ALGORITMOS ENVIESADOS E INTELIGÊNCIA ARTIFICIAL: RISCOS, DESAFIOS E RESPONSABILIDADES SOCIAIS

Nélio Do Sacramento Santiago[1]
Sabi Yari Moïse Bandiri[2]

RESUMO

A Inteligência Artificial (IA) tem avançado de forma acelerada, transformando práticas sociais, produtivas e tecnológicas em escala global. Contudo, junto aos benefícios da automação e da análise de grandes volumes de dados, emergem preocupações éticas e sociais relacionadas ao fenômeno do viés algorítmico. Os algoritmos podem se tornar verdadeiros instrumentos de exclusão social quando não são monitorados. Este trabalho tem como objetivo investigar os efeitos dos algoritmos enviesados, discutir casos reais, analisar técnicas de mitigação e apresentar um experimento computacional baseado em regressão logística. Os resultados iniciais indicaram elevada acurácia no modelo, porém com forte injustiça de gênero. Ao aplicar a técnica *Exponentiated Gradient Reduction*, implementada pela biblioteca Fairlearn, observou-se redução significativa do viés, embora com perda parcial de desempenho. Essa análise evidencia o trade-off entre precisão e justiça em modelos de decisão automatizados e destaca a urgência de adotar métricas de equidade, auditorias periódicas e regulamentações éticas. Conclui-se que a mitigação de vieses é uma responsabilidade social que ultrapassa os limites técnicos, sendo condição essencial para que a IA se consolide como instrumento de inclusão e transformação positiva da sociedade.

Palavras-chave: inteligência artificial; viés algorítmico; aprendizado de máquina; justiça algorítmica.

Universidade da Integração Internacional da Lusofonia Afro-Brasileira,, Instituto de Engenharia e Desenvolvimento Sustentável, Discente, neliosantiago.ns@aluno.unilab.edu.br[1]
Universidade da Integração Internacional da Lusofonia Afro-Brasileira, Instituto de Engenharia e Desenvolvimento Sustentável, Docente, bandiri@unilab.edu.br[2]

ISSN: 2447-6161

INTRODUÇÃO

A Inteligência Artificial (IA) tem revolucionado setores estratégicos da sociedade ao introduzir soluções baseadas em aprendizado de máquina, automação e análise de dados em larga escala. Segundo Mehrabi et al. (2021), a IA já é aplicada em saúde, finanças, transporte e segurança pública, mas ao mesmo tempo enfrenta questionamentos éticos, principalmente pela possibilidade de reproduzir preconceitos e discriminações.

O viés algorítmico ocorre quando os sistemas de IA refletem desigualdades históricas presentes nos dados de treinamento. Segundo Barocas e Selbst (2016), esse fenômeno pode provocar o que os autores chamam de "impacto discriminatório do big data", gerando exclusões sistemáticas em serviços essenciais. Esse problema tem sido comprovado por casos reais, como o da Amazon em 2018, quando a empresa descontinuou um sistema de recrutamento automatizado porque o algoritmo favorecia candidatos homens em detrimento de mulheres, devido à predominância histórica masculina no setor de tecnologia (Duarte, 2020). Da mesma forma, Buolamwini e Gebru (2018) revelaram que softwares de reconhecimento facial apresentavam taxa de erro acima de 30% em mulheres negras, enquanto para homens brancos o índice era inferior a 1%. Esses episódios demonstram que sistemas supostamente neutros podem reforçar desigualdades estruturais.

Portanto, compreender e mitigar vieses em IA é essencial não apenas como desafio técnico, mas também como compromisso ético e social. O'Neil (2019) afirma que algoritmos sem supervisão crítica tornam-se verdadeiros "algoritmos de destruição em massa", pois automatizam injustiças já existentes. Neste trabalho, investigamos teoricamente e experimentalmente o fenômeno do viés em IA, destacando alternativas para torná-la mais justa e inclusiva.

METODOLOGIA

A metodologia combinou revisão bibliográfica e experimento computacional. Na revisão de literatura, buscou-se mapear diferentes definições, impactos e soluções para o viés algorítmico. Barocas e Selbst (2016) descrevem que algoritmos enviesados podem reforçar desigualdades de maneira invisível. O'Neil (2019) ressalta que dados históricos carregam vieses e, se utilizados sem crítica, podem naturalizar preconceitos sociais. Mehrabi et al. (2021) realizaram levantamento abrangente mostrando que existem múltiplas métricas de justiça, mas nenhuma delas elimina totalmente as tensões entre precisão e equidade.

No experimento prático, implementou-se em Python um modelo de regressão logística aplicado a um conjunto sintético de 10.000 currículos. O *dataset* incluiu variáveis como nota técnica, experiência, escolaridade, gênero e resultado de contratação. Um viés proposital foi embutido: homens eram contratados com nota acima de 6,5 enquanto mulheres só com nota superior a 8.

Para avaliar o modelo, foram utilizadas métricas tradicionais de classificação, como acurácia e F1-score, e métricas de justiça, como *Disparate Impact* e *Equal Opportunity*. Segundo Barocas e Selbst (2016), o Disparate Impact é amplamente usado em estudos jurídicos por avaliar discrepâncias entre grupos protegidos.

Para mitigação, foi utilizada a técnica *Exponentiated Gradient Reduction*, disponível na biblioteca Fairlearn. Essa abordagem impõe restrições de paridade demográfica ao treinamento, buscando equilibrar as taxas de aprovação entre grupos diferentes.

RESULTADOS E DISCUSSÃO

3.1 Modelo sem mitigação

O modelo treinado diretamente com dados enviesados obteve acurácia de 99,65% e F1-score de 99,60%. Apesar do bom desempenho tradicional, os indicadores de justiça mostraram sérios problemas: *Disparate*

ISSN: 2447-6161

Impact foi de 0,62 e *Equal Opportunity* de apenas 0,005. Esses resultados confirmam a afirmação de Duarte (2020) de que alta precisão não significa justiça, pois sistemas podem ser tecnicamente eficientes e socialmente discriminatórios.

3.2 Modelo com mitigação

Após aplicar a técnica de mitigação, os resultados foram: acurácia de 82,5%, F1-score de 81,42%, *Disparate Impact* de 0,96 e *Equal Opportunity* de 0,32. Embora a performance tenha diminuído, houve uma expressiva melhoria nos indicadores de equidade. Essa evidência reforça a análise de Mehrabi et al. (2021) de que há um *trade-offt* inevitável entre precisão e justiça nos modelos de aprendizado de máquina.

3.3 Análise crítica

A comparação entre os dois modelos confirma que métricas tradicionais não são suficientes para avaliar IA em contextos sociais. O'Neil (2019) ressalta que a confiança excessiva em números de acurácia máscara desigualdades estruturais. Ao mesmo tempo, o caso europeu do *Artificial Intelligence Act* (União Europeia, 2021) mostra que políticas públicas podem estabelecer limites e exigências de transparência para reduzir riscos sociais.

Portanto, os resultados evidenciam que desenvolver IA justa exige uma visão interdisciplinar, incluindo ciência de dados, ética, direito e sociologia.

CONCLUSÕES

Este trabalho, intitulado *Algoritmos Enviesados e Inteligência Artificial: Riscos, Desafios e Responsabilidades Sociais*, teve como objetivo principal investigar como sistemas de Inteligência Artificial podem reproduzir e amplificar preconceitos sociais, caso não sejam cuidadosamente monitorados e ajustados. Por meio de fundamentação teórica e de um experimento computacional executado na plataforma Google Colab, foi possível observar com clareza o impacto do viés nos algoritmos de decisão automatizada.

O trabalho mostrou que algoritmos enviesados podem reforçar desigualdades mesmo quando atingem altos índices de acurácia. O experimento evidenciou que métricas de justiça, como *Disparate Impact* e *Equal Opportunity*, são essenciais para diagnosticar e corrigir distorções invisíveis. Constatou-se que a mitigação reduz a precisão, mas aumenta a equidade, confirmando a necessidade de equilibrar critérios técnicos e sociais.

Conclui-se que a construção de sistemas de IA responsáveis deve se basear em três pilares, como defendem Barocas e Selbst (2016): diversidade nos dados, auditorias constantes e adoção de métricas de equidade. Além disso, é fundamental o compromisso ético e regulatório (O'Neil, 2019; União Europeia, 2021) para assegurar que a IA seja utilizada como ferramenta de inclusão e não de exclusão.

AGRADECIMENTOS

Agradeço de coração à UNILAB, aos professores e colegas que me apoiaram e incentivaram durante a realização deste trabalho. Sem o auxílio e incentivo de todos, este estudo não teria sido possível.

REFERÊNCIAS

BAROCAS, Solon; SELBST, Andrew D. *Big Data's Disparate Impact.* California Law Review, v. 104, n. 3, p. 671–732, 2016.

BUOLAMWINI, Joy; GEBRU, Timnit. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, New York: PMLR, 2018. Disponível em: https://proceedings.mlr.press/v81/buolamwini18a.html. Acesso em:

25 maio 2025.

DUARTE, Ricardo. *O problema dos algoritmos enviesados*. Nexo Jornal, 2020. Disponível em: https://www.nexojornal.com.br/expresso/2020/03/02/O-problema-dos-algoritmos-enviesados. Acesso em: 25 maio 2025.

MEHRABI, Ninareh et al. *A Survey on Bias and Fairness in Machine Learning*. ACM Computing Surveys, v. 54, n. 6, p. 1–35, 2021. Disponível em: https://arxiv.org/abs/1908.09635. Acesso em: 25 maio 2025.

O'NEIL, Cathy. *Algoritmos de destruição em massa: como o big data aumenta a desigualdade e ameaça a democracia*. São Paulo: Rua do Sabão, 2019.

UNIÃO EUROPEIA. Proposal for a Regulation laying down harmonised rules on Artificial Intelligence: Artificial Intelligence Act. Brussels: European Commission, 2021. Disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206. Acesso em: 25 maio 2025.

ISSN: 2447-6161