Proposal of a Method for Identifying Unfairness in Machine Learning Models based on Counterfactual Explanations

Fernanda R. P. Cirino, Carlos D. Maia, Marcelo S. Balbino, Cristiane N. Nobre

Pontifícia Universidade Católica de Minas Gerais, Brasil
fernanda.passos.cirino@gmail.com, carlosdiasmaia@gmail.com, marcelobalbino@gmail.com, nobre@pucminas.br

Abstract. As machine learning models continue impacting diverse areas of society, the need to ensure fairness in decision-making becomes increasingly vital. Unfair outcomes resulting from biased data can have profound societal implications. This work proposes a method for identifying unfairness and mitigating biases in machine learning models based on counterfactual explanations. By analyzing the model's equity implications after training, we provide insight into the potential of the method proposed to address equity issues. The findings of this study contribute to advancing the understanding of fairness assessment techniques, emphasizing the importance of post-training counterfactual approaches in ensuring fair decision-making processes in machine learning models.

CCS Concepts: • **Supervised learning by classification**; • **Applied computing** → **Law**;

Keywords: Unfairness, Interpretability, Counterfactual Explanations, Machine Learning

1. INTRODUÇÃO

A justiça é um conceito que tem sido discutido e debatido por séculos. É uma ideia relativa que pode ser compreendida de várias maneiras, dependendo do contexto e da perspectiva da pessoa que a discute [Saxena 2019]. Em sua essência, a justiça diz respeito à equidade e à igualdade. É a ideia de que todos os indivíduos devem ser tratados com dignidade e respeito e que devem ter acesso às mesmas oportunidades e recursos.

Além de conceitos de punição e de processo legal, a justiça também envolve questões de desigualdade social. Isso inclui tópicos como distribuição de riqueza, acesso à saúde, educação e oportunidades de trabalho. Muitas pessoas argumentam que uma sociedade justa é aquela em que todos têm acesso igual a esses recursos, independentemente de sua raça, gênero ou *status* socioeconômico.

No entanto, alcançar a verdadeira justiça pode ser difícil. Muitas vezes, há interesses e perspectivas conflitantes que tornam difícil chegar a um consenso sobre o que é justo. Por exemplo, algumas pessoas argumentam que programas de ação afirmativa, que visam abordar injustiças históricas dando tratamento preferencial a grupos sub-representados, são injustos para aqueles que não fazem parte desses grupos.

Em última análise, o conceito de justiça é complexo e relativo. Embora não haja uma definição única de justiça que sirva para todos os casos, a maioria das pessoas concorda que o objetivo principal é tratar as pessoas de maneira justa, ou seja, de maneira equitativa e garantir que todos tenham acesso às mesmas oportunidades e recursos [Edor 2020].

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico do Brasil (CNPq, Código: 311573/2022-3), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, Código: APQ-03076-18), ao Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) e à Pontifícia Universidade Católica de Minas Gerais (Código: FIP-2023/29184-1S).

Com o crescente uso de técnicas de Aprendizado de Máquina (AM) em projetos de larga escala, a abordagem de justiça nesses modelos tem se tornado uma preocupação cada vez mais relevante [Mehrabi et al. 2021]. Afinal, estes sistemas são frequentemente utilizados para tomar decisões que afetam a vida das pessoas, como determinar se alguém deve receber um empréstimo ou se deve ser contratada para um emprego. No entanto, a falta de justiça nestes modelos pode resultar em discriminação, especialmente para grupos minoritários [Saxena et al. 2020]. Tendo em vista essa necessidade, diversos métodos e algoritmos começaram a ser elaborados, a fim de verificar a justiça nos modelos de AM, principalmente por meio da interpretabilidade [Gomez et al. 2021][Wexler et al. 2020].

A interpretabilidade é um fator crucial para que a justiça seja garantida em modelos de AM uma vez que a maioria dos algoritmos utilizados no presente não possuem explicabilidade [Aggarwal et al. 2019]. Sendo assim, a interpretabilidade pode ajudar a garantir que as decisões tomadas pelos modelos sejam compreensíveis e justificáveis, permitindo que os usuários entendam como as decisões são tomadas e identifiquem possíveis erros ou problemas.

Desta forma, a interpretabilidade e a justiça são temas cada vez mais relevantes na comunidade de AM, especialmente diante da popularização do uso desses modelos. A busca por modelos mais justos e éticos, que levem em consideração a diversidade e a complexidade do mundo real, é um desafio importante e necessário para garantir a confiabilidade e a equidade desses sistemas.

Métodos contrafactuais de interpretabilidade de modelos de AM visam fazer a menor quantidade possível de modificações em uma instância para que ela mude de classe [Gomez et al. 2021]. A justiça contrafactual diz que, para o modelo ser justo, essas mudanças não podem ser apenas de atributos protegidos [Kusner et al. 2018]. Esses atributos são variáveis que podem estar contaminadas por preconceitos como o gênero, raça ou etnia [Oneto e Chiappa 2020], por isso, não devem ser determinantes na tomada de decisão da classe.

Considere como exemplo um algoritmo de AM que determina se um indivíduo vai ou não ter direito a um empréstimo no banco. Se a única diferença entre dois candidatos for gênero ou etnia, o resultado deveria ser o mesmo. Contudo, normalmente estes modelos estão em contato com dados historicamente preconceituosos o que interfere em suas decisões [Kusner et al. 2018].

Desta forma, a análise de justiça em modelos de AM é essencial, pois impede que decisões injustas e discriminatórias sejam aplicadas. Ou seja, analisando e provando que o modelo seja justo, sua confiabilidade aumenta, além de proteger aqueles que o utilizam.

Diante deste contexto, este trabalho visa propor um método para identificar injustiça em modelos de AM utilizando explicações contrafactuais geradas pelo CSSE (*Agnostic Method of Counterfactual, Selected, and Social Explanations*) [Balbino et al. 2023]. Este algoritmo de interpretabilidade apresenta explicações contrafactuais para as classificações utilizando-se de algoritmo genético. É importante ressaltar que não é objetivo do CSSE reconhecer injustiça em modelos. Logo, a geração dos contrafactuais por meio do CSSE é apenas uma parte da estratégia proposta em nosso método que inclui outros procedimentos para identificação de injustiça.

Desta forma, ao contrário do CSSE, este trabalho induz a criação de contrafactuais que utilizam atributos protegidos como argumentos para a mudança de classe, com a intenção de demonstrar injustiça nas decisões do modelo. Para avaliar a efetividade do método, apresentamos os experimentos a partir da base de dados Compas, por conter atributos sociais e socioeconômicos que podem estar contaminados com preconceitos.

2. REFERENCIAL TEÓRICO

Esta seção apresenta as principais métricas utilizadas na análise de justiça em modelos de AM.

2.1 Métricas de justiça

Nos últimos anos, muitos trabalhos sobre justiça em algoritmos de AM foram realizados e algumas métricas para essa avaliação foram desenvolvidas [Chzhen et al. 2020]. Nesta seção, apresentamos algumas métricas normalmente utilizadas para avaliação de justiça em modelos de AM. Para as métricas descritas abaixo, faremos as seguintes suposições: A é o conjunto de atributos protegidos, ou seja, variáveis que não devem ser utilizadas para discriminar um indivíduo. X é o conjunto dos demais atributos de uma pessoa. Y é a saída prevista que pode estar contaminada com preconceitos. $\hat{Y}$ é a saída levando em consideração os ajustes feitos no modelo para que ele seja justo. a' representa um valor do atributo protegido, representando o grupo desprivilegiado. Por fim, a é equivalente, porém representa o grupo privilegiado. Com estas definições, temos as seguintes métricas de avaliação.

2.1.1 *Paridade Demográfica.* A paridade demográfica, Equação (1), consiste em garantir que a decisão - como por exemplo a aceitação ou negação de um empréstimo - não seja influenciado por um atributo protegido [Hardt et al. 2016].

$$P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0) \tag{1}$$

Esta equação formaliza que o modelo de aprendizado deve obter o mesmo resultado na previsão de $\hat{Y}$, independentemente de $A = 0$ ou 1.

2.1.2 *Chances iguais.* Quando dois indivíduos apresentam as mesmas características porém, um é privilegiado e o outro não, eles devem possuir a mesma probabilidade de serem classificados em determinada classe, conforme mostra a Equação 2.

$$Pr(\hat{Y} = 1|A = 0, Y = y) = Pr(\hat{Y} = 1|A = 1, Y = y), \quad y \in 0, 1 \tag{2}$$

Ou seja, nós dizemos que um preditor $\hat{Y}$ satisfaz as chances iguais em relação ao atributo protegido A e o resultado Y, se $\hat{Y}$ e A são independentes, condicionais a Y [Hardt et al. 2016].

2.1.3 *Oportunidades iguais.* Essa métrica é derivada da Equação 2 e deve ser utilizada quando não queremos discriminar o grupo privilegiado. Por exemplo, pessoas que têm uma renda alta devem ter a mesma oportunidade de conseguir empréstimos que pessoas de rendas menores. É determinada pela independência entre o conjunto de atributos protegidos e a classificação gerada pelo modelo, conforme apresentado na Equação 3 [Hardt et al. 2016].

$$Pr(\hat{Y} = 1|A = 0, Y = 1) = Pr(\hat{Y} = 1|A = 1, Y = 1) \tag{3}$$

2.1.4 *Justiça individual.* De acordo com [Kusner et al. 2018], o algoritmo é justo se apresenta saídas similares para instâncias similares.

$$\hat{Y}(X^{(i)}, A^{(i)}) \approx \hat{Y}(X^{(j)}, A^{(j)}), \ se\ i \approx j \tag{4}$$

Neste caso, é necessário determinar a distância entre os indivíduos i e j. Quando essa distância é pequena, significa que os indivíduos são parecidos e devem ser classificados similarmente.

2.2 Método CSSE para geração de explicação contrafactual

Balbino et al. [2023] desenvolveram o *Agnostic Method of Counterfactual, Selected, and Social Explanations (CSSE)*, cuja proposta é gerar explicações locais para modelos de classificação por meio de algoritmos genéticos (AG), contribuindo com a interpretabilidade desses modelos. O método apresenta múltiplas explicações contrafactuais com diversidade, sem prolixidade[1], além de outras opções para o

[1]O termo prolixidade está relacionado a contrafactuais gerados a partir de incrementos ou decrementos nos mesmos atributos sem contribuir com novas explicações [Balbino et al. 2023].

usuário comunicar suas preferências. Para alcançar explicações múltiplas com diversidade, o método gera cada explicação baseada em um conjunto diferente de atributos, o que aumenta a possibilidade de escolha do usuário. Outro recurso do método, chamado *static_ list*, permite aos usuários definirem uma lista de atributos que não devem constar nas explicações.

Segundo Balbino et al. [2023], na implementação do método, o AG busca, a cada geração, gerar exemplos contrafactuais que se aproximem da classe desejada com mudanças mínimas (minimalidade) em relação à instância original. Quando se trata de minimalidade, esta pode estar relacionada ao número de atributos modificados (esparsidade) ou à distância (usando uma função de distância) em relação à instância original. No CSSE, o usuário define a noção de minimalidade que deseja adotar, podendo inclusive compartilhar ambas as noções e atribuir pesos a cada uma.

Para os autores, um dos principais diferenciais do CSSE refere-se ao fato de ser uma abordagem voltada ao usuário final, incluindo não especialistas em Aprendizado de Máquina. Para tal, o método inclui recursos baseados nas diretrizes de Miller [2019], o qual considera que explicações contrafactuais, selecionadas e sociais tem maior capacidade de comunicação com os usuários. Tais diretrizes são baseadas em estudos que consideram aspectos computacionais e das ciências sociais.

Por fim, por meio de experimentos e análises qualitativas e quantitativas, Balbino et al. [2023] mostram que o CSSE apresenta avanços significativos quando comparado com outros métodos presentes na literatura para explicação contrafactual.

3. TRABALHOS RELACIONADOS

Kusner et al. [2018] propuseram uma estrutura para avaliar equidade contrafactual nos modelos de AM, que consiste em analisar o impacto da mudança de certas características no resultado do modelo. Para essa avaliação é necessário identificar as características sensíveis que estão relacionadas ao atributo protegido e que podem levar à injustiça. Por exemplo, se gênero é um atributo protegido, estar grávida é uma característica sensível. Os autores sugerem definir os cenários contrafactuais que serão avaliados, ou seja, um cenário hipotético no qual uma ou mais características sensíveis serão alteradas, mantendo constante as demais características. Sugerem ainda treinar o modelo usando as características sensíveis e as outras características relevantes, e prever o resultado para o cenário original e para o contrafactual. Se o modelo for justo, a diferença nos resultados deve ser mínima.

No trabalho de [Kim et al. 2018] foi desenvolvido o *framework Multiaccuracy Boost*, um algoritmo *boosting* com modificações nos limiares, com o intuito de respeitar a métrica de multi acurácia e tornar as classificações justas. É um método que garante justiça após o modelo já ter sido treinado. A abordagem *Multiaccuracy* envolve estimar várias taxas de precisão para diferentes grupos dentro do conjunto de dados. Esses grupos são geralmente definidos com base em atributos protegidos, como gênero, raça ou idade. Ao analisar as taxas de precisão entre os diferentes grupos, é possível identificar e quantificar potenciais disparidades ou vieses nas previsões do classificador. Dessa forma, este método busca encontrar limiares de decisão que equilibrem a equidade e a precisão geral. O objetivo é ajustar os limiares de forma que os resultados da classificação se tornem mais equitativos entre os diferentes grupos, mantendo níveis aceitáveis de precisão geral. Essa etapa de pós-processamento pode ajudar a mitigar o viés ou discriminação potencial introduzida pelo classificador.

Petersen et al. [2021] propõem um método prático para alcançar a justiça individual. Os autores destacam o desafio de garantir que os modelos tratem indivíduos semelhantes da mesma maneira, independentemente de sua afiliação a um grupo protegido. Segundo os autores, seu ponto de partida seria o "*Fairness Through Awarenes*", proposto em Dwork et al. [2011], porém enfatizava que esse trabalho necessitava de melhorias (escalabilidade baixa e uma troca desfavorável com a precisão). Com isso, é proposto um método que utiliza uma matriz de adjacência para representar a similaridade entre indivíduos e um regularizador baseado no *Laplaciano* do grafo, para garantir que indivíduos similares sejam tratados de forma semelhante. Os autores discutem as conexões entre o regularizador

do Laplaciano e a equidade individual e mostram que sua abordagem supera outros métodos em experimentos com dados reais.

Já Jain et al. [2020] utilizam o método SHAP, um algoritmo de explicabilidade agnóstico, para realizar o julgamento de justiça, ajudando a identificar e mitigar a discriminação causada por viéses. Para fazer isso, primeiramente o método SHAP é utilizado para calcular um valor SHAP para cada uma das predições feitas pelo modelo. Após isso, são analisados esses valores para as instâncias que contêm atributos protegidos e para as que não contêm, utilizando métricas como a paridade demográfica, chances iguais e oportunidades iguais, em diferentes partes da base de dados. Seguindo esses passos, é possível ver se o atributo protegido influencia ou não em uma decisão.

4. ESTRATÉGIA PARA AVALIAÇÃO DE JUSTIÇA UTILIZANDO CONTRAFACTUAIS

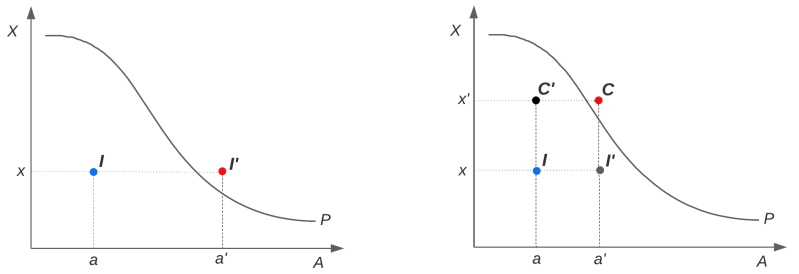
Esta seção apresenta a solução proposta para avaliação de justiça em modelos de classificação. Em suma, propomos uma estratégia para avaliar se um preditor P é injusto por meio de contrafactuais. Para tal, considere uma instância I que possui o conjunto de atributos protegidos A, o conjunto de atributos não protegidos X e que pertença à classe Y. Considerando que a e a' representam respectivamente o grupo privilegiado e o grupo desprivilegiado, tem-se que $I = (X = x, A = a)$ e $P(I) = Y$.

Existem duas maneiras de mostrar que P não é justo:

(1) Se $I' = (X = x, A = a')$ e $P(I') = Y'$. Neste caso uma simples alteração em A mudaria a classe de I;
(2) Encontrando um contrafactual $C = (X = x', A = a')$ e $P(C) = Y'$, onde x' consiste na modificação de alguns atributos de X e a' contribui para reverter a classe de I.

É trivial testar a condição 1, gerando indivíduos por uma mudança simples no atributo protegido e avaliando a classe desses indivíduos (veja Figura 1a). Para o caso 2, pode-se usar o método CSSE para encontrar este suposto indivíduo C e, em seguida, verificar se a mudança em A contribui para reverter a classe. Para tal pode-se usar a seguinte estratégia, ilustrada na Figura 1b:

(1) Executa-se o CSSE considerando como entrada para o método a instância $I' = (X = x, A = a')$, onde $P(I') = Y$ (ou seja, a simples mudança em A não foi suficiente para reverter a classe). A ideia é induzir a geração de contrafactuais com a contribuição de um atributo protegido. Deve-se ainda incluir A na *static_list* para garantir que não seja novamente modificado pelo CSSE.
(2) Se o CSSE não encontrou contrafactuais para I':
    2.1. Não foi possível mostrar que P é injusto.
(3) Se o CSSE encontrou os contrafactuais C_1, C_2, ..., C_k para I', então $P(C_i) = Y'$. Neste caso, é preciso avaliar se a mudança em A foi necessária para geração dos contrafactuais, já que esta não foi realizada pelo CSSE.
    3.1. Para cada indivíduo $C_i = (X = x_i, A = a')$, suponha um indivíduo $C_i' = (X = x_i, A = a)$.
        a. Se $P(C_i') = Y'$, não se pode afirmar que P é injusto, pois a mudança em A não é necessária para reverter a classe de I'. Em outras palavras, o que provocou a mudança de classe de I' foram as modificações em X e não no atributo protegido.
        b. Se $P(C_i') = Y$, pode-se afirmar que P não é justo, pois a mudança em A é necessária para alterar a classe de I'. Além disso, podemos usar os indivíduos C_i e C_i' para mostrar que P não é justo, uma vez que estes pertencem a classes diferentes e se diferenciam apenas pelo atributo protegido.



(a) Injustiça gerada por uma simples alteração de um atributo protegido

(b) Injustiça gerada com o auxílio do CSSE

Fig. 1: Exemplos de injustiça

Assim, a estratégia proposta busca, por meio de contrafactuais, criar dois indivíduos sintéticos que permitam evidenciar que um preditor não é justo. Ressalta-se que, caso não se consiga gerar os referidos indivíduos, não significa que o preditor seja justo, pois o AG pode simplesmente não ter encontrado um exemplo que pudesse comprovar a injustiça. Por fim, destaca-se que a solução foi apresentada considerando uma única instância, mas pode-se expandir a busca por exemplos de injustiça repetindo a estratégia para uma amostra ou todo o conjunto de dados.

É importante mencionar que a aplicação do CSSE gera explicações contrafactuais das classificações geradas por um determinado algoritmo de Aprendizado de Máquina. Destaca-se, entretanto que o CSSE, em sua execução usual, não necessariamente utilizaria um atributo protegido na geração de contrafactuais. Sendo assim, é parte da estratégia proposta neste trabalho induzir a geração de contrafactuais, por meio do CSSE, que incluam atributos protegidos e assim evidenciar a injustiça nas decisões de um modelo.

5. METODOLOGIA

A metodologia adotada neste trabalho consiste em utilizar o método proposto, e descrito na seção anterior, para avaliar a justiça em modelos de classificação. A estratégia envolve a análise de instâncias de dados, considerando conjuntos de atributos protegidos e não protegidos.

Para verificar a justiça do preditor, são aplicados dois cenários: (1) a modificação simples do atributo protegido para observar se a classe resultante é alterada e (2) a geração de contrafactuais utilizando o CSSE, modificando os atributos de forma a reverter a classe original. Caso não seja possível encontrar contrafactuais, não é possível afirmar a injustiça do preditor. No entanto, se os contrafactuais forem encontrados e a mudança no atributo protegido for necessária para reverter a classe, isso indica a existência de injustiça.

A metodologia pode ser aplicada em amostras de dados ou em conjuntos completos, permitindo a identificação de possíveis casos de injustiça em modelos de classificação. Para exemplificar o funcionamento do método proposto, foi utilizada a base de dados Compas[2], a qual inclui o atributo protegido raça.

5.1 Descrição do conjunto de dados Compas

A base de dados Compas inclui dados demográficos, histórico criminal, tempo de prisão e três pontuações *Compas* (risco de reincidência de crimes, risco de violência e o risco de não comparecer em ocasiões agendades por conta de sua situação legal) de 7214 réus do condado de Broward, Flórida. Cada réu recebe uma pontuação Compas que varia de 1 a 10. Nos experimentos, considerou-se as

[2]Disponível em https://github.com/propublica/compas-analysis.

transformações realizadas por Guidotti et al. [2019] que resultaram em 12 atributos, incluindo a classe. Para o atributo classe, os autores rotularam as pontuações de 1 a 6 como "Risco Médio-Baixo" e de 7 a 10 como "Alto Risco". A base possui 5.219 réus classificados como "Risco Médio-Baixo" e 1.995 como "Alto Risco".

Foi criado um modelo de classificação para esse conjunto de dados baseado em *Random Forest*, usando os parâmetros padrão da biblioteca *Scikit-learn*. Foram separados 20% dos dados para teste do modelo. O restante foi balanceado com *undersampling* e usado para treinamento e validação, realizado com a validação cruzada de 10 pastas. Utilizou-se as métricas *Precision*, *Recall* e *F-measure* para avaliar a qualidade do modelo.

6. RESULTADOS E DISCUSSÕES

A Tabela I apresenta o resultado do modelo com *Random Forest* para o conjunto de teste. Obteve-se *F-measure* de 65% e 83% para as classes *Alto Risco* e *Risco Médio-Baixo*, respectivamente.

Tabela I: Desempenho do modelo de classificação na etapa de teste na base Compas

Classe	Métricas (%)		
	Precisão	Sensibilidade	*F-measure*
Alto	56.0	79.0	65.0
Médio-Baixo	90.0	76.0	83.0
Acurácia	77.0		

A Tabela II apresenta a sequência de indivíduos gerados em uma aplicação da solução proposta, mantendo as mesmas identificações (*I*, *I'*, *C*, *C'*) utilizadas na Seção 4. Na tabela estão presentes somente os atributos modificados ao longo da execução. Em linhas gerais, a execução consistiu em:

(1) Foi selecionada aleatoriamente uma instância original *I* do conjunto de dados. Dentre outros atributos, trata-se de um indivíduo sem reincidência (*reincidência = 0,reincidência em dois anos = 0*), da raça caucasiana (raça = 2) e pertencente a classe *Médio-Baixo*;

(2) A partir de uma alteração simples no atributo protegido raça para afro-americano (raça = 0), gera-se o indivíduo sintético *I'*, o qual manteve-se na classe *Médio-Baixo*;

(3) Uma vez que a alteração simples da raça não foi suficiente para evidenciar uma injustiça do modelo, faz-se uso do CSSE na busca por exemplos contrafactuais. Um dos indivíduos encontrados *C* seguramente pertence a classe *Alto* já que o CSSE gera apenas contrafactuais válidos. Neste caso, *C* trata-se de um indivíduo com reincidência em dois anos (*reincidência = 1, reincidência em dois anos = 1*).

(4) Finalmente, para verificar se há injustiça nas decisões do modelo, cria-se o indivíduo *C'*, o qual é idêntico a *C*, exceto por ser da raça caucasiana. A classificação de *C'* como *Médio-Baixo* enquanto *C* pertence a classe *Alto* gera a evidência da injustiça do modelo.

Tabela II: Indivíduos gerados para avaliação de injustiça. Considere: ***reincidência***: Não (0), Sim (1), ***reincidência em dois anos***: Não (0), Sim (1), ***raça***: Afro-Americano(0), Asiático(1), Caucasiano(2), Hispânico(3), Nativo americano (4), Outros(5).

Indivíduo	**Atributos modificados**			**Risco de Reincidência (Classe)**
	reincidência	reincidência em dois anos	raça	
I	0	0	2	Médio-Baixo
I'	0	0	0	Médio-Baixo
C	1	1	0	Alto
C'	1	1	2	Médio-Baixo

Desta forma, a estratégia proposta aplicada ao modelo gerado para o conjunto de dados Compas gerou dois indivíduos sintético C e C' classificados de formas diferentes em relação ao risco de reincidência criminal devido a pertencerem a raças diferentes. Destaca-se a contribuição essencial do método CSSE na geração do contrafactual que permitiu demonstrar a evidência de injustiça.

7. CONSIDERAÇÕES FINAIS

Com o avanço do uso do Aprendizado de Máquina nos dias atuais, torna-se cada vez mais crucial abordar a questão da justiça nesses modelos, a fim de garantir que não sejam enviesados em relação a qualquer grupo minoritário. Neste artigo, buscamos apresentar métricas de justiça que indicam se um modelo é justo ou não, bem como discutir os métodos que podem ser empregados para assegurar a justiça em uma base de dados. Além disso, abordamos como a questão da justiça tem sido tratada em outros estudos.

Assim, e como objetivo principal deste trabalho, contribuímos com um método de identificação de injustiça em modelos de Aprendizado de Máquina utilizando o CSSE. Para tal, testamos esse método com a base Compas e mostramos a potencialidade do método na tarefa de identificação de injustiça.

Como proposta para trabalhos futuros, pretendemos realizar uma análise mais aprofundada do comportamento do método em diferentes bases de dados e com modificações em diferentes atributos protegidos.

REFERÊNCIAS

Aggarwal, A., Lohia, P., Nagar, S., Dey, K., and Saha, D. Black box fairness testing of machine learning models. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ESEC/FSE 2019. Association for Computing Machinery, New York, NY, USA, pp. 625–635, 2019.

Balbino, M. d. S., Zárate, L. E. G., and Nobre, C. N. Csse - an agnostic method of counterfactual, selected, and social explanations for classification models. *Expert Systems with Applications*, 2023.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Leveraging labeled and unlabeled data for consistent fair binary classification, 2020.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness, 2011.

Edor, J. John rawls's concept of justice as fairness. *PINISI Discretion Review* vol. 4, pp. 179, 12, 2020.

Gomez, O., Holter, S., Yuan, J., and Bertini, E. Advice: Aggregated visual counterfactual explanations for machine learning model validation. *2021 IEEE Visualization Conference (VIS)*, 2021.

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., and Turini, F. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems* 34 (6): 14–23, 2019.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning, 2016.

Jain, A., Ravula, M., and Ghosh, J. Biased models have biased explanations, 2020.

Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification, 2018.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. Counterfactual fairness, 2018.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54 (6), jul, 2021.

Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* vol. 267, pp. 1–38, 2019.

Oneto, L. and Chiappa, S. pp. 155–196. In L. Oneto, N. Navarin, A. Sperduti, e D. Anguita (Eds.), *Fairness in Machine Learning*. Springer International Publishing, Cham, pp. 155–196, 2020.

Petersen, F., Mukherjee, D., Sun, Y., and Yurochkin, M. Post-processing for individual fairness, 2021.

Saxena, N. A. Perceptions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. Association for Computing Machinery, New York, NY, USA, pp. 537–538, 2019.

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., and Liu, Y. How do fairness definitions fare? testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artificial Intelligence* vol. 283, pp. 103238, 2020.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., and Wilson, J. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics* 26 (1): 56–65, 2020.