Submission date: 07/05/2025
Resubmission date: 07/09/2025
Camera ready submission date: 22/01/2026

1st round notification: 29/08/2025
2nd round notification: 10/12/2025
Available online: 03/02/2026
Publication date: 03/02/2026

Section: regular article

Equilíbrio entre Justiça e Desempenho em Sistemas de Aprendizado de Máquina: Uma Revisão da Literatura

Fairness and Performance Trade-off in Machine Learning Systems: A Literature Review

Daniel Bolsonaro Vaz Filho[1], **Luciano Antonio Digiampietri**[2]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
São Paulo, SP – Brasil

[2]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
São Paulo, SP – Brasil

danielfilho@usp.br, digiampietri@usp.br

Abstract. *This study examines algorithmic unfairness in automated decision making systems, emphasizing equity and transparency given their impact. Machine learning algorithms often reproduce data biases, leading to discriminatory outcomes. Several mitigation strategies have been proposed, including fairness evaluation methods and bias detection tools. This review analyzes these approaches individually and in combination, assessing their ability to balance performance and fairness. The analysis identifies two main findings: the trade-off between accuracy and fairness, and the lack of standardized fairness metrics, which limits comparisons across studies. By systematizing these results, this study advances the debate on ethics and fairness in automated systems.*

Keywords. *Algorithmic Fairness; Machine Learning; Decision Systems; Algorithmic Unfairness; Fairness Metrics;*

Resumo. *Este estudo analisa a injustiça algorítmica em sistemas automatizados de tomada de decisão, destacando a necessidade de equidade e transparência diante de seu impacto social. Algoritmos de aprendizado de máquina frequentemente reproduzem vieses dos dados, resultando em decisões discriminatórias. Diversas estratégias de mitigação têm sido propostas, incluindo métodos de avaliação de justiça e ferramentas para detecção de vieses. Este trabalho revisa essas abordagens de forma individual e combinada, avaliando sua capacidade de equilibrar desempenho e justiça. Os resultados evidenciam dois aspectos centrais: o recorrente trade-off entre acurácia e justiça e a falta de padronização das métricas utilizadas, o que dificulta comparações entre estudos. Ao sistematizar esses achados, o estudo contribui para o avanço do debate sobre ética e justiça em sistemas automatizados.*

Palavras-Chave. *Justiça Algorítmica; Aprendizado de Máquina; Sistemas de Decisão; Injustiça Algorítmica; Métricas de Justiça;*

Cite as: Vaz Filho, D. B., & Antonio Digiampietri, L. (2025). Equilíbrio entre Justiça e Desempenho em Sistemas de Aprendizado de Máquina: Uma Revisão da Literatura (Fairness and Performance Trade-off in Machine Learning Systems: A Literature Review). iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), Vol 18(1), 11:1-11:21. DOI: 10.5753/isys.2025.6007

1. Introdução

Atualmente, os sistemas de recomendação estão assumindo um papel cada vez mais influente na vida das pessoas em todo o mundo, especialmente em sistemas que utilizam e processam *big data*. Esses sistemas fornecem recomendações personalizadas para os usuários e se aplicam em diversas áreas, desde situações mais simples, como recomendações de filmes ou músicas baseados nos gostos do usuário, até aplicações em sistemas mais complexos, como por exemplo, recomendações de crédito em instituições bancárias. Esses sistemas têm mostrado um impacto significativo no comportamento e nas decisões dos indivíduos [Ashokan and Haas 2021].

Com o avanço na relevância desses sistemas para a sociedade, surge a necessidade urgente de soluções para mitigar o problema da injustiça algorítmica. Este é um campo de pesquisa relativamente novo, que busca garantir que as decisões tomadas por sistemas automatizados sejam as mais justas possíveis, proporcionando oportunidades equitativas para todos os grupos de pessoas, independentemente de cor, gênero ou condição socioeconômica. A injustiça algorítmica pode se manifestar de várias maneiras, incluindo discriminação direta ou indireta, em que certos grupos podem ser sistematicamente desfavorecidos pelas decisões automatizadas [Ashokan and Haas 2021].

O desafio de alcançar a equidade nos sistemas de aprendizado de máquina está intrinsecamente ligado ao balanço entre a acurácia dos modelos e a igualdade entre diferentes grupos e classes. A busca por maior precisão em modelos preditivos muitas vezes pode intensificar vieses preexistentes nos dados de treinamento, levando a resultados que perpetuam ou até ampliam desigualdades sociais. A transparência das decisões algorítmicas é, portanto, uma área crítica de preocupação. Embora algoritmos possam ser projetados para tomar decisões de maneira consistente, eles também podem replicar e amplificar os vieses presentes nos dados de entrada, muitas vezes de forma menos transparente do que em um processo decisório humano [Lee and Floridi 2021].

Diversos estudos têm destacado a complexidade e os desafios associados à mitigação de vieses em algoritmos de aprendizado de máquina. Abordagens recentes incluem o desenvolvimento de métricas de equidade, técnicas de pré-processamento para balanceamento de dados, algoritmos de aprendizado adaptativo, e métodos de pós-processamento para ajuste de resultados. Além disso, ferramentas de auditoria e *frameworks* de justiça (*fairness*) têm sido criados para testar e garantir a justiça nos sistemas implementados [Biswas and Rajan 2020].

Dessa forma, o campo da justiça algorítmica não só é essencial para a evolução ética dos sistemas de recomendação, mas também representa um compromisso contínuo com a igualdade e a justiça social. À medida que a dependência de sistemas automatizados cresce, a necessidade de soluções robustas para mitigar a injustiça algorítmica torna-se cada vez mais premente, exigindo esforços contínuos de pesquisa e desenvolvimento. Nesse cenário, esta revisão sistemática se justifica por analisar uma área recente e em rápida expansão, ainda com poucas revisões disponíveis. Assim, torna-se relevante reunir e organizar informações sobre métricas de discriminação algorítmica, bem como sobre modelos e benchmarks utilizados, a partir de artigos científicos alinhados ao objetivo da pesquisa: a definição e avaliação da justiça em algoritmos de aprendizado de máquina.

Para situar este trabalho no campo de Sistemas de Informação, a próxima seção apresenta um enquadramento teórico que relaciona justiça algorítmica à ética em TI, à governança de dados e aos sistemas de apoio à decisão (Seção 2). Em seguida, descrevemos a metodologia adotada para a condução desta revisão (Seção 3).

2. Fundamentação teórica em Sistemas de Informação

Em complemento ao enquadramento delineado na Introdução, esta seção situa a discussão de justiça algorítmica no campo de Sistemas de Informação, destacando como resultados técnicos dependem também de processos organizacionais, normas internas e percepções dos usuários.

A discussão sobre justiça algorítmica pode ser situada no campo de Sistemas de Informação ao compreender sistemas de IA como sociotécnicos: seus resultados decorrem não apenas de modelos e dados, mas também de processos organizacionais, papéis e responsabilidades, normas internas e percepções dos usuários [Baxter and Sommerville 2011, Selbst et al. 2019]. Essa perspectiva não substitui a análise computacional, mas a complementa ao evidenciar que escolhas métricas e operacionais são influenciadas pelos contextos de uso.

No eixo da ética em TI, destacam-se princípios de transparência e explicabilidade, justiça procedimental e responsabilização institucional por decisões automatizadas. Na prática, isso envolve documentar pressupostos e limitações dos modelos, comunicar métricas de equidade em linguagem acessível e oferecer mecanismos de prestação de contas e contestação quando as decisões afetarem pessoas ou grupos [Ananny and Crawford 2018]. Tais práticas reforçam confiança e legitimidade sem alterar o núcleo técnico da avaliação.

Quanto à governança de dados e de TI, a justiça depende da qualidade e da rastreabilidade dos dados, de critérios claros para coleta e uso e da definição de papéis e responsabilidades ao longo do ciclo de vida informacional [Khatri and Brown 2010]. Em nível de TI, referenciais de governança, como *ISO/IEC 38500* e *COBIT 2019*, definem princípios, objetivos e práticas para uso eficaz, eficiente e aceitável da tecnologia, com controles de risco e indicadores [ISO/IEC 38500:2015 2015, ISACA 2018]. Integrar métricas de equidade a indicadores, políticas internas e auditorias periódicas reduz riscos de enviesamento e facilita a conformidade.

Por fim, a literatura de sistemas de apoio à decisão indica que decisões organizacionais são, em geral, multicritério. Ao trazer a justiça como objetivo explícito - ao lado de acurácia, custo e risco - requerem-se painéis e relatórios que tornem visíveis os compromissos entre métricas, permitam ajustes de ponderação coerentes com a política interna e registrem justificativas de decisão, conectando o nível técnico ao gerencial [Shim et al. 2002].

Em síntese, esse enquadramento sociotécnico não redefine a pesquisa: fornece o contexto de Sistemas de Informação no qual as evidências computacionais ganham sentido organizacional, favorecendo a interpretação, a adoção e a auditoria dos resultados sobre justiça algorítmica.

Concluída esta contextualização, a Seção 3 descreve os procedimentos metodológicos empregados nesta revisão, incluindo critérios de seleção, extração e análise dos estudos.

3. Método

Para a revisão da literatura, foi realizada uma Revisão Sistemática (RS) [Kitchenham 2004] do tema relacionado à discriminação algorítmica, observando pontos importantes como métricas de discriminação ou justiça utilizadas, conjuntos de dados disponibilizados, além de modelos de IA usados nos *benchmarks* e as métricas de avaliação utilizadas.

É importante ressaltar que a RS foi escolhida a fim de padronizar o método de pesquisa para uma possível reprodução dos resultados, diferenciando-se assim, de uma simples revisão bibliográfica.

3.1. Perguntas

Diante do tema abordado, quatro perguntas foram elaboradas:

1. *Quais são as métricas utilizadas para avaliar a justiça algorítmica ou a discriminação algorítmica?*
2. *Quais são os algoritmos de inteligência artificial testados?*
3. *Quais são as principais bases de dados utilizadas?*
4. *Quais são os métodos para promover justiça no aprendizado de máquina?*

3.2. Fontes

Para esta revisão, utilizou-se como critério selecionar fontes disponíveis via web, em bases de dados científicas relevantes para a área de computação.

Em virtude dos critérios para seleção de fontes, as fontes presentes na tabela 1 foram selecionadas.

Tabela 1. Fontes da Revisão Sistemática

Fonte
IEEE Xplore
ACM Digital Library
Scopus

3.3. Critérios de inclusão e exclusão

Foram especificados alguns critérios de inclusão e exclusão, listados a seguir.

Critérios de inclusão:

1. Artigos publicados e disponíveis em bases de dados científicas digitais que passem pelo processo de revisão por pares.

2. Estudos primários com contribuição para a área de discriminação algorítmica. Revisões não serão consideradas para esta RS.
3. Artigos publicados a partir de 2018.
4. Artigos que abordam métodos de avaliação de algoritmos de IA.
5. Artigos que adotam medidas ou métricas de justiça (ou de discriminação) algorítmica.
6. Artigos que disponibilizam ou referenciam a base de dados utilizada.

Critérios de exclusão:

1. Trabalhos que não se relacionam à área de discriminação algorítmica.
2. Trabalhos focados no reconhecimento facial ou na geração de imagens.
3. Trabalhos que não apresentam metodologia bem fundamentada e explicada.

Para serem incluídos, os trabalhos devem satisfazer a todos os critérios de inclusão e a nenhum de exclusão.

3.4. Protocolo de busca

Foi definido um protocolo de busca com a seguinte *string* de busca:

(“All Metadata”: ((bias OR discrimi*)*
AND (fair OR ethic*)*
AND (AI OR Machine Learning OR Artificial Intelligence OR algorithm)))*

3.5. Resultados da busca primária

Após realizar a busca nas fontes especificadas com a *string* de busca aplicada, foram encontrados 468 artigos, conforme apresentado na tabela 2

Tabela 2. Resultados da busca primária da RS

Fonte	Número de trabalhos
ACM	34
IEEE	201
SCOPUS	233
Total	468

3.6. Refinamento utilizando critérios de inclusão e exclusão

A partir desses resultados, foram extraídos e analisados todos os títulos, resumos e palavras-chave e então aplicados os critérios de inclusão e exclusão. Ao todo, **37** artigos foram selecionados para leitura inicial e extração de dados.

3.7. Extração de dados e refinamento de resultados

Extração de dados: Foram definidas quatro características a serem extraídas:

- **Bases de dados:** Bases de dados disponibilizadas pelo autor.
- **Modelos de IA utilizados:** Quais os modelos de IA usados para metrificar a justiça aplicada no projeto.
- **Métodos ou métricas de justiça utilizados:** Métricas de justiça aplicadas e explicadas no artigo.
- **Método ou métricas de avaliação:** Métricas de avaliação utilizadas para a construção de *benchmarks* ou avaliar nível de precisão do algoritmo utilizado.

Exclusão de artigos: Após uma análise criteriosa dos artigos, foram escolhidos apenas aqueles que apresentam todas as informações para extração, restando assim apenas 13 resultados plausíveis para a análise profunda. Esses artigos foram ordenados da seguinte forma:

1. Quantidade de métodos de justiça
2. Uso do método de justiça mais frequente
3. Quantidade de modelos de aprendizado de máquina utilizados

A tabela 3 lista os artigos selecionados.

4. Análise dos Resultados

Dado o número reduzido de artigos selecionados, uma descrição individual de cada artigo foi feita, seguindo a ordem presente na tabela 3. A análise será feita de acordo com os dados extraídos dos 13 artigos selecionados, separados por parágrafos no texto a seguir.

4.1. Fairness metrics and bias mitigation strategies for rating predictions

O propósito do primeiro artigo é apresentar uma extensa análise de métricas de justiça e propor um método de mitigação da injustiça em modelos de recomendação. Essa abordagem foi explorada em dois cenários distintos: uma base sintética de recomendação de cursos, com o gênero como atributo sensível, e a base real do MovieLens, composta por um milhão de entradas de filmes, também considerando o gênero como atributo sensível e categorizando filmes em Ação, Criminal, Musical, Romance e Ficção Científica.

Dois modelos de inteligência artificial foram utilizados para gerar recomendações: um modelo item-item baseado em KNN padrão e um modelo ALS (*Alternating Least Squares*). As métricas RMSE e MAE foram empregadas para avaliar o desempenho dos modelos, enquanto diversas métricas de injustiça foram exploradas, destacando-se *Value Unfairness*, *Absolute Unfairness*, *Overestimation Unfairness*, *Underestimation Unfairness*, *Non-Parity*, *Theil Index* e GEI (*Generalized Entropy Index*) com alpha = 2.

Os resultados apontam para a viabilidade prática da utilização de um conjunto específico de métricas de justiça, especialmente em contextos de recomendação. Dentre os métodos de mitigação considerados, dois se destacam por realizar ajustes após o treinamento do modelo: *Value-based adjustment* e *Parity-based adjustment*. Essas estratégias visam a atenuar a injustiça percebida nos resultados das recomendações, contribuindo para a construção de modelos mais equitativos.

Tabela 3. Resultados da extração

#	Título do artigo	Quantidade de métricas de justiça
1	Fairness metrics and bias mitigation strategies for rating predictions	10
2	Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs	8
3	Do the Machine Learning Models on a Crowd Sourced Platform Exhibit Bias? An Empirical Study on Model Fairness	7
4	Mitigating Bias by Optimizing the Variance between Privileged and Deprived Data Using Post Processing Method	7
5	Fairness in Semi-Supervised Learning: Unlabeled Data Help to Reduce Discrimination	4
6	FAWOS: Fairness-Aware Oversampling Algorithm Based on Distributions of Sensitive Attributes	3
7	The fairness-accuracy Pareto front	3
8	A Framework for Benchmarking Discrimination-Aware Models in Machine Learning	2
9	Fair Contrastive Learning on Graphs	2
10	Tuning Fairness by Balancing Target Labels	1
11	Improving fairness of artificial intelligence algorithms in Privileged-Group Selection Bias data settings	1
12	How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness	1
13	Non-Discriminatory Machine Learning through Convex Fairness Criteria	1

4.2. Algorithmic Fairness in Mortgage Lending from Absolute Conditions to Relational Trade-offs

No segundo artigo, os autores optaram por abordar uma perspectiva distinta, revelando a limitação das métricas de justiça isoladamente. O estudo realizado teve como objetivo explicar algumas métricas 'ex post', como *group fairness*, também conhecida como *demographic parity*. Eles exploraram o conceito de equalização de métricas de avaliação de justiça, utilizando a matriz de confusão para explicar métricas como *equal opportunity*, *predictive equality*, *equal odds*, *positive predictive parity*, *positive class balance* e *negative class balance*.

Além disso, os autores discutiram métricas 'ex ante', que visam a avaliar o resultado individual instantâneo, em oposição a um resultado de um conjunto de dados previamente avaliado, como é o caso de 'ex post'. Tais métricas incluem *individual fairness* e *counterfactual fairness*. Para analisar essas métricas de justiça, os autores utilizaram o conjunto de dados do Home Mortgage Disclosure Act (HDMA) dos Estados Unidos

referente ao ano de 2011, uma escolha que se alinhava com suas expectativas.

Empregando os algoritmos de aprendizado de máquina, como *Logistic Regression* (LR) sem regulagem, classificador *K-Nearest Neighbors* (KNN) com K=5, *Classification and Regression Tree* (CART), *Gaussian Naive Bayes* (NB) e *Random Forest* (RF), os autores produziram resultados gráficos notáveis em relação à precisão dos modelos. O Random Forest foi o mais bem-sucedido, enquanto o KNN foi o menos eficaz para os dados analisados, que foram adaptados para conter 50% de empréstimos aprovados e 50% de não aprovados.

O artigo também aborda a ideia de que os atributos sensíveis não podem ser previstos pelo modelo de IA. Os autores conduziram um teste para isolar o atributo sensível (raça) e executaram a mesma bateria de testes. Os resultados obtidos revelaram-se insignificantes e inexpressivos em comparação aos resultados anteriores, o que contradiz a hipótese inicial.

4.3. Do the Machine Learning Models on a Crowd Sourced Platform Exhibit Bias An Empirical Study on Model Fairness

O terceiro artigo segue uma metodologia simples ao testar modelos de aprendizado de máquina em bases de dados abertas disponíveis no Kaggle. Os autores selecionaram conjuntos de dados com atributos sensíveis, como raça, idade e sexo, além de conjuntos de dados reconhecidos por seu uso em estudos sobre justiça em IA. A escolha dos modelos de avaliação baseou-se em critérios específicos: os *kernels* deveriam conter um modelo preditivo, ter mais de 5 votos e uma precisão superior a 65%, sendo selecionado o melhor modelo de cada *kernel*. No total, foram analisados cinco conjuntos de dados e 40 modelos de aprendizado de máquina.

O objetivo do projeto, a partir dessa seleção, foi identificar os melhores modelos entre os disponíveis, treiná-los e avaliar sua acurácia e F1-Score. A avaliação dos modelos seguiu sete métricas específicas, incluindo *Disparate Impact* (DI), *Statistical Parity Difference* (SPD), *Equal Opportunity Difference* (EOD), *Average Odds Difference* (AOD), *Error Rate Difference* (ERD), *Consistency* (CNT) e *Theil Index* (TI), utilizando uma biblioteca Python desenvolvida pela IBM para esse propósito.

Os autores chegaram às seguintes conclusões. A busca por otimização do desempenho do modelo muitas vezes levou a resultados injustos. As bibliotecas de criação de modelos raramente tratam explicitamente questões de justiça, destacando a necessidade de uma abordagem mais atenta. A padronização dos recursos antes do treinamento revelou-se eficaz na redução da disparidade entre grupos sensíveis. Além disso, a exclusão de determinados atributos teve um impacto significativo na imparcialidade dos modelos. Observou-se também a necessidade de métricas diversas para compreender o viés em diferentes contextos. Embora a maioria das medidas de justiça tenha se mantido consistente em múltiplos treinamentos e previsões, os resultados apontaram para a complexidade de garantir justiça nos modelos. Os algoritmos de mitigação de pré-processamento foram identificados como preferíveis, com os algoritmos de pós-processamento muitas vezes comprometendo a precisão do modelo e a pontuação F1. Isso evidenciou um equilíbrio delicado entre o desempenho e a justiça, com os algoritmos de pós-processamento reve-

lando a necessidade de substitutos mais competitivos.

4.4. Mitigating Bias by Optimizing the Variance between Privileged and Deprived Data Using Post Processing Method

O quarto artigo tem como objetivo principal avaliar um método de mitigação de injustiça por meio de um modelo próprio, baseando-se em quatro tipos diferentes de modelos amplamente reconhecidos: SVM, *Logistic Regression*, XGBOOST e KNN. O autor empregou diversas medidas de justiça para avaliar esses modelos, incluindo *Disparate Impact* (DI), *Recall Difference* (RD), *Difference in Positive Proportions* (DPP), *Difference in Rejection Rates* (DRR), *Average Odd Difference* (AOD), *Accuracy Difference* (AD) e *Equality Opportunity* (EO). Além disso, realizou uma comparação entre o equilíbrio entre justiça e acurácia, uma temática explorada em outros artigos.

Entre os quatro *baselines*, o XGBOOST destacou-se como o mais equilibrado em termos de acurácia e justiça algorítmica ao ser analisado no conjunto de dados do *German Credit*. Por fim, os autores destacam que o modelo proposto apresenta uma justiça superior em comparação ao melhor *baseline*, ao custo de menos de meio por cento em acurácia.

4.5. Fairness in Semi-Supervised Learning Unlabeled Data Help to Reduce Discrimination

No quinto artigo, os autores implementaram uma estrutura de amostragem aprimorada denominada *Fairness-Enhanced Sampling* (FS), que integra pseudo-rotulagem, reamostragem e aprendizado justo de conjuntos em um contexto de aprendizado semi-supervisionado. Os autores validaram a eficácia dessa estrutura utilizando conjuntos de dados reais e sintéticos.

O propósito dos experimentos era evidenciar como a estrutura utilizando dados não rotulados pode alcançar um equilíbrio mais efetivo entre acurácia e discriminação, explorando o impacto de fatores como o tamanho do conjunto e a dimensão da amostra nos resultados do treinamento.

Para comparação com a estrutura proposta, os autores empregaram alguns métodos de mitigação de discriminação no pré-processamento:

- Original (ORI): Utilização do conjunto de dados original.
- *Uniform Sampling* (US): Equalização do número de pontos de dados em cada grupo por meio de *oversampling* e/ou *undersampling*.
- *Preferential Sampling* (PS): Equalização do número de pontos de dados em cada grupo pela coleta de amostras próximas aos pontos de dados limítrofes.

Após a aplicação do pré-processamento, os conjuntos de dados reais de Saúde, Bancos e Censo foram avaliados quanto à acurácia e à discriminação, esta última medida pela diferença de *Demographic Parity* entre os grupos, utilizando os algoritmos de Regressão Logística e SVM.

Tanto nos resultados dos testes reais quanto nos sintéticos, observou-se uma redução na discriminação ao utilizar o modelo proposto em comparação com os *baselines*, além de uma melhor compensação entre acurácia e discriminação.

4.6. FAWOS: Fairness-Aware Oversampling Algorithm Based on Distributions of Sensitive Attributes

No sexto artigo, é explorado o *Framework FAWOS*, capaz de gerar pontos de dados sintéticos por meio de *oversampling* em pontos topológicos, como "Safe", "Borderline", "Rare" e "Outlier", com ênfase nos últimos três, que são mais desafiadores para os modelos de aprendizado de máquina aprenderem.

O objetivo central é demonstrar a viabilidade de aumentar a justiça sem comprometer significativamente a acurácia, utilizando diversos modelos de aprendizado de máquina, tais como:

- *Support Vector Machines* (SVM)
- *Gaussian Naive-Bayes* (GNB)
- *Decision Trees* (DT)
- *Logistic Regression* (LR)
- *K-Nearest-Neighbors* (KNN)

Os autores empregaram conjuntos de dados do RICCI e do *German Credit*, separando dados de treino e teste. Os dados de treinamento foram submetidos ao algoritmo FAWOS, gerando um novo conjunto de dados *oversampled* com um balanceamento aprimorado. Em seguida, conduziram treinamentos e testes utilizando os modelos mencionados anteriormente, comparando os conjuntos que passaram pelo FAWOS com aqueles que permaneceram originais.

Os resultados indicaram que, entre todos os modelos, KNN e GNB foram os mais significativamente impactados pelo balanceamento adequado dos dados, dado que são modelos altamente dependentes desse aspecto. Os demais modelos também apresentaram melhorias perceptíveis.

4.7. The fairness-accuracy Pareto front

O objetivo deste estudo é empregar o conceito de otimalidade de Pareto na otimização multiobjetivo para identificar a "frente de Pareto" entre equidade e precisão em um classificador de rede neural. Os autores adotam o esquema de escalarização de Chebyshev, teoricamente superior ao esquema linear, para recuperar soluções ótimas de Pareto.

A arquitetura da rede neural utilizada compreende camadas totalmente conectadas, intercaladas por uma camada de abandono. A ativação ReLU é aplicada nas camadas intermediárias, enquanto a função sigmoide é utilizada na camada de saída. O treinamento da rede é conduzido com o otimizador Adam, empregando uma taxa de aprendizado inicial de 0,001. O ajuste da taxa de aprendizado é realizado através do agendador *ReduceLROnPlateau* no PyTorch, ao longo de 500 épocas com tamanhos de minibatch.

O artigo destaca as limitações do esquema de escalarização linear, propondo o esquema de Chebyshev como uma abordagem comprovadamente superior na busca por pontos ótimos de Pareto. Entretanto, ressalta que o esquema Chebyshev pode gerar pontos dominados, dada a complexidade prática da otimização, especialmente em redes neurais profundas. Sugere-se que melhorias adicionais podem ser alcançadas por meio do indicador de hiper volume, embora reconhecendo os desafios de seu cálculo em dimensões superiores.

O artigo aborda métricas de justiça, como paridade demográfica, paridade condicional, medida de justiça causal e estatística de média-variância. Os conjuntos de dados utilizados incluem dados de renda adulta da UCI e o conjunto de dados de reincidência ProPublica. A métrica de base para compensação é a acurácia, mensurando a acurácia global do algoritmo de aprendizado de máquina.

Como classificadores de aprendizado de máquina, é empregado um classificador de rede neural, focado na identificação da "frente de Pareto" entre equidade e precisão, e uma técnica de aprendizagem adversária, representando a abordagem de regularização.

4.8. A Framework for Benchmarking Discrimination-Aware Models in Machine Learning

O oitavo artigo segue uma metodologia semelhante ao artigo 3, utilizando principalmente duas métricas: *Disparate Impact* e *Disparate Mistreatment*, esta última observada em duas formas: *overestimation* e *underestimation*. Os autores empregaram uma rede Bayesiana para construir um conjunto de dados com atributos alterados, aumentando a parcialidade em um atributo sensível. Os conjuntos de dados *Adult Census Income*, *Pro Publica COMPAS* e *Dutch Census* foram utilizados para esse propósito.

Após a configuração dos conjuntos de dados, o modelo Waka, uma implementação da árvore de decisão C4.5, foi utilizado para avaliações de Acurácia, *Disparate Impact* e *Disparate Mistreatment*. Vários algoritmos de mitigação de pré-processamento, como *Massaging*, *Re-weighting*, *Uniform Sampling* e *Preferential Sampling*, foram testados. Além disso, o modelo de auditoria *Black Box* (Auditor), uma implementação do *Gradient Feature Auditing* (GFA), foi aplicado para pré-processar os dados e obter resultados menos tendenciosos.

Posteriormente, as análises foram refeitas para comparar especificamente os modelos avaliados.

4.9. Fair Contrastive Learning on Graphs

O nono artigo adota uma abordagem distinta dos anteriores, examinando o problema da discriminação algorítmica por meio de grafos. Utiliza um sistema de aprendizado construtivo não supervisionado, empregando representações nodais com reconhecimento de justiça por meio de um design adaptativo de aumento de grafos. Diversas técnicas são aplicadas para mitigar a injustiça e melhorar a acurácia dos modelos, tais como:

1. *Feature Masking*
2. *Edge Deletion* com variantes como:
    (a) *Adaptive Edge Deletion With Dyadic Fairness*
    (b) *Parity-Aware Adaptive Edge Deletion*
    (c) *Adaptive Edge Deletion With Counterfactual Fairness*
    (d) *Triangle-Based Adaptive Edge Deletion*
    (e) *Degree-Aware Adaptive Edge Deletion*

Os autores utilizam conjuntos de dados reais, incluindo o Pokec, uma rede social eslava semelhante ao Facebook, além dos conjuntos de dados convencionais Racidivism e Credit

Defaulter. O conjunto de dados Pokec é segmentado em dois subconjuntos com base na localização geográfica, que constitui o atributo sensível do conjunto.

Neste estudo, duas métricas de justiça algorítmica foram empregadas: paridade estatística e oportunidade igual. Os resultados foram comparados com sete *baselines* distintos, como:

- *DeepWalk*
- *Node2Vec*
- *FairWalk* (baseado em *random-walk*)
- *Deep Graph Infomax* (DGI)
- *Deep Graph Contrastive Representation Learning* (GRACE)
- *Graph Contrastive Learning with Adaptive Augmentations* (GCA)
- *Fair and Stable Graph Representation Learning* (NIFTY) (modelos de grafos com aprendizado contrastivo).

Os principais resultados indicaram que a implementação das duas estratégias de aprimoramento de grafos resultou em melhorias na acurácia e em uma justiça mais robusta e consistente.

4.10. Tuning Fairness by Balancing Target Labels

No décimo artigo, os autores verificam a hipótese de que um modelo adaptável com base na taxa de positivos supera o modelo tradicional de redução de injustiças/vieses. Eles adaptam dois modelos, o *Non-Parametric Gaussian Process Classifier* (GPC) e o *Parametric Logistic Regression* (LR), resultando em dois modelos distintos para classificação, denominados FairGP e FairLR, respectivamente. A intenção é demonstrar a aplicabilidade desse modelo tanto para classificadores parametrizáveis quanto não parametrizáveis.

Os autores utilizam dois conjuntos de dados amplamente reconhecidos na literatura para classificar modelos com base na justiça: Adult Income e ProPublic Recidivism. Nos conjuntos de dados, os autores realizam classificações utilizando os dois modelos propostos, além dos *baselines*, que incluem SVM (*Support Vector Machine*), o método de *Reweighting* de [Calders et al. 2009], o método baseado em LR de [Agarwal et al. 2018], e alguns métodos de [Zafar et al. 2017], como ZafarFairness, ZafarAccuracy, e ZafarEqOpp. Cada método é avaliado 10 vezes para garantir conjuntos diferentes de treino e teste.

Para o conjunto de dados de Adult Income, a justiça dos modelos é avaliada usando a noção de paridade demográfica, que mede a diferença nas taxas de aceitação entre diferentes grupos. Tanto o FairGP quanto o FairLR demonstram maior justiça em comparação com os modelos básicos, alcançando alta imparcialidade e precisão, especialmente no conjunto de dados de adultos com raça como atributo sensível. Ambos superam os modelos basais de GPC e regressão logística (LR) em termos de justiça e precisão.

No caso do conjunto de dados ProPublica, a igualdade de oportunidade é medida entre diferentes grupos de dados, usando raça e gênero como atributos sensíveis. Os resultados indicam que a taxa de positivos determinada em seu modelo representa um *trade-off* entre acurácia e justiça, em que uma taxa mais alta de positivos resulta em uma maior taxa de verdadeiros positivos, porém com uma consequente redução acentuada na acurácia.

4.11. Improving fairness of artificial intelligence algorithms in Privileged-Group Selection Bias data settings

No décimo primeiro artigo, o objetivo principal é inicialmente destacar que algoritmos podem ser enviesados pelos dados. Em seguida, os autores propõem métodos de pré-processamento e *in-process* para reduzir esses vieses, fundamentando-se em algoritmos supervisionados que utilizam dados não rotulados do grupo sem privilégios para otimizar a precisão.

Foram utilizadas duas bases de dados no projeto, um conjunto de dados de recrutamento proprietário adaptado de [Pessach et al. 2020], e o conhecido conjunto de dados ProPublica Risk.

A avaliação comparou nove métodos diferentes, em que cada método é uma combinação distinta de processamento de dados enviesados (pré-processamento, *in-processing* e nenhum) e três tipos de algoritmos: *Logistic Regression*, Rede Neural e *semi-supervised Association Learning*. Todos os métodos foram comparados considerando um *trade-off* entre *Equalized Odds* e Acurácia.

Os resultados demonstraram que os algoritmos de aprendizado de máquina treinados usando os dados disponíveis na configuração *Privileged Group Selection Bias* (PGSB) atingiram níveis elevados de precisão e justiça quando testado em instâncias de grupos privilegiados e não privilegiados.

4.12. How fair can we go in machine learning Assessing the boundaries of accuracy and fairness

O 12º artigo propõe uma abordagem inovadora na análise de modelos de inteligência artificial, introduzindo um algoritmo genético para identificar os melhores parâmetros de dois modelos (Regressão Logística e Árvore de Decisão). O objetivo é criar um modelo multiobjetivo que maximize as funções de acurácia e justiça simultaneamente.

Para a medida de acurácia, os autores buscam minimizar a função de erro, utilizando a função 1 - G-mean. Essa abordagem visa a assegurar que os custos de falsos-negativos e falsos-positivos sejam mantidos baixos. Na avaliação da justiça, os autores correlacionam o *Disparate Mistreatment* com a taxa de falsos positivos para os grupos sensíveis.

O estudo utiliza conjuntos de dados comumente empregados para avaliar a justiça de algoritmos, incluindo conjuntos como Adult Census, German Credit, ProPublica, ProPublica Violent e Ricci. Os autores desenvolvem uma metodologia que aplica um algoritmo evolutivo para otimizar os parâmetros dos modelos, detalhando todos os parâmetros utilizados e o algoritmo em questão.

Ao final, é realizada uma comparação entre os dois modelos, utilizando o modelo de Pareto para visualizar as funções em termos de acurácia versus justiça. O artigo oferece ponderações significativas sobre o equilíbrio entre essas medidas.

4.13. Non-Discriminatory Machine Learning through Convex Fairness Criteria

No último artigo, é apresentada uma noção de justiça no aprendizado de máquina denominada “proportional fairness”. Introduz-se também a ideia de justiça proporcional pon-

derada, chamada “Weighted Sum of Logs”, que atribui mais peso ao grupo que foi tratado injustamente no passado. A técnica é inovadora por combinar o uso de um critério de justiça convexo com a soma ponderada dos logaritmos para alcançar a não discriminação.

Além dessa técnica, os autores comparam dois outros *baselines* conceituais retirados do artigo de [Zafar et al. 2017], denominados pelos autores como AISTATS-17a e AISTATS-17b. Essas técnicas minimizam a perda sujeita a restrições de covariância e minimizam a covariância sujeita a restrições de perda, respectivamente.

O artigo demonstra, por meio de experimentos nos conjuntos de dados de ProPublic Recidivism e Adult Income, que a técnica proposta alcança a não discriminação sem uma perda significativa de acurácia.

5. Considerações sobre a revisão

Esta seção sumariza de forma crítica os resultados da presente revisão.

5.1. Extração de métricas

A partir dos artigos selecionados foi possível sintetizar uma tabela com as principais métricas de justiça que permeiam esse tema. Esse resultado é apresentado nas tabelas 4 , 5 e 6 localizadas no apêndice A.

5.2. Convergências Temáticas

Os artigos analisados revelam convergências temáticas notáveis. Destaca-se a prevalência de um *trade-off*, uma troca delicada entre a acurácia, ou métricas de efetividade do modelo, e os princípios de justiça abordados em cada texto. Este padrão reflete o desafio recorrente enfrentado pelos pesquisadores ao equilibrar a acurácia do modelo com as preocupações éticas inerentes à justiça algorítmica.

Outro ponto de convergência é o objetivo comum de abordar e mitigar o problema da injustiça por meio da manipulação estratégica de dados e resultados. Os autores, de maneira uníssona, buscam um equilíbrio nos grupos sensíveis conforme delineado pela base de dados utilizada. Este alinhamento de objetivos ressalta a importância da busca por métodos que possam harmonizar a eficácia preditiva do modelo com a equidade desejada.

A recorrência dos mesmos conjuntos de dados entre os artigos destaca a relevância e robustez dos conjuntos explorados. Essa repetição reflete a consciência dos pesquisadores sobre a importância de utilizar fontes de dados coerentes e reconhecidas para avaliar as propostas de mitigação.

5.3. Soluções de Mitigação

A maioria dos artigos não apenas identifica os desafios, mas também propõe soluções para mitigar a injustiça em modelos preditivos. Seja por meio de algoritmos inovadores apresentados nos próprios textos ou empregando bibliotecas públicas com parâmetros especificados pelos autores. Há uma clara intenção de contribuir para a evolução de ferramentas e métodos mais justos.

5.4. Implicações práticas e gerenciais

A presença recorrente do *trade-off* entre desempenho e justiça exige leitura cuidadosa por parte da gestão de SI. Duas formas de enquadramento são particularmente úteis: tratar a justiça como restrição mínima a ser atendida (estabelecendo limites internos para diferenças aceitáveis entre grupos) ou incorporá-la como objetivo adicional no processo decisório, com pesos definidos pelas diretrizes da organização. Em ambos os casos, a análise deve explicitar a fronteira de soluções disponíveis e os compromissos assumidos entre métricas.

A seleção e a comunicação das métricas devem refletir o risco do domínio e os públicos envolvidos. Isso inclui: indicar quais métricas capturam melhor o potencial de dano (por exemplo, desequilíbrios em verdadeiros positivos versus distribuição de decisões positivas), apresentar indicadores por coortes relevantes, reportar incerteza e registrar premissas e limitações de dados e modelos. Relatórios claros, com histórico de versões e justificativas de escolha, favorecem a compreensão e o acompanhamento das consequências práticas.

Para além do aspecto técnico, a implementação sustentável de medidas de equidade depende de condições organizacionais: qualidade, representatividade e rastreabilidade dos dados; definição de papéis e responsabilidades no ciclo de vida informacional; políticas internas que incorporem métricas de equidade a indicadores e auditorias; capacitação de áreas de negócio para leitura das métricas; aderência regulatória e canais de contestação para pessoas afetadas; e rotinas operacionais de versionamento, monitoramento de deriva e critérios de reversão.

No plano operacional, um encadeamento objetivo apoia a decisão e a prestação de contas:

1. Diagnosticar o risco e identificar grupos afetados no contexto de uso.
2. Selecionar métricas de justiça e de desempenho alinhadas aos objetivos institucionais.
3. Definir limites mínimos e/ou pesos para as métricas escolhidas.
4. Escolher e parametrizar a técnica de mitigação, validando em piloto controlado.
5. Implantar com monitoramento contínuo, registro das decisões e revisão periódica planejada.

Por fim, a heterogeneidade observada nas métricas e definições reforça a necessidade de convenções internas de reporte e de uma lista mínima comum para comparação entre iniciativas, tema aprofundado na subseção seguinte sobre limitações e padronização.

5.5. Limitações e Necessidade de Padronização

Uma limitação comum entre os estudos é a definição variada do conceito de injustiça a ser aplicado nos modelos. A emergência recente dessa área de estudo dificulta, ainda, a existência de um consenso sobre quais métricas são mais apropriadas para problemas específicos. A falta de padronização representa um desafio significativo, tornando complexa a comparação direta entre os diferentes estudos.

5.6. Desafios em Aberto

O principal desafio identificado continua sendo a busca por métodos de mitigação que conciliem justiça e acurácia de maneira equilibrada. Poucos estudos exploram a delicada interação desse *trade-off*, destacando a necessidade de investigações adicionais nessa área. Além disso, a literatura carece de análises comparativas que permitam compreender em quais contextos determinadas métricas ou técnicas produzem melhores resultados, o que limita a consolidação de práticas padronizadas.

Para além das questões técnicas, emergem também desafios de natureza sociotécnica que podem orientar uma agenda futura de pesquisa em Sistemas de Informação. Entre eles, destaca-se a necessidade de compreender como os usuários percebem sistemas considerados "justos" e de que forma essa percepção influencia sua confiança e adoção de soluções baseadas em inteligência artificial. Outro aspecto relevante é a análise de casos concretos de implementação de governança algorítmica em organizações, investigando não apenas a viabilidade técnica, mas também barreiras culturais, regulatórias e éticas que afetam sua institucionalização. Finalmente, a integração de perspectivas interdisciplinares - combinando ciência da computação, direito, sociologia e gestão - configura-se como uma direção promissora para avançar na construção de sistemas mais equitativos e transparentes.

Assim, os desafios em aberto não se restringem à esfera técnica, mas apontam para uma agenda de pesquisa mais ampla, que envolva aspectos humanos, organizacionais e regulatórios da justiça algorítmica em sistemas de informação.

Referências

[Agarwal et al. 2018] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR.

[Ananny and Crawford 2018] Ananny, M. and Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3):973–989.

[Ashokan and Haas 2021] Ashokan, A. and Haas, C. (2021). Fairness metrics and bias mitigation strategies for rating predictions. *Information Processing and Management*, 58(5). Cited by: 20.

[Baxter and Sommerville 2011] Baxter, G. and Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*, 23(1):4–17.

[Biswas and Rajan 2020] Biswas, S. and Rajan, H. (2020). Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2020, page 642–653, New York, NY, USA. Association for Computing Machinery.

[Calders et al. 2009] Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18.

[Goel et al. 2018] Goel, N., Yaghini, M., and Faltings, B. (2018). Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 116, New York, NY, USA. Association for Computing Machinery.

[ISACA 2018] ISACA (2018). Cobit 2019 framework: Introduction and methodology. `https://www.isaca.org/resources/cobit`. Acesso em: 07 set. 2025.

[ISO/IEC 38500:2015 2015] ISO/IEC 38500:2015 (2015). ISO/IEC 38500:2015 information technology — governance of it.

[Kehrenberg et al. 2020] Kehrenberg, T., Chen, Z., and Quadrianto, N. (2020). Tuning fairness by balancing target labels. *Frontiers in Artificial Intelligence*, 3.

[Khatri and Brown 2010] Khatri, V. and Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, 53(1):148–152.

[Kitchenham 2004] Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

[Lee and Floridi 2021] Lee, M. S. A. and Floridi, L. (2021). Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds and Machines*, 31(1):165–191.

[Mandhala et al. 2022] Mandhala, V. N., Bhattacharyya, D., Midhunchakkaravarthy, D., and Kim, H.-J. (2022). Mitigating bias by optimizing the variance between privileged and deprived data using post processing method. *Rev. D Intell. Artif.*, 36(1):87–91.

[Pessach and Shmueli 2021] Pessach, D. and Shmueli, E. (2021). Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings. *Expert Systems with Applications*, 185:115667.

[Pessach et al. 2020] Pessach, D., Singer, G., Avrahami, D., Chalutz Ben-Gal, H., Shmueli, E., and Ben-Gal, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, 134:113290.

[Rodrigo L. Cardoso 2019] Rodrigo L. Cardoso, Wagner Meira Jr., V. A. M. J. Z. (2019). A framework for benchmarking discrimination-aware models in machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA. ACM.

[Salazar et al. 2021] Salazar, T., Santos, M. S., Araujo, H., and Abreu, P. H. (2021). FAWOS: Fairness-aware oversampling algorithm based on distributions of sensitive attributes. *IEEE Access*, 9:81370–81379.

[Selbst et al. 2019] Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. ACM.

[Shim et al. 2002] Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., and Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision Support Systems*, 33(2):111–126.

[Valdivia et al. 2021] Valdivia, A., Sánchez-Monedero, J., and Casillas, J. (2021). How fair can we go in machine learning ? assessing the boundaries of accuracy and fairness. *Int. J. Intell. Syst.*, 36(4):1619–1643.

[Wei and Niethammer 2022] Wei, S. and Niethammer, M. (2022). The fairness-accuracy pareto front. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(3):287–302.

[Zafar et al. 2017] Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1171–1180, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[Zhang et al. 2022] Zhang, T., Zhu, T., Li, J., Han, M., Zhou, W., and Yu, P. S. (2022). Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination. *IEEE Transactions on Knowledge and Data Engineering*, 34(4):1763–1774.

Apêndice A: Tabelas Complementares

Tabela 4. Métricas relacionadas à justiça algorítmica — Parte I

Conceito	Definição	Referências
Equal opportunity	Grupos privilegiados e não privilegiados devem possuir as mesmas taxas de verdadeiros e falsos positivos.	[Ashokan and Haas 2021], [Lee and Floridi 2021]
Demographic parity	A probabilidade de resultados positivos deve ser igual entre os grupos privilegiados e não privilegiados.	[Ashokan and Haas 2021], [Lee and Floridi 2021], [Zhang et al. 2022], [Wei and Niethammer 2022], [Kehrenberg et al. 2020], [Pessach and Shmueli 2021], [Goel et al. 2018]
Disparate impact (di)	Razão entre a probabilidade de o grupo não privilegiado receber predição favorável e a do grupo privilegiado.	[Ashokan and Haas 2021], [Biswas and Rajan 2020], [Mandhala et al. 2022], [Salazar et al. 2021], [Rodrigo L. Cardoso 2019], [Pessach and Shmueli 2021]
Equalized odds	Requer que as predições sejam independentes do atributo sensível, equalizando taxas de verdadeiros e falsos positivos.	[Ashokan and Haas 2021], [Biswas and Rajan 2020], [Mandhala et al. 2022], [Salazar et al. 2021], [Rodrigo L. Cardoso 2019], [Pessach and Shmueli 2021], [Goel et al. 2018]
Theil index	Medida de iniquidade entre indivíduos.	[Ashokan and Haas 2021], [Biswas and Rajan 2020]
Predictive equality	Também chamada de “False positive error rate balance”; balanceia falsos positivos entre os grupos.	[Lee and Floridi 2021]
Average odds difference (aod)	Média das diferenças de falso-positivo e verdadeiro-positivo entre os grupos.	[Biswas and Rajan 2020], [Mandhala et al. 2022]

Tabela 5. Métricas relacionadas à justiça algorítmica — Parte II

Conceito	Definição	Referências
Disparate mistreatment	Diferenças na taxa de erro de classificação entre grupos durante o treinamento.	[Ashokan and Haas 2021], [Rodrigo L. Cardoso 2019], [Kehrenberg et al. 2020], [Valdivia et al. 2021]
Non-parity	Diferença absoluta nas taxas preditas entre grupos; adaptação de *statistical parity*.	[Ashokan and Haas 2021]
Positive predictive parity	Em grupos favoráveis, as predições devem ser consistentes independentemente do atributo sensível.	[Lee and Floridi 2021]
Positive class balance	Em grupos favoráveis, deve haver igualdade entre indivíduos com diferentes atributos sensíveis.	[Lee and Floridi 2021]
Negative class balance	Em grupos não favoráveis, deve haver igualdade entre indivíduos com diferentes atributos sensíveis.	[Lee and Floridi 2021]
Individual fairness	Indivíduos que diferem apenas no atributo sensível devem receber o mesmo resultado preditivo.	[Ashokan and Haas 2021], [Lee and Floridi 2021], [Biswas and Rajan 2020], [Salazar et al. 2021], [Goel et al. 2018]
Counterfactual fairness	A decisão deve ser a mesma no mundo real e no cenário "contrafactual" com atributo sensível favorável.	[Ashokan and Haas 2021], [Lee and Floridi 2021]

Tabela 6. Métricas relacionadas à justiça algorítmica — Parte III

Conceito	Definição	Referências
Statistical parity difference (spd)	Diferença entre as probabilidades de resultados positivos entre grupos.	[Biswas and Rajan 2020]
Equal opportunity difference (eod)	Diferença nas taxas de verdadeiros positivos entre grupos.	[Biswas and Rajan 2020], [Mandhala et al. 2022]
Error rate difference (erd)	Soma das taxas de falso-positivo e falso-negativo.	[Biswas and Rajan 2020]
Consistency (cnt)	Similaridade das predições entre instâncias semelhantes.	[Biswas and Rajan 2020]
Recall difference (rd)	Diferença de revocação entre grupos privilegiado e não privilegiado.	[Mandhala et al. 2022]
Difference in positive proportions (dpp)	Diferença entre as proporções de predições positivas entre grupos.	[Mandhala et al. 2022]
Difference in rejection rates (drr)	Diferença nas taxas de rejeição entre classes privilegiada e não privilegiada.	[Mandhala et al. 2022]