Justiça nas previsões de modelos de Aprendizado de Máquina: um estudo de caso com dados de reincidência criminal

Ronaldo Lopes Inocêncio Júnior[1], **Leonardo Silveira**[1], **Victor Castro Nacif de Faria**[1,2], **Ana Carolina Lorena**[1]

[1]Instituto Tecnológico de Aeronáutica - ITA
Praça Marechal Eduardo Gomes, 50, São José dos Campos-SP

[2]Universidade Federal de São Paulo - UNIFESP
Instituto de Ciência e Tecnologia - ICT
Avenida Cesare Mansueto Giulio Lattes, 1201, São José dos Campos-SP

{ronaldo.junior,leonardo.silveira,victor.faria}@ga.ita.br,

aclorena@ita.br

Abstract. *The use of Data Science and Artificial Intelligence techniques is permeating several crictical areas nowadays. This includes law and justice, where some data-based decision models give support to determine the risk of recidivism of convicts. This paper analyzes the COMPAS criminal recidivism database with Machine Learning techniques. This dataset contains data from offenders in the USA along a risk of recidivism. Specifically, we focus on the fairness of the decisions in relation to the race attribute. As the race bias was found to be embedded within the dataset as a consequence of the skeweness of this attribute in relation to the target variable, we also experimented a simple balancing of the dataset in regard to the race attribute.*

Resumo. *O uso de técnicas de Ciência de Dados e Inteligência Artificial permeia diversas áreas críticas hoje em dia. Isso inclui a lei e justiça, onde alguns modelos de decisão baseados em dados dão suporte para determinar o risco de reincidência de condenados. Este artigo analisa os dados de reincidência criminal do conjunto de dados COMPAS com técnicas de Aprendizado de Máquina. Esse conjunto de dados contém dados de criminosos nos EUA e seu risco de reincidência. Especificamente, focamos na justiça das decisões dos modelos gerados a partir desses dados em relação ao atributo de etnia dos indivíduos. Como o viés racial foi encontrado dentro do conjunto de dados como uma consequência da assimetria deste atributo em relação à variável alvo, também experimentamos um balanceamento simples do conjunto de dados em relação ao atributo etnia.*

1. Introdução

O uso de algoritmos de Aprendizado de Máquina (AM) tem se intensificado em diversas áreas [Storm et al. 2020, Hart et al. 2021, Meuwly 2021]. Modelos de AM têm sido responsáveis por auxiliar a tomada de decisões em diferentes situações [Bunker and Thabtah 2019, Mohammad-Rahimi et al. 2021, AlQuraishi 2021]. A utilização desses algoritmos muitas vezes leva a resultados com um alto grau de

impacto para uma certa área, como na detecção de fraudes em cartões de crédito [Awoyemi et al. 2017], em detecção de problemas cardíacos por meio de sinais de eletrocardiograma [Salgueiro 2020], na previsão do comportamento da bolsa de valores [Usmani et al. 2016], entre outros exemplos.

Porém, tal avanço no uso de técnicas de AM também pode trazer sérias consequências que não podem ser negligenciadas. Um exemplo é o aumento da disseminação de notícias falsas nas redes sociais pela manipulação de vídeos utilizando *DeepFake* [Botha and Pieterse 2020]. Assim como o caso de *Tay*, um *chatbot* desenvolvido pela empresa *Microsoft* em 2016 que interagia com usuários da rede social *Twitter* e passou a exibir atitudes racistas e sexistas apenas 16 horas após o seu lançamento [Fuchs 2018]. Outro caso similar ocorreu em 2015, na qual os usuários descobriram que o *software Google Photos* classificava pessoas de pele negra como gorilas [Garcia 2017].

Muitos desses problemas estão relacionados ao viés intrínseco à base de dados utilizada no treinamento do algoritmo de AM [Prates et al. 2020], devido à falta de representatividade dos grupos envolvidos. Este tipo de viés embutido nos dados pode ocasionar em problemas no desempenho dos modelos de AM, prejudicando sua capacidade de decisão para determinadas populações [Khosla et al. 2012]. Esse é um problema relevante, pois vidas serão afetadas em alguma medida com a adoção de tais modelos para decisões de impacto social.

Nos Estados Unidos, houve uma grande adoção nos tribunais de justiça de algoritmos capazes de gerar pontuações para réus em julgamento, sendo o mais utilizado dentre eles o algoritmo desenvolvido pela *Northpointe*, uma empresa privada com fins lucrativos, denominado COMPAS ou *Correctional Offender Management Profiling for Alterative Sanctions*. Tal algoritmo visa definir uma pontuação para o risco que o réu tem de cometer outros delitos (reincidir), pontuação essa conhecida como *avaliação de risco*. Essa pontuação é utilizada na tomada de decisões pelo sistema de justiça criminal de alguns estados, desde a atribuição de valores de fiança até a decisão sobre liberdade do réu [Angwin et al. 2016]. Porém, não é revelado como o processo decisório dessas pontuações é realizado, comportando-se, assim, como uma caixa preta [Rudin et al. 2020].

Neste trabalho, pretende-se avaliar quais características da base de dados podem ser exploradas ao produzir um modelo de AM que tenha um melhor aspecto de *fairness* (justiça) comparado com o apresentado pela *Northpointe*. Propomos como hipótese a implicação em aspectos de *fairness* da interpretabilidade do modelo e do balanceamento dos dados de treinamento quanto à variável alvo.

Para alcançar os objetivos, realizamos uma análise do desempenho de modelos de AM gerados a partir de dados da base COMPAS, focando em uma análise dos erros por grupo de populações de etnias distintas. Essa análise segue o conceito de *fairness* nas decisões dos modelos de AM, que não devem ser afetadas por atributos sensíveis dos dados. Dada a sub-representação da população de pele negra em uma das classes do problema de reincidência, foi ainda experimentado o uso de um balanceamento dos dados segundo a etnia dos indivíduos para o treinamento dos modelos de AM. Este balanceamento equilibrou os erros e acertos por classe dos modelos de AM testados, sendo efetivo na diminuição de vieses em suas decisões.

Este trabalho está estruturado como segue. Na Seção 2 são apresentados e discutidos alguns trabalhos e conceitos relacionados à temática do artigo. Na Seção 3 é apresentada a metodologia seguida nos experimentos realizados neste artigo. Na Seção 4 os resultados desses experimentos são apresentados. O trabalho é concluído na Seção 5.

2. Trabalhos Relacionados

O trabalho de [Angwin et al. 2016] concluiu, por meio de análise exploratória da base COMPAS, que existe um viés do algoritmo proprietário responsável pela classificação de risco de reincidência em relação a detentos de pele negra e branca. Entre os resultados encontrados, a taxa de falsos positivos (detentos classificados como reincidentes que, posteriormente, não reincidiram) se mostrou maior entre indivíduos de pele negra (44,9%) do que entre indivíduos de pele branca (23,5%), enquanto que a taxa de falsos negativos (detentos que foram classificados como de baixo risco, mas reincidiram) entre brancos (47,7%) supera aquela observada entre negros (28,0%).

Outros trabalhos chegaram a conclusões similares àquelas de [Angwin et al. 2016] por meio de ferramentas estatísticas de testes de hipóteses. Khademi e Honavar [Khademi and Honavar 2020] concluíram que a hipótese de que não existiria uma relação causal entre etnia do detento e sua pontuação foi rejeitada, indo ao encontro das conclusões de [Angwin et al. 2016]. Já Rudin et al. [Rudin et al. 2020] argumentam que existe um viés sobre a base de dados, mas relacionado ao atributo idade dos detentos, discordando dos resultados de [Angwin et al. 2016] quanto ao viés racial dos dados.

Corbett-Davies e Goel [Corbett-Davies and Goel 2018] apontam uma gama de definições oriundas de pesquisadores para justiça quanto ao projeto de ferramentas de avaliação de riscos (*Risk assessment tools*). Para seu entendimento, considera-se a definição de *atributos protegidos* ou *atributos sensíveis* como características dos dados que não devem ser considerados durante o processo decisório a depender do contexto do algoritmo. Exemplos comuns de atributos protegidos são etnia e gênero de indivíduos. Dentre aspectos relacionados ao desempenho justo dos algoritmos, os autores destacam: (i) *anticlassificação*, em que as ferramentas não devem determinar previsões baseadas em atributos protegidos; (ii) paridade de classificação, em que medidas de desempenho do algoritmo devem ser similares para observações pertencentes a grupos com atributos protegidos idênticos; e (iii) calibração, em que a distribuição de pontuações de risco como saídas dos algoritmos devem ser idênticas para diferentes grupos de observações.

Em outro trabalho, Mehrabi et al. [Mehrabi et al. 2021] mencionam a necessidade de paridade estatística (*statistical parity*), que prega similaridade de proporção de verdadeiros positivos oriundos do algoritmo para grupos de observações com diferentes valores para os atributos protegidos. Além deste, há o aspecto justiça por desconhecimento (*fairness through unawareness*), que trata da remoção de características sensíveis durante o processo de treinamento dos algoritmos de AM.

Todavia, obstáculos estatísticos relativos aos dados de entrada complicam a eficiente obtenção e aplicação de tais princípios sobre os algoritmos de AM e suas saídas. Como exemplo de tais limitações observáveis na base COMPAS, a distribuição de pontuações de risco varia entre grupos étnicos, de maneira que a adoção de um critério de limiar de risco para as pontuações pode causar diferentes taxas de falsos positivos ou negativos por população [Corbett-Davies and Goel 2018]. Outra limitação observada pe-

los autores surge de vieses incorporados no processo de coleta de dados, como vieses de decisões judiciais oriundas de juízes e a diferença em atividade criminal já observada entre grupos como reflexo de desigualdades sociais. Em relação aos atributos e valores contidos na base de dados COMPAS, autores identificaram ainda evidências de inconsistências ou erros em atributos de detentos avaliados pelo algoritmo [Rudin et al. 2020]. Por fim, Paiva et al. [Paiva et al. 2022] analisaram o perfil de dificuldade de classificação de observações de diferentes grupos étnicos no conjunto COMPAS. Com ou sem o uso do atributo sensível, o perfil de erros de classificação de uma combinação de diferentes algoritmos de AM se mantém similar e tendencioso em relação aos grupos étnicos envolvidos. Contudo, também foram verificadas dificuldades de classificação de algumas observações de acordo com o padrão verificado para o número de infrações juvenis cometidas.

3. Materiais e Métodos

A base de dados COMPAS conta com 6396 observações, cada uma descrevendo informações de um detento e rotulada com um risco de reincidência. Desses detentos, 45,9% cometeram reincidência, enquanto 54,1% não cometeram e a base pode ser considerada balanceada quanto ao atributo meta.

Originalmente há 53 atributos descrevendo cada um dos detentos, incluindo informações processuais e crimes prévios. Neste trabalho foram mantidos sete dos atributos para a geração dos modelos de AM, incluindo o atributo alvo. Os atributos excluídos compreendem dados nominais e processuais, como nome, sobrenome, número de identificação do processo criminal e datas de entrada na polícia. Adicionalmente, a base de dados original contém o atributo Pontuação de Risco de Reincidência (risco de o acusado cometer um crime de qualquer natureza no período de até dois anos após o crime atual), calculado pelo algoritmo comercial COMPAS e utilizada pelos juízes Norte-Americanos. Esse atributo foi utilizado por [Angwin et al. 2016] de modo a investigar se havia preconceito ou *viés* relacionado à etnia dos condenados na decisão do algoritmo comercial, referente à pontuação dada a cada indivíduo da população. Uma vez que o objetivo do nosso estudo não é avaliar o algoritmo comercial, mas gerar um novo modelo de predição de reincidência, esse atributo foi excluído da base de dados. A variável resposta utilizada em nosso estudo foi a informação se o indivíduo cometeu ou não reincidência no intervalo de dois anos após sua prisão.

Por fim, o atributo estado civil foi integrado à base de dados por meio de outro conjunto fornecido por [Angwin et al. 2016], e os três atributos que descrevem as diferentes infrações juvenis passíveis de serem cometidas por um indivíduo da população (*Juvenile Fellony Count*, *Juvenile Misdemeanor Count* e *Juvenile Other Count*) foram somados, resultando em apenas um atributo. A descrição dos atributos empregados, bem como sua classificação por tipo e escala, é mostrada na Tabela 1.

Na etapa de pré-processamento, as variáveis idade do acusado e número total de delitos juvenis foram normalizadas para valores entre 0 e 1, e foi empregada a codificação *one-hot encoding* das variáveis categóricas.

Devido ao nosso interesse em investigar como os atributos sensíveis se relacionam a variável resposta, na Tabela 2 é apresentada uma descrição do porcentual total de cada grupo étnico na base COMPAS e a porcentagem de reincidência observada em cada uma delas. Nota-se que afro-americanos e caucasianos compõem mais de 85% da população,

Tabela 1. Descrição dos atributos da base de dados COMPAS utilizados, seu tipo e escala.

Nome do Atributo	Descrição	Tipo	Escala
Age	Idade do Acusado	Quantitativo Discreto	Racional
Race	Etnia do Acusado	Qualitativo	Nominal
Sex	Sexo do Acusado	Qualitativo	Nominal
Marital Status	Estado civil	Qualitativo	Nominal
Juvenile Summed Offenses	Número total de delitos juvenis	Quantitativo Discreto	Racional
Charge Degree	Grau do delito cometido	Qualitativo	Nominal
Prior Count	Contagem de antecedentes criminais na vida adulta	Quantitativo Discreto	Racional
Two Year Recidivism	Variável binária indicando se ocorreu reincidência	Qualitativo Discreto	Nominal

com o restante dividido em quatro etnias sobressalentes. Destaca-se também que as diferentes etnias encontram-se desbalanceadas frente à variável resposta de reincidência: a população afro-americana, por exemplo, tem porcentagem de reincidência de 52,7%, enquanto a população caucasiana apresenta porcentagem de reincidência de 39,5%.

Tabela 2. Percentual da população e taxa de reincidência por etnia.

Etnia	Percentual da População	Taxa Reincidência
Afro-Americana	51,2%	52,7%
Caucasiana	34,0%	39,5%
Hispânica	8,5%	37,5%
Outras	5,5%	36,6%
Asiática	0,4%	25,8%
Nativo-Americana	0,1%	45,4%

Foram selecionados diferentes modelos de AM a serem avaliados nessa base de dados, sendo o critério de seleção a sua interpretabilidade. O foco na interpretabilidade tem como intenção fornecer transparência nas decisões dos modelos de AM nessa área crítica, seguindo o tripé FAT (*fairness, accountability and transparency*) [Ahmad et al. 2020]. As técnicas escolhidas foram: Naïve-Bayes (NB), método probabilístico em que a influência de cada variável isolada na previsão das classes pode ser aferida; K-vizinhos mais próximos (KNN), em que as observações mais próximas dão indicativos das principais características influenciando a decisão para uma nova observação; e árvore de decisão (AD), uma técnica de AM naturalmente simbólica. Dois modelos de AD foram explorados, sendo eles o AD com profundidade limitada (AD-PD) e o AD sem limite de profundidade (AD-PND).

Os hiperparâmetros avaliados para cada modelo foram:

- KNN: 1, 5, 10, 20, 30 e 40 vizinhos;
- Árvore de decisão com profundidade limitada: 5, 10 e 20 níveis de profundidade, métodos de entropia e índice de Gini;

- Árvore de decisão sem limite de profundidade: métodos de entropia e índice de Gini.
- Naïve-Bayes: somente a variação convencional do algoritmo foi utilizada.

Os modelos foram gerados e avaliados em um procedimento de validação cruzada estratificada com 10 pastas. Os hiperparâmetros das técnicas de AM foram selecionados utilizando um procedimento de validação cruzada interno nas partições de treinamento, com três pastas. Após isso, os resultados preditivos dos métodos foram comparados através do teste de hipótese *Wilcoxon signed-rank* [Demšar 2006], com o objetivo de verificar a existência de uma diferença significativa entre eles quanto a sua acurácia.

O melhor modelo entre os comparados teve em seguida seus resultados analisados sob o aspecto de justiça frente à variável sensível *etnia*. Devido às pessoas *brancas* e *negras* comporem a maioria da população analisada, as análises focaram nesses subconjuntos, e foram aplicadas da seguinte forma:

1. Análise da base de dados em sua estrutura original;
2. Análise da base de dados retirando o atributo sensível etnia;
3. Análise da base de dados contendo o atributo etnia, com suas classes balanceadas frente à variável resposta.

Para avaliação da justiça para as diferentes instâncias do modelo, foi utilizada a métrica de paridade estatística [Mehrabi et al. 2021], bem como a influência da justiça por desconhecimento quanto aos resultados. Essas análises são feitas considerando as taxas de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos por grupo étnico. Verdadeiros positivos ocorrem quando o modelo de AM prediz corretamente que um indivíduo irá reincidir. Verdadeiros negativos correspondem aos não reincidentes que o modelo de AM classifica corretamente. Falsos positivos ocorrem quando o modelo prevê que um indivíduo irá reincidir, mas isso não ocorre, o que pode resultar em uma condenação injusta. Por outro lado, falsos negativos envolvem soltar indivíduos potencialmente perigosos na sociedade, pois representa o caso de reincidentes erroneamente classificados como não reincidentes.

4. Resultados

O resultado em termos de acurácia dos modelos de AM gerados pelo procedimento de validação cruzada usando a base com todos os seis atributos de entrada, incluindo a etnia, são apresentados na Figura 1. No caso das ADs, há duas variantes testadas, árvore com profundidade determinada (AD-PD) e árvore sem profundidade determinada (AD-PND). Para o método K-vizinhos mais próximos, o hiperparâmetro com melhor resultado na maioria das partições de validação cruzada foi de 30 vizinhos, e para árvore de decisão com profundidade determinada, a variante do algoritmo com cinco níveis e usando método de entropia obteve o melhor resultado médio. Nota-se na Figura 1 que os métodos K-vizinhos mais próximos com 30 vizinhos e a variante da árvore de decisão com 5 níveis de profundidade apresentaram acurácia média maior que os métodos Naïve-Bayes e árvore de decisão sem profundidade máxima determinada.

Os valores de acurácia média atingidos e o desvio padrão dessa taxa são mostrados na Tabela 3. Quantificando o que foi observado nos boxplots da Figura 1, a árvore

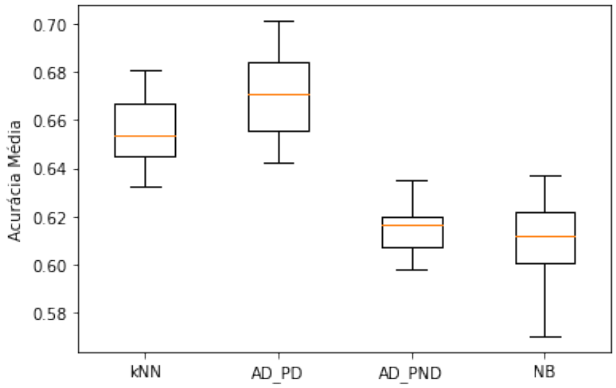


Figura 1. Boxplot da acurácia dos modelos em treinamento por validação cruzada. kNN: K-vizinhos mais próximos; AD-PD: árvore de decisão com profundidade determinada; AD-PND: árvore de decisão sem profundidade determinada; NB: Naïve-Bayes.

Tabela 3. Acurácia dos modelos selecionados, após treinamento por validação cruzada.

Modelo	Acurácia Média	Desvio Padrão
K-vizinhos mais próximos	0,66	0,01
Árvore de Decisão PD	**0,67**	0,02
Árvore de Decisão PND	0,62	0,01
Naïve-Bayes	0,61	0,02

de decisão com profundidade limitada e o algoritmo k-vizinhos mais próximos apresentaram as maiores acurácias médias, com um baixo desvio-padrão. A melhor acurácia, apresentada pela árvore com profundidade limitada, é destacada em negrito.

De modo a verificar a existência de diferença significativa na acurácia dos modelos, foi realizado o teste estatístico de *Wilcoxon signed-rank* para cada par de modelos, com significância de 95%. Como resultado, foi observado que não há uma diferença estatística significativa entre os modelos K-vizinhos mais próximos e árvore de decisão com profundidade determinada, mas que essa diferença existe entre cada um desses modelos e os demais modelos treinados. Os p-valores e estatísticas dos testes estão dispostos na Tabela 4, sendo os significativos destacados em negrito. Devido aos dois melhores modelos serem equivalentes quanto à sua acurácia preditiva, os demais experimentos foram realizados utilizando a árvore de decisão com profundidade limitada.

Tabela 4. Valores da estatística do teste de *Wilcoxon signed-rank* e p-valor para a comparação de diferentes pares de algoritmos.

Algoritmos comparados	Estatística do teste	p-valor
kNN e AD-PD	0,9	0,064
kNN e AD-PND	0,0	**0,002**
kNN e NB	1,0	**0,004**
NB e AD-PD	0,0	**0,002**
AD-PND e AD-PD	0,0	**0,002**

Foi calculado o resultado de previsão desse modelo para a população total e também para a parcela da população branca e negra, separadamente. Como os valores de desvio-padrão foram baixos, eles são omitidos nessa análise. As melhores taxas por linha são destacadas em negrito (maiores valores são melhores para acurácia, taxa de verdadeiros positivos e de verdadeiros negativos; o raciocínio contrário aplica-se para as taxas de falsos positivos e falsos negativos).

Tabela 5. Avaliação do resultado do modelo para população em geral, população negra e população branca.

Métrica	População Geral	População negra	População branca
Acurácia	**0,67**	0,66	0,66
Verdadeiro Positivo	0,66	**0,69**	0,63
Falso Positivo	0,34	**0,31**	0,37
Verdadeiro Negativo	**0,68**	0,64	**0,68**
Falso Negativo	0,32	0,36	**0,32**

A acurácia geral foi semelhante entre as populações consideradas. Contudo, os resultados para a população negra são melhores em termos de taxas de verdadeiros positivos e de falsos positivos. Por outro lado, as taxas de verdadeiros negativos e falsos negativos são melhores para a população de pele branca. Dessa forma, tem-se uma precisão maior para a previsão correta de não reincidentes entre indivíduos brancos e de reincidência entre indivíduos negros. Destaca-se que não existe paridade estatística entre o grupo de pessoas negras e brancas da população, uma vez que existe uma diferença de 6 pontos percentuais entre o percentual de verdadeiros positivos dos dois grupos.

De modo a verificar se o modelo de aprendizado está utilizando a variável sensível etnia como parte do seu processo decisório, foi realizada a exclusão dessa variável da base de dados e o modelo foi novamente treinado. O resultado pode ser visto na Tabela 6. Novamente as melhores taxas por métrica são destacadas em negrito. Nota-se pouca diferença entre ambos os resultados, embora haja um incremento de desempenho na classe positiva (reincidência) e um decremento de desempenho na classe negativa (não reincidência). Disso pode ser inferido que o modelo não está usando o atributo etnia para explicar o fenômeno de reincidência, de modo que não é a presença do atributo como entrada explícita dos modelos que causa viés em seus resultados.

Tabela 6. Métricas de avaliação do modelo ao ser treinado com e sem a presença do atributo etnia.

Métrica	Com atributo etnia	Sem atributo etnia
Acurácia	0,67	**0,69**
Verdadeiro Positivo	0,63	**0,69**
Falso Positivo	0,37	**0,31**
Verdadeiro Negativo	**0,71**	0,68
Falso Negativo	**0,29**	0,32

Na Tabela 2 apresentou-se que a população negra tem uma porcentagem de 51,2% de reincidentes na base de dados, maior que a porcentagem de 39,5% de reincidentes da população branca. Com o objetivo de investigar o efeito desse desbalanceamento das

categorias do atributo sensível frente à variável resposta, foi realizado o balanceamento da quantidade de pessoas reincidentes e não reincidentes para cada uma das etnias. Para fazer esse balanceamento, foram retirados, de maneira aleatória, exemplos de pessoas brancas não reincidentes da base de dados. Essa remoção foi repetida 10 vezes e o resultado médio é reportado. O desvio padrão das métricas acurácia, verdadeiros e falsos positivos, e verdadeiros e falsos negativos não ficou superior a 0,04 para nenhuma população.

A avaliação do modelo com as classes balanceadas segundo o atributo etnia pode ser vista na Tabela 7. Destaca-se a melhora da capacidade do modelo para predição de reincidência para pessoas brancas, o que é notável nas métricas de verdadeiros e falsos positivos. Ademais, o desempenho geral do algoritmo para a população negra teve melhoras na predição de não reincidência, refletindo nas taxas de verdadeiros negativos e falsos negativos para essa população.

Tabela 7. Avaliação do resultado do modelo para população em geral, população negra e população branca, após rebalanceamento do atributo etnia.

Métrica	População Geral	População negra	População branca
Acurácia	**0,68**	**0,68**	**0,68**
Verdadeiro Positivo	**0,68**	0,67	0,65
Falso Positivo	**0,32**	0,33	0,35
Verdadeiro Negativo	0,68	**0,69**	**0,69**
Falso Negativo	0,32	**0,31**	**0,31**

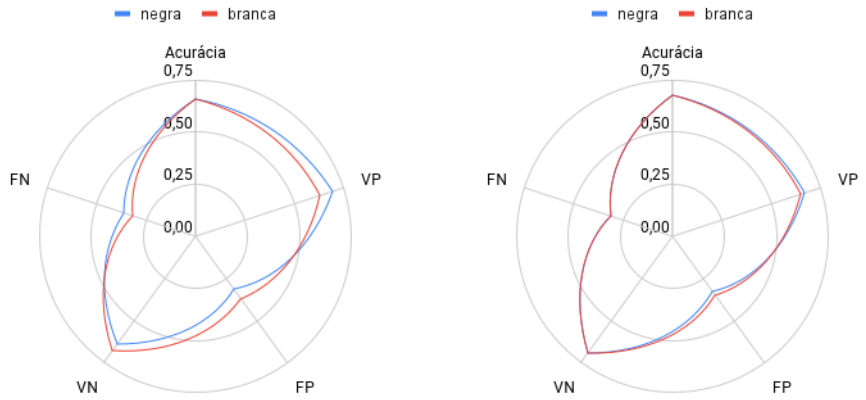
No que tange à métrica paridade estatística, o balanceamento das classes quanto ao atributo etnia reduziu a diferença entre os valores de verdadeiros positivos para pessoas negras e brancas, de 6 pontos percentuais para 2 pontos percentuais, mostrando uma redução no viés do modelo. De fato, de maneira geral as taxas de acerto e erro por tipo de etnia se equiparam quando o balanceamento é realizado, como mostrado nos gráficos radar da Figura 2. No primeiro radar, as taxas se distanciam para as duas etnias, principalmente para verdadeiros positivos (VP) e falsos positivos (FP). No segundo radar, todas as taxas se tornam próximas, incluindo os verdadeiros negativos (VN) e falsos negativos (FN), independente do grupo populacional considerado, o que torna a classificação mais justa e independente do grupo étnico considerado.

Portanto, medidas simples como um balanceamento dos dados por classe em relação a um atributo sensível são mecanismos efetivos para melhorar a paridade estatística do modelo de AM, atenuando o viés inicial da base de dados, que estava se refletindo no desempenho do modelo por população.

5. Conclusões

Neste trabalho foi realizada uma análise de aspecto de justiça nas previsões em relação aos acertos e erros alcançados por modelos de AM para diferentes grupos étnicos em um conhecido conjunto de dados de reincidência criminal. O conjunto de dados original naturalmente possui uma quantidade muito maior de reincidentes de uma população e não reincidentes de outra. Essa assimetria se refletiu nos resultados, que apresentavam baixa paridade estatística.

Notou-se a importância do balanceamento do conjunto de dados considerando a variável sensível e sua representatividade nas classes presentes no conjunto de dados.



(a) Sem balanceamento (b) Com balanceamento

Figura 2. Métricas de desempenho do modelo de AM para o conjunto COMPAS em duas situações: com e sem balanceamento das classes por etnia.

Esse desbalanceamento ao nível dos atributo sensíveis pode influenciar diretamente no poder decisório do algoritmo, principalmente quanto ao aspecto de justiça almejado. As consequências desse desbalanceamento foram observadas nos experimentos, nos quais o algoritmo treinado não era capaz de decidir quando uma pessoa branca iria reincidir com a mesma precisão que uma pessoa negra (e o complemento ocorria em relação à classe de não reincidência). Somente após a correção de distribuição entre as etnias em relação à variável resposta (quanto à composição de reincidentes), o algoritmo apresentou uma característica de maior justiça entre os diferentes grupos étnicos, evidenciado pelo aumento de paridade estatística. Em outras palavras, para obter um classificador com viés mínimo, não importa analisar apenas a proporção da classe alvo, mas também investigar a proporção das classes alvo dada a variável sensível.

Conclui-se também que, mesmo após a remoção do atributo sensível etnia, o resultado do algoritmo manteve-se similar, evidenciando que a presença explícita dessa variável não está explicando o fenômeno em questão. Isto é, não é levado em consideração o atributo protegido durante o processo decisório, endossando o aspecto de justiça por desconhecimento.

Para trabalhos futuros, sugere-se a criação de uma medida de justiça que seja capaz de analisar o desbalanceamento das classes ao nível do atributo sensível, assim como um método para balanceamento dos atributos sensíveis do conjunto de dados, sem que haja perda considerável de informação. Ainda, sugere-se que outros algoritmos de AM sejam treinados sobre os dados e tenham seus resultados preditivos comparados com os algoritmos previamente avaliados. Outra possível investigação é detectar a presença de vieses que estão implícitos em outros atributos em relação ao atributo sensível de etnia.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e da FAPESP (processo 2022/10917-8). A Profa Ana C. Lorena também agradece ao apoio da FAPESP (processo 2021/06870-3).

Referências

Ahmad, M. A., Teredesai, A., and Eckert, C. (2020). Fairness, accountability, transparency in ai at scale: Lessons from national programs. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 690–690.

AlQuraishi, M. (2021). Machine learning in protein structure prediction. *Current opinion in chemical biology*, 65:1–8.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: there's software used across the country to predict future criminals and it's biased against blacks. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

Awoyemi, J. O., Adetunmbi, A. O., and Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI)*, pages 1–9.

Botha, J. and Pieterse, H. (2020). Fake news and deepfakes: A dangerous threat for 21st century information security. In *ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and publishing limited*, page 57.

Bunker, R. P. and Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied computing and informatics*, 15(1):27–33.

Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30.

Fuchs, D. J. (2018). The dangers of human-like bias in machine-learning algorithms. *Missouri S&T's Peer to Peer*, 2(1):1.

Garcia, M. (2017). Racist in the machine: The disturbing implications of algorithmic bias. retrieved december 03, 2017.

Hart, G. L., Mueller, T., Toher, C., and Curtarolo, S. (2021). Machine learning for alloys. *Nature Reviews Materials*, 6(8):730–755.

Khademi, A. and Honavar, V. (2020). Algorithmic bias in recidivism prediction: A causal perspective (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13839–13840.

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A., and Torralba, A. (2012). Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer.

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

Meuwly, M. (2021). Machine learning for chemical reactions. *Chemical Reviews*, 121(16):10218–10239.

Mohammad-Rahimi, H., Nadimi, M., Ghalyanchi-Langeroudi, A., Taheri, M., and Ghafouri-Fard, S. (2021). Application of machine learning in diagnosis of covid-19

through x-ray and ct images: a scoping review. *Frontiers in cardiovascular medicine*, 8:185.

Paiva, P. Y. A., Moreno, C. C., Smith-Miles, K., Valeriano, M. G., and Lorena, A. C. (2022). Relating instance hardness to classification performance in a dataset: a visual approach. *Machine Learning*, 111(8):3085–3123.

Prates, M. O., Avelar, P. H., and Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10):6363–6381.

Rudin, C., Wang, C., and Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1). https://hdsr.mitpress.mit.edu/pub/7z10o269.

Salgueiro, A. T. F. (2020). *Detecção de problemas cardíacos usando sinais de electrocardiograma (ECG)*. PhD thesis, Universidade de Coimbra.

Storm, H., Baylis, K., and Heckelei, T. (2020). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*, 47(3):849–892.

Usmani, M., Adil, S. H., Raza, K., and Ali, S. S. A. (2016). Stock market prediction using machine learning techniques. In *2016 3rd international conference on computer and information sciences (ICCOINS)*, pages 322–327. IEEE.