Transparência e Justiça em Aprendizado de Máquina: Um Estudo com Bibliotecas para Explicabilidade e Detecção e Mitigação de Vieses

Thiago Severo Garbin
Porto Alegre, Brasil
tsgarbin@gmail.com

***Resumo*—A crescente utilização de modelos de aprendizado de máquina por parte de entes empresariais e governamentais com vistas à automatização de decisões trouxe consigo preocupações relacionadas a questões práticas. Maior necessidade de explicabilidade dos modelos e preocupações relacionadas à ética se tornaram centrais, uma vez que algoritmos são suscetíveis a vieses que podem representar cenários injustos a determinados grupos da sociedade. Nesse sentido, este trabalho buscou explorar a aplicação de técnicas de explicabilidade e detecção e mitigação de vieses através do uso de aprendizado de máquina e da base de dados da Relação Anual de Informações Sociais (RAIS). Com a utilização de uma biblioteca voltada a auxiliar no processo de explicabilidade, conseguiu-se alcançar maior transparência no entendimento das predições, enquanto que com uma biblioteca para detecção e mitigação de vieses foi possível gerar, em geral, modelos com métricas de *fairness* em melhores patamares, explicitando a relevância das referidas aplicações.**

***Palavras-Chave*—transparência, justiça, explicabilidade, viés**

I. Introdução

Fatores como a disponibilidade de dados em larga escala e novas arquiteturas de algoritmos impactaram na constante ampliação da adoção de sistemas baseados em aprendizado de máquina, que tomam ou suportam decisões. Tal movimento pode ser notado tanto no setor público quanto no setor privado, e em diferentes áreas, dentre as quais estão a educacional, a financeira, a empregatícia e a jurídica [1, 2].

Essa automatização, no entanto, trouxe consigo uma gama de preocupações, que incluem tanto a capacidade de explicar as decisões dos modelos, bem como a ética inerente a estas. Sobre esse último ponto, tornaram-se notórios casos em que algoritmos cometeram erros reiterados em desfavor de determinados indivíduos ou grupos. No estado da Flórida, foi descoberto que o algoritmo que indicava a probabilidade de reincidência de réus do sistema penal atribuía erroneamente o rótulo de alto risco a réus afro-americanos a uma taxa quase duas vezes maior que aquela relacionada a réus de cor branca [1]. A empresa *Amazon* deixou de utilizar uma versão de seu sistema de recrutamento pois este penalizava currículos de mulheres [3]. Ao criar miniaturas para imagens de grandes dimensões, a ferramenta de recorte automático da rede social *Twitter* privilegiava pessoas brancas em relação a negras e mulheres em relação a homens [4].

A necessidade de explicar as saídas dos modelos e de assegurar que estes não apresentem vieses se encontra expressa também em dispositivos elaborados em níveis governamentais. A Estratégia Brasileira para Inteligência Artificial (EBIA), coloca como uma das ações estratégicas de seus eixos a necessidade de promover a produção de uma Inteligência Artificial ética com foco em *fairness*, *accountability* e *transparency* [5]. De maneira semelhante, a *General Data Protection Regulation* (GDPR), a lei que versa sobre a proteção de dados nos países da União Europeia, dispõe como um dos seus princípios relacionados ao processamento de dados pessoais que este ocorra de maneira legal, justa e transparente [6].

A partir do treinamento de um modelo de aprendizado supervisionado com um conjunto de dados público, o presente estudo foca nos campos de transparência e justiça (*fairness*). Por meio de uma biblioteca de explicabilidade, buscou-se analisar de maneira mais clara como um modelo produz as suas predições. Já com a aplicação de uma biblioteca voltada para detecção e mitigação de vieses, foram analisadas métricas, de modo a se verificar a necessidade e a possibilidade de correção dos modelos no sentido destes representarem cenários sem discriminação[1].

O artigo é organizado da seguinte forma. Na Seção II, é feita uma revisão dos principais conceitos relacionados a explicabilidade e detecção e mitigação de vieses. Os experimentos são descritos na Seção III, enquanto os resultados são discutidos na Seção IV. Por fim, na Seção V são apresentadas as conclusões.

II. Explicabilidade e Detecção e Mitigação de Vieses

A. Explicabilidade

As pesquisas em torno da explicabilidade em inteligência artificial (*Explainable AI* - XAI) ganharam impulso com a difusão de algoritmos de aprendizado de máquina, em especial aqueles de aprendizado profundo. Tais algoritmos são considerados modelos caixa-preta em virtude da alta complexidade nas suas estruturas, não linearidade e dificuldade de explicação para pessoas leigas [7]. O campo da XAI, então, é aquele que proporciona ferramentas, algoritmos e técnicas que levam a explicações intuitivas e interpretáveis por humanos das decisões de inteligência artificial, deste modo

[1]Os arquivos do estudo estão disponíveis em https://github.com/thiago-sg/Explicabilidade-e-Mitigacao-de-Vieses--RAIS

procurando satisfazer a busca por transparência e confiança em tais decisões [8].

A biblioteca selecionada para uso neste trabalho é chamada de SHAP, um acrônimo para *Shapley Additive Explanations*. Segundo os autores, buscou-se com essa ferramenta unificar outras abordagens da classe dos métodos *additive feature importance*, uma vez que se trata de uma única solução que satisfaz propriedades desejadas como *Local accuracy*, *Missingness* e *Consistency*. Dessa forma, por meio dos chamados SHAP *values*, é possível analisar a contribuição dos atributos para as predições de modelos, sejam eles mais simples como os lineares, ou mais complexos como os de aprendizado profundo [9].

B. Detecção e Mitigação de Vieses

A proposição de soluções para vieses discriminatórios em modelos de aprendizado de máquina evoluiu ao mesmo tempo em que casos de discriminação foram trazidos à tona. Nesse sentido, houve o surgimento de abordagens para mitigação automática de vieses que atuam em diferentes estágios do ciclo de um trabalho de aprendizado de máquina. Nos métodos chamados *pre-processing*, ocorre a remoção dos vieses encontrados no conjunto de dados antes do treinamento do modelo, enquanto nos chamados *in-processing* se busca melhores métricas de *fairness* em conjunto com a maximização da acurácia durante o treinamento do modelo. Já nos métodos conhecidos como *post-processing*, é realizada uma modificação das predições após o treinamento do modelo, de forma a fazer com que estas se adequem a uma métrica de *fairness* [10, 11].

Em [11] os autores traçam um panorama das ferramentas disponíveis no campo de *fairness*, com vistas a avaliar as potencialidades e recursos disponibilizados por elas. Conforme comparativo de funcionalidades elaborado no estudo, é possível verificar que a biblioteca *AI Fairness 360* (AIF360) é a que apresenta maior número de possibilidades perante às outras cinco consideradas no que se refere a métricas de *fairness* em nível de grupo e quantidade de algoritmos de mitigação de vieses. Por essa razão, decidiu-se escolher essa biblioteca para os experimentos referentes a *fairness* neste trabalho.

Desenvolvido pela IBM, o AIF360 faz parte do conjunto de ferramentas do *Trusted AI*, o qual abarca bibliotecas que permitem explorar a confiabilidade de sistemas de aprendizado de máquina, e que foi doado posteriormente para a Fundação Linux [12]. Uma das intenções era que a biblioteca pudesse contemplar os esforços de ferramentas de *fairness* anteriores [10].

A ferramenta se vale de conceitos como rótulo favorável, que é aquele que confere uma vantagem ao seu portador, grupo privilegiado e atributo protegido, que é aquele o qual se deseja analisar métricas relativas a *fairness* [13]. Vale ressaltar que a simples exclusão do atributo protegido não é suficiente para que a discriminação seja eliminada e pode resultar no chamado efeito *redlining*, uma vez que demais atributos podem estar relacionados àquela variável, levando à sua identificação. Como exemplo, é possível mencionar um atributo relativo ao código postal da área residencial de uma pessoa, que pode estar ligado à sua etnicidade [14].

O AIF360 atualmente conta com 14 algoritmos de mitigação [15], tendo sido elaborado com mais de 71 métricas para detecção de vieses [10]. Apesar das múltiplas definições existentes para *fairness*, sejam estatísticas ou baseadas nas predições do modelo [16], neste estudo serão analisadas quatro métricas relativas a fairness em nível de grupos disponíveis na ferramenta AIF360, descritas a seguir:

- *Statistical Parity Difference*: tem seu valor obtido por meio da diferença entre a probabilidade de obtenção de rótulos positivos no grupo não privilegiado e a probabilidade de obtenção de rótulos positivos no grupo privilegiado. Deste modo, para não haver qualquer discriminação esta métrica condiciona que ambos os grupos devem ter a mesma probabilidade de obtenção do rótulo positivo, resultando em um valor 0. Ainda assim, valores entre -0,1 e 0,1 são aceitáveis em relação a *fairness* [10, 13].
- *Disparate Impact*: é auferida através da divisão da taxa de rótulos favoráveis para determinado grupo considerado não privilegiado, pela taxa de rótulos favoráveis para outro grupo considerado privilegiado, levando em conta um rótulo binário. O valor 1 indica ausência de discriminação e os valores considerados aceitáveis ficam entre 0,8 e 1,25, sendo que aqueles abaixo de 1 indicam vantagem para o grupo considerado privilegiado e acima de 1 o contrário [10, 13].
- *Average Odds Difference*: é calculada a partir da diferença média entre a taxa de falsos positivos (FPR) e a taxa de verdadeiros positivos (TPR), considerando os grupos privilegiados e não privilegiados. Com um valor 0 indicando completa igualdade, valores entre -0,1 e 0,1 são considerados aceitáveis em termos de *fairness*. Valores abaixo de 0 apontam vantagem para o grupo privilegiado e acima de 0 o inverso [10, 13].
- *Equal Opportunity Difference*: apura a diferença na taxa de verdadeiros positivos (TPR) entre o grupo não privilegiado e o grupo privilegiado. Valores abaixo de 0 indicam benefício para o segundo grupo e valores acima de 0 indicam benefício para o primeiro, enquanto um valor 0 indica perfeita igualdade. Valores entre -0,1 e 0,1 estão dentro do limite considerado justo [10, 13].

III. Experimentos

A. Contexto

Os métodos de explicabilidade explicitam como as variáveis impactam a saída do modelo, assim como auxiliam em uma análise inicial de possíveis discriminações relacionadas a variáveis de interesse. Com base nisso e levando em consideração a base de dados utilizada (Seção III B), optou-se inicialmente pela aplicação das técnicas da biblioteca SHAP.

Posteriormente, com a utilização de uma biblioteca de detecção e mitigação de vieses, procedeu-se a análise de

métricas de *fairness* em nível de grupos, antes e após a aplicação de três algoritmos de mitigação, sendo um da fase *pre-processing*, um da fase *in-processing* e o último da fase *post-processing*. Para esta análise foram colocadas como variáveis protegidas as de sexo e raça, assumindo, para os casos dos algortimos *pre* e *post-processing*, que o sexo masculino é privilegiado, bem como a raça branca.

B. Base de Dados

O conjunto de dados utilizado no estudo é originado da Relação Anual de Informações Sociais (RAIS), a qual se trata de um documento a ser submetido anualmente por empresas e empregadores, subsidiando o controle da atividade trabalhista e as estatísticas do mercado de trabalho para o governo brasileiro [17]. Os microdados não identificados da RAIS trabalhador podem ser obtidos via *File Transfer Protocol* (FTP) em formato csv [18]. Tais dados contém informações pessoais e relativas ao emprego de cada trabalhador que tenha possuído vínculo empregatício em determinado ano. Os dados obtidos para posterior pré-processamento são relativos aos estados da região Sul do Brasil no ano de 2019, correspondendo a 12284030 registros.

C. Pré-processamento

Com o intuito de realizar os experimentos considerando profissionais da área de TI no estado do Rio Grande do Sul, foram excluídos da base os dados de trabalhadores cujo empregador não era do estado gaúcho e aqueles que não tinham sua ocupação pertencente a uma das seguintes famílias da classificação brasileira de ocupações (CBO 2002): Analistas de Tecnologia da Informação, Administradores de Tecnologia da Informação, Engenheiros em Computação, Diretores de Tecnologia da Informação e Gerentes de Tecnologia da Informação.

Fixou-se como rótulo a variável que remete à remuneração do mês de dezembro. Essa remuneração foi binarizada com base na média verificada no conjunto de treino correspondente a 80% dos dados, resultando em um conjunto de profissionais rotulados com remuneração acima ou abaixo da média, que corresponde a um valor de R$7.066,59. Para as variáveis explicativas foram mantidas aquelas relativas a faixa etária, faixa de horas mensal do contrato de trabalho, faixa de tempo no emprego, faixa de escolaridade, indicador de deficiência, raça, sexo, faixa de tamanho do estabelecimento empregador e ocupação.

Com base nas variáveis explicativas, foram excluídas as instâncias cujo valor era inválido ou tido como "ignorado". Além disso, a variável raça foi binarizada entre brancos e não brancos, bem como cada ocupação se tornou uma variável binária por meio de processo de *one-hot encoding*. Assim, o conjunto de dados resultante ficou com 16532 instâncias, das quais, na variável protegida sexo, 20,70% são mulheres e 79,30% são homens, e na variável protegida raça, 7,08% são não brancos e 92,92% são brancos.

D. Modelo e Algoritmos para Mitigação de Vieses

Para os experimentos, optou-se pela utilização de um modelo base *XGBoost Classifier* da linguagem *Python*, com fixação de *random_state* de valor 0 para reproducibilidade. Na biblioteca de explicabilidade SHAP e no algoritmo de mitigação de vieses da fase *in-processing*, foram utilizadas apenas as partições de treino, correspondente a 80% dos dados, e teste, correspondente a 10% dos dados. Para os casos dos algoritmos de mitigação de vieses *pre* e *post processing*, fez-se uso também do conjunto de validação, correspondente a 10% dos dados. Este foi base para a busca pelos melhores limiares de classificação, levando em conta a acurácia balanceada, que é obtida pela divisão por 2 da soma entre a taxa de verdadeiros positivos (sensibilidade) e taxa de verdadeiros negativos (especificidade). O uso da acurácia balanceada, nos casos em que foi possível, deu-se em virtude do suave desbalanceamento dos dados, que apresentou 63,28% de rótulos para salários abaixo da média, e 36,72% para salários acima da média, no conjunto de treino. Ressalta-se, por fim, que para a aplicação dos algoritmos de mitigação de vieses é necessário antes instanciar os dados como objetos da classe *Binary Label Dataset* do AIF360, assinalando neles os atributos protegidos.

Como mencionado, os algoritmos para mitigação de vieses selecionados atuam em diferentes estágios. O *Reweighing* (RW), algoritmo *pre-processing*, se utiliza da atribuição de pesos para as instâncias do conjunto de treino com base em suas probabilidades esperadas e reais, de modo que este conjunto se torne balanceado e o modelo nele treinado seja livre de discriminação. Assim, o peso atribuído a uma instância será dependente do valor em seu atributo sensível e a classe a qual ela pertence [19].

O *Exponentiated Gradient Reduction* (EGR), do estágio *in-processing*, recebe um estimador, de classificação ou regressão, e, considerando um parâmetro de restrição de *fairness* escolhido, reduz o problema de classificação justa a uma sequência de problemas de classificação sensíveis a custo. Deste modo, é gerado um classificador randomizado capaz de otimizar métricas de performance ao mesmo tempo em que satisfaz a restrição [20].

Por fim, o *Reject Option Classification* (ROC), da fase *post-processing*, atua modificando os labels obtidos após a classificação. Isto ocorre por meio da estimação de uma região crítica sobre os limites de decisão, que faz as instâncias dos grupos não privilegiados e privilegiados serem rotuladas, respectivamente, de maneira favorável e não favorável [21].

IV. Resultados

A. Explicabilidade

Optou-se pela utilização da biblioteca de explicabilidade SHAP em virtude da sua gama de visualizações intuitivas, como é o caso da Figura 1, que sumariza os impactos das doze variáveis mais importantes para o modelo simplificado, que se utilizou somente das partições de treino e teste. No topo se encontra a variável com maior impacto médio, enquanto as cores representam o valor da variável em cada instância e a posição horizontal indica para que lado se dá o impacto.

Foi possível averiguar que variáveis como as relativas a tamanho do estabelecimento, escolaridade, idade e tempo no emprego direcionam os rótulos para salários acima da média, quando seus valores são altos, bem como direcionam para valores abaixo salários abaixo da média quando seus valores são baixos. Nas variáveis relativas às profissões, destaca-se que há um indicativo de que Engenheiro de Sistemas Operacionais e Gerente de Desenvolvimento de Sistemas se apresentam como ocupações que geram salários acima da média, enquanto Analista de Suporte Computacional e Analista de Redes e Comunicação de Dados proporcionam o contrário.

Outro ponto a ser destacado é que há indício de que a variável relacionada ao sexo do trabalhador produz rótulos em direção a salários abaixo da média em caso de sexo feminino. Deste modo, considera-se que houve uma contribuição positiva no que se refere à transparência do modelo, a qual também possibilitou um primeiro indicativo a respeito de vieses indesejados e que serão discutidos mais profundamente a seguir.

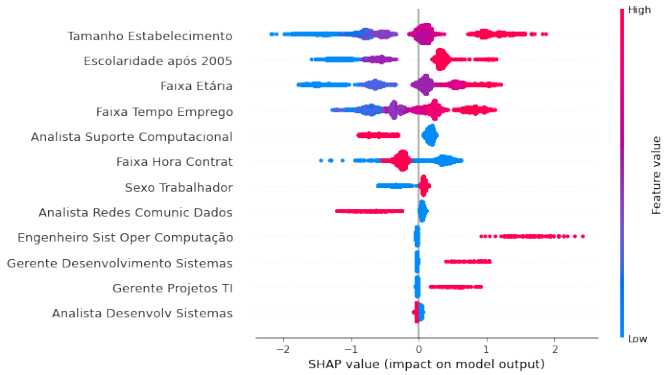

Figura 1. Impacto das variáveis no modelo

B. Detecção e Mitigação de Vieses

Para o algoritmo RW, inicialmente foi realizada a busca pelo limiar ótimo que correspondia à melhor acurácia balanceada com base no conjunto de dados de validação original. O limiar ótimo encontrado foi 0,3862, com acurácia de 79,34%. Na sequência, procedeu-se a execução do algoritmo de mitigação separadamente para os atributos protegidos escolhidos, sexo e raça.

Após a aplicação do algoritmo de mitigação sobre o conjunto de treino, verificou-se melhora nas métricas de *Disparate Impact* e *Statistical Parity Difference* relativas a este conjunto, pois estas assumiram valores virtualmente iguais aos ideais, passando de 0.7944 para 1 e -0.0788 para 0, na variável sexo, e 0,7603 para 1 e -0,0895 para 0, no atributo raça.

O modelo treinado neste conjunto com os novos pesos e considerando o mesmo limiar de classificação ótimo obtido no modelo original, também apresentou expressiva melhora em quase todas as métricas de fairness para ambos atributos protegidos nas predições do conjunto de teste, conforme a Tabela I. A acurácia apresentou pequena melhora para o atributo sexo e pequena queda para o atributo raça. A *Statistical Parity Difference* passou de -0,0823 para -0,0290 no atributo sexo e de -0,1799 para -0,1211 no atributo raça, ao passo que a *Disparate Impact* saiu de 0,8149 para 0,9337 no atributo sexo e de 0,5915 para 0,7251 no atributo raça, evidenciando pior desempenho de mitigação no atributo raça nesses casos, pois permaneceram com valores fora dos aceitáveis. A *Average Odds Difference* passou de -0,0347 para 0,0223 no atributo sexo e de -0,1126 para -0,0398 no atributo raça, enquanto na métrica *Equal Opportunity Difference* não houve melhora na variável sexo, que era o único caso de viés originalmente para o grupo não privilegiado, uma vez que se passou de 0,0032 para 0,0710, ao contrário do atributo raça, que passou de -0,1337 para 0,0320.

TABELA I
MÉTRICAS REWEIGHING

Atributo	Sexo		Raça	
	Antes	Após	Antes	Após
Acurácia	79,34%	79,51%	79,34%	79,25%
Statistical Parity Difference	-0,0823	-0,0290	-0,1799	-0,1211
Disparate Impact	0,8149	0,9337	0,5915	0,7251
Average Odds Difference	-0,0347	0,0223	-0,1126	-0,0398
Equal Opportunity Difference	0,0032	0,0710	-0,1337	-0,0320

De maneira diferente dos demais algoritmos, para o EGR foi tomado como base de comparação apenas o modelo treinado e testado de maneira simplificada, com conjuntos de treino e teste, sem busca pela melhor acurácia balanceada, pois o algoritmo não dispõe de método para se obter a probabilidade das predições. Além disso, a execução do algoritmo se deu apenas uma vez, já que este recebe como hiperparâmetros o estimador e uma restrição escolhida, que nesse caso foi a de *True Positive Rate Difference*, para então ser ajustado aos dados de treino, sem a possibilidade de indicar separadamente grupos privilegiados como ocorre nos outros métodos. Por fim, se procedeu a verificação das suas predições no conjunto de teste.

Desse modo, as métricas expostas na Tabela II demonstram queda na acurácia, de 81,80% para 80,83%, porém com melhora na *Statistical Parity Difference*, de -0,0820 para -0,0618 no atributo sexo e de -0,1914 para -0.1021 no atributo raça, que seguiu fora do limiar aceitável. Na *Disparate Impact*, apesar da evolução, as métricas seguiram com valores ruins, pois se passou de 0,7686 para 0,8196 no atributo sexo e de 0,4547 para 0,6972 no atributo raça, enquanto a *Average Odds Difference* saiu de -0,0399 para -0,0243 no atributo sexo e de -0,1357 para -0,0469 no atributo raça, que ficou dentro do aceitável em termos de *fairness*. Por fim, na *Equal Opportunity Difference* houve melhora com relação à métrica para o atributo sexo, de -0,0203 para -0,0144, e também no atributo raça, de -0,1824 para -0,0886.

No algoritmo ROC, a base de comparação se dá com as métricas obtidas no limiar ótimo correspondente à melhor acurácia balanceada. Entretanto, o algoritmo calcula um novo limiar ótimo conforme os hiperparâmetros informados, que são relativos à uma restrição de fairness selecionada, que para esse caso foi a *Statistical Parity Difference*, e a banda de crítica de classificação.

TABELA II
MÉTRICAS EXPONENTIATED GRADIENT REDUCTION

Atributo	Sexo		Raça	
	Antes	Após	Antes	Após
Acurácia	81,80%	80,83%	81,80%	80,83%
Statistical Parity Difference	-0,0820	-0,0618	-0,1914	-0,1021
Disparate Impact	0,7686	0,8196	0,4547	0,6972
Average Odds Difference	-0,0399	-0,0243	-0,1357	-0,0469
Equal Opportunity Difference	-0,0203	-0,0144	-0,1824	-0,0886

Assim, com o novo limiar ótimo relativo à acurácia balanceada obtido pelo algoritmo, de valor 0,3466, e conforme os dados da Tabela III, nota-se melhora nessa métrica, a qual era inicialmente de 79,34% mas passou para 79,50% no atributo sexo e para 79,61% no atributo raça. Quanto às métricas de fairness, a *Statistical Parity Difference* se modificou de -0,0823 para 0,0041 no atributo sexo, e de -0,1799 para -0,0564 no atributo raça, que entrou nos limiares aceitáveis de *fairness*. A *Disparate Impact* evoluiu de 0,8149 para 1,0092 no atributo sexo e de 0,5915 para 0,8703 no atributo raça. A *Average Odds Difference* passou de -0,0347 para 0,0546 no atributo sexo, e de -0,1126 para 0,0378 no atributo raça, enquanto a *Equal Opportunity Difference* saiu de 0,0032 para 0,1005 no atributo sexo, e de -0,1337 para 0,0697 no atributo raça. Os retrocessos verificados nas últimas duas métricas para a variável sexo possivelmente estão atrelados ao parâmetro de restrição escolhido.

TABELA III
MÉTRICAS REJECT OPTION CLASSIFICATION

Atributo	Sexo		Raça	
	Antes	Após	Antes	Após
Acurácia	79,34%	79,50%	79,34%	79,61%
Statistical Parity Difference	-0,0823	0,0041	-0,1799	-0,0564
Disparate Impact	0,8149	1,0092	0,5915	0,8703
Average Odds Difference	-0,0347	0,0546	-0,1126	0,0378
Equal Opportunity Difference	0,0032	0,1005	-0,1337	0,0697

Dessa forma, nota-se que, em geral, os algoritmos de mitigação de vieses utilizados de fato são capazes de proporcionar avanço no que se refere às métricas de fairness e assim, consequentemente, podem auxiliar na construção de modelos mais justos. Entretanto, ao mesmo tempo se percebe que não é possível se obter em todos os casos uma evolução conjunta de todas as métricas.

V. CONCLUSÃO

O presente estudo se propôs a realizar experimentos com bibliotecas de explicabilidade e detecção e mitigação de vieses, de modo a demonstrar possibilidades e a importância de transparência e justiça em aprendizado de máquina. Para isso, foi treinado um modelo *XGBoost Classifier* na base de dados pública RAIS. Tal modelo foi avaliado com a biblioteca SHAP, proporcionando entendimento acerca das variáveis e seus impactos nas predições.

Com a biblioteca AIF360, ainda que os autores mencionem que essa deve ser usada apenas como ponto de partida para análises mais profundas [10], a verificação de métricas de *fairness* evidenciou discriminação relacionada aos atributos de interesse, sexo e raça, na maior parte dos casos contra os grupos não privilegiados. Com os algoritmos de mitigação de vieses, essa discriminação exposta nas métricas pôde ser diminuída na maior parte dos casos, levando a modelos mais justos. Diante disso, mostra-se importante que ocorra discussão de modo reiterado acerca de *fairness* em aprendizado de máquina, uma vez que algoritmos podem ajudar a perpetuar padrões injustos expressos nos dados de treinamento, ainda que estes possam refletir a realidade da sociedade.

Para trabalhos futuros, coloca-se a possibilidade da utilização de modelos mais complexos para análise na biblioteca SHAP, como os de aprendizado profundo, além da avaliação de outras variáveis sensíveis do conjunto de dados na biblioteca AIF360, como portabilidade de deficiência ou idade.

REFERÊNCIAS

[1] T. Mahoney, K. R. Varshney e M. Hind, "AI Fairness: How To Measure and Reduce Unwanted Bias in Machine Learning," O'Reilly, Sebastopol, CA, Estados Unidos, 2020. Acessado em: 08 Fev., 2022. [Online]. Disponível: https://community.ibm.com/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=09903149-3c91-6208-feaf-95004d77a8eeforceDialog=0

[2] L. Edwards e M. Veale, "Slave to the algorithm? why a 'right to an explanation' is probably not the remedy you are looking for," *Duke Law Technology Review*, vol. 16, no. 1, pp. 18-84, 2017. Acessado em: 08 Fev., 2022. [Online]. Disponível: https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=1315context=dltr

[3] J. Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, Out. 2018. Acessado em: 08 Fev., 2022. [Online]. Disponível: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[4] "Twitter finds racial bias in image-cropping AI," 20 Maio, 2021. Acessado em: 08 Fev., 2022. [Online]. Disponível: https://www.bbc.com/news/technology-57192898

[5] Ministério da Ciência, Tecnologia e Inovação, *Estratégia Brasileira de Inteligência Artificial*, 2021. Acessado em 08 Fev., 2022. [Online]. Disponível: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/arquivosinteligenciaartificial/ia_estrategia_documento_referencia_4-979_2021.pdf

[6] General Data Protection Regulation, *Principles relating to processing of personal data*. Acessado em 08 Fev., 2022. [Online]. Disponível: https://gdpr-info.eu/art-5-gdpr/

[7] G. Vilone e L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, 2021, pp. 89-106. Acessado em 08 Fev., 2022. [Online]. Disponível: https://www.sciencedirect.com/science/article/pii/S1566253521001093

[8] A. Das e P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): a survey," 2020. Acessado em 08 Fev., 2022. [Online]. Disponível: https://arxiv.org/pdf/2006.11371.pdf

[9] S. M. Lundberg e S. Lee, "A unified approach to interpreting model predictions," em *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777. Acessado em 08 Fev., 2022. [Online]. Disponível: https://dl.acm.org/doi/pdf/10.5555/3295222.3295230

[10] R. K. E. Bellamy et al., "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," em *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1-4:15, Jul-Set. 2019. Acessado em 08 Fev., 2022. [Online]. Disponível: https://ieeexplore.ieee.org/document/8843908

[11] M. S. A. Lee e J. Singh, "The landscape and gaps in open source fairness toolkits," em *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021,

pp. 1-13. Acessado em 08 Fev., 2022. [Online]. Disponível: https://dl.acm.org/doi/pdf/10.1145/3411764.3445261

[12] T. Moore, S. Raghavan e A. Mojsilovic, *IBM and LFAI move forward on trustworthy and responsible AI*, IBM Developer Blog, Jun. 2020. Acessado em 08 Fev., 2022. [Online]. Disponível: https://developer.ibm.com/blogs/ibm-and-lfai-move-forward-on-trustworthy-and-responsible-ai/

[13] AI Fairness 360, IBM Research Trusted AI. Acessado em 08 Fev., 2022. [Online]. Disponível: https://aif360.mybluemix.net/

[14] F. Kamiran e T. Calders, "Classifying without discriminating," *2009 2nd International Conference on Computer, Control and Communication*, 2009, pp. 1-6. Acessado em 08 Fev., 2022. [Online]. Disponível: https://ieeexplore.ieee.org/document/4909197

[15] AI Fairness 360 documentation. Acessado em 08 Fev., 2022. [Online]. Disponível: https://aif360.readthedocs.io/en/stable/

[16] S. Verma e J. Rubin, "Fairness Definitions Explained, *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 2018, pp. 1-7. Acessado em 08 Fev., 2022. [Online]. Disponível: https://dl.acm.org/doi/pdf/10.1145/3194770.3194776

[17] Ministério da Economia, *O que é RAIS?*. Acessado em 08 Fev., 2022. [Online]. Disponível: http://www.rais.gov.br/sitio/sobre.jsf

[18] Ministério do Trabalho, *Microdados RAIS e CAGED*. Acessado em 08 Fev., 2022. [Online]. Disponível: http://pdet.mte.gov.br/microdados-rais-e-caged

[19] F. Kamiran e T. Calders, "Data pre-processing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, 2012, pp. 1–33. Acessado em 08 Fev., 2022. [Online]. Disponível: https://link.springer.com/content/pdf/10.1007/s10115-011-0463-8.pdf

[20] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford e H. Wallach, "A reductions approach to fair classification," *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 2018, pp. 60-69. Acessado em 08 Fev., 2022. [Online]. Disponível: https://proceedings.mlr.press/v80/agarwal18a/agarwal18a.pdf

[21] F. Kamiran, A. Karim e X. Zhang, "Decision theory for discrimination-aware classification," *2012 IEEE 12th International Conference on Data Mining*, 2012, pp. 924-929. Acessado em 08 Fev., 2022. [Online]. Disponível: https://mine.kaust.edu.sa/Documents/papers/ICDM_2012.pdf