Atividade 4 MO810/MC959

[*] O link contendo o jupyter notebook e outros conjuntos de dados e também o datacar estão no drive:

https://drive.google.com/drive/folders/1XbJe1t7qfeV_G1yKNZamV8UuZNwLuPZP?usp=sharing

1st Betania E R da Silva
IC-Unicamp

2nd Decio Miranda Filho
IC-Unicamp

3rd Felipe Scalabrin Dosso
IMECC-Unicamp

I. Introdução

Este trabalho utiliza o *Adult Dataset*, amplamente conhecido por sua aplicação em estudos sobre justiça algorítmica, para avaliar técnicas de mitigação de vieses em aprendizado de máquina. A base de dados, rica em informações socioeconômicas, apresenta desafios relacionados à equidade, pois reflete desigualdades históricas. Para abordar essas questões, as métricas de justiça utilizadas incluem *Statistical Parity Difference*(SPD), *Disparate Impact* (DI) e *Equalized Odds Difference* (EOD), que permitem mensurar os impactos das técnicas de mitigação sobre os grupos sensíveis.

As técnicas analisadas são divididas em três níveis principais de intervenção. No pré-processamento, métodos como *Reweighting* ajustam pesos das instâncias para promover paridade entre grupos. No processamento, abordagens como *Fairness Penalty* adicionam termos de regularização na função de perda, equilibrando desempenho preditivo e justiça. No pós-processamento, utilizamos *Equalized Odds* para ajustar as predições finais do modelo para garantir igualdade de oportunidades.

II. Trabalhos Relacionados

A mitigação de vieses em aprendizado de máquina é um tema de grande importância, especialmente porque os dados frequentemente replicam padrões observados na sociedade, perpetuando desigualdades históricas. Diversos estudos têm utilizado o *dataset Adult*, amplamente adotado para avaliar técnicas de mitigação de vieses devido à sua relevância em contextos socioeconômicos sensíveis. Essas técnicas incluem abordagens de pré-processamento, in-processamento, intra-processamento, pós-processamento e até combinações dessas estratégias.

Maliheh Heidarpour Shahrezaei et al. [1] analisaram técnicas de pré-processamento para reduzir o viés em aprendizado de máquina usando o *dataset Adult*. As técnicas *Reweighting*, *Uniform Sampling*, *Preferential Sampling* e *Massaging* foram aplicadas a classificadores de regressão logística (LR) e árvores de decisão (DT). Os resultados mostraram que o *Massaging* foi mais eficaz com LR para as métricas *Disparate Impact* e *Statistical Parity*, enquanto *Reweighting* e *Uniform Sampling* tiveram melhor desempenho com DT. O estudo conclui que as abordagens de pré-processamento são flexíveis e úteis para melhorar a justiça dos modelos, com impactos variados dependendo da técnica e do classificador utilizado.

O artigo de Mingyang Wan et al. [2] é um survey que apresenta técnicas de mitigação de viés em aprendizado de máquina, com foco em métodos de *in-processing*. Usando o *Adult Dataset* como exemplo, ele detalha estratégias como regularização explícita, aprendizado adversarial, contrastivo e de representações, que reduzem a dependência de atributos sensíveis no modelo. Os autores concluem que essas técnicas são promissoras, mas enfrentam desafios relacionados à escolha de métricas apropriadas e à complexidade de cenários reais.

Yash Savani et al. [3], propõe o paradigma intra-processamento para mitigar vieses em redes neurais, posicionado entre os métodos *in-processing* e *post-processing*. Eles introduzem três algoritmos aplicados a *datasets* como COMPAS, *Adult*, Bank Marketing e CelebA. As técnicas incluem perturbação aleatória, otimização por camadas e ajuste fino adversarial. Os experimentos mostram que essas abordagens superam significativamente os métodos pós-processamento na redução de viés, mantendo bom desempenho, especialmente em contextos complexos.

Pranay K. Lohia et al. [4] propõem o *Individual+Group Debiasing* (IGD), um algoritmo de pós-processamento para mitigar viés individual e grupal em modelos de machine learning. Utilizando *datasets* como *Adult*, *German Credit* e COMPAS, eles demonstram que o IGD melhora métricas de justiça individual e grupal sem comprometer a precisão da classificação. Ao identificar amostras propensas ao viés individual e ajustar seus rótulos, o método se destaca por ser aplicável em configurações de caixa-preta e por não requerer rótulos verdadeiros na validação. Os resultados indicam que o IGD é uma solução promissora para melhorar a equidade em decisões algorítmicas.

Tal Feldman e Ashley Peake [5] propõem um framework de mitigação de viés ”end-to-end”que combina técnicas de pré-processamento, in-processamento e pós-processamento, focando na redução do viés de gênero em modelos de aprendizado profundo. Usando o *Adult Dataset*, eles empregam métodos como *Disparate Impact Remover*, *Adversarial Debiasing* e *Calibrated Equalized Odds*. Os resultados mostram que o *framework* melhora as métricas de justiça (como paridade estatística e igualdade de oportunidades) sem sacrificar significativamente a precisão do modelo, indicando que

a combinação dessas técnicas é promissora para aplicações futuras.

Faisal Kamiran et al. [6] propõem dois métodos, ROC e DAE, para mitigar discriminação em classificadores de aprendizado de máquina sem modificar dados ou algoritmos. Utilizando os *datasets Adult* e *Communities and Crimes*, as técnicas ajustam decisões em regiões de baixa confiança ou em discordâncias de ensembles para beneficiar grupos desfavorecidos. Os resultados demonstram que essas soluções reduzem discriminação com impacto mínimo na precisão, proporcionando controle direto sobre os níveis de justiça e flexibilidade em múltiplos atributos sensíveis.

III. Metodologia

A. Métricas de Fairness

Com a base de dados devidamente limpa e processada, iniciaremos a fase de modelagem.

As métricas de ***fairness*** utilizadas foram ***Statistical Parity Diference (SPD)***, ***Disparate Impact (DI)*** e ***Equal Opportunity Difference (EOD)*** para observar o impacto dos modelos considerando. Complementarmente, foram utilizadas ***AUC-score (Area Under the Curve-score)*** e acurácia como medidas de precisão do modelo.

O roc-AUC-score ou AUC é amplamente utilizado para conjuntos de dados desbalanceados por diversos trabalhos científicos [7] [8] [9] [10] [11] muito por conta de efetivamente capturar a área sobre a curva ROC. As métricas de *fairness* adotadas, SPD e *Disparate Impact* (DI), são recomendadas pelo próprio Varshney (2022) [12].

O SPD é definido como:

$$\text{SPD} = P(\hat{y}(X) = \text{fav} \mid Z = \text{desfav}) - P(\hat{y}(X) = \text{fav} \mid Z = \text{priv})$$

Por sua vez, o *Disparte Impact*, de forma análoga ao SPD quantifica a mede a diferença na probabilidade de resultados favoráveis entre grupos desfavorecidos e privilegiados por meio de um *ratio*. Neste, Grupos privilegiados e desprivilegiados são calculados, independentemente da intenção do tomador de decisão e do procedimento de tomada de decisão [12].

Por fim, o EOD é dado por:

$$\text{equal opportunity} = P(\hat{y}(X) = \text{fav} \mid y = \text{fav}, Z = \text{unpr}) - P(\hat{y}(X) = \text{fav} \mid y = \text{fav}, Z = \text{priv}) \tag{1}$$

O ***Equal Opportunity Difference*** (EOD) mede a diferença entre as taxas de verdadeiros positivos para dois grupos distintos: um grupo desfavorecido e um grupo privilegiado. O objetivo é garantir que ambos os grupos tenham igual oportunidade de receber um resultado favorável, considerando apenas as instâncias em que o rótulo verdadeiro é positivo (Y=1).

B. Mitigação de vieses

1) Pré-processamento: O método de ***Reweighting*** foi escolhido para a parte de pré-processamento dos dados e foi aplicado na *feature sex*. Seu objetivo é melhorar a equidade de grupos entre os rótulos e os atributos protegidos, como raça ou gênero, promovendo a paridade estatística. Nesse método, cada instância do conjunto de dados recebe um peso calculado como o produto das probabilidades marginais dos rótulos e atributos protegidos dividido pela probabilidade conjunta observada.

Os pesos são definidos da seguinte forma:

$$w_j = \frac{P(Y = y_j) \cdot P(Z = z_j)}{P(Y = y_j, Z = z_j)}$$

Os pesos ajustados são utilizados para modificar a influência de cada instância durante o treinamento do modelo. Isso assegura que as amostras sejam representadas de forma proporcional em relação às suas categorias sensíveis.

2) Processamento: Para a mitigação de vieses na fase de processamento dos modelos, foi utilizado a metodologia de ***Fairness Penalty*** no atributo sensível *native-country*. Essa técnica aplica penalizações adicionais na função de perda, forçando o modelo a equilibrar desempenho preditivo e equidade. Ao introduzir termos de regularização baseados em métricas de justiça, busca-se minimizar a dependência entre o atributo sensível Z e as predições $\hat{Y}$, promovendo igualdade de oportunidades.

Para um modelo preditivo $f_\theta(X)$, onde θ são os parâmetros do modelo, a função de otimização com penalidade de justiça é definida como:

$$\min_\theta \left[L(f_\theta(X), Y) + \lambda \cdot R(f_\theta, Z) \right]$$

em que $L(f_\theta(X), Y)$ representa a função perda tradicional, $R(f_\theta, Z)$ é o termo de regularização para equidade e λ é um hiperparâmetro que controla o *trade-off* entre a acurácia do modelo e a equidade.

Uma forma comum de definir $R(f_\theta, Z)$ é usando o índice de preconceito (*prejudice index*), que mede a divergência entre a distribuição conjunta das predições e do atributo sensível e suas distribuições marginais:

$$PI = \sum_{(y,z) \in \mathcal{D}} P(\hat{Y} = y, Z = z) \ln \left(\frac{P(\hat{Y} = y, Z = z)}{P(\hat{Y} = y) P(Z = z)} \right)$$

Ao minimizar PI, a penalidade de justiça reduz a dependência entre as predições $\hat{Y}$ e os atributos sensíveis Z, promovendo maior equidade nas decisões do modelo.

3) Pós-processamento: A técnica de ***Equalized Odds Fairness Post-Processing*** foi utilizada para mitigar vieses na fase após o treinamento do modelo para o atributo *race*. Essa metodologia foca na igualdade de odds, que exige que as taxas de verdadeiros positivos (TPR) e falsos positivos (FPR) sejam equivalentes para todos os grupos definidos pelos atributos sensíveis.

O processo consiste em aplicar ajustes às predições de saída $\hat{Y}$ do modelo já treinado. Esses ajustes são realizados alterando as probabilidades condicionais de decisão para cada grupo sensível Z, com base nas seguintes condições de igualdade:

$$P(\hat{Y} = 1 \mid Y = 1, Z = z) = P(\hat{Y} = 1 \mid Y = 1, Z = z')$$

$$P(\hat{Y} = 1 \mid Y = 0, Z = z) = P(\hat{Y} = 1 \mid Y = 0, Z = z')$$

onde Z é o atributo sensível, com z e z' representando diferentes valores (como grupos protegidos e privilegiados).

O método utiliza as probabilidades preditivas S produzidas pelo modelo e ajusta os limiares de decisão para cada grupo Z, de modo que as condições acima sejam atendidas. O ajuste minimiza a divergência entre as taxas TPR e FPR dos grupos, enquanto busca preservar o desempenho geral do modelo.

C. Modelagem

Após a fase de **Processamento**, o modelo de Regressão Logística foi treinado sobre os dados resultante da aplicação de técnicas de mitigação de vieses. As três metodologias foram empregadas de forma independente, e uma em cada um dos atributos sensíveis de Sexo, Raça e País Nativo.

IV. Resultados

A Tabela IV apresenta os valores de SPD (*Statistical Parity Difference*) e DI (v) para três atributos sensíveis: Sexo, Raça e País Nativo. Os resultados indicam que o atributo sexo apresenta a maior discrepância, com um SPD de 0.199504 e um DI de 0.364009. O atributo país nativo apresenta a menor discrepância (SPD = 0.058710, DI = 0.769073), sugerindo um impacto menos significativo no modelo sem mitigação.

Atributo sensível	SPD	DI
Sexo	0.199504	0.364009
Raça	0.104809	0.602482
País nativo	0.058710	0.769073

Tabela I
MÉTRICAS DE SPD E DE DI SEM MITIGAÇÃO DE VIESES NA BASE DE DADOS

A Tabela IV compara diferentes técnicas de mitigação no atributo sensível sexo. As métricas SPD, DI, EOD (*Equalized Odds Difference*), ROC-AUC e acurácia foram utilizadas para avaliar o impacto das metodologias ***Reweighting***, ***Fairness Penalty*** e ***Equalized Odds***.

A metodologia ***Reweighting*** sobre o atributo Sexo alcançou valores ideais de SPD (próximo de zero) e DI (1.0), indicando um balanceamento quase perfeito. No entanto, a ausência de valores de ROC-AUC e uma acurácia relativamente baixa (0.7028) sugerem uma possível limitação no desempenho preditivo do modelo.

Apesar de um SPD maior (0.111076), ***Fairness Penalty*** sobre a *feature* Raça apresenta um DI ligeiramente acima de 1 (1.175856), o que é razoável. Contudo, a acurácia (0.8097) e o ROC-AUC (0.8543) foram superiores aos do método Reweighting, demonstrando um melhor equilíbrio entre precisão e justiça.

Equalized Odds apresentou o melhor valor de EOD (0.047045) e valores muito bons de SPD (0.047045) e DI (1.055792), indicando uma maior igualdade entre grupos sensíveis. A acurácia (0.8309) foi a mais alta entre os métodos, reforçando sua eficácia.

Método	SPD	DI	EOD	ROC-AUC	Acurácia
Reweighting	2.045033e-10	1.0	0.088329	N/A	0.7028
Fairness Penalty	0.111076	1.175856	0.111076	0.8543	0.8097
Equalized Odds	0.047045	1.055792	0.047045	0.6899	0.8309

Tabela II
MÉTRICAS DE *fairness* E DE DESEMPENHO COM METODOLOGIAS DE MITIGAÇÃO DE VIESES

V. Discussão

Os modelos sem mitigação evidenciam disparidades significativas nas métricas de *fairness*, com sexo sendo o atributo mais impactado. O SPD e o DI indicam que, no estado original dos dados, há desigualdade no tratamento de grupos sensíveis. Essas disparidades reforçam como os dados podem replicar padrões sociais injustos, levando a decisões algorítmicas enviesadas. Nesse contexto, a ausência de intervenções pode perpetuar ou até agravar desigualdades históricas.

Reweighting conseguiu praticamente eliminar os vieses em termos de SPD e DI, mas comprometeu a acurácia do modelo, mostrando-se mais adequado para cenários onde a justiça é prioridade absoluta.

Fairness Penalty equilibrou bem a redução de vieses com o desempenho preditivo, sendo uma solução interessante para aplicações que exigem algum nível de compromisso entre fairness e precisão.

Equalized Odds apresentou o melhor equilíbrio geral, reduzindo os vieses de forma eficiente (com SPD, DI e EOD próximos de valores ideais) e mantendo a maior acurácia entre as técnicas avaliadas.

Os resultados sugerem que ignorar a mitigação de vieses pode levar a decisões enviesadas e injustas, enquanto a aplicação dessas técnicas pode corrigir desigualdades, ainda que isso possa vir a custo de desempenho em alguns casos. A escolha da abordagem de mitigação deve ser informada pelo tipo de viés a ser tratado, pela sensibilidade dos atributos envolvidos e pelo *trade-off* aceitável entre justiça e precisão, garantindo que os modelos sejam não apenas eficazes, mas também éticos e responsáveis.

Referências

[1] M. H. Shahrezaei, R. Loughran, and K. M. Daid, "Pre-processing techniques to mitigate against algorithmic bias," in *2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, 2023, pp. 1–4.

[2] M. Wan, D. Zha, N. Liu, and N. Zou, "In-processing modeling techniques for machine learning fairness: A survey," *ACM Transactions on Knowledge Discovery from Data*, vol. 17, no. 3, pp. 35:1–35:27, March 2023. [Online]. Available: https://doi.org/10.1145/3551390

[3] Y. Savani, C. White, and N. S. Govindarajulu, "Intra-processing methods for debiasing neural networks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 2798–2810. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1d8d70dddf147d2d92a634817f01b239-Paper.pdf

[4] P. K. Lohia, K. Natesan Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri, "Bias mitigation post-processing for individual and group fairness," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2847–2851.

[5] T. Feldman and A. Peake, "End-to-end bias mitigation: Removing gender bias in deep learning," *arXiv preprint arXiv:2104.02532v3*, 2021, submitted on 6 Apr 2021 (v1), last revised 21 Jun 2021 (this version, v3).

[6] K. T. Hufthammer, T. H. Aasheim, S. Ånneland, H. Brynjulfsen, and M. Slavkovik, "Bias mitigation with aif360: A comparative study," in *NIK Norsk informatikkonferanse*. NIK, 2020. [Online]. Available: https://hdl.handle.net/11250/2764230

[7] J. seok Lee, "Auc4.5: Auc-based c4.5 decision tree algorithm for imbalanced data classification," *IEEE Access*, vol. 7, pp. 106 034–106 042, 2019.

[8] Asniar, N. Maulidevi, and K. Surendro, "Smote-lof for noise identification in imbalanced data classification," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, pp. 3413–3423, 2021.

[9] B. Zhou, Y. Ying, and S. Skiena, "Online auc optimization for sparse high-dimensional datasets," *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 881–890, 2020.

[10] Y. Liu, Y. Li, and D. Xie, "Implications of imbalanced datasets for empirical roc-auc estimation in binary classification tasks," *Journal of Statistical Computation and Simulation*, vol. 94, pp. 183 – 203, 2023.

[11] Y.-C. Wang and C.-H. Cheng, "A multiple combined method for rebalancing medical data with class imbalances," *Computers in biology and medicine*, vol. 134, p. 104527, 2021.

[12] K. R. Varshney, *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published, 2022.