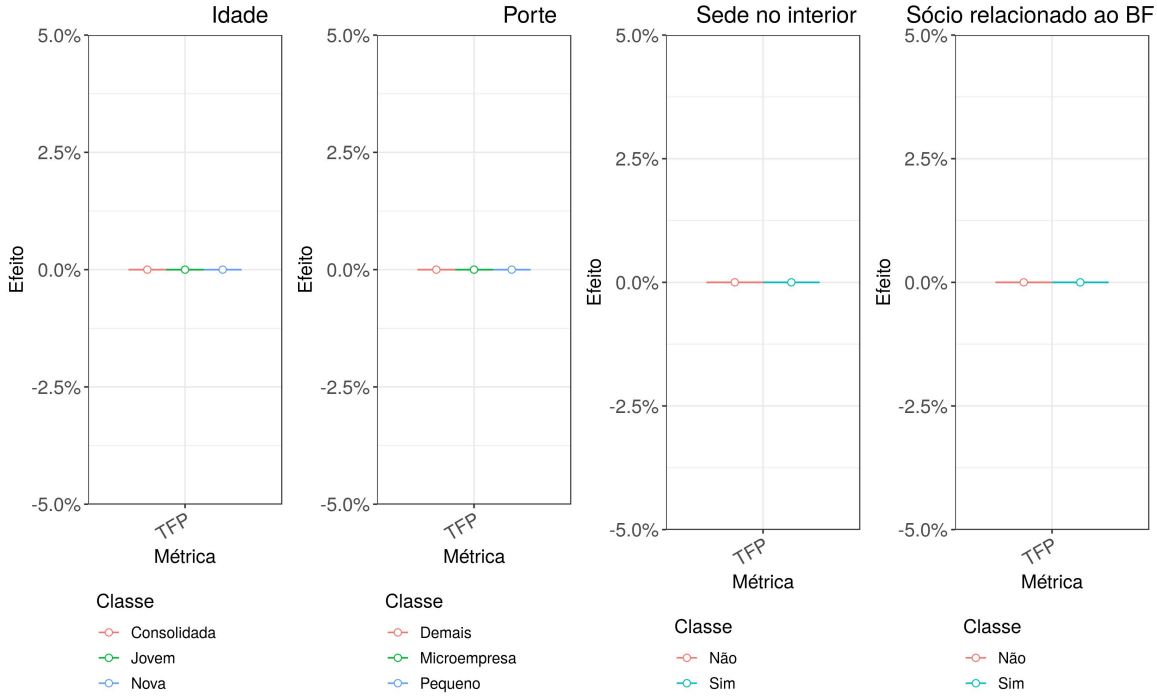

Figura 6.5: Disparidade das métricas após mitigação de injustiças das empresas da esfera municipal (abordagem AM) através do calibrated equalized odds.

como pode ser visto nas Figuras 6.5 e 6.6.

Mitigação através do pré-processamento Disparate Impact Remover

O algoritmo *Disparate Impact Remover* aplicado sobre o conjunto de dados de empresas da esfera municipal obteve uma melhora na justiça das microempresas entre 55,5% e 99,6%, como pode ser visto na Figura 6.7. Além disso, obteve uma melhora tanto na precisão, entre 54% e 251,4%, quanto no f1-score, entre 35,6% e 204,5%, como pode ser visto na Figura 6.8.

Isso significa que este algoritmo de pré-processamento aplicado sobre o cenário de empresas na esfera municipal com aprendizagem de máquina reduz as injustiças contra microempresas. Além disso, a melhora na justiça é seguida de uma melhora na eficácia, de forma que o modelo acerta mais empresas estimadas como de alto risco.

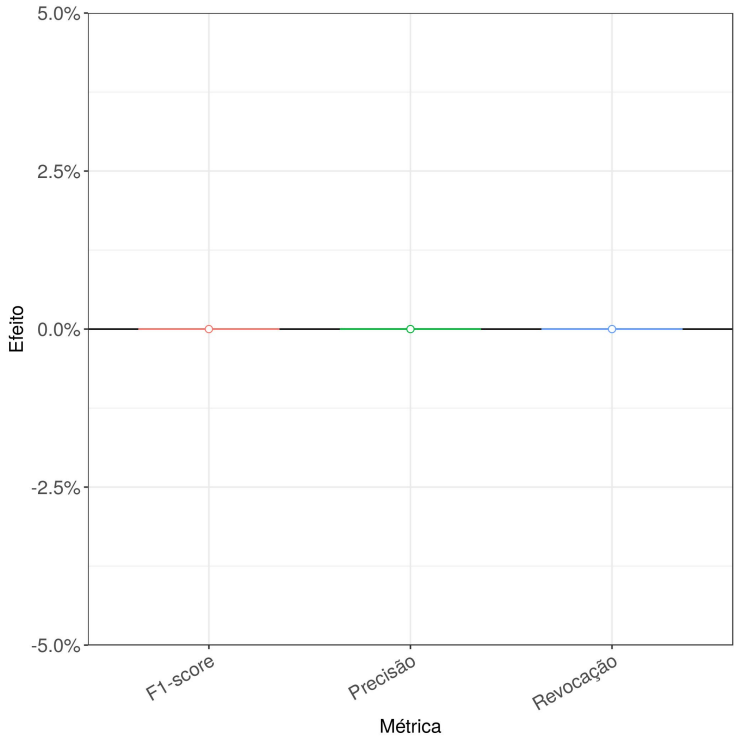

Figura 6.6: Eficácia após mitigação de injustiças das empresas da esfera municipal (abordagem AM) através do calibrated equalized odds.

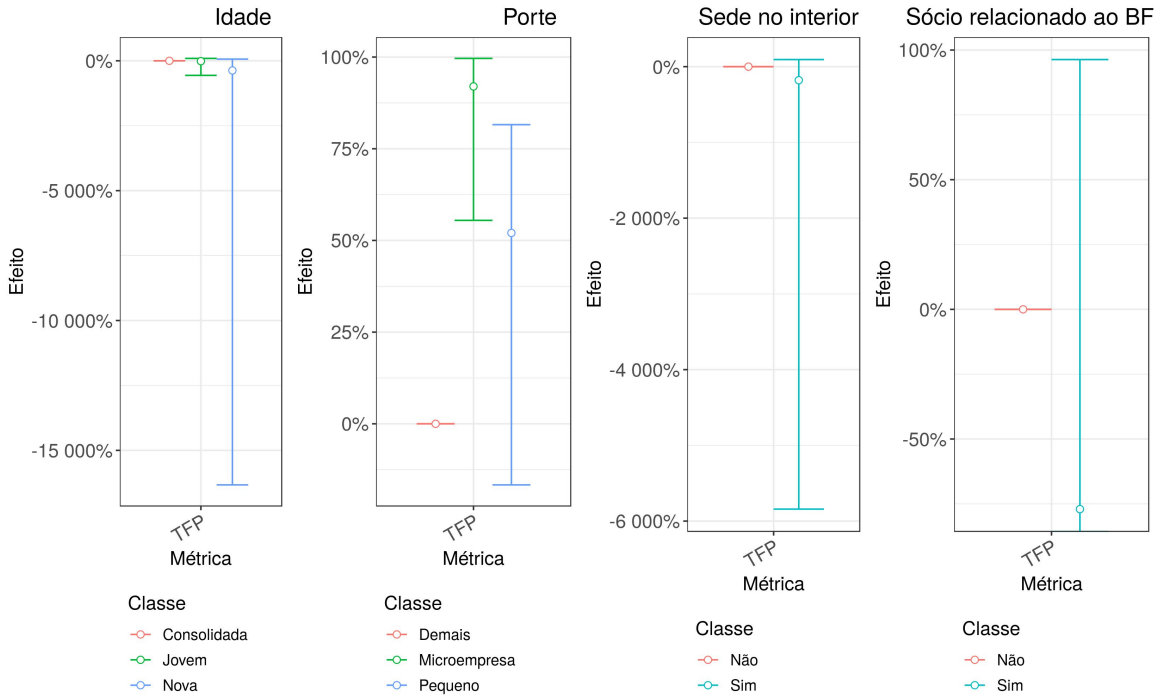

Figura 6.7: Disparidade das métricas após mitigação de injustiças das empresas da esfera municipal (abordagem AM) através do disparate impact remover.

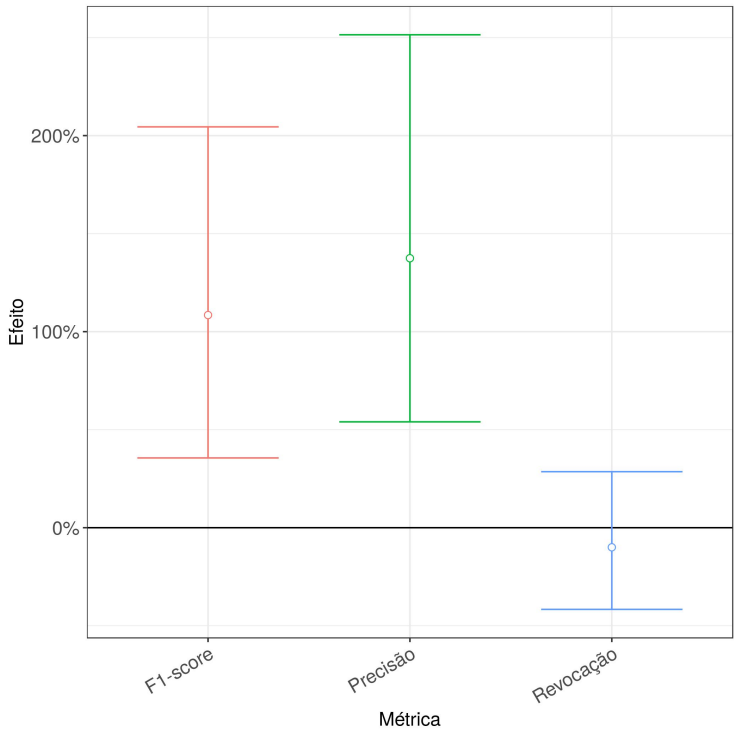

Figura 6.8: Eficácia após mitigação de injustiças das empresas da esfera municipal (abordagem AM) através do disparate impact remover.

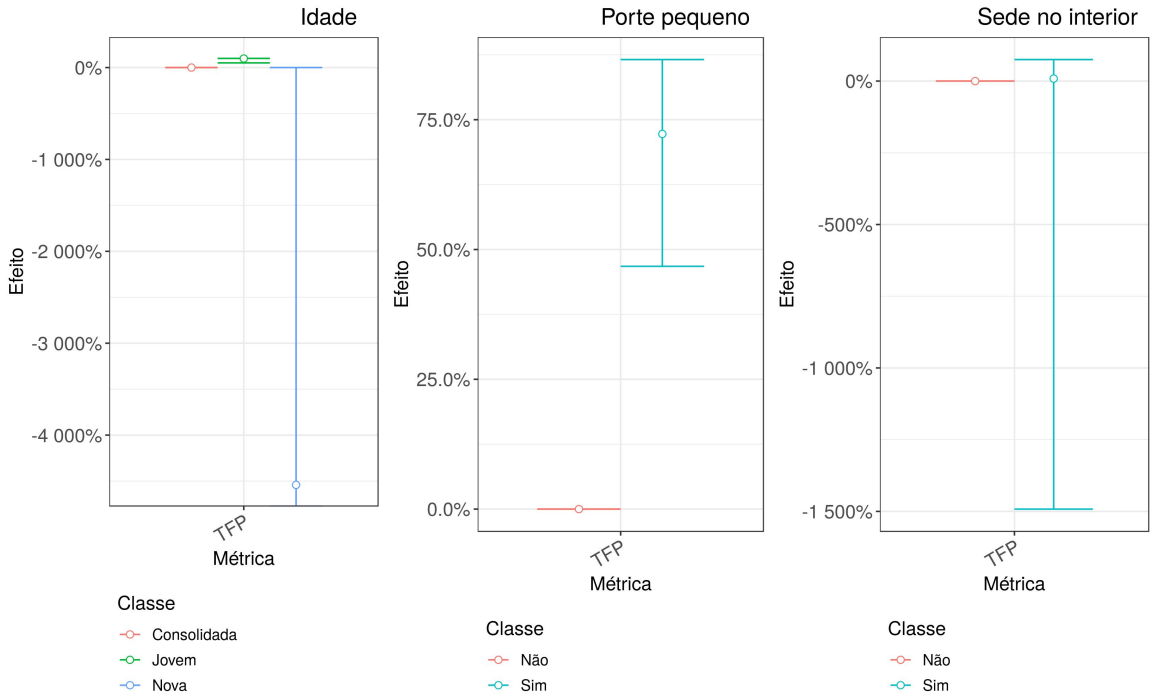

Figura 6.9: Disparidade das métricas após mitigação de injustiças das empresas da esfera federal através do adversarial debiasing.

6.2.3 Empresas na esfera federal com aprendizagem de máquina

Mitigação através do in-processing Adversarial Debiasing

Para o conjunto de dados de empresas na esfera federal, utilizando o algoritmo de mitigação *Adversarial Debiasing*, os resultados, presentes na Figura 6.9 mostram que houve uma melhora na justiça de empresas de pequeno porte, entre 46,8% e 86,6%. Já para a idade da empresa houve uma melhora da justiça entre 50,7% e 99,6%.

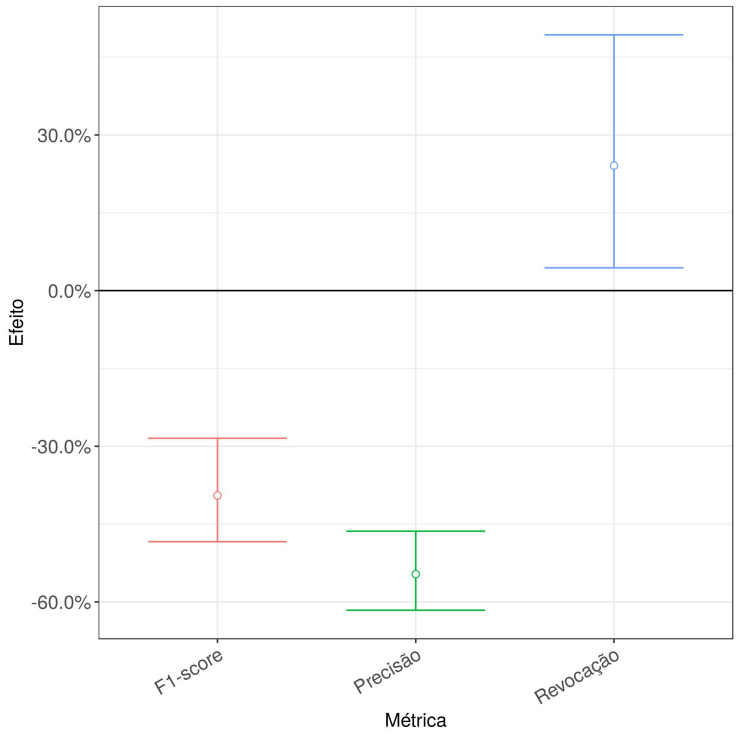

Figura 6.10: Eficácia após mitigação de injustiças das empresas da esfera federal através do adversarial debiasing.

Ao observar a Figura 6.10, é possível identificar que existe um grande *trade-off* entre a revocação, e a precisão e f1-score. A revocação obteve uma melhora entre 4,4% e 49,3%, enquanto a precisão obteve uma piora entre 46,4% e 61,6%, e o f1-score obteve uma piora entre 28,5% e 48,4%.

Este algoritmo aplicado a este conjunto de dados possui um grande *trade-off*. Apesar de pelo menos uma classe sensível de duas das três categorias sensíveis obter uma melhora na justiça e haver uma melhora na precisão, houve também uma piora tanto na precisão, quanto no f1-score. Ou seja, os resultados apontam que caso os órgãos de controle apliquem esse

método de mitigação de injustiça no cenário de empresas da esfera federal, uma melhora na justiça deve ser observada para empresas de pequeno porte e empresas jovens. Porém, apesar deste método de mitigação proporcionar que mais empresas de alto risco sejam detectadas, ele também erra mais das estimadas como de alto risco.

Mitigação através do pós-processamento Calibrated Equalized Odds

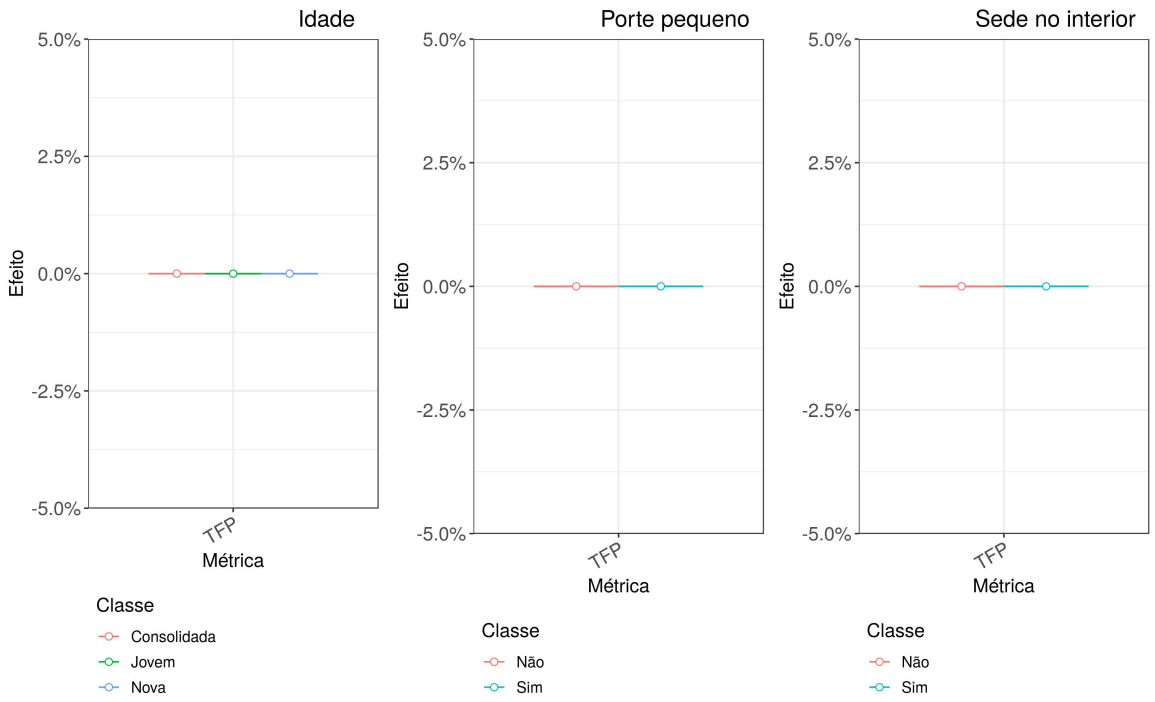

Figura 6.11: Disparidade das métricas após mitigação de injustiças das empresas da esfera federal através do calibrated equalized odds.

Como pode ser visto tanto na Figura 6.11, quanto na Figura 6.12, o algoritmo *Calibrated Equalized Odds* não produziu efeitos sobre o conjunto de dados de empresas da esfera federal. Logo, não houveram melhoras ou pioras, tanto na justiça, quanto na eficácia.

Mitigação através do pré-processamento Disparate Impact Remover

O algoritmo *Diparate Impact Remover*, aplicado sobre o conjunto de dados de empresas da esfera federal, apresentou um resultado curioso. Ele não produziu melhoras na justiça, mas produziu uma piora para empresas jovens, entre 9,4% e 156,9%, como pode ser visto na

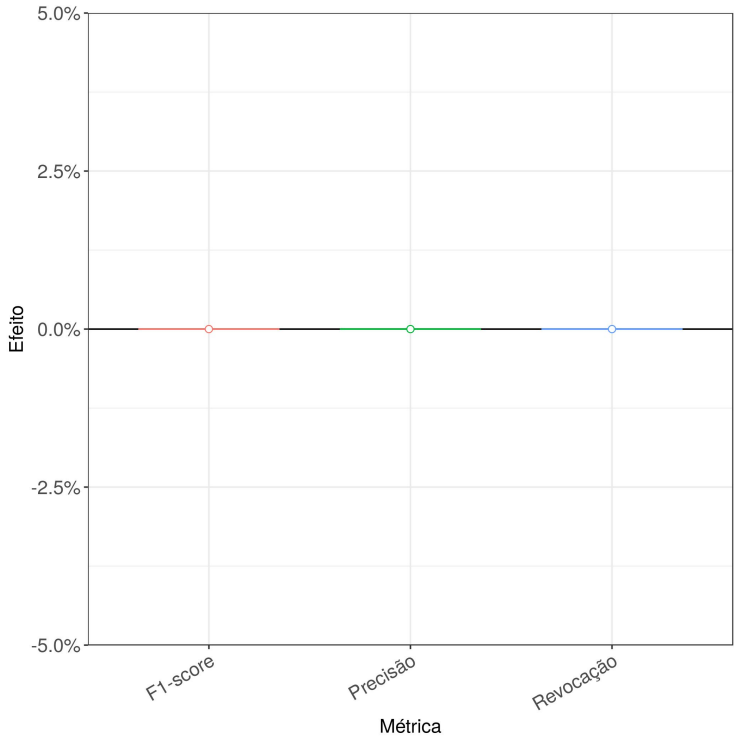

Figura 6.12: Eficácia após mitigação de injustiças das empresas da esfera federal através do calibrated equalized odds.

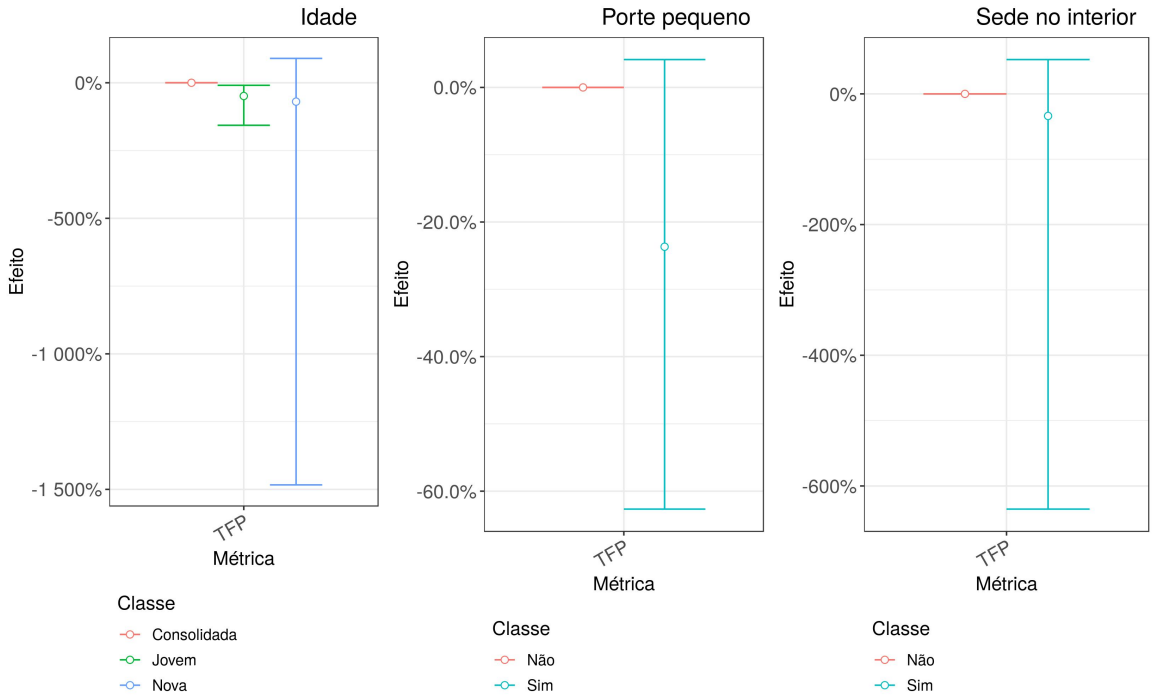

Figura 6.13: Disparidade das métricas após mitigação de injustiças das empresas da esfera federal através do disparate impact remover.

Figura 6.13. Apesar disso, produziu uma melhora na precisão, entre 1,3% e 16,6%, como pode ser visto na Figura 6.14.

Desta forma, o algoritmo de pré-processamento aplicado ao cenário de empresas da esfera federal não produziram melhoras na justiça, mas produziram uma piora na justiça para empresas jovens.

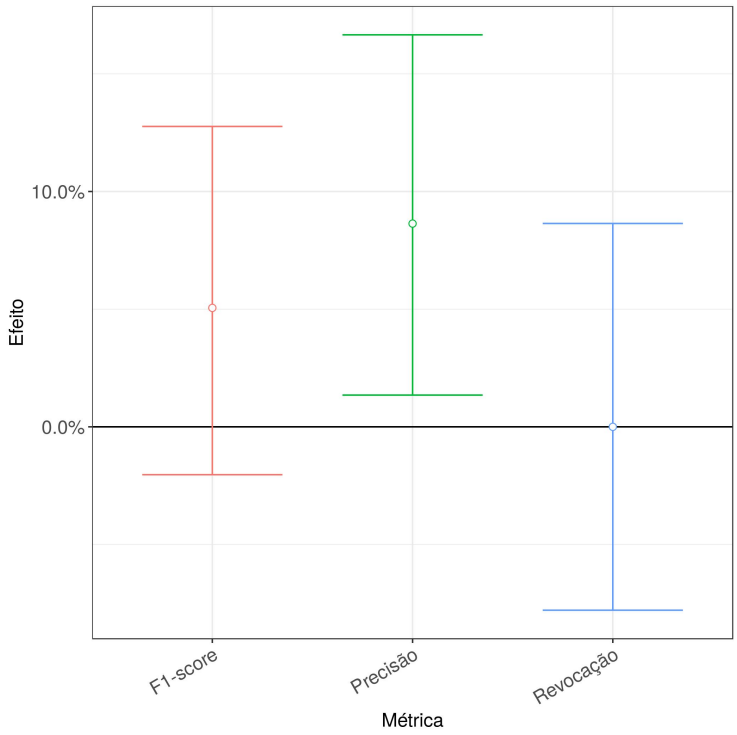

Figura 6.14: Eficácia após mitigação de injustiças das empresas da esfera federal através do disparate impact remover.

6.2.4 Contratos na esfera municipal com aprendizagem de máquina

Mitigação através do in-processing Adversarial Debiasing

A mitigação através do algoritmo *Adversarial Debiasing* aplicada sobre a base de dados de contratos na esfera municipal pode ser observada na Figura 6.15. Nessa figura é possível identificar que empresas jovens tiveram uma melhora na justiça entre 93,5% e 99,5%.

Ao observar a melhora na eficácia, presente na Figura 6.16, é possível observar que existe uma grande *trade-off*. A revocação apresentou uma melhora entre 25% e 212,5%, porém a

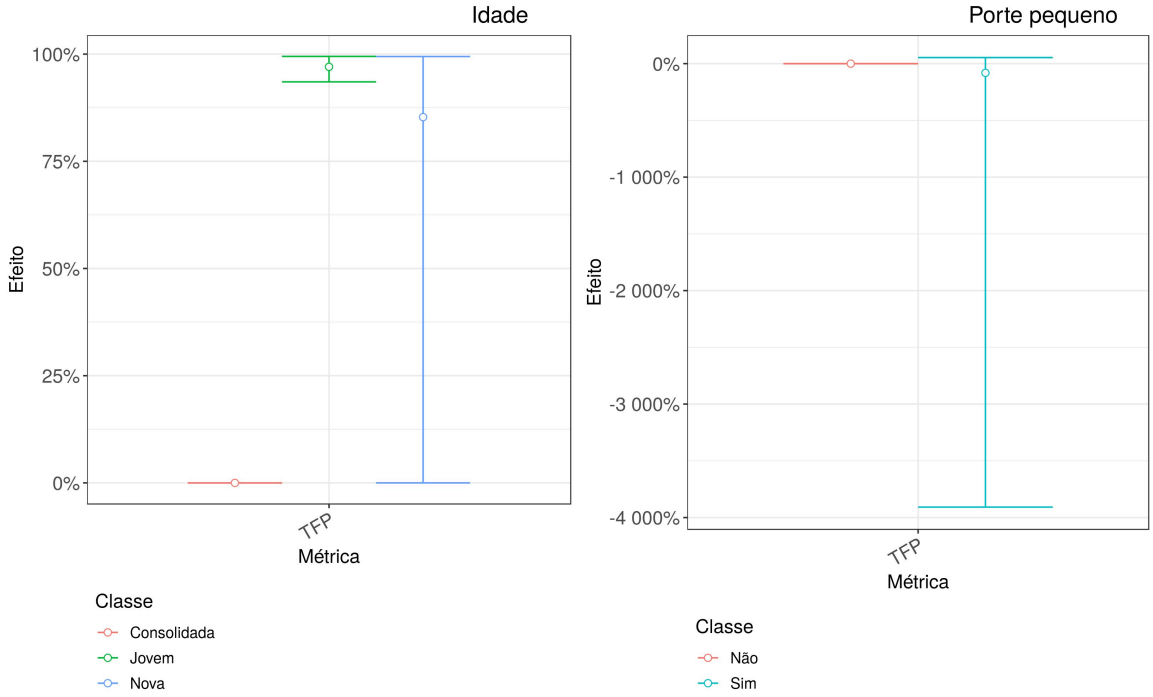

Figura 6.15: Disparidade das métricas após mitigação de injustiças dos contratos da esfera municipal através do adversarial debiasing.

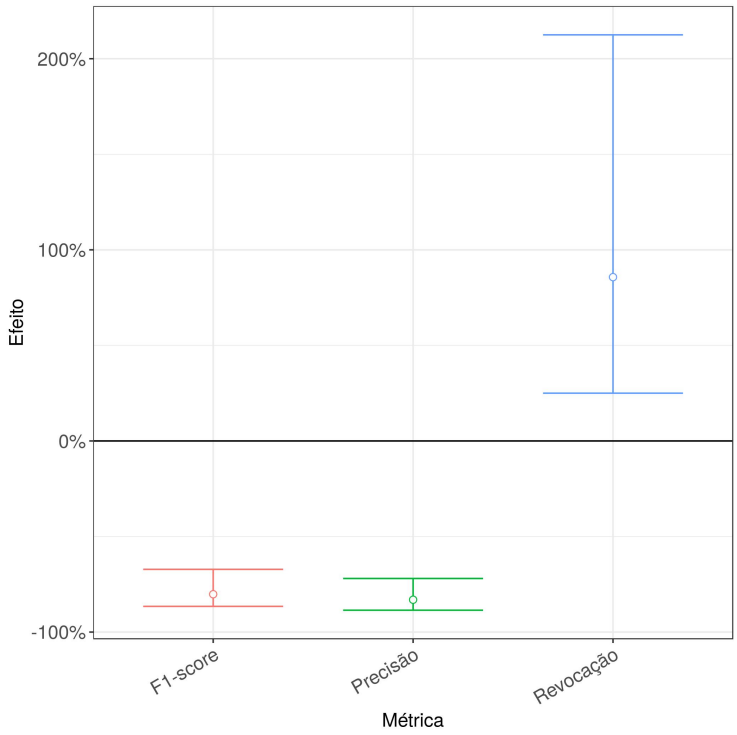

Figura 6.16: Eficácia após mitigação de injustiças dos contratos da esfera municipal através do adversarial debiasing.

precisão apresentou uma piora entre 72% e 88,5%, enquanto o f1-score apresentou uma piora entre 67,2% e 86,6%.

Os resultados juntos indicam um grande trade off. Apesar de apresentar uma melhora muito grande na justiça em uma das classes sensíveis seguido de uma melhora na revocação, houve também uma grande piora na precisão e f1-score. Ou seja, os resultados apontam que, caso órgãos de controle apliquem esse método de mitigação, haverá uma melhora na justiça para empresas jovens. Porém, apesar do método proporcionar que mais empresas de alto risco sejam detectadas, ele também acarreta que o modelo tenda a errar mais a proporção de empresas estimadas como de alto risco.

Mitigação através do pós-processamento Calibrated Equalized Odds

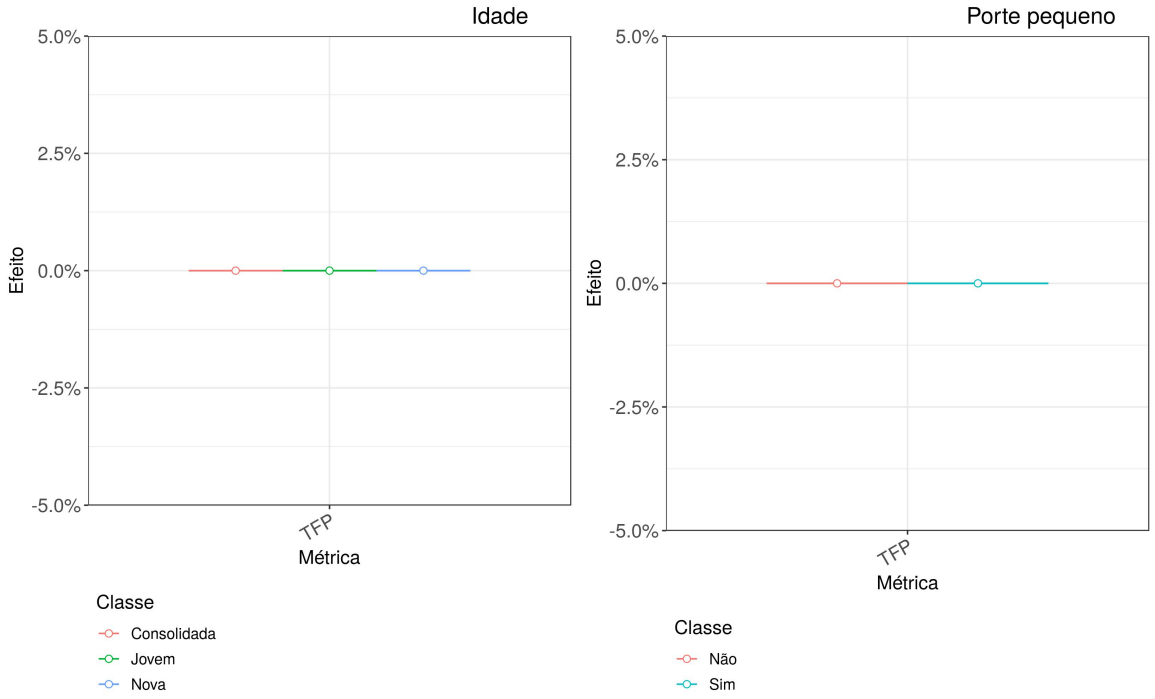

Figura 6.17: Disparidade das métricas após mitigação de injustiças dos contratos da esfera municipal através do calibrated equalized odds.

O algoritmo *Calibrated Equalized Odds* novamente não proporcionou nenhuma mudança nos resultados, visto nas Figuras 6.17 e 6.18. Logo, não houveram melhoras ou pioras na justiça ou eficácia.

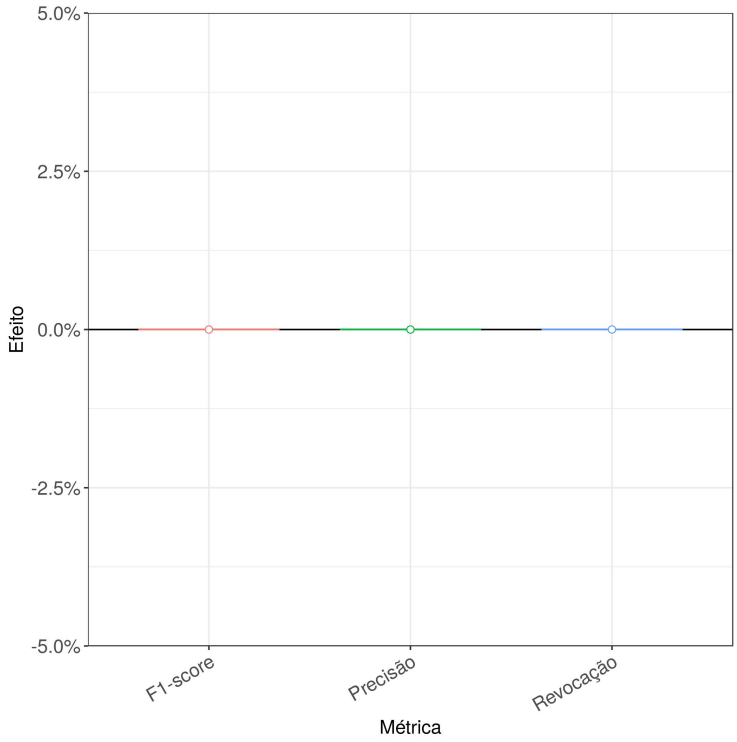

Figura 6.18: Eficácia após mitigação de injustiças dos contratos da esfera municipal através do calibrated equalized odds.

Mitigação através do pré-processamento Disparate Impact Remover

Por último, o algoritmo *Disparate Impact Remover* apesar de ter produzido resultados diferentes dos resultados sem aplicar o algoritmo de mitigação, não houve diferenças visíveis na justiça, como pode ser visto na Figura 6.19. Apesar disso, pode não haver nenhuma mudança na revocação, mas pode haver uma melhora de até 27,3%, como visto na Figura 6.20, onde mais empresas de alto risco seriam detectadas.

6.3 Discussão

Nosso experimento mostra que ao aplicar os modelos de mitigação em diferentes contexto, é possível encontrar uma melhora na justiça em pelo menos uma classe sensível dos dados para pelo menos um algoritmo de mitigação. Além disso, exceto no contexto de empresas da esfera municipal via aprendizagem de máquina, os algoritmos de mitigação melhoraram a justiça para empresas jovens. A injustiça contra empresas jovens já havia sido detectada

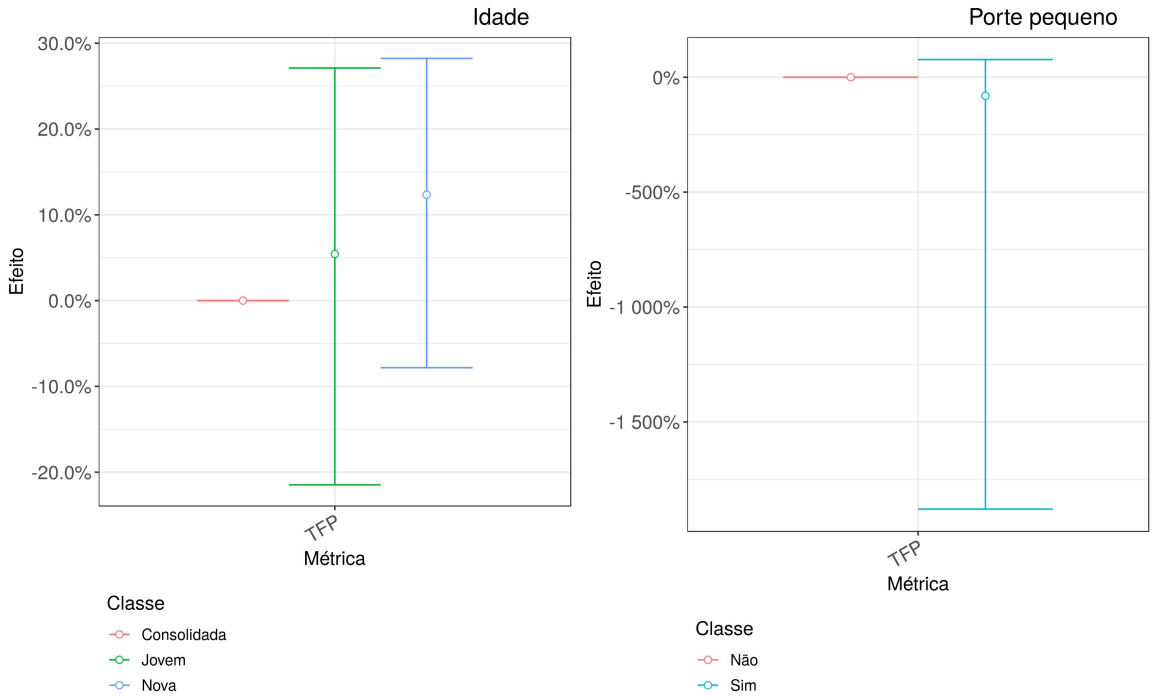

Figura 6.19: Disparidade das métricas após mitigação de injustiças dos contratos da esfera municipal através do disparate impact remover.

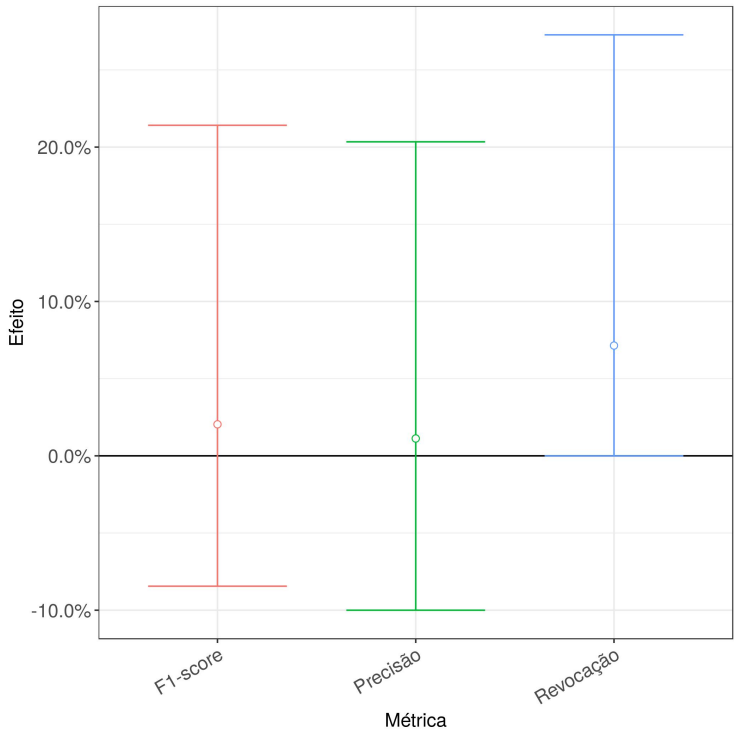

Figura 6.20: Eficácia após mitigação de injustiças dos contratos da esfera municipal através do disparate impact remover.

Tabela 6.1: Melhores algoritmos de mitigação para cada cenário

Cenário	**Melhor algoritmo de mitigação**	**Porcentagem de características sensíveis que melhoraram**	**Porcentagem de classes sensíveis que melhoraram**	**Houve apenas melhoras na eficácia?**
Empresas na esfera municipal com abordagem ad-hoc	Calibrated Equalized Odds	75%	50%	Sim
Empresas na esfera municipal via aprendizagem de máquina	Disparate Impact Remover	25%	16,7%	Sim
Empresas na esfera federal	Adversarial Debiasing	66,7%	50%	Não
Contratos na esfera municipal	Adversarial Debiasing	50%	33,3%	Não

anteriormente no experimento de análise de justiça, detalhado no Capítulo 5. Além disso, em dois desses contextos houve uma grande melhora também na eficácia, enquanto nos dois restantes houve um *trade-off* entre justiça e eficácia.

O melhor resultado de mitigação foi obtido a partir do contexto de empresas da esfera municipal com abordagem ad-hoc dos especialistas ao aplicar o algoritmo *Calibrated Equalized Odds*. Houve uma melhora na justiça em três das quatro características sensíveis, enquanto na restante houve uma piora. Esse modelo proporcionou além de uma considerável melhora na justiça, uma grande melhora no recall. Ou seja, além de melhorar a justiça para empresas sediadas no interior, empresas jovens e empresas relacionadas ao BF, o modelo de mitigação de pós-processamento proporcionou uma detecção mais eficaz das empresas de alto risco.

Apesar de obter uma melhora em duas classes sensíveis através do *Adversarial Debiasing*, consideramos que no contexto de empresas na esfera municipal com abordagem via aprendizagem de máquina o algoritmo que apresentou os melhores resultados foi o *Disparate Impact Remover*. Apesar de obter uma melhora em apenas uma classe sensível dos

dados, houve também uma melhora considerável para a eficácia. Ou seja, o algoritmo de pós-processamento *Disparate Impact Remover* proporcionou uma melhora na justiça para microempresas, ao passo que também acertou proporcionalmente mais empresas estimadas como alto risco.

No contexto de empresas na esfera federal, consideramos que o algoritmo que obteve melhores resultados foi o *Adversarial Debiasing*. Houve uma melhora na justiça em duas das três características sensíveis, em pelo menos uma classe sensível. Além desta melhora substancial na justiça, houve uma melhora também na revocação. Porém, além disso, houve também uma piora considerável no f1-score e precisão. Ou seja, o método de in-processing proporcionou uma melhora na justiça tanto para empresas de pequeno porte quanto para empresas jovens. Além disso, detectou mais empresas de alto risco, porém acertou proporcionalmente menos empresas das estimadas como de alto risco.

No contexto de contratos na esfera municipal, consideramos que o algoritmo de mitigação que apresentou melhores resultados foi o *Adversarial Debiasing*. A melhora na justiça ocorreu em apenas uma classe sensível - empresas jovens -, onde haviam duas características sensíveis. Além disso, houve também uma melhora considerável no recall, porém uma piora considerável no precisão e f1-score. Ou seja, apesar do modelo ter proporcionado uma melhor detecção das empresas de alto risco, ele também acarretou em mais erros na proporção das empresas estimadas como de alto risco.

Um fato curioso é que o algoritmo *Calibrated Equalized Odds* não proporcionou mudanças nos resultados, exceto no contexto de empresas na esfera municipal com abordagem ad-hoc. Provavelmente esse contexto foi o que mais teve melhoras na justiça porque era o contexto mais tendencioso, como visto no Capítulo 5.

Para o contexto de empresas na esfera municipal, seja na abordagem ad-hoc ou via aprendizagem de máquina, além de apresentar melhoras na justiça, houveram grandes melhoras na eficácia dos modelos. Nos demais contextos houveram *trade-offs* entre a justiça e a eficácia dos modelos. Cada contexto apresentou um algoritmo que melhor se encaixou, ou seja, não houve uma concordância de qual o melhor algoritmo de mitigação. Isso significa que os resultados apontam que caso os órgãos de controle apliquem modelos de mitigação de injustiças na estimativa de risco haverá uma melhora na justiça para pelo menos uma classe sensível. Além disso, metade dos cenários apontam para uma melhora da eficácia dos mode-

los seguidos da melhora na justiça, ou seja, os modelos proporcionam um melhor acerto das empresas estimadas como de alto risco ou detectam mais empresas de alto risco.

Identificamos que o algoritmo que mais melhor performou foi o Calibrated Equalized odds, tendo em vista que melhorou a justiça para 75% das características. Porém foi um dos piores algoritmos em termos de abrangência, tendo em vista que em 3 dos quatro cenários não produziu diferenças na justiça. Identificamos também que o melhor algoritmo com relação à eficácia é o Disparate Impact Remover, pois em todos os cenários em que pôde ser aplicado permaneceu com a mesma eficácia ou melhorou. Porém, ao analisar a melhora da justiça, este algoritmo não performou muito bem, pois apresentou melhoras apenas em 1 dos 3 cenários em que o algoritmo pôde ser utilizado. Por fim, identificamos que o algoritmo que melhorou a justiça em maior quantidade de cenários foi o Adversarial Debiasing, pois melhorou a justiça para pelo menos uma classe sensível em todos os cenários que pôde ser aplicado. Porém essa melhora da justiça acarretou em um grande trade-off na eficácia, pois, apesar de ter melhorado a revocação em todos os cenários aplicados, também piorou o F1-score e precisão.

Capítulo 7

Conclusões

7.1 Discussão

O objetivo deste trabalho foi avaliar a justiça em aprendizagem de máquina para estimativa de risco e mitigação de injustiças no cenário de investigação de gastos públicos, mais especificamente na estimativa de risco de contratos públicos e empresas. Até então nenhuma de avaliação da justiça neste cenário foi encontrada, logo percebemos a necessidade de tratar deste assunto. Desta forma, esse trabalho contribui tanto para o cenário de justiça em aprendizagem de máquina, como também inicia uma análise da justiça sobre a investigação de gastos públicos.

Nossas análises mostram que a estimativa de risco de contratos públicos e empresas é injusta contra pelo menos uma classe sensível em todos os cenários analisados. Os resultados indicam que empresas jovens são mais falsamente acusadas quando comparadas com empresas consolidadas em todos os cenários, além de outras classes sensíveis também serem em contextos específicos. Uma estimativa de risco que seja enviesada contra empresas jovens pode desencorajar a participação no processo licitatório mesmo que elas tenham plena capacidade de execução do objeto da licitação.

As disparidades foram observadas tanto em estimativas de risco em uma abordagem ad-hoc quanto via aprendizagem de máquina. Elas existem tanto no âmbito federal quanto no municipal, e tanto no nível de contrato quanto no nível de empresa. Além disso, tanto as características criadas por especialistas em controle quanto por pesquisadores em aprendizagem de máquina levam a resultados semelhantes. Isso mostra que esse é um cenário em

que a avaliação da justiça na estimativa de risco por parte dos órgãos de controle é de grande importância.

Ao analisar os resultados percebemos que as disparidades que indicam injustiças são seguidas, por vezes, da ausência de disparidade de eficácia. Quando esse caso ocorre o modelo está sendo totalmente injusto, pois ele acusa falsamente empresas como arriscadas, sem nenhum ganho com isso. Isso pode levar a auditores investigarem sem necessidade contratos ou empresas, tendo em vista que a estimativa de risco acusa falsamente como de alto risco essas classes sensíveis, havendo assim um desperdício dos recursos públicos. Já quando existe tanto a disparidade que indica injustiça quanto a de eficácia temos um caso crítico, pois descobrir mais irregularidades sem se importar se está prejudicando empresas é imoral. Empresas jovens ainda estão se estabelecendo no mercado, mas mesmo assim os órgãos de controle desconfiam excessivamente delas. Caso o auditor venha acusar falsamente a empresa tendenciado pela estimativa de risco, ela pode ser sobrecarregada financeiramente para provar sua inocência. Isso pode prejudicar também a máquina pública, por exemplo: caso uma construtora jovem esteja prestando serviço para o estado e uma falsa acusação prejudique financeiramente a empresa, pode acarretar em falta de recursos para completar a obra. Desta forma pode acarretar em um atraso da obra, ou até mesmo em uma paralisação da mesma.

Nosso experimento de mitigação de injustiças mostra que ao aplicar modelos com este propósito é possível encontrar uma melhora na justiça em pelo menos uma classe sensível dos dados para pelo menos um algoritmo de mitigação em diferentes contextos. Além disso, os algoritmos puderam mitigar injustiças contra empresas jovens, visto anteriormente, em três dos quatro cenários abordados.

Os algoritmos de mitigação que apresentaram melhores resultados variaram em função do contexto. Em dois dos quatro contextos houve uma melhora na eficácia dos modelos seguido da melhora na justiça. Nos dois cenários restantes houve um grande *trade-off* entre a melhora na eficácia e na justiça. Apesar de obter uma melhora na revocação e melhorar a justiça para pelo menos uma classe sensível de pelo menos metade das características sensíveis, houve uma piora considerável na precisão e f1-score. A melhora na justiça pode ajudar empresas que estão tentando se estabelecer no mercado, tendo em vista que a taxa de erro da estimativa de risco não será tão desigual quando comparadas classes sensíveis com

o grupo de referência. Empresas jovens, por exemplo, serão menos falsamente acusadas como de alto risco caso esses modelos de mitigação sejam aplicados nos órgãos de controle. A melhora na eficácia ajudará os órgãos de controle a investigar empresas que têm maior chance de rescindir os contratos, reduzindo assim os recursos públicos que são utilizados para investigar empresas que têm menor chance de rescindir contratos.

Como o objetivo do trabalho é analisar injustiças e mitigá-las, concluímos que o melhor algoritmo de mitigação é o Adversarial Debiasing, tendo em vista que melhorou a justiça de pelo menos uma classe sensível em todos os cenários aplicados. Porém essa melhora da justiça acarretou em um grande trade-off de eficácia, tendo em vista que o algoritmo proporciona uma melhora na revocação, mas uma piora no F1-score e precisão. Caso o requisito principal seja permanecer ou melhorar a eficácia ao passo que tenta melhorar a justiça, concluímos que o melhor algoritmo é o Disparate Impact Remover. Esse algoritmo permaneceu com a mesma eficácia ou melhorou em todos os cenários estudados, ao passo que apresentou melhora na justiça em um terço dos cenários aplicados. Por último, concluímos que o Calibrated Equalized Odds pode ser uma boa opção caso o cenário tenha uma injustiça mais evidente, tendo em vista que esse algoritmo apresentou melhoras na justiça e eficácia no cenário com maior quantidade de características sensíveis injustiçadas.

Desta forma, a análise de justiça aponta que o cenário de estimativa de risco para fiscalização de contratos públicos é tendenciosa. Além disso, nossos resultados mostram uso de algoritmos de mitigação de injustiças funcionam bem em certos cenários, sendo essa uma abordagem promissora para serem utilizadas por órgãos de controle em futuras investigações de contratos públicos.

7.2 Limitações

Este estudo inicia a análise e debate da justiça na estimativa de risco para gastos públicos. Como as características sensíveis identificadas são contingente da estrutura dos dados, as características utilizadas nesse trabalho não são as únicas existentes para todos os cenários de investigação de contratos. Além disso, assumimos que as características sensíveis dos dados são relacionados ao tamanho da empresa, ou seja, entendemos que as empresas que possuem características sensíveis são mais frágeis. Uma outra investigação sobre o mesmo

cenário pode utilizar outras características sensíveis.

Para a realização deste trabalho assumimos que a estimativa de risco é realizada de modo semelhante ao descrito por Sun e Sales [22] ou de maneira ad-hoc, através de ponderações de características da empresa. Caso um órgão de controle não realize a estimativa de risco através destes dois modos, a avaliação de justiça, bem como a mitigação de injustiças, pode ter de ser feita de maneira diferente da proposta. Além disso, assumimos que a melhor métrica para avaliar a justiça neste cenário é a disparidade da TFP. Caso outra interpretação da justiça seja avaliada para o cenário, outras conclusões poderiam ser tomadas, tendo em vista que outra métrica de avaliação de justiça seria utilizada.

Nos dados abordados, as características sensíveis são uma pequena porção dos dados. Desta forma, mesmo ao realizar o bootstrap estratificado, não foi possível encontrar intervalos de confiança pequenos. Isso acarretou em incertezas, pois houveram alguns casos que não se pode afirmar se existe ou não injustiça para o experimento de análise de justiça. Da mesma forma, para o experimento de mitigação de injustiças, houveram casos que não se pode afirmar e se existe melhora, piora ou mesmo se não existe diferença ao aplicar determinado método.

Por último, este cenário tem uma característica particular, pois o rótulo para um contrato ser considerada de alto risco é quando um contrato é rescindido. Desta forma, apenas contratos que foram investigados que podem ser considerado de alto risco. Ou seja, ainda que um contrato esteja rotulado como de baixo risco, existe a possibilidade de na verdade ele ser de alto risco, caso uma investigação tivesse sido realizada.

7.3 Trabalhos Futuros

Trabalhos futuros podem investigar métodos para mitigar a injustiça no uso de AM com as características sensíveis que identificamos, bem como outras características identificadas como sensíveis contingente do novo conjunto de dados. Novos experimentos com dados de outros estados e outros órgãos de controle também são necessários para avaliar quão generalizáveis são os nossos resultados. Experimentos com bases maiores, incluindo todo o Brasil também melhorariam a precisão dos intervalos de confiança que estimamos. Por último, seria interessante utilizar novas técnicas de mitigação de injustiças para experimentar

novas abordagens.

Bibliografia

[1] Lei 8666. In *Constituição Federal*, Junho 1993.

[2] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018.

[3] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. 2017.

[4] Matheus Carvalho. *Manual de Direito Administrativo*. 2017.

[5] Rommel Carvalho, Leonardo Sales, Henrique Rocha, and Gilson Mendes. Using bayesian networks to identify and prevent split purchases in brazil. volume 1218, 01 2014.

[6] Geoff Cumming and Robert Calin-Jageman. *Introduction to the New Statistics: Estimation, Open Science, and Beyond.* Routledge, New York, NY, 10001, 2016.

[7] Victor Aguiar Jardim de Amorim. *Licitações e contratos administrativos: teoria e jurisprudência*. 2017.

[8] Marcelo Alexandrino e Vicente Paulo. *Direito Administrativo Descomplicado*. 2017.

[9] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA, 2015. Association for Computing Machinery.

[10] J. Galindo and P. Tamayo. Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. volume 15, pages 107–143, Apr 2000.

[11] T. Alencar Gomes, R. Novaes Carvalho, and R. Silva Carvalho. Identifying anomalies in parliamentary expenditures of brazilian chamber of deputies with deep autoencoders. In *IEEE ICMLA*, 2017.

[12] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. 2016.

[13] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. 2016.

[14] Hossein Mojaddadi, Biswajeet Pradhan, Haleh Nampak, Noordin Ahmad, and Abdul Halim bin Ghazali. Ensemble machine-learning-based geospatial approach for flood risk assessment using multi-sensor remote-sensing data and gis. *Geomatics, Natural Hazards and Risk*, 8(2):1080–1102, 2017.

[15] Jerzy Neyman. *On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection*. Springer New York, New York, NY, 1992.

[16] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5680–5689. Curran Associates, Inc., 2017.

[17] Célia Ghedini Ralha and Carlos Vinícius Sarmento Silva. A multi-agent data mining system for cartel detection in brazilian government procurement. volume 39, pages 11642 – 11656, 2012.

[18] Pedro Saleiro, Abby Stevens Benedict Kuester, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. In *eprint arXiv:1811.05577*, 2018.

[19] Itiel E.Dror Saul M. Kassin and Jeff Kukucka. The forensic confirmation bias: Problems, perspectives, and proposed solutions. In *Journal of Applied Research in Memory and Cognition*, volume 2, pages 42–52, 2013.

[20] Ricardo S. Carvalho Silvio L. Domingos, Rommel N. Carvalho and Guilherme N. Ramos. Identifying it purchases anomalies in the brazilian government procurement system using deep learning. In *15th IEEE ICMLA*, pages 722–727, 2016.

[21] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual group unfairness via inequality indices. In *ACM KDD '18*, 2018.

[22] Ting Sun and Leonardo J. Sales. Predicting public procurement irregularity: An application of neural networks. In *Journal of Emerging Technologies in Accounting: Spring 2018*, volume 15, pages 141–154, 2018.

[23] Jenna Reps Joe Kai Jonathan M. Garibaldi Weng, Stephen F. and Nadeem Qureshi. Can machine-learning improve cardiovascular risk prediction using routine clinical data? volume 12, Apr 2017.

[24] Chia-Jung Hsu Wun-Hwa Chen Soushan Wu Zan Huang, Hsinchun Chen. Credit rating analysis with support vector machines and neural networks: a market comparative study. In *Decision Support System*, volume 37, pages 543–558, 2004.

[25] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340. Association for Computing Machinery, 2018.

[26] Órion Darshan Winter De Lima and Nazareno Andrade. Fairness in risk estimation of brazilian public contracts. In *Anais do VII Symposium on Knowledge Discovery, Mining and Learning*, pages 57–64. SBC, 2019.