Instituto Superior
de Contabilidade
e Administração

Politécnico de Coimbra

Núbia Stein Kuhn

Técnicas de aprimoramento de equidade em aprendizado de máquina: Uma revisão sistemática de literatura

Coimbra, agosto de 2022

Técnicas de aprimoramento de equidade em aprendizado de máquina: Uma revisão sistemática de literatura
Núbia Stein Kuhn
ISCAC | 2022

Instituto Superior
de Contabilidade
e Administração
Politécnico de Coimbra
COIMBRA BUSINESS SCHOOL
ISCAC.pt

Núbia Stein Kuhn

Técnicas de aprimoramento de equidade em aprendizado de máquina: Uma revisão sistemática de literatura

Dissertação submetida ao Instituto Superior de Contabilidade e Administração de Coimbra para cumprimento dos requisitos necessários à obtenção do grau de **Mestre em Análise de Dados e Sistemas de Apoio a Decisão** realizada sob a orientação do Professor Doutor António Rui Trigo Ribeiro e coorientação do Professor Doutor Fernando Paulo dos Santos Rodrigues Belfo.

Coimbra, agosto de 2022

TERMO DE RESPONSABILIDADE

Declaro ser a autora desta dissertação, que constitui um trabalho original e inédito, que nunca foi submetido a outra Instituição de ensino superior para obtenção de um grau académico ou outra habilitação. Atesto ainda que todas as citações estão devidamente identificadas e que tenho consciência de que o plágio constitui uma grave falta de ética, que poderá resultar na anulação da presente dissertação.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

PENSAMENTO

"O microscópio nos mostrou que há mais em uma gota de água de um lago do que vemos. O telescópio nos mostrou que existe mais no céu noturno do que pensamos que vemos. E os novos dados digitais, agora, nos mostram que existe mais na sociedade humana do que pensamos ver. Eles podem ser o microscópio ou o telescópio de nossa era, possibilitando percepções importantes e até revolucionárias" (Davidowitz, 2018).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

AGRADECIMENTOS

Agradeço à Deus, por me conceder sabedoria e força de vontade, quando achei que não fosse capaz de aprender tantos conceitos nunca ouvidos, ou quando duvidei que poderia escrever uma dissertação seguindo rigorosas regras acadêmicas, para finalizar um curso de tamanha relevância para minha carreira, em um país diferente do qual cresci, e em um período de muitas notícias tristes, em que o confinamento nos obrigou a desenvolver interações humanas via tela de computador.

Agradeço a minha família que com toda a simplicidade, nunca deixaram de me apoiar e de acreditar em mim.

Agradeço ao meu namorado e aos meus amigos que foram minha rede de apoio e de companheirismo.

E agradeço aos professores doutores que me orientaram e conduziram nesta trajetória para realização desta pesquisa, sempre com calma, profissionalismo, atenção e gentileza.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

RESUMO

Decisões baseadas em algoritmos de aprendizado de máquina podem reproduzir tendências ou preconceitos presentes em dados históricos enviesados, oferecendo graves consequências sociais a grupos de indivíduos mal representados.

Tendo como motivação principal poder contribuir para combater os preconceitos e vieses em aprendizado de máquina, esta dissertação tem como questão de investigação central a identificação de técnicas para o aprimoramento da equidade em aprendizado de máquina, através da realização de uma revisão sistemática de literatura sobre este tema.

Com base no método PRISMA, apresentamos a revisão sistemática de literatura efetuada, a qual identificou 15 estudos recentes, com propostas e abordagens técnicas focadas em aprimorar a equidade em sistemas algorítmicos. Os trabalhos selecionados foram publicados em inglês nos últimos cinco anos e foram classificados e discutidos de acordo com as métricas de equidade e desempenho, os estágios de intervenção (pré-processamento, processamento ou pós-processamento), os conjuntos de dados e algoritmos utilizados nos experimentos.

Foram identificadas 48 técnicas para aprimorar a equidade em aprendizado de máquina, as quais se apresentam neste trabalho. Verificou-se que a maioria das publicações procura a equidade através da métrica de paridade demográfica, e no estágio de pré-processamento dos dados. Uma parte significativa das referidas técnicas foram desenvolvidas através de algoritmos de regressão logística e de *Random Forest*. Relativamente aos *datasets* utilizados, o *UCI Adult* e COMPAS são largamente explorados em experimentações sobre o tema, o que, se por um lado, permite uma maior uniformização quanto à interpretação dos resultados, por outro lado, pode representar uma limitação quanto aos grupos envolvidos e as variáveis exploradas. A identificação de padrões sobre experimentações em equidade algorítmica, tais como em relação a métrica, técnicas e *datasets* utilizados, contribui com uma visão mais unificada que ajuda a democratizar e a promover o tema.

Palavras Chaves: equidade, técnicas de aprimoramento da equidade, métricas de equidade, aprendizado de máquina, revisão sistemática de literatura, PRISMA.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

ABSTRACT

Decisions based on machine learning algorithms may reproduce biases or prejudices present in biased historical data, offering serious social consequences to groups of misrepresented individuals.

Having as its main motivation to be able to combat prejudices and biases in machine learning, this dissertation has as its central investigation question, the identification of techniques for the improvement of fairness in machine learning, through a systematic literature review on this topic.

Based on the PRISMA method, we presented the systematic literature review carried out, which identified 15 recent studies with proposals and technical approaches focused on improving fairness in algorithmic systems. The selected works were published in English in the last five years and were classified and discussed according to the, the fairness and performance metrics, the intervention stages (pre-processing, processing, or post-processing), the datasets and algorithms used in the experiments.

A total of 48 techniques were identified to improve fairness in machine learning, which are presented in this work. It was found that most researchers look for fairness through demographic parity metric and on the pre-processing stage. A significant part of these techniques was developed through Logistic Regression and Random Forest algorithms. Regarding the datasets used, the UCI Adult and COMPAS are widely explored in experiments on the subjection the topic, which in one hand, allows for greater uniformity in terms of interpretation of results, on the other hand, it may represent a limitation in relation to the groups involved and the variables explored. The identification of patterns identification on experiments in algorithmic fairness, such as in relation to metrics, techniques and datasets used, contribute to a more unified view that helps democratize and promote the theme.

Keywords: fairness, fairness enhancement techniques, fairness metrics, machine learning, systematic literature review, PRISMA.

ÍNDICE GERAL

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

ÍNDICE DE TABELAS

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

ÍNDICE DE FIGURAS

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Lista de abreviaturas, acrónimos e siglas

ACM .....*Association for Computing Machinery*

ASR.....*Adaptive Sensitive Reweighting*

CNN .....*Convolutional Neural Network*

COMPAS .....*Correctional Offender Management Profiling for Alternatives Sanctions*

CULEP .....*Convex Underlying Label Error Perturbation*

DFGR .....*Distributed Fair Logistic Regression*

DFKRR .....*Distributed Fair Kernel Ridge Regression*

DFPCA.....*Distributed Fair Principal Component Analysis*

DFRR .....*Distributed Fair Ridge Regression*

DI .....*Disparate Impact*

ETL .....*Extract, Transform and Load*

FAcct.....*Fairness Accountability and Transparency*

FAT/ML .....*Fairness, Accountability and Transparency in Machine Learning*

FATIT .....*Fairness Transparency and Identification data usage traffic*

FN .....Falsos Negativos

FP .....Falsos Positivos

FRM .....*Fair Recommendation Matrix*

HTML .....*Hyper Text Markup Language*

IA .....Inteligência Artificial

IEEE .....*Institute of Electrical and Electronics Engineers*

IOTs .....*Internet of things*

KNN.....*K Nearest Neighbors*

LR .....*Logistic Regression*

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

MAIRST.....*Minimize AI bias applying Random Sampling Technique*

MATLAB .....*Matrix Laboratory*

ML.....*Machine Learning*

MNIST .....*Modified National Institute of Standards and Technology*

NN.....*Neural Network*

NSGA II .....*Non dominated sorting genetic algorithm II*

PRISMA.....*Preferred Reporting Items for Systematic Reviews and Meta Analyses*

PTFA.....*Privacy Transparency Fairness Agreement*

RF.....*Random Forest*

RGPD.....Regulamento Geral de Proteção de Dados

RIS .....*Research Information Systems*

RISKSLIM ....*Risk-calibrated Super Sparse Linear Integer Model*

RMSE.....*Root Means Square Error*

ROC .....*Receiver Operating Characteristic Curve*

ROUGE .....*Recall-Oriented Understudy for Gisting Evaluation*

RSL .....Revisão Sistemática de Literatura

SAT .....*Scholastic Aptitude Test*

SLIM .....*Super Sparse Linear Integer Model*

STEM.....*Science Technology Engineering and Mathematics*

SVM.....*Support Vector Machines*

TFD .....Taxa de Falso Descobrimento

TFN.....Taxa de Falsos Negativos

TFP.....Taxa de Falsos Positivos

TFO .....Taxa de Falsa Omissão

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

TVN .....Taxa de Verdadeiros Negativos

TVP .....Taxa de Verdadeiros Positivos

UCI.....*University of California Irvine*

UE .....União Europeia

VN.....Verdadeiros Negativos

VP .....Verdadeiros Positivos

VPN .....Valor Predito Negativo

VPP .....Valor Predito Positivo

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

1 INTRODUÇÃO

O crescente volume de dados disponíveis combinado a processamento e armazenamento computacional mais poderosos e acessíveis, impulsionaram a utilização de sistemas baseados em aprendizado de máquina, subcampo da ciência da computação, e grande pilar da tecnologia da informação, cada dia mais consolidado e difundido em todo mundo.

O aprendizado de máquina, utiliza os grandes dados, para identificar os padrões presentes ali, e então aprender comportamentos futuros, permitindo o escalonamento e a aceleração do processo de tomada de decisões, a automação de tarefas, a identificação de fraudes financeiras, o monitoramento de safras, e a previsão de comportamentos humanos.

Esta tecnologia provocou novas e significativas mudanças na forma como vivemos, trabalhamos e interagimos (Belfo et al., 2022), ajudando as organizações a definir estratégias que lhes permitem aumentar o seu desempenho, tanto em setores públicos, tais como na área fiscal (Seiça et al., 2019), na área educacional (Pimenta et al., 2018) e na área médica (Brandão et al., 2021; Pimenta et al., 2022), quanto no setor privado, na indústria de mídia e entretenimento (Sereday & Cui, 2017), na indústria de eventos (Loureiro et al., 2014), no turismo (Esteves et al.,2021; Pimenta et al., 2011) e em muitas outras áreas.

Porém, esta mesma tecnologia, também pode estimular desejos, manipular preços e necessidades, catalisar discursos de ódio nas redes sociais e interferir no curso de processos democráticos, impactando a vida de milhares de pessoas, moldando oportunidades e personalizando a forma como enxergamos e compreendemos o mundo de maneira silenciosa e invisível.

Como elucida Cathy O´Neil (2020), em seu livro “Algoritmos de destruição em massa”, somos somados e organizados de todas as formas possíveis, formando uma salada de dados com nossos códigos postais, histórico de navegação na internet, compras recentes e conexões em redes sociais. Essa salada, alimenta sentenças algorítmicas, que quando mal concebidas, podem criar ciclos que se retroalimentam e perpetuam preconceitos antigos como racismo, sexismo e discriminação de classe (O’Neil, 2020).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Grandes empresas de tecnologia já estiverem envolvidas em polêmicas de discriminação algorítmica, como a Amazon, quando fez uso de um sistema de recrutamento que penalizava currículos de mulheres para cargos de alto escalão, e o Google, quando o seu serviço de compartilhamento e armazenamento de fotos, Google Fotos, identificou pessoas negras como gorilas (Pessach & Shmueli, 2020).

O preconceito contra pessoas de pele escura em particular, parece ser comumente reproduzido. Neste sentido, Cardenas e Vallejo (2019) mencionam alguns exemplos curiosos, como pedestres negros com menor probabilidade de identificação por carros autônomos, e consequentemente maior propensão a possíveis acidentes, falhas em sensores que dispensam sabão, funcionando com maior dificuldade em peles negras, e sistemas de reconhecimento facial instalados em lojas de varejo para detecção de ladrões, que identificam pretos como potenciais agressores em maior frequência do que brancos.

Provavelmente o caso mais famoso sobre o tema seja o COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*), sistema americano que realizava avaliações de risco de reincidência em crimes, no qual observou-se que negros eram duas vezes mais inclinados a reincidir ao crime do que brancos, levando assim muitas pessoas negras a receberem penas injustas (Mehrabi et al., 2021).

1.1 Justificativa

O crescimento acelerado do emprego de algoritmos em contextos sociais e econômicos, e da consequente preocupação pública sobre os impactos da tecnologia digital na coletividade, levaram à necessidade de pesquisas específicas sobre equidade, responsabilidade e transparência (Barocas et al., 2019) formando um sub campo específico em *machine learning* (ML).

Neste sentido, importantes conferências debatem especificamente o tema e reúnem pesquisadores académicos, indústrias, e praticantes interessados em explorar maneiras de construir sistemas mais justos transparentes e éticos, como por exemplo a *Conference on Fairness, Accountability and Transparency* (FAcct) da *Association of Computing Machinery* (ACM) (ACM, 2022), a *Fairness, Accountability and Transparency in ML* (FAT/ML, 2022), e a *FairWare´22: International Workshop on Equitable Data and*

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Technology (*Fairware 22*, 2022). Além destas, revistas científicas e conferências em áreas como comunicação, política, direito, saúde e sociologia demostram interesse pelo tema e publicam artigos sobre a equidade algorítmica e seus impactos.

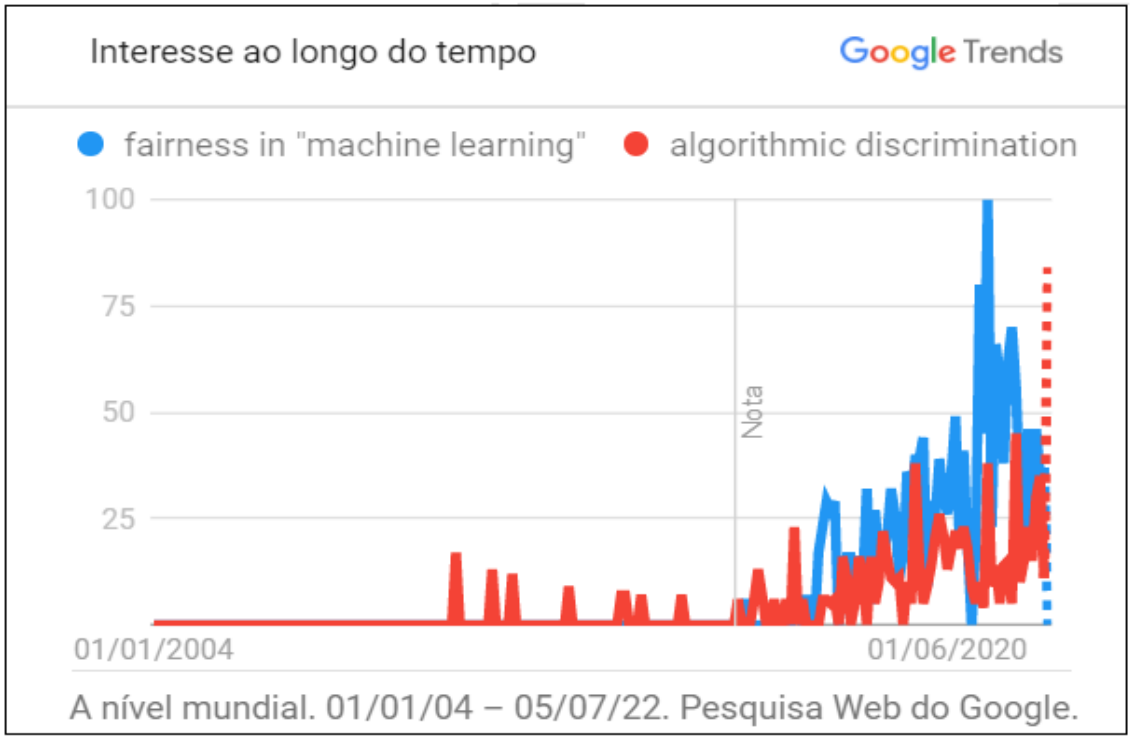

Figura 1.1. Interesse pelo tema ao longo dos últimos anos (01/01/2004-05/07/2022)
Fonte: Google Trends (2022).

Uma simples pesquisa das expressões em inglês *"fairness in machine learning"* e *"algorithmic discrimination"* no google trends, evidencia também o crescimento do interesse popular pelo tema nos últimos anos, como mostra a Figura 1.1 (Google, 2022).

1.2 Objetivo

Dada a relevância do tema, seu impacto social, e o crescimento do interesse acadêmico e popular, constatamos que além de contextualizar a discriminação em sistemas algorítmicos e explicar as terminologias inerentes a ela, seria importante também realizar uma pesquisa objetiva e clara, capaz de identificar soluções apresentadas na literatura e desenvolvidas com rigor científico que ofereçam opções técnicas capazes de melhorar a equidade em sistemas de aprendizado de máquina.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

O objetivo desta dissertação é selecionar e analisar, por meio de revisão sistemática de literatura (RSL), artigos científicos que contenham propostas técnicas de aprimoramento de equidade em aprendizado de máquina, descrevendo-as por meio de síntese narrativa, sumarizando-as em forma de tabela e classificando-as de acordo com suas particularidades, procurando identificar possíveis padrões comuns e discuti-los, buscando assim facilitar pesquisas futuras.

1.3 Estrutura do relatório

Essa dissertação compreende seis capítulos. Para além deste capítulo de introdução, contemplamos ainda o enquadramento teórico, a metodologia, os resultados, a discussão e pôr fim a conclusão.

No segundo capítulo, enquadramento teórico, abordamos os conceitos necessários para a compreensão do tema, como possíveis explicações sobre como a discriminação acontece em sistemas algorítmicos, as métricas de equidade mais conhecidas, as fases de atuação no processo de desenvolvimento dos sistemas, e alguns exemplos iniciais de técnicas já propostas por importantes autores, além da introdução sobre perspectivas mais amplas, que ultrapassam a atuação da ciência da computação e interagem com outras áreas da sociedade, a que denominamos aqui de abordagens sociotécnicas.

No capítulo de metodologia apresentamos as motivações e os detalhes da RSL, como as fontes de informação, os critérios de inclusão e de exclusão das publicações, e o fluxograma seguido nesta revisão, baseado nas fases que compreendem o método *Preferred Reporting Items for Systematic Reviews and Meta Analyses* (PRISMA).

No capítulo de resultados, descrevemos a proposta de cada publicação selecionada, de forma narrativa, resumimos as 15 publicações elegíveis em forma de tabela, de acordo com o objetivo, as métricas de equidade e desempenho, os estágios de intervenção (pré-processamento, processamento ou pós-processamento), os conjuntos de dados e algoritmos utilizados nos experimentos, seguido então do capítulo de discussão dos resultados, onde verificamos quais destas características foram mais utilizadas, procurando suporte na literatura para tal preferência.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Por fim, no capítulo da conclusão apresenta-se uma síntese do trabalho efetuado, contribuições para a indústria e academia, algumas limitações sentidas durante o desenvolvimento da pesquisa e sugestões de trabalhos futuros.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2 ENQUADRAMENTO TEÓRICO

Muita coisa mudou desde a publicação do artigo "*Computing machinery and intelligence*" de Alan Turing em 1950, considerado o marco do surgimento da inteligência artificial (IA), (Medeiros, 2019) até hoje. Somente em 2021, com a era do *Big Data* e da Internet das coisas, do inglês, *Internet of things* (IOTs), foram gerados cerca de 79 zetabbytes de dados, com projeção de 180 zetabbytes até 2025 (Domo, 2022).

Esta explosão de dados, como sabemos, alimenta algoritmos programados com regras estatísticas em linguagem de computador, para reconhecer padrões e revelar regularidades, pelas quais o processo de tomada de decisão pode, supostamente, confiar. Ao relacionamento aprendido entre as variáveis sustentadas nos dados, chamamos comumente de modelo de *machine learning* (Barocas & Selbst, 2016).

2.1 Discriminação em aprendizado de máquina

Raça, sexo, religião, e idade são exemplos das chamadas variáveis protegidas, ou atributos sensíveis, e o desfavorecimento de grupo de indivíduos em razão da utilização de tais variáveis, conceitua a discriminação direta, ou tratamento díspar (Mehrabi et al., 2021).

Algumas decisões, porém, não são explicitamente baseadas em variáveis sensíveis, mas, ainda assim, geram resultados que prejudicam ou beneficiam desproporcionalmente certos grupos de indivíduos. Este desfavorecimento é chamado de impacto díspar, ou discriminação indireta. Certas decisões podem estar sendo baseadas em variáveis fortemente correlacionadas com alguma variável sensível, como por exemplo código postal e classe social, ou salário e gênero. Estas variáveis correlacionadas que se apresentam no lugar das reais variáveis de interesse, as quais não estão disponíveis, ou não podem ser usadas, são chamadas de *proxies* (Mehrabi et al., 2021).

A contrário do tratamento díspar, o impacto díspar em si, não é ilegal e algumas permissões sistêmicas podem ser feitas, de acordo com as necessidades do negócio ou por exemplo em decisões de contratação, sendo assim, o principal objeto de investigação em equidade algorítmica (Feldman et al., 2015).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Um algoritmo é tão bom quanto os dados utilizados por ele, uma vez que, um modelo de ML é treinado para se comportar conforme os exemplos a que foi exposto, logo, quando dados de treinamento do modelo contêm vieses e preconceitos, a regra aprendida será a reprodução deles (Barocas & Selbst, 2016). Vieses humanos provenientes de razões históricas são introduzidos em sistemas de aprendizado de máquina, prejudicando aqueles que estão sub ou super representados em tais dados (Barocas & Selbst, 2016).

Inúmeros tipos de vieses podem ser encontrados em sistemas algorítmicos (Srinivasan & Chander, 2021) e a categorização de tais vieses é relevante, pois pode motivar futuras soluções de acordo com cada tipo específico de viés identificado (Mehrabi et al., 2021).

Como esclarece Barocas e Selbst (2016), as cinco principais razões que podem possibilitar discriminação em modelos de ML são: 1) a incorreta definição da variável objetivo e dos rótulos das classes; 2) a falta de representatividade de grupos, e os erros de rotulação de classes nos dados de treino; 3) a falta de compreensão e seleção das variáveis envolvidas; 4) a falta de compreensão e seleção dos *proxies* das variáveis envolvidas; 5) e o mascaramento de visões preconceituosas de tomadores de decisões que podem ser intencionais ou não.

Todas essas razões, são fases de um processo subjetivo de compreensão de negócios e definição de problema, onde os dados utilizados são representações redutivas de um fenômeno do mundo real, que é infinitamente mais complexo e específico, e estas representações podem não captar todos os detalhes envolvidos na questão que desejamos resolver (Barocas & Selbst, 2016).

Outro fator importante a ser observado, é a minimização de erros médios que tendem a se ajustar ao grupo majoritário. Ou seja, quando a distribuição das variáveis é diferente entre grupos, estas variáveis terão diferentes relacionamentos com a variável objetivo, logo, ao treinar um classificador que não distingue grupos para minimizar o erro geral, ele se ajustará somente ao grupo de população majoritária, isso leva a uma diferente e maior distribuição de erros médios no grupo minoritário, fazendo que o modelo aprenda menos sobre este grupo. Por exemplo, ao prever o desempenho universitário de estudantes utilizando dados do ensino médio, em que um grupo majoritário é formado por estudantes tutores que fazem o exame *Scholastic Aptitude Test* (SAT) várias vezes e reportam apenas

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

as notas mais altas, e o grupo minoritário não o repetem, treinar um classificador cego para grupos resultará em melhores desempenho na universidade no grupo majoritário (Chouldechova & Roth, 2018).

2.2 Aprimoramento da equidade em aprendizado de máquina

Em ML, a equidade é reconhecida e imposta através do estabelecimento de métricas, operacionalizadas através da definição de equações matemáticas, que devem ser atendidas em determinadas fases do processo do desenvolvimento dos modelos, como veremos abaixo.

É importante nos atentarmos que ao buscar alcançar sistemas mais justos, podemos comprometer o desempenho ou a precisão do sistema, uma vez que a natureza do modelo que o define passa justamente por descobrir e reproduzir os padrões identificados nos dados, os quais por sua vez, podem estar carregando preconceitos e vieses. Ao atender a métrica de equidade determinada, o sistema acabará admitindo alguns erros em seu desempenho. Esses erros, porém, não devem comprometer significativamente a precisão do modelo e vice e versa, logo, devemos procurar sempre as melhores compensações geradas entre equidade e precisão em um modelo.

2.2.1 Métricas de equidade

Existe um amplo debate sobre quais são as melhores definições ou métricas de equidade algorítmica, e para sintetizar as principais ideias, utilizamos e estendemos o trabalho realizado por Verma e Rubin (2018), abrangendo ao todo seis grupos de métricas: métricas estatísticas, métrica de distância, equidade individual, equidade individual e grupal, raciocínio causal e representação justa.

2.2.1.1 Métricas estatísticas

As métricas estatísticas, também chamadas de métricas de equidade de grupo são as mais conhecidas, e fazem uma distinção entre um grupo protegido de pessoas, por exemplo negros e mulheres, e um grupo não protegido, como brancos e homens respectivamente,

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

e verificam a similaridade aproximada de algumas medidas nestes grupos (Chouldechova & Roth, 2018).

As métricas estatísticas podem ainda ser baseadas na matriz de confusão, nos resultados preditos e nos rótulos reais.

2.2.1.1.1 Baseadas na matriz de confusão

A matriz de confusão, apresentada na Tabela 2.1, é utilizada para descrever a acurácia dos modelos, onde as linhas representam os resultados preditos, e as colunas os rótulos reais (Verma & Rubin, 2018).

Tabela 2.1. Matriz de confusão

	Rótulos reais positivos	**Rótulos reais negativos**
Resultados preditos positivos	Verdadeiros positivos (VP): Quando o resultado predito e o rótulo real são positivos	Falsos positivos (FP): Quando o resultado predito é positivo, mas rótulo real é negativo
Resultados preditos negativos	Falsos Negativos (FN): Quando o resultado predito é negativo, mas o rótulo real é positivo	Verdadeiros Negativos (VN): Quando o resultado predito e o rótulo real são negativos

Fonte: (Verma & Rubin, 2018)

As métricas estatísticas baseadas nos resultados da matriz de confusão, consideram as probabilidades (P) de resultados preditos pelo sistema como positivos (d=1) ou negativos (d=0), estarem corretos ou incorretos em relação aos rótulos reais positivos (Y=1) ou negativos (Y=0) presentes no *dataset.* Sendo assim temos as seguintes métricas:

- Taxa de Verdadeiros Positivos (TVP): Também chamada de sensibilidade ou *Recall*, e representada através da equação (1), essa métrica indica a probabilidade de um resultado predito positivo realmente ser um rótulo real positivo:

$$\frac{VP}{VP+FN} = P(d = 1|Y = 1) \qquad (1)$$

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- Taxa de Falsos Positivos (TFP): Também chamada de taxa de falso alarme, essa métrica aponta a probabilidade de um resultado predito positivo, ser na verdade um rótulo real negativo, representada pela equação (2):

$$\frac{FP}{FP+VN} = P(d = 1|Y = 0) \quad (2)$$

- Taxa de Falsos Negativos (TFN): Mostra a probabilidade de um resultado predito como negativo ser um resultado real positivo, representada pela equação (3):

$$\frac{FN}{VP+FN} = P(d = 0|Y = 1) \quad (3)$$

- Taxa de Verdadeiros Negativos (TVN): Indica a probabilidade de um caso predito como negativo ser mesmo um rótulo real negativo, representada pela equação (4):

$$\frac{VN}{FP+VN} = P(d = 0|Y = 0) \quad (4)$$

- Valor Predito Positivo (VPP): Também chamada de precisão, indica a probabilidade de um resultado predito positivo ser verdadeiramente um rótulo real positivo, representada pela equação (5):

$$\frac{VP}{VP+FP} = P(Y = 1|d = 1) \quad (5)$$

- Taxa de Falso Descobrimento (TFD): Indica a probabilidade de um rótulo real negativo ser incorretamente predito como positivo, representada pela equação (6):

$$\frac{FP}{VP+FP} = P(Y = 0|d = 1) \quad (6)$$

- Taxa de Falsa Omissão (TFO): Indica a probabilidade de um rótulo positivo ser incorretamente predito como negativo, representada pela equação (7):

$$\frac{FN}{VN+FN} = P(Y = 1|d = 0) \quad (7)$$

- Valor Predito Negativo (VPN): Indica a probabilidade de um rótulo negativos ser predito como negativo, representada pela equação (8):

$$\frac{VN}{VN+FN} = P(Y = 0|d = 0) \quad (8)$$

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2.2.1.1.2 Baseadas no resultado predito

Para as métricas estatísticas baseadas no resultado predito (d), consideramos as probabilidades de resultados preditos como positivo (d=1) ou negativo (d=0) estarem corretos ou incorretos levando em consideração o grupo a que pertencem, representado nas equações abaixo pela letra G.

- Paridade estatística: Também chamada de equidade de grupo, essa métrica é satisfeita quando pessoas de grupos protegidos e não protegidos tem a mesma probabilidade de receber um resultado predito positivo, de acordo com a equação (9):

$$P(d = 1|G = protegido) = P\ (d = 1|\ G = desprotegido \quad (9)$$

Obs.: A regra conhecida como Regra 80%, ou apenas Regra de %, representada na equação (10), seria a fórmula legalmente aceite para quantificar o impacto díspar, suportada pelo *U.S Equal Employment Opportunity Comission*. A regra afirma que a fração entre sujeitos do grupo protegido com resultado positivo e sujeitos do grupo não protegido com resultado positivo deve ser menor que 80% sendo uma derivação da equidade de grupo, ou paridade estatística (Feldman et al., 2015).

$$\frac{P(d = 1|G = protegido)}{p(d = 1|G = desprotegido)} \leq 0.8 \quad (10)$$

- Paridade estatística condicional: Esta definição é satisfeita se sujeitos dos grupos protegidos e desprotegidos tiverem a mesma probabilidade de terem resultado positivo controlado por um conjunto de fatores (L), ou seja, esta medida permite que um conjunto de variáveis afetem o resultado, de acordo com a equação (11):

$$P(d = 1|L = l, G = protegido) = P(d = 1|\ L = l, G = desprotegido \quad (11)$$

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2.2.1.1.3 Baseadas no resultado predito e no rótulo real

Métricas estatísticas também podem ser baseadas no resultado predito (d), e no rótulo real (Y), presente no conjunto de dados, sendo elas:

- Paridade preditiva ou teste de resultado: Esta definição requer que o grupo protegido e o desprotegido tenham a mesma precisão ou VPP, e expressa a probabilidade de um sujeito com resultado positivo, realmente pertencer a uma classe positiva, conforme equação (12):

$$P(Y = 1|d = 1, G = protegido) = P(Y = 1|d = 1, G = desprotegido) \quad (12)$$

- Igualdade preditiva: Também chamada de taxa balanceada de erro de falsos positivos, esta métrica busca que os grupos protegido e desprotegido tenham a mesma TFP, e mostra a probabilidade de um sujeito na classe negativa ter um resultado positivo, ou seja, o classificador deve dar resultados similares em ambos os grupos para pessoas com rótulo real negativo, conforme equação (13):

$$P(d = 1|Y = 0, G = protegido) = P(d = 1|Y = 0, G = desprotegido) \quad (13)$$

- Probabilidades equalizadas: Também chamada de *disparate mistreatment*, esta métrica proposta por Hardt et al. (2016) em citação de Verma e Rubin (2018), requer que ambos os grupos, protegidos e não protegidos tenham a mesma TVP e a mesma TFP. Matematicamente é o mesmo que taxa balanceada de erro de falsos positivos e taxa balanceada de erro de falsos negativos. Implica que a probabilidade de um aplicante da classe real positiva, ou da classe real negativa devem apresentar classificações similares, independentemente do grupo a que pertencem, e está representada na equação (14):

$$P(d = 1|Y = i, G = protegido) = P(d = 1|Y = i, G = desprotegido) \quad (14)$$

$$com\ i \in 0,1$$

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- Oportunidades iguais: Também chamada de taxa balanceada de erro de falsos negativos, esta métrica foi proposta por Hardt et al. (2016) em citação de Verma e Rubin (2018), e requer que ambos os grupos tenham a mesma TFN e mostra como representada na equação (15), a probabilidade de um sujeito da classe positiva ter um resultado predito negativo, ou seja, busca que o classificador de resultados similares em ambos os grupos para pessoas com rótulo real positivo. Essa métrica é considerada um relaxamento da métrica probabilidades igualadas, para casos em que a TVP seja mais importante, como em casos de contratação por exemplo:

$$P(d = 0|Y = 1, G = protegido) = P(d = 0|Y = 1, G = desprotegido) \quad (15)$$

2.2.1.2 Métricas de distância

O segundo grupo de métricas apresentado, foca na distância observada entre variáveis nas matrizes de classificação (Zafar et al., 2017) ou recomendação (Edizel et al., 2020), compreendendo as métricas de:

- Covariância do limite da decisão de injustiça: Esta métrica proposta por Zafar et al. (2017), mede a covariância entre o atributo sensível do usuário e a distância sinalizada dos vetores de recurso dos usuários para limite de decisão. Representada na equação (16), onde, X = variáveis não sensíveis, Z= variáveis sensíveis, θ =parâmetros de limite de decisão e $d_\theta(x)$ = distância da margem. Esta medida mostra o grau de independência do resultado predito da variável sensível com a decisão limite do classificador, ou seja, o grau de aproximação tolerado, para atuar como *proxy* que captura a relação entre o atributo sensível e as previsões em nível de grupo.

$$Cov\left(z, d_\theta(x)\right) = E\left[(z - \bar{z})d_\theta(x)\right] - E\left[(z - \bar{z})\right]\bar{d}_\theta(x) \approx \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i) \quad (16)$$

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- Métricas μ_1 *e* μ_2: Estas métricas foram apresentadas por Edizel et al. (2020), para medir a equidade em sistemas de recomendação de filmes, onde: μ_1 representada na equação (17), induz a distância entre matrizes do número de usuários afetados de qualquer forma quando substituímos C (matriz de recomendação original) por C´(matriz de recomendação justa), e μ_2 representada pela equação *(18)*, induz a distância entre matrizes do número de itens de recomendação alterados quando substituímos C (matriz de recomendação original) por C´(matriz de recomendação justa).

$$\mu_1(y,z) := \begin{cases} 0 & y = z \\ 1 & y \neq z \end{cases} \quad e \qquad (17)$$

$$\mu_2(y,z) = \frac{1}{2} \cdot \|y - z\|_1 \qquad (18)$$

2.2.1.3 Equidade individual

O terceiro grupo, são as métricas de equidade individual, ou métricas baseadas em similaridade. Essas métricas ao contrário das métricas estatísticas de grupo, possuem restrições que comparam pares de indivíduos específicos e não média de grupos distintos (Chouldechova & Roth, 2018), sendo elas:

- Discriminação causal: Segundo Verma e Rubin (2018), para atender essa métrica, a classificação de dois indivíduos com os mesmos atributos (X), deve ser igual, com exceção apenas do atributo sensível. Dessa forma, aplicantes para liberação de empréstimos do sexo feminino e masculino que possuem os mesmos atributos X, por exemplo, devem ter a mesma classificação para uma nota de crédito boa ou ruim, caso essa classificação seja diferente entre eles, significa que a métrica não foi atendida. Representação na equação (19):

$$(X\ desprotegido = X\ protegido \wedge G\ desprotegido\ ! = G\ protegido) \rightarrow d\ protegido = d\ desprotegido \qquad (19)$$

- *Fairness through unawareness:* Definição que requer que o classificador não utilize explicitamente nenhum atributo sensível, e que a classificação dos

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

indivíduos i e j que têm os mesmos atributos, seja a mesma, como segue na equação (20), (Verma & Rubin, 2018):

$$X: X_i = X_j \rightarrow d_i = d_j \qquad (20)$$

- *Fairness through awareness*: Métrica proposta por Dwork et al. (2012) parte do princípio de que indivíduos similares devem ser tratados de forma similar, ou seja, ter classificações similares. Chouldechova e Roth (2018) dizem que tal semelhança é definida em relação a uma métrica específica da tarefa que deve ser determinada conforme o caso. Verma e Rubin (2018) explicam que, para um conjunto de candidatos V, uma métrica de distância entre os candidatos k: V ×V → R, um mapeamento de um conjunto de candidatos a distribuições de probabilidade sobre os resultados, onde M: V → δA, e uma distância D métrica entre a distribuição de saídas, a justiça é alcançada conforme representação na equação (21) :

$$D(M(x), M(y)) \leq k\,(x, y) \qquad (21)$$

- Controle de discriminação ou *Sample Distortion*: Proposta por Calmon et al. (2017), esta métrica calcula a distância entre o mesmo ponto individual no *dataset* original e no *dataset* transformado. Para limitar a dependência do resultado transformado $\hat{Y}$ na variável protegida *D*, são propostas duas formulações:
  a) Requer que a distribuição condicional $P_{\hat{y}|D}$ seja próxima a distribuição justa PY_T para todos os valores *D.* Onde J denota a função distância, conforme equação *(22)*:

$$J\left(P_{\hat{y}|D}(y|d), PY_T(y)\right) =< \epsilon_{y,d}\ \forall d\ \epsilon\ D, y\ \epsilon\{0,1\} \qquad (22)$$

  b) Limita-se que a probabilidade condicional $P_{\hat{y}|D}$ deve ser similar para quaisquer dos dois valores de *D*, conforme equação *(23)*:

$$J\left(P_{\hat{y}|D}\ (y|d1), P_{\hat{y}|D}\ (y|d2)\right) =< \epsilon_{y,d1,d2}\ \forall\ d1, d2\ \epsilon\ D, y\ \epsilon\{0,1\} \qquad (23)$$

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

A escolha da distribuição justa PY_T em a, e J em a e b devem ser informadas por aspetos sociais, especialistas de domínio partes interessadas, e considerações legais como a regra 80%.

2.2.1.4 Métrica de equidade individual e grupal

O quarto grupo de métricas, proposto por Speicher et al. (2018), argumenta que a equidade buscada pode ser decomposta entre os grupos e entre os componentes dentro de um grupo, sendo individual e grupal ao mesmo tempo. Esta métrica pode ser alcançada através de uma família de índices, generalizadamente chamadas de índice de entropia, que inclui o coeficiente de variação e o índice *Theil* (Speicher et al., 2018). Os índices de entropia generalizada possuem a propriedade de decomposição do subgrupo, que permite quantificar a injustiça a nível do indivíduo e a nível do grupo, representados na equação *(24)*:

$$\varepsilon^{\alpha}(b_1, b_2, \cdots b_n) = \frac{1}{n\alpha(\alpha-1)} \sum_{i=1}^{n} \left[\left(\frac{b_i}{\mu}\right)^{\alpha} - 1\right] \qquad (24)$$

2.2.1.5 Raciocínio causal

O quinto grupo de métricas, assume um gráfico cíclico causal representado na Figura 2.1, em que os eles representam atributos e as arestas as relações entre estes atributos G (Verma & Rubin, 2018).

- Equidade contra factual: Esta noção de equidade é fundamentada na ideia de que uma decisão é justa para um indivíduo, se for a mesma tanto no mundo real quanto no mundo contra factual, mundo este em que relativamente ao atributo sensível, o indivíduo pertenceria a um grupo diferente (Pessach & Shmueli, 2020). Um gráfico causal é contra factualmente justo se o resultado previsto d no gráfico não depender de um descendente do atributo protegido G (Verma & Rubin, 2018).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

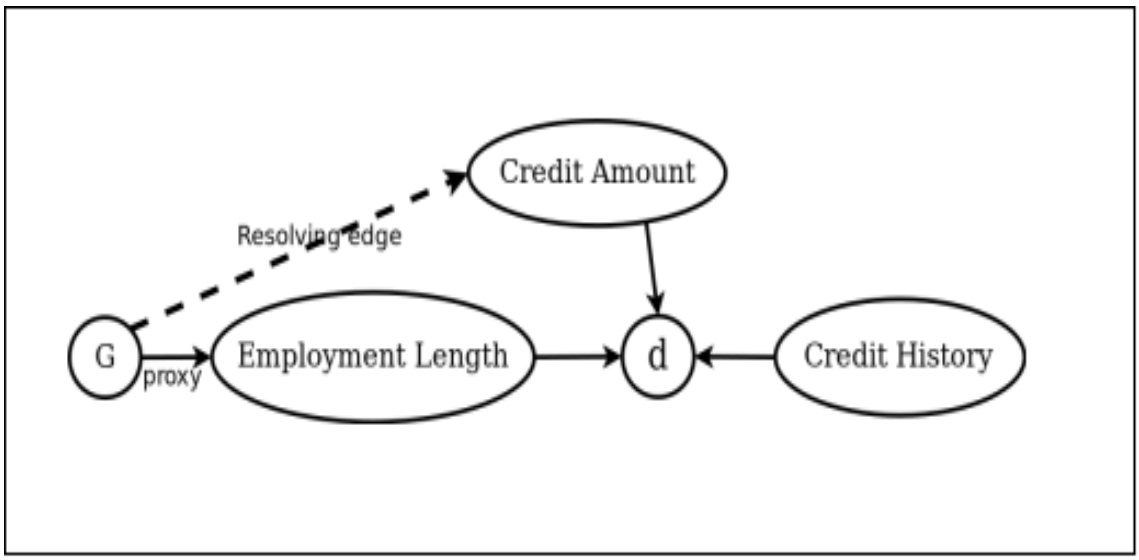

Figura 2.1. Gráfico Causal

Fonte: Verma & Rubin (2018)

2.2.1.6 Representação justa

No sexto grupo de métricas, é demonstrado um processo de despolarização de dados, que reproduz transformações dos dados originais tendenciosos, onde a participação do grupo pode ser inferida por outras variáveis, para um espaço onde os atributos protegidos são estatisticamente independentes de outras características (Chouldechova & Roth, 2018).

- Consistência: Essa métrica compara a classificação prevista de um dado item de dados x para seu k-mais próximo vizinho. Aplicamos a função *K- nearest Neighbors* (kNN) ao conjunto completo de exemplos para obter a estimativa mais precisa de vizinhos mais próximos de cada ponto, conforme equação (25), (Zemel et al., 2013).

$$yNN = 1 - 1NkXn|y\hat{}\ n - Xj\epsilon NN(xn)y\hat{}\ j| \qquad (25)$$

2.2.2 Equidade nas etapas do processo de aprendizado de máquina

Geralmente as técnicas para mitigar preconceitos e vieses sugerem uma possível atuação em um dos três momentos: pré-processamento, processamento e pós-processamento.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Na fase de pré-processamento, o objetivo é trabalhar o conjunto de dados de treino, através de redistribuição dos dados, alteração de rótulos de classes, atribuição de pesos diferentes ou realização de reamostragem a fim de equilibrar os grupos protegidos e desprotegidos. Procura-se ajustar as variáveis para não se correlacionarem com o atributo sensível (Barocas et al., 2019).

As abordagens que ocorrem durante a fase de processamento, atuam no algoritmo em si, reformulando o problema e incorporando o comportamento de discriminação do modelo na função objetivo através de regularizações e restrições (Ntoutsi et al., 2020).

Já na fase de pós processamento, o foco se concentra em atuar nos resultados produzidos pelo modelo, ajustando o classificador treinado a satisfazer as restrições de equidade. Em um modelo de caixa preta por exemplo, uma atuação seria manter a proporcionalidade das decisões entre grupos protegidos e desprotegidos, promovendo ou rebaixando classificações (Ntoutsi et al., 2020).

2.2.3 Técnicas de aprimoramento da equidade

As técnicas para aperfeiçoamento de equidade variam de acordo com a combinação das métricas de *fairness*, dos algoritmos base, e dos estágios de processamento em que são executadas. Sintetizamos em seguida importantes técnicas propostas por vários autores.

Para Feldman et al. (2015), o pré-processamento dos dados deve ser realizado sem modificar os rótulos de treinamento. Eles alteram os atributos para que as distribuições marginais de grupos privilegiados e não privilegiados sejam semelhantes, e utilizam a métrica de impacto díspar para tal comparação. Desta forma, algoritmos de classificação não fazem diferenciação entre os grupos.

Por outro lado, Kamiran e Calders (2012), pré-processam os dados alterando e reavaliando os pesos dos rótulos que estão relacionados a amostras mais próximas da decisão limite no conjunto de treino, pois estas amostras têm mais probabilidade de serem discriminadas. A técnica se desenvolve com base em qualquer algoritmo de pontuação e mede a equidade pela métrica de paridade demográfica.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Já Calders e Verwer (2010), utilizam um algoritmo *Naive Bayes* que treina dois modelos diferentes para os valores dos atributos sensíveis, e através de pequenas alterações nas probabilidades observadas, busca reduzir a medida de paridade demográfica. Os modelos são então reciclados, formando o terceiro modelo, ou seja, ele altera o funcionamento do algoritmo para atingir a paridade demográfica.

Como exemplo de aprimoramento da equidade no pós-processamento, Hardt et al. (2016) propõem inverter algumas decisões finais de algoritmos classificadores para aumentar as chances equalizadas e as oportunidades iguais dos grupos, métricas estas propostas no próprio estudo.

2.3 Abordagens sociotécnicas

Apesar da lógica matemática e computacional necessária no desenvolvimento de algoritmos mais justos, o aprimoramento da equidade demanda também interações interdisciplinares com outras esferas da sociedade, como normatizações e implementações guiada por princípios éticos.

O Regulamento Geral de Proteção de Dados (RGPD) da União Europeia (UE), interpela o assunto equidade ao dizer que as empresas precisam dar atenção ao efeito do processamento de dados nos indivíduos, devendo justificar qualquer impacto adverso, e se responsabilizar por possíveis decisões injustas de um sistema algorítmico (Mokhtari et al., 2021).

Vários autores também detalham a relação conflituosa entre sistemas de IA com a legislação e a justiça (Altman et al., 2018; Lupo, 2019; Russell, 2020), com direitos humanos (Završnik, 2020) e com o sistema habitacional (Schneider, 2020). No entanto, como lembra Rivas (2020), marcos legais internacionais podem ser desafiadores de serem cumpridos, uma vez que requerem cooperação significativa entre as nações, por isso a tentativa de normatização muitas vezes é conduzida pelas próprias empresas, como forma de autopoliciamento.

Iniciativas como o grupo de normas P7000 estabelecidas pelo *Institute of Electrical and Electronics Engineers* (IEEE), buscam definir padrões para sistemas inteligentes e eticamente alinhados e sugerem a utilização de guias para ajudar desenvolvedores a

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

identificar e evitar potenciais fontes de vieses (Mokhtari et al., 2021). Ainda nesta perspectiva, Adams e Hagras (2020) revisam as orientações regulatórias sobre a adoção de IA ética e segura nos principais mercados mundiais, e explicam porque *fairness*, *accountability* e *transparency* são as palavras chaves para orientar o desenvolvimento de sistemas algorítmicos.

Considerando a série de partes afetadas pelos comportamentos dos processos algorítmicos, e levando em conta os princípios de *awareness, access, redress, accountability, explanation, data provenance, auditability, validation and testing*, proposto pela ACM, Tal et al. (2019) desenvolveram um *framework* integrativo que visa garantir a equidade em sistemas através da atuação de um regulador, que define especificações, requisitos e audita o sistema, e um desenvolvedor, que deve seguir as regras do regulador e certificar-se através de um processo iterativo, que o sistema é justo e transparente.

Para Desmoulin-Canselier e Le Métayer (2018), certificação e explicação são as apostas para decisões de sistemas algorítmicos mais confiáveis. Os autores sugerem que a certificação é a obrigação de mostrar para um auditor ou autoridade responsável, que o sistema atende a certos critérios, como por exemplo a relevância das decisões (acurácia) e a ausência de fatores de discriminação, e que deve garantir tanto a explicação local da lógica de decisão específica que impacta indivíduos sem conhecimento técnico, quanto a explicação da lógica do modelo global para especialistas.

Outro *framework* integrativo foi proposto por Kristiadi et al. (2020), cujo foco é dividido em três principais categorias, *group fairness, individual fairness e data fairness*. A estrutura permite visualizar o problema de forma incremental e compreender como algumas questões podem cruzar ou convergir com várias categorias. O estudo propõe indagações desafiadoras como “quem deveria ser a autoridade a tomar a decisão de intervenção de um sistema?” e “como podemos fazer predições precisas quando as circunstâncias desses dados coletados não são justas?”.

Na esfera da saúde, McCradden et al. (2020) propõem uma abordagem regulatória para segurança de pacientes na prevenção de danos não intencionais e a melhoria da qualidade dos serviços de saúde com aprendizado de máquina. Nos serviços sociais, Gillingham

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

(2019) discute como assistentes sociais podem identificar recomendações mal feitas ou tendenciosas e desenvolver estratégias para ajudar os usuários finais que podem ser vítimas de decisões algorítmicas injustas.

Já Cardenas e Vallejo-Cardenas (2019), debatem como vieses de aprendizado de máquina afetam comunidades marginalizadas e apontam que o incentivo à educação e à diversidade deve ser promovido, principalmente em profissões de *Science, Technology, Engeneering and Mathematics* (STEM), para que mais pessoas com diferentes características participem do processo de desenvolvimento tecnológico.

Considerando a falta de padronização no processo de documentação de conjunto de dados, Gebru et al. (2021) recomendam que os conjuntos de dados deveriam ser acompanhados de uma ficha que explique a motivação, composição, processo de coleta, recomendação de uso, e possíveis implicações e danos, para encorajar a reflexão cuidadosa da utilização de dados por desenvolvedores, usuários, políticos, advogados, jornalistas, e indivíduos que podem ser impactados.

De forma semelhante, Mitchell et al. (2019) recomendam a utilização de documentos curtos para modelos de ML treinados, que contenham avaliação de referência de condições em diferentes grupos culturais, demográficos ou fenótipos, o contexto em que os modelos se destinam a serem usados, detalhes de procedimento de avaliação e outras informações relevantes.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

3 METODOLOGIA

Para realização deste trabalho, seguimos os critérios de revisão sistemática de literatura, que busca responder a uma pergunta formulada de forma clara, utilizando métodos sistemáticos explícitos para identificar e selecionar pesquisas relevantes, e coletar e analisar dados dos estudos incluídos na revisão (Freire et al., 2015) a fim de fornecer um alto nível de evidência e reduzir os potenciais vieses na identificação de estudos que suportariam a opinião do próprio autor (Impellizzeri, 2012).

Para documentar as evidências encontradas na pesquisa de forma transparente e reproduzível, nos guiamos no essencial da diretriz para revisões sistemáticas de literatura designada por *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) (Page et al., 2021).

O guia de redação de revisões sistemáticas PRISMA, fornece em um diagrama de fluxo, as fases mais importantes da RSL, identificação, triagem e inclusão, e mapeia as publicações elegíveis e excluídas em cada fase, e suas respetivas razões de elegibilidade e exclusão, demonstrando de forma visual as informações de todas as fases do processo de elaboração da revisão sistemática.

3.1 Estratégia de busca

A estratégia de busca foi baseada no uso de palavras-chaves em inglês, onde os três principais termos identificados para a pesquisa foram: *technique*, *fairness* e *machine learning*.

Como possíveis alternativas e variações da palavra *technique* utilizamos: *method, tool, way, action, framework, approach, how, strategy, system.* Como possíveis variações da palavra *fairness*, consideramos *bias* e *justice* e como possíveis variações do termo *machine learning*, consideramos *artificial intelligence, algorithm, IA, Big Data* e *data mining.*

Aplicamos os operadores lógicos *OR* e *AND* do inglês, e o símbolo asterisco * para truncar o sufixo e captar as variações destas palavras, e fixamos a estratégia de busca apenas no campo título.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Por fim, a expressão pesquisada na b-on foi exatamente:

(technique OR method* OR tool* OR way* OR action* OR framework* OR approach* OR how OR strateg* OR syste*) AND (fairness OR fair* OR *bias* OR *justice*) AND ("machine learning" OR "artificial intelligence" OR "algo*" OR "AI" OR " Big Data" OR "data mining")

3.2 Fontes de informação e gerenciamento dos dados

A pesquisa foi realizada no dia 9 de dezembro de 2021, através da plataforma b-on, a biblioteca de conhecimento online de Portugal, que reúne textos integrais de periódicos e *ebooks* de uma extensa lista de instituições de investigação científica e tecnológica e algumas das principais editoras científicas internacionais (B-On, 2022).

Os resultados foram extraídos dos diversos fornecedores de conteúdo, de acordo com o especificado na Tabela 3.1.

Tabela 3.1. Fontes de informações

Fornecedor	**Número de Registos**
Complementary Index	101
Science Citation Index Expanded	83
Scopus	69
Academic Search Complete	62
IEEE Xplore Digital Library	41
Directory of Open Access Journals	36
Social Sciences Citation Index	35
Medline	35
Business Source Complete	20
Science Direct	17
Supplemental Index Library, Information	8
Science & Tecnology Abstracts	5
Gale In Context: Science	3
Total	515

Fonte: Elaboração própria

Os resultados provenientes das bases de dados da b-on foram exportados em arquivo do tipo *Research Information System* (RIS), arquivados pelo sistema gerenciador de

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

referências Zotero, e manuseados e explorados por Excel, onde implementamos os critérios de seleção e triagem através do uso de filtros.

3.3 Critérios de elegibilidade

Para esta revisão, foram eleitos os artigos que descrevem estudos primários, com abordagens técnicas que buscam aprimorar a equidade em aprendizado de máquina.

Dada a contemporaneidade do assunto, pesquisamos artigos publicados entre os anos de 2017 até 2021, de qualquer tipo metodológico, e sem restrição de área. Porém, devido a particularidade e relevância do tema, o tipo de publicação foi limitado a artigos de revistas científicas e revisado por pares (*peer reviewed*), descartando qualquer literatura cinzenta como relatórios ou artigos de imprensa tradicional.

A pesquisa compreendeu apenas artigos em idioma inglês, uma vez que este é o idioma predominante das publicações.

Em um segundo momento, foram adicionados quatro estudos provenientes de citações identificadas em artigos elegíveis integralmente lidos e analisados.

3.4 Critérios de exclusão

Foram excluídos títulos referentes a pesquisas secundárias ou outras revisões narrativas e sistemáticas, publicações em duplicidade, e artigos que não estavam disponíveis nas bases de dados da b-on e do *google scholar*.

Durante a fase de triagem de relevância (leitura de títulos e resumos), foram também identificados e excluídos os registros que abordavam equidade e aprendizado de máquina em outros contextos que não são interesse deste estudo, como por exemplo alocação de recursos ou tarefas de componentes de sistemas de computador.

Outro critério de exclusão utilizado foi de títulos e resumos que abordavam e debatiam o assunto, mas que não sugeriam nenhuma intervenção ou proposta para o aprimoramento do problema.

Por fim, após a leitura completa dos artigos selecionados, excluímos também aqueles em que a proposta para promoção de equidade algorítmica se concentrava em esforços

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

sociotécnicos, como normatizações, diretrizes e princípios éticos, cuja atuação extrapolam a ciência da computação e foram brevemente contemplados no capítulo 2.3.

3.5 Diagrama de fluxo

Os detalhes de cada fase estão apresentados na Figura 3.1 e descritos através de síntese narrativa, sumarizados em forma da Tabela 4.1. Por fim, as principais características dos estudos foram evidenciadas através de gráficos, discutidas pela visão da autora e confrontados com a literatura.

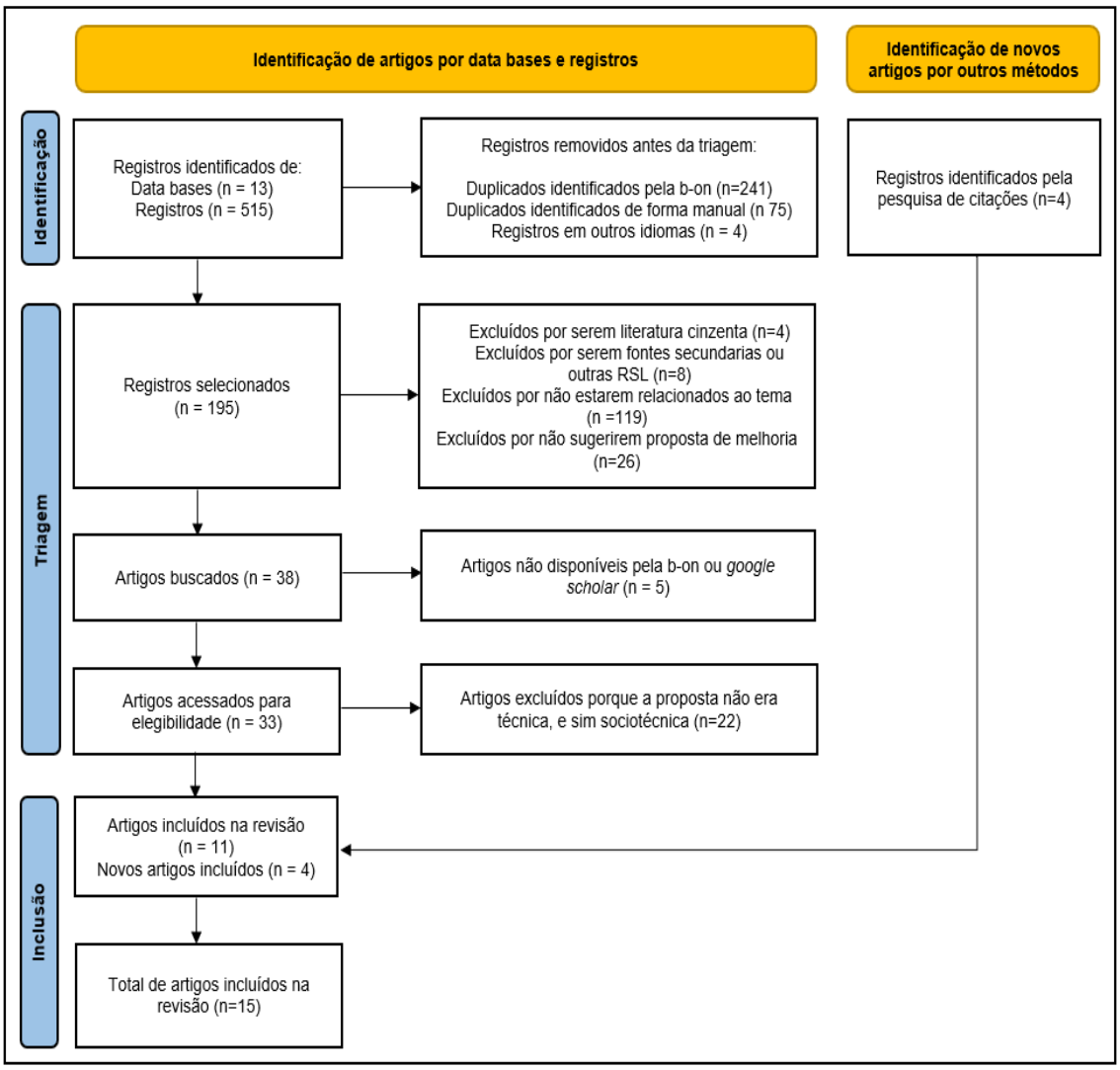

Figura 3.1. Diagrama de fluxo PRISMA

Fonte: Elaboração própria.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

4 RESULTADOS

A fase de identificação resultou em 515 registros provenientes de 13 fontes de informações diferentes, dentre os quais, 316 eram registros duplicados e 4 estavam em outros idiomas. A fase de triagem (leitura de títulos e resumos) contou com 195 registros, onde 4 não eram artigos científicos, 8 não eram fonte primária, 119 não estavam diretamente relacionados ao tema de equidade em aprendizado de máquina e 26 não indicavam propostas de melhoria do problema no título ou no resumo.

Passando para a fase de acesso aos artigos completos com 38 artigos elegíveis, dos quais, 5 não estavam disponíveis online nem na b-on e nem no google acadêmico. Durante a fase final, 33 artigos foram lidos por completo, de onde 22 foram excluídos pois suas propostas de melhoria para o tema de equidade em aprendizado de máquina se davam em um contexto sociotécnico, finalizando com 11 artigos considerados elegíveis.

Adicionamos ainda na revisão, 4 artigos provenientes de citações dos 11 artigos inicialmente lidos, totalizando 15 artigos finais na revisão. Os detalhes de cada fase podem ser conferidos na Figura 3.1.

4.1 Descrição dos resultados

Lin et al. (2021), avaliam intervenções existentes assistidas por IA e exploram novas e promissoras abordagens de IA na área de contratação de pessoas. A publicação aponta 11 ferramentas já em utilização para redução de riscos de vieses implícitos no momento de recrutamento e seleção de candidatos, sendo elas: *Eightfold, Entelo, IBM Watson, Blendoor, Interviewing.io, Pymetrics, Textio Hire, Hire Vue, Tengai, Equal Reality* e *Vantage Point*. O estudo ainda classifica cada ferramenta de acordo com suas informações descritivas, preditivas, e prescritivas, e também de acordo com as intervenções baseadas em *inputs* em *output* ou intervenções baseadas na cognição, desenhando assim um mapa para identificação da melhor ferramenta de acordo com cada caso específico (Lin et al., 2021).

Visando promover a popularidade do uso das métricas de equidade e das técnicas de mitigação de vieses, Bellamy et al. (2019) oferecem um pacote de ferramentas *open*

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

source em *Python* chamado AI 360. O pacote disponibiliza *datasets*, métricas, e algoritmos de pré, in e pós-processamento, que mantêm a qualidade do código e facilitam a compreensão da validação dos modelos. O AI 360 oferece também uma experiência web interativa para explicação de conceitos, documentação, guias e tutoriais para desenvolvedores e pesquisadores do tema (Bellamy et al., 2019).

Também nesta lógica, a pesquisa de Ahmed et al. (2021) dá continuidade ao trabalho anterior, utilizando o AI 360 para comparar a equidade alcançada antes e depois da experimentação das técnicas, no conjunto de dados *US Employment Demografics*, e confirma a eficácia do pacote demonstrando bons resultados atingidos principalmente com os algoritmos de pós-processamento baseados em *Random Forest* (RF) e *Logistic Regression* (LR).

O estudo de Dash et al. (2019) também apresenta técnicas de pré, in e pós-processamento para mitigação de vieses, porém, especificamente para sistemas de sumarização de textos. O algoritmo *FairSumm* é otimizado para buscar qualidade na sumarização dos textos e atender aos critérios de equidade aplicados como restrições de matróides (conceito usado para generalizar a noção de independência linear das matrizes), durante o processamento do modelo. O algoritmo *ClasswiseSumm* pré-processa os dados agrupando os textos com base em diferentes classes e em seguida resume cada grupo separadamente. O algoritmo *RefaSumm* realiza a reclassificação justa de textos para duas classes, ou dois grupos sociais diferentes, com base em alguma medida de equidade definida na configuração dos parâmetros do algoritmo. O objetivo do trabalho é realizar a sumarização de textos com qualidade e garantir que todos os grupos tenham as opiniões representadas no resumo (Dash et al., 2019).

Preocupado com a forma de utilização de dados pessoais sensíveis, Mokhtari et al. (2021) apresentam uma estrutura de monitoramento de uso de dados pessoais em plataforma de *Big Data*, para controle de conformidade com o *Privacy transparency fairness Agreement* (PTFA). A estrutura coleta os dados pessoais via o processo *extract, transform and load* (ETL), realiza serialização, fusão, anonimização e agrupamento, e disponibiliza para utilização de usuários, de forma regulamentada por um contrato contendo as regras do PTFA em forma de *checkbox*. A estrutura então confronta os *logs* e as cláusulas do

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

contrato, acompanha o comportamento do tráfico de dados, sabendo quais regras estão sendo violadas, e quais estão conformes, podendo atuar contra a má utilização de dados pessoais (Mokhtari et al., 2021).

Também buscando a proteção de dados sensíveis, Hu et al. (2019) desenvolvem uma estrutura para aprendizado distribuído, e acesso restrito a dados demográficos. O framework conta com a participação de duas instituições diferentes, uma que detém a propriedade privada dos dados demográficos e outra que gerencia o *data center* com dados não demográficos e desenvolve o modelo de aprendizado de máquina. Funcionando da seguinte forma: O *data center* gera hipóteses aleatórias de distribuições gaussianas dos dados não demográficos, obtém previsões nos conjuntos de treinamento e envia estas previsões para a instituição que detém os dados demográficos, que por sua vez, estima a correlação entre as previsões e esses dados. Se a correlação for pequena, a hipótese é justa, e então é confirmada. O próximo passo então é utilização destas hipóteses justas para geração dos modelos privados. O estudo testou o método proposto para redesenhar quatro tipos de algoritmos não privados já utilizados anteriormente, em versões privadas que atendem a métrica de paridade estatística, sendo elas: 1) *Distributed Fair Ridge Regression* (DFRR); 2) *Distributed Fair Kernel Ridge Regression* (DFKRR); 3) *Distributed Fair Logistic Regression* (DFGR); e 4) *Distributed Fair Principal Component Analysis* (DFPCA).

A proposta de Zafar et al. (2017) segue outra linha, através da utilização de uma nova medida de limite de decisão de injustiça que modifica os classificadores *Logistic Regression* (LR) e *Support Vector Machines* (SVM) para penalizarem a discriminação. Tal medida é calculada pela covariância entre o atributo sensível e a distância sinalizada entre os vetores de características dos sujeitos. Esta medida deriva de duas formulações complementares de restrições para treinamento de classificadores, uma formulação que procura maximizar a precisão sob restrição de equidade para cumprimento da política ou lei de discriminação (*p% rule)* e a outra que procura maximizar a equidade sob restrição de precisão para garantir a necessidade do negócio. A medida garante a equidade em relação a um ou mais atributos sensíveis, para tratamento simultâneo de discriminação direta e indireta (Zafar et al., 2017).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Outra proposta que tem por objetivo a exploração dos limites de compensações entre equidade, medida pela taxa de FP e acurácia medida pelo erro médio, foi feita por Valdivia et al. (2021). A publicação sugere um método baseado no algoritmo multiobjectivo, *Non dominated Sorting Genetic Algorithm* (NSGA-II) para orientar um classificador, neste caso árvores de decisão, por serem compreensíveis e transparentes. Este algoritmo busca obter as árvores com melhores compensações de acurácia e equidade aprendendo a melhor combinação dos hiper parâmetros de critério, profundidade máxima, número mínimo de amostras para dividir um nó, número total de folhas e o peso de cada classe, e oferece as melhores soluções viáveis através de uma frente de Pareto (Valdivia et al., 2021).

Os algoritmos *Super Sparse Linear Integer Model* (SLIM) e *Risk-calibrated Super Sparse Linear Integer Model* (RiskSLIM) de Rudin e Ustunb (2018), também trabalham com restrições de equidade. O algoritmo SLIM é otimizado para a compensação entre a taxa de verdadeiros positivos (TVP) e a taxa de falsos positivos (TFP) e as previsões são baseadas em se a pontuação excede um valor limite ou não (se pontuação total > 1 = sim, e se pontuação total é <1 = não). A escolha da taxa de VPs ou FPs dependem da aplicação, por exemplo, para a triagem médica, é mais indicado buscar uma maior taxa de FPs, ou seja, uma maior taxa de falso alarme, onde o usuário estipula a taxa máxima de FP que pode tolerar, e o SLIM otimiza a taxa de VP sujeitos a esta restrição.

O algoritmo RiskSLIM é um sistema de pontuação de risco, calibrado pelo risco, ou seja, o risco previsto pelo modelo, e o mesmo que existe nos dados. O RiskSLIM não procura compensações entre TVP e TFP, ao invés disso ele busca alcançar a melhor taxa de VP. Os autores argumentam que os modelos indicados atingem bons resultados de desempenho, e apresentam ainda a vantagem de serem transparentes e explicáveis, sendo boas opções em relação a modelos caixa preta, em que as regras de decisão não são explícitas (Rudin & Ustunb, 2018).

Uma abordagem diferente é oferecida por Zhang et al. (2018), que utilizaram a atuação de múltiplas redes com objetivos concorrentes para mitigação de vieses. O modelo preditor busca prever a variável alvo Y baseado nas variáveis independentes X, modificando os pesos e minimizando as perdas, usando métodos baseados em gradiente

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

como regressão e classificação. O resultado deste modelo é então usado como entrada na rede concorrente que tenta prever Z, a variável protegida, se baseando em outras entradas, que dependerão da métrica de equidade a ser alcançada. Por exemplo, se a métrica a ser alcançada for a de probabilidades equalizadas (*equalized odds*), o modelo adversário terá acesso também aos rótulos verdadeiros e aprenderá a relação entre Y e Z independentemente do que o preditor faz. Quando a métrica estipulada é alcançada o treinamento do modelo adversário termina. Os autores detalham quais são os *inputs* de acordo com a métrica que se busca atingir (Zhang et al., 2018).

Os estudos de Sahu e Singh (2019) e de Edizel et al. (2020) trataram especificamente de sistemas de recomendação de filmes. O modelo de Sahu e Singh (2019) traz um algoritmo baseado em filtragem colaborativa, que utiliza as variáveis de usuários e as variáveis do filme, e mistura dois tipos de recomendação. Uma que reforça as preferências do usuário e outra que é inversamente correlacionada a essas preferências, ou seja, o algoritmo recomenda filmes que o usuário gosta, e outros que não tem tanto interesse, mas que, porém, são de qualidade. Estas variáveis são aprendidas por uma extensão de gradiente de descida estocástica e utilizada para medir o *Root Means Square Error* (RMSE). A ideia é promover a diversidade e tentar furar a "bolha" de viés pessoal no conteúdo recomendado, expondo o usuário a novas possibilidades (Sahu & Singh, 2019).

Já o modelo de Edizel et al. (2020), concentra-se em um algoritmo de pós processamento da matriz de recomendação. O FaiRecSys, procura resolver problemas de *Fair Recomendation Matrix* (FRM) utilizando um vetor de atributo sensível binário e um nível de equidade estipulado para computar uma nova matriz de recomendação, que deve respeitar as métricas μ_1 e μ_2 propostas no estudo (Edizel et al., 2020).

Como sistemas de IA não utilizam regras de códigos convencionais, mas sim dados, para prever comportamentos futuros, Obaidat et al. (2021) dizem que tais sistemas são mais suscetíveis a adulteração de dados de adversários, que podem inundar o sistemas com dados falsos e consequentemente gerar decisões não confiáveis. Para abordar este problema, o autor propõe o método *Minimize AI bias applying Random Sampling Technique* (MAIRST) que combina a amostragem aleatória para treinamento de dados em algoritmos de redes neurais convolucionais, do inglês, *Convolutional Neural Network*

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

(CNN). Na dinâmica proposta, uma rede neural utiliza dados de teste que foram adulterados propositalmente de forma aleatória para simular um ataque adversário que conseguiu se infiltrar no sistema, trazendo um efeito negativo no desempenho do modelo. Em seguida, o método MAIRST é aplicado a este conjunto de teste para filtrar ao máximo os dados modificados, trabalhando como uma segunda linha de defesa, e produzindo um novo conjunto de teste final que será usado para avaliar o modelo, que deve reconhecer e classificar imagens de peças de vestuário (Obaidat et al., 2021).

Em resumo, trata-se de um modelo de simulação funcional que utiliza amostragem aleatória para minimizar efeitos de adulteração de dados e ataques falsos sem que seja necessária intervenção humana (Obaidat et al., 2021).

Calmon et al. (2017) sugerem um *framework* para transformação probabilística de dados para redução de discriminação. Através de mapeamento aleatório, o conjunto de dados original é transformado em um novo conjunto, que é utilizado para treinar o modelo e similarmente transformar os dados em que o modelo for aplicado. Este mapeamento aleatório, portanto, deve satisfazer a métrica de controle de discriminação, o limite de distorção estabelecido e a preservação da utilidade detalhados no estudo. Os autores experimentaram a estratégia em dois classificadores padrões, respetivamente o LR e o RF, e em dois *datasets* reais (Calmon et al., 2017).

A publicação de Krasanakis et al. (2018) propõe um esquema de pré processamento, designado por *Adaptive Sensitive Reweighting* (ASR), o qual utiliza um modelo *Convex Underlying Label Error Perturbation* (CULEP) para estimar distribuições de rótulos subjacentes com os quais adapta os pesos para alcançar boa compensação entre precisão e eliminação de discriminação direta e indireta. O método assume que existe um conjunto não observável de rótulos de classe no treinamento, que se previstos podem produzir classificação parcial, respeitando um objetivo de equidade. É realizada então uma procura pelos pesos das amostras que fazem o treinamento ponderado nos dados originais, e treinam visando estes rótulos sem conhecê-los explicitamente. Para obter esses pesos, utilizou-se o CULEP, um modelo de inferência de probabilidade não linear, que pode ser treinado para converter o erro de classificação em uma probabilidade de que os rótulos estimados se aproximem dos rótulos desejados. O método então é utilizado para inferir

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

pesos de treinamento com base nas saídas do classificador e treinar novamente o classificador nestes novos pesos. O classificador utilizado pelos autores foi o LR (Krasanakis et al., 2018).

4.2 Sumário de resultados

A Tabela 4.1 apresenta as principais características das publicações, como nome dos autores, ano de publicação, objetivo do estudo, estágio do processamento, métricas de equidade e de desempenho, as técnicas propostas para melhorar a equidade e os conjunto de dados utilizados.

Tabela 4.1. Sumário dos resultados da RSL[1]

Autor(s) (ano)	**Objetivo**	**Estágio**	**Métrica de equidade**	**Métrica de desempenho**	**Técnicas**	***Datasets* utilizados**
Zafar et al. (2017)	Projetar classificadores mais justos	In	Covariância do Limite de Decisão de injustiça	Acurácia	*LR e SVM* customizados pela Covariância do Limite de Injustiça	*UCI Adult UCI Bank marketing*
Calmon et al. (2017)	Prevenir discriminação em modelos de ML	Pré	Distorção da Amostra ou Controle da discriminação	ROC	Transformação probabilística dos dados em *LR e RF*	*ProPublica; COMPAS; UCI Adult;*
Rudin e Ustunb (2018)	Apresentar alternativas confiáveis e transparentes para modelos de ML utilizados na saúde e na justiça	In	Paridade Estatística	TVP TFN *ROC*	*SLIM* *RiskSLIM*	*Obstructive Sleep Apnea: Seizure Prediction Recidivism of Prisoners Released in 1994*

[1] [1] Onde foi possível, traduzimos para português, onde não fez sentido, manteve-se os termos originais em inglês.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Autor(s) (ano)	**Objetivo**	**Estágio**	**Métrica de equidade**	**Métrica de desempenho**	**Técnicas**	***Datasets* utilizados**
Zhang et al. (2018)	Mitigar introdução de vieses em sistemas de ML	In	Paridade Estatística; Probabilidades equalizadas e Oportunidades Iguais	TFP TFN	Aprendizado adversarial em LR	*UCI Adult*
Krasanaki et al. (2018)	Mitigação de vieses em sistemas de classificação	Pré	Regra P%; Diferença da TFP e TFN entre grupos	Acurácia	*ASR + CULEP*	*UCI Adult; UCI Bank marketing; ProPublica COMPAS*
Hu et al. (2019)	Proteção de Dados demográficos	Pré	Paridade Estatística	Classificador de erro	*DFRR;* *DFKRR;* *DFGR;* *DFPCA*	*ProPublica COMPAS; UCI default of credit card; Community Crime*
Bellamy et al. (2019)	Detectar e mitigar vieses e compreender e investigar equidade em sistemas algorítmicos, em ambientes industriais e acadêmicos	Pré; In; Pós	Paridade Estatística; Probabilidades equalizadas e Oportunidades Iguais; Consistência; Distorção da amostra; Índice *Theil*	Matriz de Confusão	*LR e RF Reweighting; NN Adversarial debiasing; LR Prejudice remover; LR e RF Optimized Preprocessing; LR e RF Equal odds postprocessing ; LR e RF Disparate Impact Remover; LR e RF Calibrated equal odds postprocessing ; LR e RF Learning Fair Representations; LR e RF Reject option classification;*	*UCI Adult; UCI Stat log German Credit Data; ProPublica COMPAS*

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Autor(s) (ano)	**Objetivo**	**Estágio**	**Métrica de equidade**	**Métrica de desempenho**	**Técnicas**	***Datasets* utilizados**
Sahu e Singh (2019)	Reduzir vieses pessoais em recomendação de filmes	Pós	Diversidade na Recomendação	Qualidade da Recomendação	Modelo de Filtragem Colaborativa	*Movie Lens*
Dash et al. (2019)	Sumarização de textos que representa corretamente os diferentes grupos sociais envolvidos	Pré; In; Pós	Paridade Estatística	*ROUGE; Recall; F1*	*ClasswiseSumm; FairSumm; RefaSumm*	*Claritin tweets; US Election 2016 tweets; Me Too tweets*
Mokhtari et al. (2021)	Monitoramento do uso de dados para garantir conformidade com regras de equidade e transparência em plataformas de *Big Data*	-	Equidade X Transparência X Contratos	Eficiência e Escalabilidade	FATIT	*Dataset* sintético
Edizel et al. (2020)	Mitigar vieses algorítmicos em sistemas de recomendação	Pós	μ_1 e μ_2	Precisão e *Recall*	FaiRecSys	*Movie Lens; Reddit*
Obaidat et al. (2021)	Minimizar os efeitos da adulteração de dados em sistemas de ML hackeados	Pré	-	Acurácia	MAIRST + CNN (*Tensor Flow)*	*Fashion MNIST*
Lin et al. (2021)	Identificar intervenções baseadas em IA para redução de danos de vieses implícitos no processo de recrutamento e seleção de candidatos	-	-	-	*Eightfold, Entelo, IBM Watson, Blendoor, Interviewing.io, Pymetrics, Textio Hire, Hire Vue, Tengai, Equal Reality e Vantage Point*	-

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Autor(s) (ano)	**Objetivo**	**Estágio**	**Métrica de equidade**	**Métrica de desempenho**	**Técnicas**	***Datasets* utilizados**
Valdivia et al. (2021)	Explorar limites entre acurácia e equidade em classificadores	In	TFP	Erro médio G	*NSGA-II +* Arvores de decisão	*UCI Adult; UCI Stat log; German Credit Data; ProPublicCOMPAS; ProPublicViolent; Ricci*
Ahmed et al. (2021)	Mitigar vieses presentes no conjunto de dados US Employment Demographics com técnicas do AI 360	Pós	Paridade Estatística, Probabilidades equalizadas e Oportunidades Iguais, Índice Theil	Acurácia	LR e RF Equal odds postprocessing ; LR e RF Calibrated equal odds postprocessing	US Employment Demographics

Fonte: Elaboração própria.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

5 DISCUSSÃO

Embora o objetivo dos estudos apresentados seja de forma geral o aprimoramento da equidade em sistemas algorítmicos, após a leitura de cada publicação podemos verificar que algumas delas foram ainda mais específicas em seus objetivos, em relação ao tipo de sistema algorítmico estudado ou em relação a ênfase em determinadas áreas de negócios.

Neste sentido, Rudin e Ustunb (2018) procuram melhorar a equidade em sistemas de ML para as áreas de negócios relacionadas à saúde e justiça enquanto Lin et al. (2021) apresentam ferramentas que podem ser utilizadas especificamente para tratar a equidade durante o processo de recrutamento e seleção.

Em relação ao tipo de sistemas, destacamos Edizel et al. (2020) e Sahu e Singh (2019), que buscaram a melhoria da equidade na recomendação de conteúdo pós processando os resultados das recomendações iniciais, e Dash et al. (2019), que oferecem solução para sumarização de textos que representem igualmente todos os grupos presentes nos dados. Já Obaidat et al. (2021), buscam atingir um objetivo ainda mais particular, para minimizar vieses e preconceitos provenientes especificamente de invasões em sistemas de aprendizado de máquina.

Percebemos também que algumas publicações buscaram soluções para melhorar especificamente a utilização e disponibilização de dados demográficos, oferecendo soluções de aprendizado distribuído (Hu et al., 2019), e monitoramento de utilização de dados pessoais em plataformas de *Big Data* (Mokhtari et al., 2021).

5.1 Métricas de equidade e desempenho

Acerca das métricas, ou noções de equidade empregues em cada proposta, de acordo com a Figura 5.1, podemos perceber uma preferência pela utilização da métrica paridade estatística.

Seis publicações buscaram atingir esta métrica, também chamada de *Disparated impact, Equal acceptance rate*, *demografic parity* ou ainda apenas equidade de grupo e uma publicação buscou ainda atender à P*% rule*, que consideramos uma derivação da paridade estatística.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Esta medida busca igualar os resultados entre grupos protegidos e não protegidos, e é independente dos rótulos reais, o que se mostra uma vantagem quando tais rótulos não estão disponíveis, como nos casos dos atributos sensíveis.

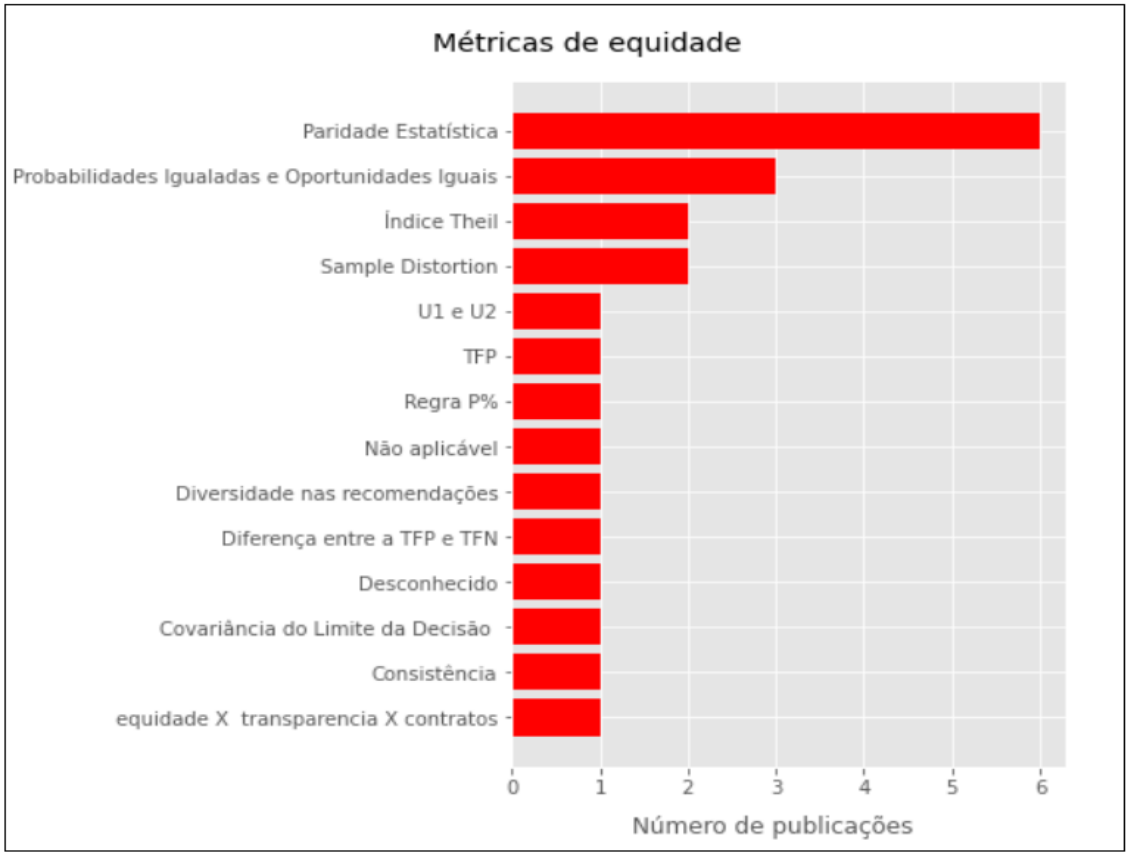

Figura 5.1. Métricas de equidade
Fonte: Elaboração própria.

Para Friedler et al. (2019), as métricas de *fairness* tendem a se correlacionar, e de uma forma geral, com as taxas de erro de cada grupo, e a métrica *Disparate Impact* (correspondente a paridade estatística), já se pode garantir um bom trabalho ao buscar impor a equidade em sistemas de aprendizado de máquina.

Porém, embora esta métrica pareça ser desejável e simples, ela foi criticada no famoso artigo da área, "*Fairness Through Awareness*" de Dwork et al. (2012), que argumenta a inadequação da medida, justificando que embora possamos garantir a paridade estatística entre grupos, se analisarmos do ponto de vista de um indivíduo específico destes grupos, o resultado pode ser injusto. Para explicar a crítica, a publicação cita a profecia do

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

autorrealizável, onde em um caso de recrutamento e seleção por exemplo, membros não qualificados de um grupo protegido são escolhidos a fim de justificar futura discriminação contra aquele mesmo grupo.

A segunda métrica mais utilizada, presente em três publicações foi *Equalized Odds*, ou em português, probabilidades equalizadas. Esta métrica estipula que ambos os grupos tenham a mesma taxa de Verdadeiros Positivos (TVP) e de Falsos Positivos (TFP), e a derivação dela, chamada de *Equal opportunities* ou oportunidades iguais, onde os grupos devem ter apenas a mesma taxa de VP, ambas propostas por Hardt et al. (2016).

Segundo Pessach e Shmueli (2020), a eficácia da medida *Equalized Odds* foi comprovada na utilização do *dataset* Pro Publica do famoso caso COMPAS, onde observou que apesar de a acurácia ser similar entre os dois grupos (afro americanos e caucasianos) a taxa de falsos positivos (TFP) entre afro americanos era duas vezes maior que a taxa de falsos positivos entre caucasianos, provando que o sistema errou duas vezes mais ao prever a reincidência em crimes para pretos do que para brancos.

Para Speicher et al. (2018), qualquer noção de equidade de grupo, que engloba tanto a paridade estatística quanto a *equalized odds*, não considera o tamanho dos diferentes grupos e por isso não deve ser considerada ideal.

Dentre as 15 publicações elegíveis nesta pesquisa, apenas a publicação de Calmon et al. (2017) utilizou métrica do grupo de equidade individual com objetivo de equidade algorítmica em sua respetiva proposta.

A covariância do limite de decisão, proposta por Zafar et al. (2017), foi considerada relevante no levantamento do estado da arte feito por Krasanakis et al. (2018) que a utilizaram para efeito de comparação em seu estudo. Da mesma forma, observa-se que a proposta de Bellamy et al. (2019) incorporou a ideia de Calmon et al. (2017) através da métrica de distorção da amostra.

Em relação a métricas de desempenho, observamos uma certa preferência pela acurácia, pois quatro publicações optaram por medir o desempenho do modelo desta forma, conforme apresentado na Figura 5.2. Esta medida reflete qual foi a porcentagem de

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

acertos do sistema, e é calculada pelo total das predições corretas (VPs + VNs) dividido pelo número total de predições (VPs + FPs + VNs + FNs).

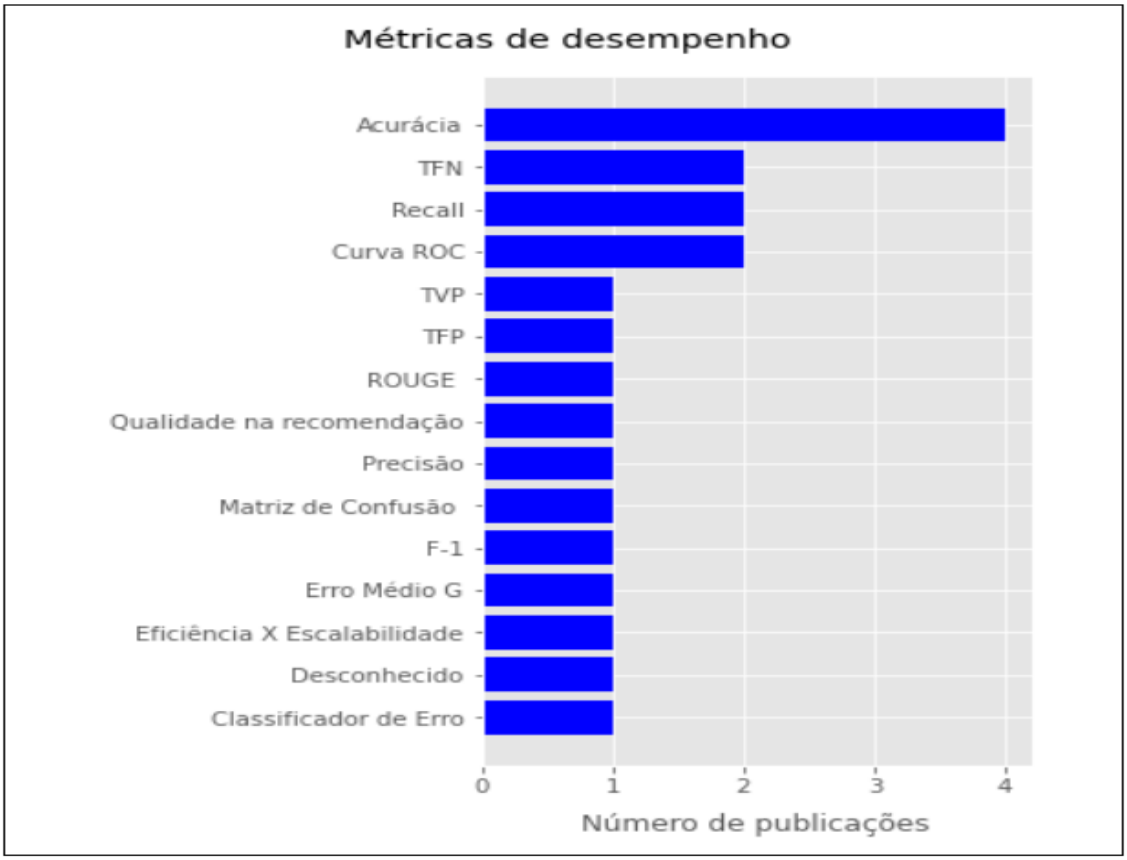

Figura 5.2. Métricas de desempenho
Fonte: Elaboração própria.

Enquanto a maioria das publicações utilizou métricas de desempenho relacionadas com a matriz de confusão, Hu et al. (2019) e Valdivia et al. (2021) focaram em medir os erros dos sistemas para avaliar a sua precisão e desempenho.

Três publicações utilizaram ainda métricas de equidade e de desempenho específicas e inerentes aos objetivos de cada estudo. No estudo de Dash et al. (2019), enquanto a equidade é imposta pela métrica de paridade estatística, o desempenho utiliza a medição *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE), que avalia a qualidade de resumos produzidos por algoritmos. No estudo de Sahu e Singh (2019), a métrica de desempenho do sistema, é a qualidade nas recomendações, que leva em consideração a avaliação de usuários, entre outros pontos. E a medição de equidade é entendida neste

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

caso como a diversidade introduzida nas recomendações finais, que buscam aliviar o viés pessoal de cada usuário durante a recomendação de filmes. Já para Mokhtari et al. (2021), a eficiência em termos de tempo de execução ao detectar violações de conformidade e a escalabilidade em termos de quantidade de dados analisados medem o desempenho do sistema.

O estudo de Lin et al. (2021) não trouxe detalhes de como as ferramentas apresentadas foram avaliadas, ou quais foram as métricas de desempenho ou de equidade utilizadas durante o desenvolvimento.

5.2 Técnicas e estágios de processamento

Verificamos que algumas publicações apresentam mais do que uma técnica em mais do que um estágio de intervenção para o aprimoramento da equidade. Portanto, minuciamos cada estudo de acordo com as propostas e algoritmos utilizados, e desenvolvemos a Figura 5.3 para melhor visualização das soluções apresentadas e o estágio do processamento em que foram empregadas.

Através das 15 publicações selecionadas, identificamos 48 técnicas diferentes para aprimoramento de equidade em aprendizado de máquina, dentre as quais, 9 delas atuam adaptando os algoritmos e impondo restrições de equidade em suas funções objetivo durante o estágio de processamento, 9 reclassificam os resultados para que as métricas sejam cumpridas no pós-processamento, e 17 modificam o conjunto de dados utilizados para treinar o modelo durante o estágio de pré-processamento.

Neste sentido, Pessach e Shmueli (2020) ponderam que enquanto as técnicas de pré e pós processamento podem ser aplicadas a qualquer tipo de algoritmo, elas podem prejudicar a transparência e a explicação dos resultados. Já os mecanismos implementados no momento do processamento podem explicitamente impor a compensação que desejam fazer entre acurácia e equidade na função objetivo, porém estão intimamente presos ao algoritmo utilizado sendo um pouco mais complexos de serem desenvolvidos. A atuação durante o processamento também pode comprometer a explicação dos resultados, pois deve explicar com clareza, para compreensão de humanos, como o algoritmo realiza a busca por equidade e por desempenho e ainda as compensações envolvidas.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Especificamente durante o pós-processamento, dois indivíduos que são semelhantes em todas as características, exceto no grupo a que pertencem, por exemplo raça, podem ser tratados de forma diferente, exigindo que o tomador de decisões possua informações sensíveis de cada indivíduo, o que é proibido por lei (Pessach & Shmueli, 2020).

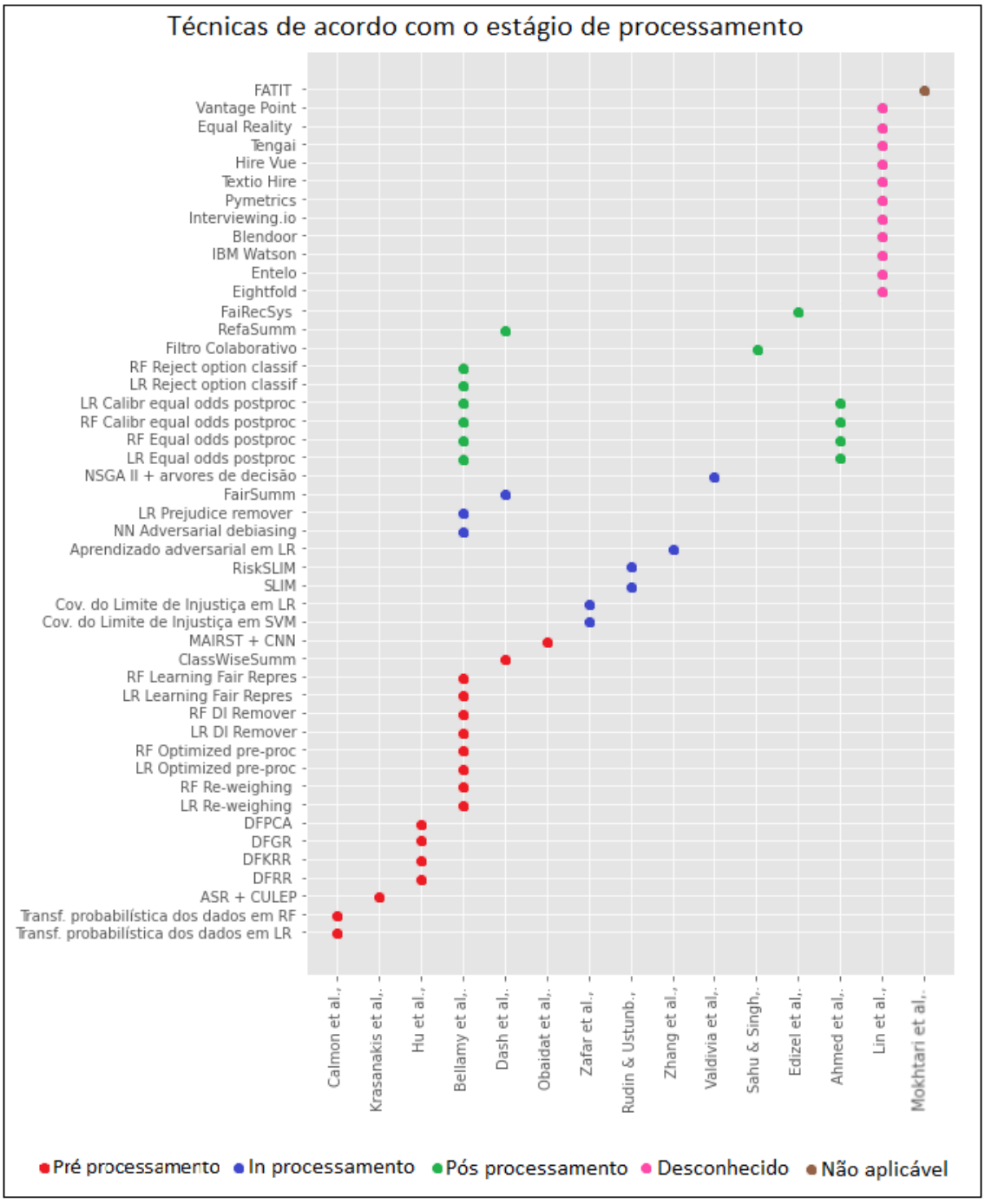

Figura 5.3. Técnicas e estágios de processamento

Fonte: Elaboração própria.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Continuando a análise da Figura 5.3, observamos também que algoritmos de *Logístic Regression* (LR) e *Random Forest (RF)*, adequados de acordo com a proposta de cada estudo foram os mais utilizados, 11 e 8 vezes respectivamente, ambos os algoritmos de aprendizado supervisionado.

Modelos baseados em LR têm a vantagem de serem interpretáveis e explicáveis, com menor probabilidade de *overfitting* e aplicáveis para predições multiclasse, enquanto modelos baseados em RF apresentam alta acurácia quando comparado a outros algoritmos, mas podem ser mais difíceis de interpretar e exigem maior complexidade no treinamento, (Data Camp, 2022).

Apenas duas publicações, Obaidat et al. (2021) e Bellamy et al. (2019) citam a utilização de técnicas de aprendizado não supervisionado como redes neurais (NN) e redes neurais convolucionais (CNN).

A publicação de Lin et al. (2021), apresentou 11 ferramentas comerciais já em utilização para aprimoramento da equidade durante o processo de recrutamento e seleção, porém estas não foram classificadas segundo o momento de atuação, pois o estudo não apresentou detalhes de como elas foram desenvolvidas. Já a plataforma FATIT proposta por Mokhtari et al. (2021) não se relaciona ao desenvolvimento de modelos, mas sim a monitoramento da utilização de dados, por isso foi considerada como não aplicável a classificação em relação aos estágios de desenvolvimento.

5.3 Conjuntos de dados

Sendo a escolha do conjunto de dados fundamental para o aprendizado do modelo, desenvolvemos também um gráfico com os conjuntos mais utilizados nas publicações elegíveis, que se apresenta na Figura 5.4.

Podemos verificar que dentre as publicações que compreenderam a RSL, os *datasets UCI Adult* e *ProPublica COMPAS* foram utilizados seis e quatro vezes respectivamente, sendo os mais utilizados, demonstrando um padrão de referência na exploração de resultados de compensação em equidade e desempenho.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

O *dataset* UCI Adult compreende informações extraídas do censo de 1994 dos Estados Unidos, e é formado por 14 atributos, como idade, gênero, ocupação, educação, raça, e renda por exemplo, sendo principalmente utilizado nos estudos de equidade para comparar a interferência de raça e gênero na renda.

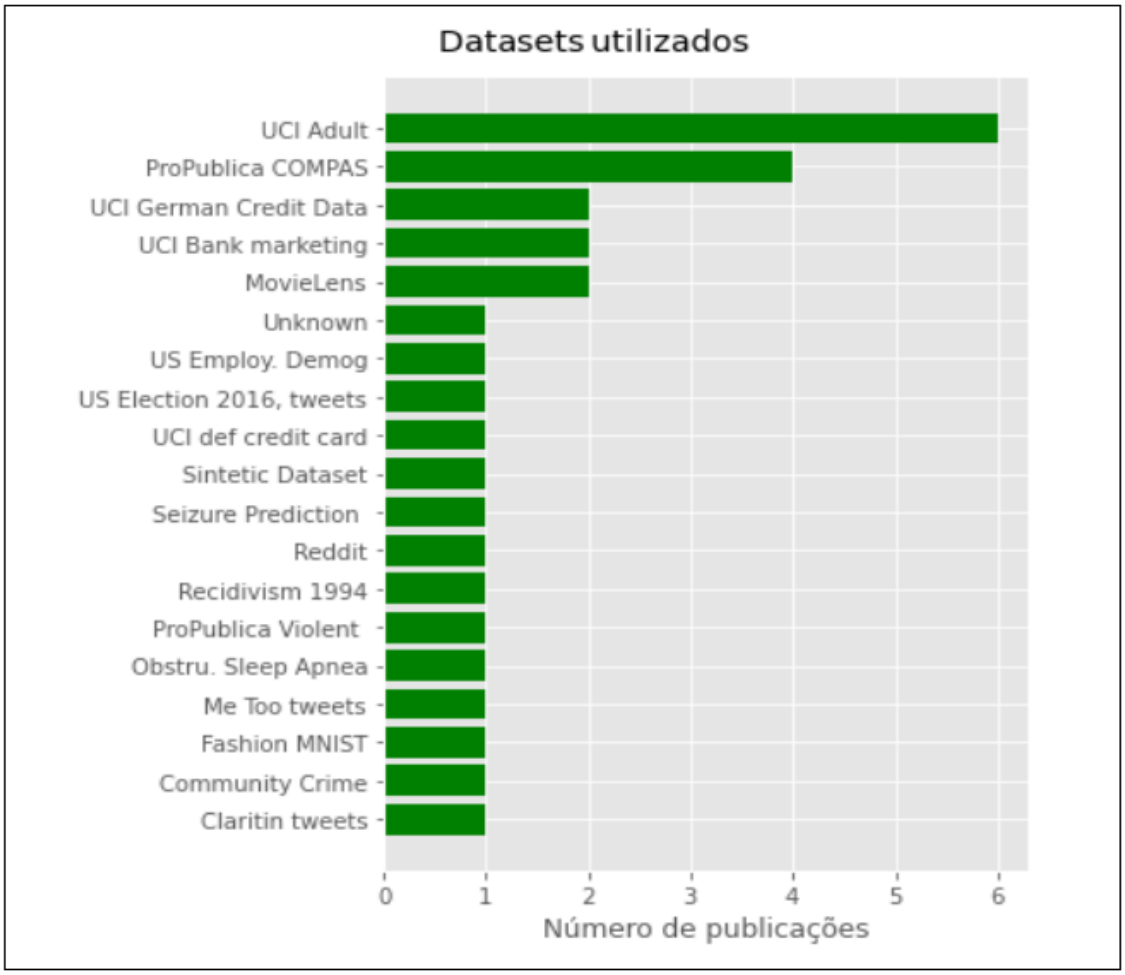

Figura 5.4. Datasets utilizados

Fonte: Elaboração própria.

Já o *dataset* Pro Publica COMPAS, traz o histórico criminal de réus do condado de Broward no estado da Florida nos Estados Unidos, durante 2013 e 2014 e contém atributos como tempo de prisão, idade, endereço, raça, gênero, renda etc., e a pontuação de risco do sistema COMPAS, utilizado nos estudos principalmente para compreender a relação da raça nos riscos atribuídos pelo sistema.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Apenas Obaidat et al. (2021) fizeram uso de um *dataset* com imagens, o Fashion MNIST, um conjunto de dados que contém 70.000 imagens de peças de vestuário, como camisetas, calças e botas. A publicação procura minimizar efeitos da adulteração de dados em sistemas invadidos por adversários, e para isso treinaram uma rede neural convolucional com a dinâmica MAIRST proposta no estudo, que deveria reconhecer e classificar as peças.

A publicação de Lin et al. (2021) não detalhou os *datasets* em que as ferramentas citadas foram treinadas, por isso são desconhecidas para esta classificação.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

6 CONCLUSÃO

O aprendizado de máquina que possibilita evoluções indiscutíveis, também possui efeitos colaterais que podem repercutir de maneira negativa ao convívio social. A tecnologia que desencadeia estes efeitos, é a mesma capaz de revelar tais desequilíbrios e injustiças para possibilitar a busca pela correção de erros do passado.

Como diz Kahneman (2013), algoritmos nem sempre são os vilões das decisões, e especialistas erram mais do que os algoritmos, pois "seres humanos são incorrigivelmente inconsistentes em fazer julgamentos sumários de informações complexas". O objetivo, portanto, deve estar em "eliminar o viés humano, e não apenas camuflá-lo com tecnologia" (O'Neil, 2020).

O debate sobre os impactos sociais das decisões algorítmicas tendenciosas é crescente e diversas iniciativas vêm sendo desenvolvidas para atenuar prejuízos gerados em grupos e subgrupos específicos, mas como citam Simon et al. (2020), o dever coletivo não está em apenas detectar e combater vieses em sistemas computadorizados, mas também em abordar e remediar suas origens sociais.

6.1 Síntese do trabalho

Nesta pesquisa discorremos sobre como preconceitos são introduzidos em aprendizado de máquina, apresentamos diversas definições e métricas de equidade algorítmica, e indicamos algumas abordagens socio técnicas sugeridas para atenuação dos impactos gerados em esferas que extrapolam o campo de atuação da ciência da computação.

Dentre tais abordagens socio técnicas, destacamos a necessidade de regulamentações mais específicas, em forma de legislação, que pressionem empresas a adotarem uma conduta ética que norteiem a implementação consciente da inteligência artificial.

O objetivo principal desta dissertação foi apresentar soluções técnicas para este problema, e através de uma revisão sistemática da literatura, conhecemos as propostas de 15 publicações recentes, que sugerem ferramentas e abordagens técnicas de como melhorar a equidade algorítmica. Estes estudos foram descritos de forma narrativa, sumarizados

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

em tabela e discutidos de acordo com seus objetivos, métricas, conjunto de dados, técnicas utilizadas e o estágio do desenvolvimento em que foram implementadas.

6.2 Contribuições para a indústria e para a academia

Consideramos que a correta compreensão das métricas de equidade apresentadas é de extrema importância para o desenvolvimento de novos sistemas algorítmicos e de melhoria dos já existentes. É através delas que se será capaz de medir a discriminação presente nos dados e de buscar resultados diferentes para o futuro.

Mapeamos 15 publicações que se propõem a solucionar diferentes objetivos, ainda mais afunilados dentro do subcampo de equidade em ML, como a mitigação de vieses em sistemas de classificação, de recomendação, de sumarização de textos, a mitigação de vieses inseridos através de invasões destes sistemas, e a atenção a correta utilização de dados demográficos.

Identificamos 48 técnicas de modificação de dados, de algoritmos e de resultados, que atendam a métricas de equidade algorítmica, por exemplo ferramentas para exploração e compreensão de métricas como o AI 360, ferramentas comerciais já implementadas para o processo de seleção e recrutamento sem preconceitos, e plataformas para desenvolvimento de modelos que preservem a privacidade de dados e monitorem a utilização deles de acordo com diretrizes de privacidade.

Através da análise e sumarização destas propostas, observamos que a maioria das publicações buscou atingir a equidade através da paridade demográfica, ou equidade de grupo como principal métrica estatística.

Observamos também que a maioria das publicações aplicou suas técnicas durante o tratamento dos dados, no estágio de pré-processamento do modelo, e que houve considerável utilização dos algoritmos LR e RF nas técnicas identificadas.

Os conjuntos de dados *UCI Adult*, e o *ProPublica COMPAS*, que contém dados demográficos da sociedade americana de 28 e 8 anos atrás respectivamente, foram amplamente utilizados nas pesquisas. Uma grande parte dos trabalhos anteriores eleitos,

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

utilizou, de entre outros, um destes dois conjuntos de dados, apresentando diversas abordagens possíveis de melhoria da equidade em sistemas de ML.

Os caminhos que apresentam poderão ser seguidos pela indústria em novas aplicações e pela academia em novas investigações de forma idêntica ou adaptada. Um entendimento aprofundado de como cada um dos referidos trabalhos passado abordaram cada conjunto de dados poderá ser a base de partida para a definição de um caminho com vista a uma melhor equidade.

Pretendemos difundir e popularizar o tema e as soluções e definições aqui expostas, posto que a execução de tais soluções, necessitam de profunda compreensão sobre o contexto em que o sistema será utilizado, e dos grupos e indivíduos afetados por ele.

Desta forma, acreditamos que esta RSL pode ajudar a estabelecer padrões sobre experimentações em equidade algorítmica como em relação a métrica, técnicas e *datasets* utilizados, trazendo uma visão mais unificada de conceitos e métodos, no qual pesquisas futuras podem se basear para compreender o que já está estabelecido e o que ainda merece refinamento em relação ao tema.

6.3 Limitações

O tema ainda é relativamente recente e, por isso, o número de publicações final pode não ser considerado muito significativo. Ainda, no decorrer da pesquisa reconhecemos que o resultado dos estudos selecionados se mostrou não homogêneo em relação ao objetivo específico dentro de equidade em ML que cada autor buscou resolver. Isso nos trouxe dificuldade para classificar os estudos em relação aos parâmetros que pré-definimos como métricas, fases do desenvolvimento e algoritmos em que procuramos reconhecer referências.

6.4 Trabalhos futuros

Para seguir a linha de pesquisa do tema, sugerimos a experimentação prática das técnicas apresentadas nesta dissertação, porém com a utilização de *datasets* diferentes dos

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

utilizados nas publicações originais selecionadas, que apresentem varáveis e atributos de outros países, e explorem a equidade algorítmica em diferentes grupos sociais.

Recomendamos também a experimentação das métricas estatísticas aqui relacionadas, para comparação de resultados obtidos em cada uma delas, também em *datasets* diferentes dos apresentados nesta dissertação.

Relacionando os inúmeros exemplos encontrados na literatura sobre a dificuldade de identificação de pessoas de pele negra, percebemos também a necessidade de uma maior exploração de dados de reconhecimento facial em estudos de aprimoramento de equidade algorítmica uma vez que esta tecnologia se mostra cada vez mais inserida em situações rotineiras de milhares indivíduos em diversos países.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

REFERÊNCIAS

ACM. (2021). *ACM FAccT*. https://facctconference.org/index.html

Adams, J., & Hagras, H. (2020). A Type-2 Fuzzy Logic Approach to Explainable AI for regulatory compliance, fair customer outcomes and market stability in the Global Financial Sector. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Fuzzy Systems (FUZZ-IEEE), 2020 IEEE International Conference on* (pp. 1–8). IEEE. https://doi.org/10.1109/FUZZ48607.2020.9177542

Ahmed, S., Athyaab, S. A., & Muqtadeer, S. A. (2021). Attenuation of Human Bias in Artificial Intelligence: An Exploratory Approach. In *2021 6th International Conference on Inventive Computation Technologies (ICICT), Inventive Computation Technologies (ICICT), 2021 6th International Conference on* (pp. 557–563). IEEE. https://doi.org/10.1109/ICICT50816.2021.9358507

Altman, M., Wood, A., & Vayena, E. (2018). A Harm-Reduction Framework for Algorithmic Fairness. *IEEE Security and Privacy*, *16*(3), 34–45. https://doi.org/10.1109/MSP.2018.2701149

Aslaoui Mokhtari, K., Benbernou, S., Ouziri, M., Lahmar, H., & Younas, M. (2021). A monitoring framework for transparency and fairness in big data plataform. *WILEY INTERDISCIPLINARY REVIEWS-DATA MINING AND KNOWLEDGE DISCOVERY*. https://doi.org/10.1002/cpe.6069

b-on. (2022). *Quem somos*. https://www.b-on.pt/quem-somos/

Barocas, S., Hardt, M., & Narayanan, A. (2019). Limitations and Opportunities Solon Barocas, Moritz Hardt, Arvind Narayanan. *Fairness and Machine Learning: Limitation and Oppotunities*. https://fairmlbook.org

Barocas, Solon, & Selbst, A. D. (2016). Big Data' S Disparate Impact. *California Law Review*, *104*(671), 671–732. https://doi.org/http://dx.doi.org/10.15779/Z38BG31

Belfo, F. P., Silva, P. R. da, & Afonso, C. M. (2022). O impacto organizacional e social da inteligência artificial. In *Sistemas de Informação: Diagnósticos e Prospectivas* (1ª, pp. 146–164). Edições Sílabo.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovie, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM JOURNAL OF RESEARCH AND DEVELOPMENT*, *63*(4–5), 4. https://doi.org/10.1147/JRD.2019.2942287

Bellamy, R. K. E., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., Zhang, Y., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., & Mehta, S. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4–5). https://doi.org/10.1147/JRD.2019.2942287

Brandão, L., Belfo, F. P., & Silva, A. (2021). Wavelet-based cancer drug recommender system. *Procedia Computer Science, Communications in Computer and Information Science*, *181*, 487–494. https://doi.org/https://doi.org/10.1016/j.procs.2021.01.194

Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292. https://doi.org/10.1007/s10618-010-0190-x

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems*, *2017-Decem* (Nips), 3993–4002.

Camp, D. (2022). *Machine Learning Cheat Sheet*. https://s3.amazonaws.com/assets.datacamp.com/email/other/ML+Cheat+Sheet_2.pdf

Cardenas, S., & Vallejo-Cardenas, S. F. (2019). Continuing the Conversation on How Structural Racial and Ethnic Inequalities Affect AI Biases. In *2019 IEEE International Symposium on Technology and Society (ISTAS), Technology and Society (ISTAS), 2019 IEEE International Symposium on* (pp. 1–7). IEEE. https://doi.org/10.1109/ISTAS48451.2019.8937853

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Chouldechova, A., & Roth, A. (2018). *The Frontiers of Fairness in Machine Learning*. 1–13. http://arxiv.org/abs/1810.08810

Dash, A., Shandilya, A., Biswas, A., Ghosh, K., Ghosh, S., & Chakraborty, A. (2019). Summarizing User-generated Textual Content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW). https://doi.org/10.1145/3359274

DOMO. (2022). *Data Never Sleeps 9.0*. https://www.domo.com/learn/infographic/data-never-sleeps-9

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255

Edizel, B. (1), Bonchi, F. (2), Panisson, A. (2), Hajian, S. (3), & Tassa, T. (4). (2020). FaiRecSys: mitigating algorithmic bias in recommender systems. *International Journal of Data Science and Analytics*, *9*(2), 197–213. https://doi.org/10.1007/s41060-019-00181-5

Esteves, R., Belfo, F. P., & Trigo, A. (2021). Fatores mais valorizados no turismo rural: Uma análise de comentários de clientes portugueses numa plataforma de reservas. *CAPSI 2021 Proceedings*, 26.

Fairware 22. (2022). https://fairwares.github.io/

FAT/ML. (2022). https://www.fatml.org/

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *2015-Augus*, 259–268. https://doi.org/10.1145/2783258.2783311

Freire, G., Pansani, T. de S., & Harrad, D. (2015). Principais itens para relatar Revisões sistemáticas e Meta-análises: A recomendação PRISMA. *Epidemiologia e Serviços de Saúde*, *24*(2), 335–342. https://doi.org/10.5123/s1679-49742015000200017

Friedler, S. A., Choudhary, S., Scheidegger, C., Hamilton, E. P., Venkatasubramanian,

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

S., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 329–338. https://doi.org/10.1145/3287560.3287589

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Gillingham, P. (2019). Decision Support Systems, Social Justice and Algorithmic Accountability in Social Work: A New Challenge. *Practice*, *31*(4), 277–290. https://doi.org/10.1080/09503153.2019.1575954

Google. (2022). *Google Trends*. https://trends.google.com.br/trends/?geo=BR

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *Nips*, 3323–3331.

Hu, H., Liu, Y., Wang, Z., & Lan, C. (2019). A Distributed Fair Machine Learning Framework with Private Demographic Data Protection. In *2019 IEEE International Conference on Data Mining (ICDM), Data Mining (ICDM), 2019 IEEE International Conference on* (pp. 1102–1107). IEEE. https://doi.org/10.1109/ICDM.2019.00131

Impellizzeri, F. M. (2012). Systematic review and meta-analysis: A primer. *International Journal of Sports Physical Therapy*, *7*(5), 493–503. https://www.researchgate.net/publication/232612227_Systematic_review_and_meta-analysis_A_primer

Kahneman, D. (2013). *Thinking, fast and slow* (F. S. Giroux (Ed.)).

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. In *Knowledge and Information Systems* (Vol. 33, Issue 1). https://doi.org/10.1007/s10115-011-0463-8

Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., & Kompatsiaris, Y. (2018). Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. *The*

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018, *2*, 853–862. https://doi.org/10.1145/3178876.3186133

Kristiadi, D. P. (1), Sunarya, P. A. (2), Ismanto, M. (3), Dylan, J. (3), Santoso, I. R. (3), & Warnars, H. L. H. S. (4). (2020). Framework for developing algorithmic fairness. *Bulletin of Electrical Engineering and Informatics*, *9*(4), 1550–1557. https://doi.org/10.11591/eei.v9i4.2028

Lin, Y.-T., Hung, T.-W., & Huang, L. T.-L. (2021). Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias. *Philosophy & Technology*, *34*(1), 65–90. http://10.0.3.239/s13347-020-00406-7

Loureiro, A., Lourenço, J., Costa, E., & Belfo, F. (2014). Indução de Árvores de Decisão na Descoberta de Conhecimento: Caso de Empresa de Organização de Eventos. In *VI Congresso Internacional de Casos Docentes em Marketing Público e Não Lucrativo*.

Lupo, G. (2019). Regulating (Artificial) Intelligence in Justice: How Normative Frameworks Protect Citizens from the Risks Related to Ai Use in the Judiciary. *European Quarterly of Political Attitudes and Mentalities*, *8*(2).

McCradden, M. D., Joshi, S., Anderson, J. A., Mazwi, M., Goldenberg, A., & Shaul, R. Z. (2020). Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *Journal of the American Medical Informatics Association*, *27*(12), 2024–2027. https://doi.org/10.1093/jamia/ocaa085

Medeiros, N. R. F. V. (2019). USO DA INTELIGÊNCIA ARTIFICIAL NO PROCESSO DE TOMADA DE DECISÕES JURISDICIONAIS: Uma análise sob a perspectiva da teoria normativa da compartição. In *PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS*. PUC Minas.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *FAT* 2019* -

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, *Figure 2*, 220–229. https://doi.org/10.1145/3287560.3287596

Ntoutsi, E. (1), Gadiraju, U. (1), Iosifidis, V. (1), Nejdl, W. (1), Staab 12,13), S. (1, Fafalios, P. (2), Vidal, M.-E. (3), Ruggieri, S. (4), Turini, F. (4), Papadopoulos, S. (5), Krasanakis, E. (5), Kompatsiaris, I. (5), Kinder-Kurlanda, K. (6), Wagner, C. (6), Karimi, F. (6), Fernandez, M. (7), Alani, H. (7), Berendt 9), B. (8, Kruegel, T. (10), … Tiropanis, T. (12). (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3). https://doi.org/10.1002/widm.1356

O'Neil, C. (2020). *Algoritmos de destruição em massa: como o big data aumenta a desigualdade e ameaça a democracia* (Editora Rua do Sabão (Ed.); 1st ed.).

Obaidat, M., Singh, N., & Vergara, G. (2021a). Artificial Intelligence Bias Minimization Via Random Sampling Technique of Adversary Data. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference, CCWC 2021*, 1226–1230. https://doi.org/10.1109/CCWC51732.2021.9375929

Obaidat, M., Singh, N., & Vergara, G. (2021b). Artificial Intelligence Bias Minimization Via Random Sampling Technique of Adversary Data. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), Computing and Communication Workshop and Conference (CCWC), 2021 IEEE 11th Annual* (pp. 1226–1230). IEEE. https://doi.org/10.1109/CCWC51732.2021.9375929

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., Mcdonald, S., … Mckenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *The BMJ*, *372*. https://doi.org/10.1136/bmj.n160

Pessach, D., & Shmueli, E. (2020). Algorithmic Fairness. *ArXiv:2001.09784*. https://doi.org/10.1257/pandp.20181018

Pimenta, C., Ribeiro, R., Sá, V., & Belfo, F. P. (2018). Fatores que Influenciam o Sucesso

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Escolar das Licenciaturas numa Instituição de Ensino Superior Portuguesa. In *Atas da 18ª Conferência da Associação Portuguesa de Sistemas de Informação (CAPSI 2018) Associação Portuguesa de Sistemas de Informação*. Associação Portuguesa de Sistemas de Informação.

Pimenta, D., Teles, M., Belfo, F. P., & Trigo, A. (2022). Medication recommendation in cancer treatment based on cell line similarity. *Book of Abstracts of the CENTERIS 2022, Conference on ENTERprise Information Systems*.

Pimenta, P., Belfo, F., & Trigo, A. (2011). Study the Impact of Booking. com User Scores and Reviews in Hotel Management. In M. M. Cruz-Cunha, J. Varajão, P. Powell, & R. Martinho (Eds.), *Book of abstracts of the CENTERIS 2011, Conference on ENTERprise Information Systems* (Vol. 30, pp. 8–9).

Rivas, P. (2020). AI Orthopraxy: Towards a Framework for That Promotes Fairness. *ISTAS*, 80–84.

Rudin, C., & Ustunb, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, *48*(5), 449–466. https://doi.org/10.1287/inte.2018.0957

Russell, J. (2020). Machine Learning Fairness in Justice Systems: Base Rates, False Positives, and False Negatives. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Machine Learning and Applications (ICMLA), 2020 19th IEEE International Conference on, ICMLA* (pp. 817–820). IEEE. https://doi.org/10.1109/ICMLA51294.2020.00133

Sahu, S., & Singh, S. K. (2019). Ethics in AI: Collaborative filtering based approach to alleviate strong user biases and prejudices. In *2019 Twelfth International Conference on Contemporary Computing (IC3), Contemporary Computing (IC3), 2019 Twelfth International Conference on* (pp. 1–6). IEEE. https://doi.org/10.1109/IC3.2019.8844875

Schneider, V. (2020). Locked out by big data: How big data, algorithms and machine learning may undermine housing justice. *Columbia Human Rights Law Review*, *52*(1), 251–305.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Seiça, A., Trigo, A., & Belfo, F. P. (2019). LexiNB - Uma Abordagem Bietápica de Classificação de Sentimentos em Tweets Relacionados com as Autoridades Fiscais Portuguesas. *Proceedings of the 19.ª Conferência Da Associação Portuguesa de Sistemas de Informação (CAPSI'2019) Held in Lisboa, Portugal, 11-12 October 2019. Paper 5.*

Sereday, S., & Cui, J. (2017). Using Machine Learning to Predict Future TV Ratings. *Nielsen Journal of Measurement*, *1*(February), 1–13.

Simon, J., Wong, P.-H., & Rieder, G. (2020). Algorithmic bias and the Value Sensitive Design approach. *Internet Policy Review*, *ume 9*(Issue 4). https://doi.org/10.14763/2020.4.1534

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). *A Unified Approach to Quantifying Algorithmic Unfairness*. 2239–2248. https://doi.org/10.1145/3219819.3220046

Srinivasan, R., & Chander, A. (2021). Biases in AI systems. *Communications of the ACM*, *64*(8), 44–49. https://doi.org/10.1145/3464903

Stephens-Davidowitz, S. (2018). *Todo mundo mente: O que a internet e os dados dizem sobre quem realmente somos* (Auta Books editora (Ed.)).

Tal, A. S., Batsuren, K., Bogina, V., Giunchiglia, F., Hartman, A., Loizou, S. K., Kuflik, T., & Otterbacher, J. (2019). "End to End" Towards a Framework for Reducing Biases and Promoting Transparency of Algorithmic Systems. In *2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Semantic and Social Media Adaptation and Personalization (SMAP), 2019 14th International Workshop on* (pp. 1–6). IEEE. https://doi.org/10.1109/SMAP.2019.8864914

Valdivia, A., Sánchez-Monedero, J., & Casillas, J. (2021). How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. *International Journal of Intelligent Systems*, *36*(4), 1619–1643. https://doi.org/10.1002/int.22354

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings - International Conference on Software Engineering*, 1–7.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

https://doi.org/10.1145/3194770.3194776

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, *54*.

Završnik, A. (2020). Criminal justice, artificial intelligence systems, and human rights. *ERA Forum*, *20*(4), 567–583. https://doi.org/10.1007/s12027-020-00602-0

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *30th International Conference on Machine Learning, ICML 2013*, *28*(PART 2), 1362–1370.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. https://doi.org/10.1145/3278721.3278779

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt