XXXIII Congresso de Iniciação Científica da Unicamp

FAIRNESS EM APRENDIZADO DE MÁQUINA: FUNDAMENTOS E IMPLEMENTAÇÃO DE UM ALGORITMO

Palavras-Chave: APRENDIZADO DE MÁQUINA, FAIRNESS, VIÉS RACIAL

Autores(as):
NICOLE PETRICA ARAUJO, IC/FEEC – UNICAMP
Prof. Dr. ROMIS RIBEIRO DE FAISSOL ATTUX (orientador), FEEC – UNICAMP

INTRODUÇÃO:

Nas últimas décadas, os avanços tecnológicos impulsionaram o Aprendizado de Máquina (do inglês *machine learning* – ML), transformando diversas áreas, como reconhecimento de padrões, diagnósticos médicos, seleção de candidatos em processos seletivos, políticas públicas, entre outras. No entanto, o uso crescente dessas tecnologias trouxe preocupações com *fairness* (equidade), já que algoritmos treinados com dados históricos podem reproduzir ou até amplificar vieses. Um exemplo disso é o racismo algorítmico, abordado no documentário *Coded Bias* (KANTAYYA, 2020), que revela falhas no reconhecimento facial de grupos minoritários e decisões injustas baseadas em algoritmos de ML.

Diante desse cenário, este projeto de pesquisa busca compreender questões relacionadas à *fairness* em ML, com base em um estudo de caso abordando dados públicos de saúde. No primeiro semestre da pesquisa foram realizados estudos sobre os fundamentos de ML e *fairness*, incluindo uma revisão sistemática da literatura com base no protocolo *Preferred Reporting Items for Systematic reviews and Meta-Analyses* – PRISMA (MOHER et al., 2010). Com base na fundamentação teórica e na análise crítica da literatura, definiu-se um caso aplicado: investigar a presença de viés racial em modelos preditivos para auxílio ao diagnóstico de hipertensão.

METODOLOGIA:

O estudo foi desenvolvido com base nos microdados da Pesquisa Nacional de Saúde (PNS) 2019, realizada pelo IBGE, em parceria com o Ministério da Saúde (IGBE, 2019). Trata-se de uma pesquisa amostral importante no cenário nacional, e que fornece informações detalhadas sobre dados demográficos, socioeconômicos, de saúde, de estilo de vida e de acesso a serviços. A variável de interesse escolhida foi: “Algum médico já lhe deu o diagnóstico de hipertensão arterial (pressão alta)?”.

A base inicial continha 279.382 amostras com 816 variáveis. Foram selecionadas dez variáveis preditoras em quatro dimensões: demográfica (sexo, idade, cor ou raça), socioeconômica (faixa de rendimento domiciliar per capita), comportamental (prática de atividade física, consumo de bebidas alcoólicas, tabagismo e percepção sobre o consumo de sal) e de saúde (estado de saúde percebido, e tempo desde a última consulta médica).

O pré-processamento dos dados incluiu a padronização dos nomes das variáveis, filtragem de dados para considerar apenas indivíduos adultos, limpeza de dados, tratamento de valores ausentes ou inconsistentes, e transformação de variáveis categóricas para formato numérico.

Em seguida, foi conduzida uma análise exploratória a fim de verificar possíveis disparidades raciais no diagnóstico de hipertensão. Essa etapa permitiu observar a distribuição da variável de interesse entre os diferentes grupos étnicos, bem como verificar indícios preliminares de desbalanceamento de dados, que podem refletir desigualdades estruturais, as quais tendem a ser reproduzidas pelos algoritmos de aprendizado de máquina se não forem consideradas.

Para avaliar a presença de viés racial nos dados e nos modelos, foram adotadas métricas de *fairness* amplamente conhecidas na literatura, como *Demographic Parity Difference* (DPD), *Equalized Odds Difference* (EOD) e *True Positive Rate Difference* (TPR Diff) por grupo racial. A DPD mede a diferença na taxa de diagnósticos positivos previstos entre os grupos raciais, independentemente da condição real de saúde, sendo útil para indicar desigualdades no acesso ao diagnóstico. Já a EOD compara as taxas de verdadeiros positivos e falsos positivos entre os grupos, capturando se o modelo erra mais em determinados grupos. Por fim, a TPR Diff avalia a diferença na proporção de casos corretamente identificados entre os que realmente têm hipertensão, revelando possíveis padrões de sub-reconhecimento em grupos específicos (MEHRABI *et al.*, 2019).

A modelagem foi realizada com dois algoritmos de classificação supervisionada: Regressão Logística (RL) e Árvore de Decisão (AD). Os dados foram divididos em conjuntos de treino (80% dos dados) e de teste (20% dos dados), e o desempenho dos modelos foi avaliado por meio de métricas como acurácia, precisão, *recall* e *F1-score*. Os hiperparâmetros de cada modelo foram selecionados com a técnica de *Grid Search* usando validação cruzada (*GridSearchCV*) (BELETE; HUCHAIAH, 2021). Esses modelos iniciais, sem aplicar nenhuma metodologia de mitigação de viés, foram chamados de "modelos base".

Como parte da investigação sobre *fairness*, foram aplicadas técnicas de mitigação de viés em três abordagens: pré-processamento, durante o processamento (*in-processing*) e pós-processamento, a fim de avaliar seu impacto na redução das disparidades raciais. No pré-processamento, foram aplicadas técnicas de reamostragem (como *SMOTE*) para equilibrar a representatividade dos grupos sensíveis no conjunto de treino. No *in-processing*, utilizou-se o algoritmo *Exponentiated Gradient*, que incorpora restrições de equidade diretamente no treinamento, buscando conciliar desempenho preditivo e *fairness*. Já no pós-processamento, foram ajustados os limiares de decisão, de modo a redistribuir as predições positivas de forma mais equitativa entre os grupos.

Por fim, os resultados dos modelos mitigados foram comparados aos modelos base, tanto em termos de desempenho preditivo quanto de redução das disparidades raciais, destacando os impactos das diferentes abordagens de *fairness* e os *trade-offs* entre acurácia e justiça algorítmica em aplicações sensíveis, como o diagnóstico de hipertensão no sistema público de saúde.

RESULTADOS E DISCUSSÃO:

Após o pré-processamento dos dados, foram selecionadas 86.831 observações de indivíduos adultos, com diagnóstico autorreferido de hipertensão arterial como variável-alvo. A proporção de casos positivos (que relataram ter hipertensão) foi de 27,4%. Dentre as variáveis explicativas, observou-se predominância de indivíduos pardos

(50,3%), seguidos por brancos (36,8%), pretos (11,4%), indígenas (0,7%) e amarelos (0,6%), com maior concentração nas faixas mais baixas de renda.

A análise exploratória revelou diferenças significativas na taxa de hipertensão entre os grupos raciais: indivíduos autodeclarados pretos apresentaram a maior taxa de hipertensão (30,3%), enquanto pardos apresentaram a menor (26,1%). O teste qui-quadrado indicou dependência estatisticamente significativa entre raça/cor e diagnóstico de hipertensão ($\chi^2 = 104{,}08;\ p < 0{,}001$), evidenciando a necessidade de investigar possíveis disparidades no comportamento dos modelos com base neste atributo sensível (SEPEHRI; DICICCIO, 2020).

Foram treinados dois modelos de classificação: RL e AD. A Tabela 1 apresenta os resultados dos modelos base. A RL obteve melhor desempenho em acurácia (0,738) e F1-score (0,610), enquanto a AD teve maior sensibilidade (recall = 0,783), sendo mais eficaz na identificação de casos positivos de hipertensão.

Modelo	**Métricas**			
	Acurácia	**Precisão**	**Recall**	**F1-score**
RL	0,738	0,516	0,747	0,610
AD	0,717	0,490	0,783	0,603

Tabela 1 – Desempenho dos modelos base com métricas tradicionais.

Apesar do desempenho satisfatório quando analisado com as métricas tradicionais, ambos os modelos apresentaram disparidades raciais, como pode ser visto na avaliação dos modelos com métricas de *fairness*, na Tabela 2. Esses valores indicam que as probabilidades de receber um diagnóstico positivo (seja correto ou não), podem variar entre os grupos raciais.

Modelo	**Métricas**	
	DPD	**EOD**
RL	0,085	0,158
AD	0,117	0,168

Tabela 2 – Desempenho dos modelos base com métricas de fairness.

Ao observar as métricas por grupo, é possível identificar um viés algorítmico que penaliza especialmente os grupos indígena e pardo. No modelo de RL, esses grupos apresentaram as menores taxas de verdadeiros positivos (TPR) e de seleção, evidenciando menor acesso ao diagnóstico correto. No modelo de AD, essas diferenças foram ainda mais acentuadas, com o grupo branco sendo favorecido com maiores taxas de verdadeiros positivos e falsos positivos, o que sugere um viés que amplia desigualdades raciais ao superestimar positivamente esse grupo e subestimar outros historicamente vulneráveis.

Diante das disparidades identificadas, foram aplicadas as três abordagens de mitigação de viés. No pré-processamento, foram utilizadas as técnicas de reamostragem *SMOTE*, *ADASYN* e *BorderlineSMOTE*, sendo o SMOTE o mais eficaz com a AD. No *in-processing*, foi aplicado o algoritmo *Exponentiated Gradient*, com restrições de DPD e EOD. A AD com EOD apresentou o melhor equilíbrio entre desempenho e equidade. Já no pós-processamento, foi aplicada a técnica *Threshold Optimizer*, que ajusta os limiares de decisão do modelo com base no grupo sensível. Essa técnica não altera o modelo treinado, mas modifica suas predições com o objetivo de alcançar critérios de equidade. Os resultados indicaram perda expressiva de F1-score na AD, limitando sua aplicabilidade. O melhor desempenho nesse caso foi da RL com restrição de DPD.

Para facilitar a comparação entre todas as estratégias, foi proposto um *framework* de análise de *trade-offs*, em que os modelos foram comparados com base em duas medidas: desempenho preditivo (F1-score) e *fairness* (EOD e TPR), seguindo a fórmula:

$$Score\_Final = 0.25 * EOD\_Norm + 0.25 * TPR\_Norm + 0.5 * F1\_Norm$$

A Tabela 3 sintetiza os resultados obtidos pelos principais modelos testados, enquanto a Tabela 4 apresenta um *ranking* final das abordagens com base no equilíbrio entre essas duas medidas.

Estratégia	**Modelo**	**Métricas**				
		Acurácia	**F1-score**	**DPD**	**EOD**	**TPR**
Base	RL	0,738	0,610	0,085	0,152	0,152
Base	AD	0,717	0,603	0,117	0,168	0,168
Pré-processamento (*SMOTE*)	AD	0,725	0,595	0,058	0,120	0,120
In-processing (EOD)	AD	0,767	0,524	0,044	0,121	0,105
Pós-processamento (DPD)	AD	0,768	0,515	0,017	0,053	0,053

Tabela 3 – Comparação entre modelos e estratégias de mitigação.

Rank	**Modelo**	***Score* Final**
1	Pré-processamento (*SMOTE*) - AD	0.797
2	Pré-processamento (*ADASYN*) - AD	0.770
3	Base - RL	0.633
4	Pós-processamento (DPD) - AD	0.614
5	Pós-processamento (EOD) - RL	0.493
6	*In-processing* (EOD) - AD	0.492
7	Base - AD	0.449

Tabela 4 – Ranking baseado em trade-offs.

Dentre todos os modelos, a AD com mitigação via *SMOTE* apresentou o melhor compromisso entre justiça e desempenho, alcançando a melhor posição no *ranking* final. Por outro lado, as abordagens pós-processamento obtiveram ótimos resultados em termos de *fairness*, mas com queda notável de desempenho preditivo, especialmente nos grupos com menor representatividade. Já as técnicas de *in-processing* não apresentaram desempenho satisfatório neste estudo.

Os modelos de AD mostraram maior responsividade às estratégias de mitigação, superando o modelo base da RL com as técnicas de pré-processamento *SMOTE* e *ADASYN* de pré-processamento. Isso pode ser explicado pela sensibilidade da AD à distribuição das amostras no conjunto de treino. Assim, ao incluir amostras sintéticas, há uma correção nas fronteiras de decisão, o que favorece desempenho e equidade, especialmente em dados com grande desbalanceamento entre os grupos.

De forma geral, os resultados obtidos demonstram que estratégias de mitigação de viés são viáveis e eficazes para reduzir desigualdades raciais em algoritmos de classificação, mas devem ser escolhidas considerando o contexto de uso. Em aplicações de saúde pública, nas quais tanto a precisão quanto a equidade são essenciais, soluções baseadas em pré-processamento mostraram-se mais adequadas, ao promoverem justiça sem comprometer a capacidade de generalização dos modelos.

CONCLUSÕES:

Este estudo demonstrou que modelos de aprendizado de máquina aplicados a dados de saúde pública podem reproduzir desigualdades raciais presentes nos dados. A análise com base em dados nacionais de saúde pública revelou disparidades significativas nos diagnósticos preditivos de hipertensão entre diferentes grupos raciais, reforçando a importância de avaliar e mitigar vieses algorítmicos em contextos sensíveis como o da saúde. Dentre as estratégias avaliadas, as técnicas de pré-processamento com reamostragem, especialmente a combinação de SMOTE com AD, apresentaram o melhor equilíbrio entre desempenho preditivo e *fairness* algorítmica. Contudo, a escolha da estratégia de mitigação deve levar em conta não apenas as métricas de *fairness*, mas também a sensibilidade do modelo às características dos dados. Nesse sentido, a pesquisa contribui para o fortalecimento de práticas responsáveis no uso de aprendizado de máquina em contextos sociais, reforçando a importância de incorporar critérios de equidade na formulação e avaliação de modelos de ML aplicados a políticas públicas.

REFERÊNCIAS BIBLIOGRÁFICAS

BELETE, D. M.; HUCHAIAH, M. D. **Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results.** International journal of computers & applications, v. 44, n. 9, p. 875–886, 2022.

IBGE. **PNS - Pesquisa Nacional de Saúde 2019.** Disponível em: https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html.

KANTAYYA, S. **Coded Bias.** Estados Unidos. Netflix, 26 de janeiro de 2020. Disponível em: https://www.netflix.com/br/title/81328723.

MEHRABI, N. et al. **A Survey on Bias and Fairness in Machine Learning.** arXiv:1908.09635 [cs], 17 set. 2019.

MOHER, D.; LIBERATI, A.; TETZLAFF, J.; ALTMAN, D. G.; PRISMA GROUP*, **T. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement.** International Journal of Surgery, v. 8, n. 5, p. 336–341, 2010.

SEPEHRI, A.; DICICCIO, C. **Interpretable assessment of fairness during model evaluation.** 2020. Disponível em: http://arxiv.org/abs/2010.13782.