VERIFICAÇÃO E MITIGAÇÃO DE VIESES NO APRENDIZADO DE MÁQUINA NO DIAGNÓSTICO DE DOENÇAS

Palavras-chave: aprendizado de máquina, vieses, inteligência artificial, saúde.

Autoras:
Luisa Martins Felix de Oliveira, FT – UNICAMP
Profa. Dra. Lívia Ruback (orientadora), FT - UNICAMP
Dr. Ariel Soares Teles, UFMA

1 INTRODUÇÃO

Com o avanço de métodos e técnicas baseadas em Inteligência Artificial (IA) e Machine Learning (ML), tem-se observado uma crescente incorporação dessas abordagens em contextos científicos e médicos, sobretudo no apoio à predição de diagnósticos clínicos. No entanto, esses sistemas podem herdar e até amplificar desigualdades sociais preexistentes caso seus algoritmos apresentem vieses não identificados. Este projeto tem como objetivo diagnosticar e mitigar vieses algorítmicos em modelos de ML aplicados ao diagnóstico de diabetes, analisando disparidades de tratamento entre diferentes grupos sociais. A pesquisa utiliza métricas de fairness como *Statistical Parity Difference*, *Equal Opportunity Difference* e *Average Odds Difference* para avaliar a presença de viés. São aplicadas técnicas de mitigação ,como Reweighing e Prejudice Remover, nos vieses encontrados, com posterior análise do impacto dessas abordagens na redução das disparidades. Os resultados esperados podem orientar futuras práticas de desenvolvimento e avaliação de modelos de IA sensíveis a questões éticas e sociais.

2 METODOLOGIA

2.1 Dataset

Este estudo utiliza dados da edição 2015 do Behavioral Risk Factor Surveillance System (BRFSS), conduzido pelo

CDC [Xie et al., 2019], que reúne informações de pacientes nos EUA sobre comportamentos de saúde e condições médicas crônicas. A base contempla 441.456 instâncias, descritas por 330 atributos, coletadas nos 50 estados norte-americanos, Distrito de Columbia, Guam e Porto Rico.

Para o experimento realizado, foram selecionados 24 atributos, incluindo fatores clínicos associados ao diabetes (como pressão alta e IMC [Pias et al., 2023]) e determinantes sociais da saúde, como gênero, etnia, renda e acesso a plano de saúde. As variáveis gênero e etnia foram tratadas como protegidas, sendo usadas para avaliação de equidade e mitigação de viés no diagnóstico de diabetes tipo 2, que é a variável-alvo do estudo.

2.2 Pré-Processamento

O atributo alvo era inicialmente dividido em quatro categorias: pessoas com diabetes(1), com diabetes apenas durante a gravidez(2), sem diabetes(3) e com pré-diabetes(4). Posteriomentr, por conta da avaliação de vieses esse atributo foi transformado em um atributo binário, em que o valor 1 indica a presença de diabetes ou pré-diabetes, e o valor 0 representa a ausência de diabetes ou ocorrência apenas durante a gestação. A pré-diabetes foi considerada como caso positivo devido à sua relevância clínica, conforme orientação da American Diabetes Association (ADA), que destaca esse estado como de alto risco para o desenvolvimento de diabetes e complicações associadas, sendo fundamental sua detecção precoce [Association, 2023].

A variável de etnia foi simplificada para dois grupos — brancos (1) e não brancos (0) — visando à análise de justiça algorítmica. O gênero, por já ser binário, não exigiu ajustes.

2.3 Classificação

Para a tarefa de classificação com ML no diagnóstico de diabetes tipo 2, treinamos um algoritmo de Regressão Logística (LR) [Wright, 1995]. Embora a regressão logística nem sempre apresente o melhor desempenho em conjuntos de dados desbalanceados, ela continua sendo um modelo amplamente utilizado e interpretável em pesquisas sobre justiça algorítmica, sendo útil para identificar e mitigar viés [Mehrabi et al., 2021, H.P.Jain et al., 2022]. A LR também é particularmente compatível com diversas técnicas de mitigação de viés, já que alguns métodos de pós-processamento — como o Prejudice Remover — dependem internamente da dinâmica de aprendizagem da regressão logística.

Utilizamos a biblioteca AI Fairness 360 (AIF360) [Kush Varshney, 2018] para treinar modelos de Regressão Logística (LR), adotando o otimizador SAGA em todos eles, inclusive nos treinados com reponderação, para assegurar comparabilidade consistente entre as abordagens de mitigação — dado que a implementação do Prejudice Remover na AIF360 também utiliza LR com SAGA. A avaliação foi realizada por meio de métricas de desempenho em clas-

sificação, como acurácia, revocação (recall), precisão, F1-score e AUC-ROC. Para robustez, aplicamos tanto a divisão hold-out (70-30) quanto validação cruzada com 5 folds.

2.4 Avaliação de Fairness

Para a avaliação de vieses de raça e de gênero, foram utilizadas três métricas de fairness com o objetivo de avaliar possíveis disparidades nos resultados de predição dos modelos desenvolvidos: SPD, EOD e AOD [Hardt et al., 2016]. Todas as métricas são baseadas em diferenças, ou seja, calculam a diferença absoluta entre dois grupos em relação a determinados desfechos preditivos. Um valor ideal de 0 para cada métrica indica equidade perfeita, representando ausência de disparidade entre os grupos.

2.5 Estratégias de Mitigação

Neste estudo, foram adotadas estratégias de mitigação de viés nos estágios de pre-processing (ajuste dos dados antes do treinamento) e in-processing (intervenções durante o processo de aprendizado), com o objetivo de promover maior equidade nas predições dos modelos. As técnicas de pre-processing foram utilizadas para ajustar os dados antes do treinamento, reduzindo desigualdades na representação de grupos demográficos - como o método de Reweighting, que atribui pesos às amostras com base na distribuição dos grupos sensíveis [Mehrabi et al., 2021, Caton and Haas, 2024]. Já no in-processing, o viés foi melhor mtigado durante o próprio processo de aprendizado, por meio de modificações na função de perda, como no método Prejudice Remover, o qual adiciona uma penalização à dependência entre a variável sensível e o resultado previsto, forçando o modelo (ex. regressão logística) a reduzir associações discriminatórias [Kamishima et al., 2012, Caton and Haas, 2024].

3 RESULTADOS E DISCUSSÃO

Todo o código utilizado neste estudo está disponível publicamente no GitHub [1], incluindo o pré-processamento dos dados, a visualização gráfica e a mitigação de vieses. Gráficos adicionais que não puderam ser incluídos neste artigo também estão disponíveis lá, oferecendo uma compreensão mais aprofundada dos resultados e das distribuições dos dados.

3.1 Avaliação de Desempenho

O modelo de diagnóstico de diabetes utilizando LR obteve alta acurácia (0.84728) e ROC-AUC (0.8181), porém apresentou valores relativamente baixos de precisão (0.54950), revocação (0.18622) e F1-score (0.2781). Esses resultados de desempenho refletem, em grande parte, a distribuição desbalanceada dos casos positivos de diabetes, o qual representam apenas 15% do conjunto de dados.

[1] `https://github.com/luisafelixx/Fairness_Diabetes`

3.2 Avaliação de Fairness

A Tabela 1 apresenta os resultados das métricas de fairness antes da mitigação de viés, com foco em etnia e gênero. Os grupos privilegiados são definidos como brancos (para etnia) e homens (para gênero). Em contraste, indivíduos não brancos e do sexo feminino são considerados não privilegiados.

Métricas Fairness	**Etnia**	**Gênero**
SPD	0.075	-0.005
EOD	0.162	0.017
AOD	0.101	0.006

Tabela 1: Métricas de Fairness sem mitigação de viés para etnia e gênero.

As métricas de justiça indicam que o modelo apresenta viés mais evidente em relação à etnia, favorecendo indivíduos brancos nas previsões de diabetes. Os valores de SPD (0.075), EOD (0.162) e AOD (0.101) para etnia mostram que pessoas não brancas têm menos chances de serem corretamente diagnosticadas.

Em contrapartida, os resultados para gênero são bons: os valores de SPD (-0.005), EOD (0.017) e AOD (0.006) indicam desigualdades quase nulas entre homens e mulheres. Isso sugere que o modelo é praticamente neutro quanto ao gênero, distribuindo previsões e acertos de forma equilibrada entre os dois grupos.

3.3 Resultados Pós Mitigação de Viés

Foram aplicadas duas estratégias de mitigação de viés com foco na variável protegida etnia: Reweighting (pre-processing) e Prejudice Remover(in-processing). Como as métricas de justiça para gênero não indicaram viés significativo, nenhuma técnica foi aplicada para essa variável.

Ambas as abordagens apresentaram resultados positivos, melhorando consistentemente as três métricas de justiça (SPD, EOD e AOD) e contribuindo para a redução das disparidades observadas anteriormente, conforme mostra a Tabela 2, a que comparapara as métricas de fairness entre as estratégias aplicadas e o cenário sem mitigação para etnia.

	Sem Mit	RW	PR
SPD	0.075048	0.014945	**0.014301**
EOD	0.162955	**0.002936**	0.006501
AOD	0.101394	**0.003856**	0.004988

Tabela 2: Resultados de Mitigação de Viés para Etinia.

4 CONCLUSÕES

Este estudo avaliou a justiça em modelos de aprendizado de máquina para diagnóstico de diabetes, com base nos dados do BRFSS. Embora a regressão logística tenha apresentado boa acurácia e AUC-ROC, métricas como precisão e F1-score foram impactadas pelo desbalanceamento dos dados. As análises de justiça indicaram desigualdades, especialmente relacionadas à etnia,

desfavorecendo indivíduos não brancos.

Para mitigar esses vieses, foram aplicadas as estratégias Reweighting e Prejudice Remover, com destaque para a primeira, que apresentou melhores resultados em justiça para etnia, sem comprometer a acurácia.

Os resultados reforçam a importância de avaliar justiça e aplicar estratégias de mitigação em modelos de saúde, com o apoio de especialistas da área. Estudos futuros explorarão novas técnicas e a ampliação dos atributos protegidos analisados. Parte desses resultados está sendo publicada no Simpósio Brasileiro de Banco de Dados (SBBD) 2025, contribuindo para o debate acadêmico sobre justiça algorítmica na saúde.

Referências

[Association, 2023] Association, A. D. (2023). Standards of medical care in diabetes. *Diabetes Care*, 46(Supplement 1):S1–S291.

[Caton and Haas, 2024] Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7).

[Hardt et al., 2016] Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, volume 29, pages 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

[H.P.Jain et al., 2022] H.P.Jain, S.R.Patel, and M.V.Joshi (2022). Addressing bias in diabetes prediction models through fairness-aware learning. *Journal of Ambient Intelligence and Humanized Computing*, 13(2):1085–1098.

[Kamishima et al., 2012] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2012)*, pages 35–50. Springer.

[Kush Varshney, 2018] Kush Varshney (2018). Introducing ai fairness 360. `https://research.ibm.com/blog/ai-fairness-360`. Accessed: 2025-04-27.

[Mehrabi et al., 2021] 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).

[Pias et al., 2023] Pias, T. S., Su, Y., Tang, X., Wang, H., Faghani, S., and Yao, D. D. (2023). Enhancing fairness and accuracy in type 2 diabetes prediction through data resampling. *medRxiv*.

[Wright, 1995] Wright, R. E. (1995). Logistic regression.

[Xie et al., 2019] Xie, Z., Nikolayeva, O., Luo, J., and Li, D. (2019). Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing Chronic Disease*, 16:E130.