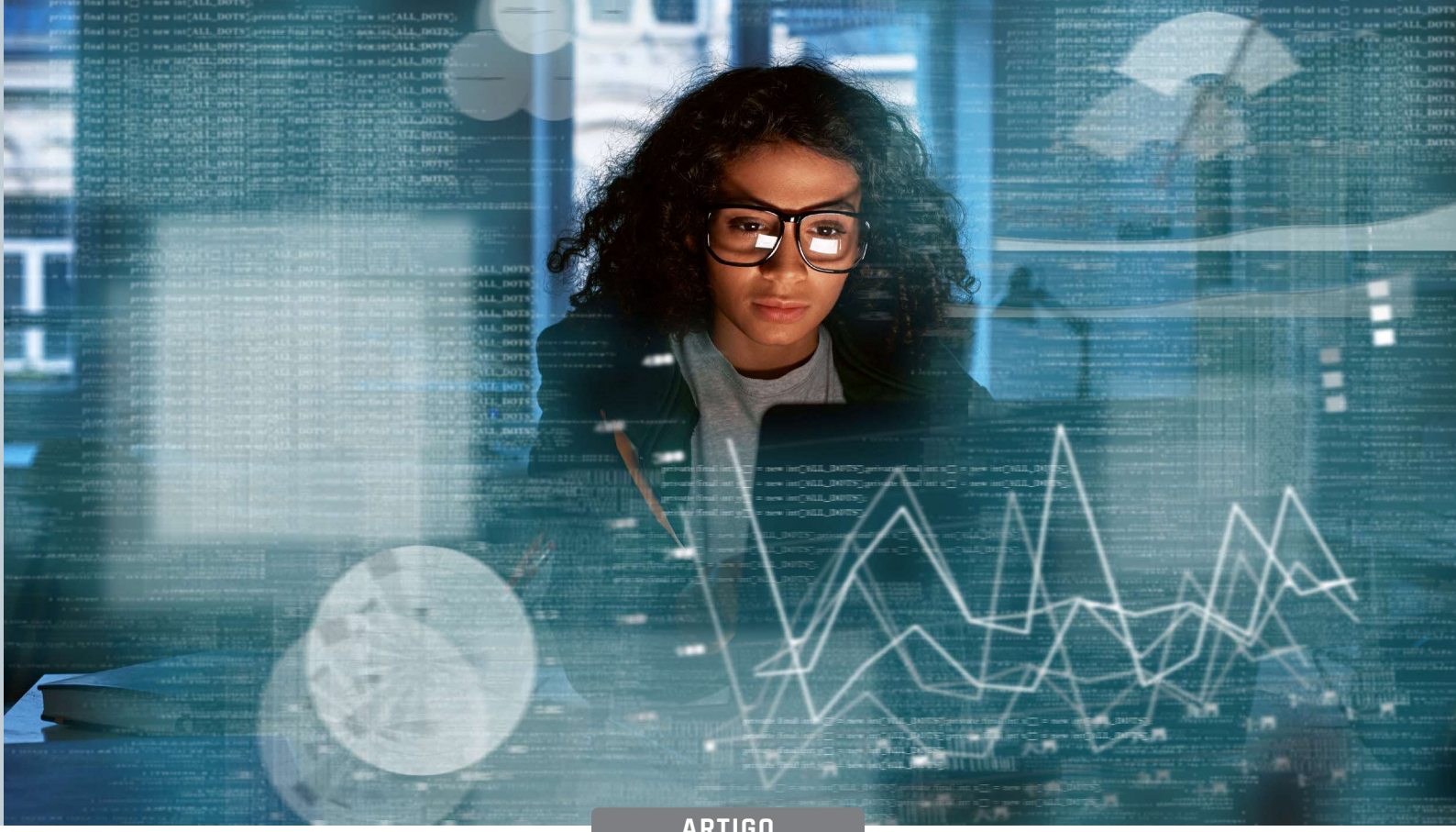


ARTIGO

# VIÉS, ÉTICA E RESPONSABILIDADE SOCIAL EM MODELOS PREDITIVOS

**POR**

Damires Yluska de Souza Fernandes e Alex Sandro da Cunha Rêgo

damires@ifpb.edu.br, alex@ifpb.edu.br

A sociedade, de um modo geral, já incorporou ao seu cotidiano o hábito de utilizar a Internet para manter-se digitalmente conectada. Essa atividade pode ser registrada em diferentes plataformas disponibilizadas na Web, sendo os dados passíveis de serem compartilhados publicamente conforme licença de uso e atendimento à Lei Geral de Proteção de Dados (LGPD). Dados disponíveis podem ser valiosos quando são usados em tarefas de análises, e estas fornecem *insights* que podem apoiar tomadas de decisão. Como ilustração, ao extrair dados de postagens em uma rede social como o Reddit[1], pode-se identificar padrões a respeito de posicionamentos políticos ou sentimentos de um público-alvo e, assim, planejar campanhas de marketing ou de cuidados à saúde com base nas percepções e sinais observados.

*Insights* de dados podem ser obtidos por meio de diversos tipos de análises de dados, como aquelas que utilizam técnicas de análise preditivas. Nestas categorias de análises, são construídos modelos

1 https://www.reddit.com/

que fazem previsões com base em padrões extraídos de dados históricos. Para treinar esses modelos, métodos de Aprendizado de Máquina (AM) são habitualmente utilizados [1]. A preparação de dados, seu uso na etapa de treinamento de modelos, assim como a interpretação de resultados obtidos dentro de determinados domínios de aplicação, compõem as tarefas trilhadas na área de Ciência de Dados.

Muitos trabalhos na área da Ciência de Dados têm perseguido a maximização da precisão e eficiência de modelos treinados; entretanto, aspectos éticos e legais associados aos dados usados na geração desses modelos têm sido cada vez mais pautados pela sociedade e pela comunidade científica atual. Nesse contexto, surgiu a denominada "Ciência de Dados Responsável", uma área que aborda questões de ética em inteligência artificial, qualidade de dados, justiça e diversidade algorítmica, transparência de dados e de algoritmos [7].

Um dos grandes desafios associados à construção de bons modelos preditivos diz respeito ao viés que pode ser induzido ao algoritmo de AM, pois os dados de treinamento são preparados por humanos, assim podendo já originalmente embutir algum tipo de parcialidade. O viés está enraizado na sociedade humana e, como resultado, reflete-se também nos dados. O impacto associado ao viés embutido pode se tornar ainda mais crítico quando o modelo treinado ultrapassa o limiar dos princípios éticos e porta-se de forma discriminatória, em que alguns grupos ou indivíduos são desfavorecidos [2].

Existem inúmeras categorias de viés, como viés temporal, comportamental e social [5]. Sua manifestação pode variar dependendo da perspectiva adotada. Como exemplo, um viés de cunho social pode ocorrer dentro de qualquer grupo social com base em atributos, muitas vezes sensíveis, como sexo, idade, etnia, orientação sexual, condição física etc. [8]. Quando há uma generalização imprecisa com base no grupo social ao qual uma instância de dados pertence, pode ocorrer uma injustiça no resultado da predição, ou seja, o modelo de AM pode desencadear uma conduta discriminatória associada ao viés. Nesse cenário, como ilustração, trabalhos apresentam ferramentas de contratação onde fica explícita a preferência por candidatos do sexo masculino em detrimento aos do sexo feminino [3]. Da mesma forma, alguns serviços de crédito tendem a oferecer linhas de crédito menores para mulheres do que para homens [2].

Eliminar completamente os vieses em modelos preditivos é uma tarefa desafiadora, haja vista que eles podem ser originados em diferentes etapas do processo de construção do modelo. Vários trabalhos têm buscado abordar aspectos e métodos para identificação e mitigação de problemas associados a vieses [2,4]. A concepção de um bom modelo preditivo exige conhecimento e habilidade em relação ao pipeline de projetos em Ciência de Dados e AM. O viés pode ser introduzido em qualquer estágio do desenvolvimento do modelo, muitas vezes de forma não intencional devido à falta de conhecimento do profissional e/ou até mesmo por ser algo habitual ao grupo social em que está inserido.

Para garantir que os modelos preditivos sejam desenvolvidos atendendo a princípios éticos, é essencial estabelecer uma equipe diversificada de desenvolvimento de produtos que seja ativa em todas as fases do processo. Uma equipe heterogênea trará uma "diversidade de pensamento" para a iniciativa, especialmente durante a seleção e limpeza de dados, o que pode ajudar a remover possíveis vieses. Além disso, é necessário atentar à coleta de dados e seus tratamentos. Neste percurso, a literatura destaca algumas práticas que norteiam a condução do processo com o intuito de mitigar o viés:

a) Partindo da fase de ingestão de dados, eles podem ser extraídos a partir de fontes diversas como websites, bancos de dados ou obtidos sob a forma de conjuntos de dados abertos. Estes devem estar diretamente relacionados ao problema o qual o modelo preditivo pretende resolver;

b) Na etapa de limpeza de dados, deve-se garantir que os atributos deles sejam precisos e confiáveis. Os dados devem ser de alta qualidade, com mínimo de erros, inconsistências, valores ausentes e outliers. Ao identificar e corrigir tais problemas, reduz-se o viés de exclusão, incrementando a precisão da análise dos dados.

c) Na etapa de seleção de dados são determinados os exemplos que irão fazer parte do conjunto de treinamento do modelo preditivo. O viés é introduzido quando os exemplos utilizados não refletem a distribuição real do problema. É importante que os dados possuam diversidade de representação, incluindo variedade de exemplos existentes no domínio do problema, além da atenção à quantidade de dados que será usada para treinar o modelo. Caso contrário, algumas categorias de exemplos podem ser super-representadas (viés de amostragem). Todos esses aspectos possuem um impacto significativo no treinamento do modelo e, desse modo, afetam seu desempenho.

d) Em problemas de classificação, os exemplos de dados precisam ser identificados por um rótulo. Pode ocorrer viés quando o dataset rotulado não é representativo quanto ao universo de rótulos possíveis, até mesmo pela dificuldade de rotular exemplos manualmente. A rotulação dos exemplos deve ser precisa, pois terá impacto na determinação da capacidade de generalização do modelo preditivo.

e) O viés de medição ocorre quando o cientista de dados realiza uma avaliação imprecisa do resultado devido à existência de ruídos no processo de seleção de dados. Por exemplo, considere um modelo concebido para predizer o risco de um paciente com pneumonia vir a óbito que utilize, também, dados de pacientes hospitalizados no período pandêmico da COVID-19. O vírus da COVID-19 desencadeia uma série de problemas respiratórios e provoca complicações generalizadas, o que potencializa o risco de morte e dificulta o real entendimento do quadro clínico

do paciente. Portanto, exemplos de pneumonia associados com a COVID-19 deveriam ser desconsiderados até que sejam explicados.

Quando o aprendizado é realizado a partir de um conjunto de dados desbalanceado, o modelo preditor é naturalmente enviesado a predizer a classe majoritária, normalmente a classe de menor interesse (maximização da acurácia). Neste caso, uma alta acurácia não significa necessariamente que o modelo é estável e consegue distinguir bem as classes envolvidas. Para desviar-se desse problema, é comum observar decisões que optam por estabelecer de forma arbitrária o equilíbrio da proporção das classes envolvidas e, assim, conseguir alcançar uma melhor taxa de acerto nas predições. Esse procedimento, entretanto, já atribui um viés automático, pois não reflete o cenário real do problema que é naturalmente desbalanceado.

Os exemplos de vieses apresentados são apenas uma amostra do que é possível encontrar ao percorrer o pipeline de desenvolvimento de um modelo preditivo. Não se olha apenas números de forma isolada em termos de desempenho, mas, especialmente, a habilidade de realizar decisões justas e equitativas, principalmente quando se trata de questões com dados sensíveis. Dados ou algoritmos enviesados podem influenciar eleições ou políticas públicas assim como incitar a violência. Modelos preditivos baseados em dados tendenciosos podem legitimar e ampliar políticas racistas ou violar de forma silenciosa e escalável leis de igualdade de oportunidades, principalmente em relação a minorias populacionais, levando à recorrente falta de diversidade e representatividade. Portanto, à medida que são desenvolvidos e implantados modelos preditivos, deve-se pensar também sobre os efeitos que esses modelos têm nos indivíduos, grupos populacionais e na sociedade em geral. Nessa perspectiva, como confiar nos resultados de modelos preditivos, entregues à sociedade e às organizações?

A constatação dessa preocupação tem encorajado a comunidade científica a definir e propor métodos para identificar e mitigar problemas com vieses em modelos de AM preditivos. Um dos conceitos estudados diz respeito à "justiça", que inclui a definição de técnicas e métricas para garantir que vieses em dados e modelos não gerem resultados desfavoráveis a grupos minoritários, reproduzindo discriminação [5] ou que, quando identificados, seja possível retificar dados ou parâmetros do modelo com o intuito de atenuar predições irresponsáveis. Alguns caminhos a serem trilhados em pesquisas e em políticas associadas são listados [2,6]:

**Curadoria dos dados**: a seleção de dados éticos e responsáveis deve iniciar com um processo de curadoria de dados abrangente e reprodutível.

**Critérios mensuráveis de justiça**: exemplos de métricas que vêm sendo experimentadas e verificadas incluem Paridade Demográfica (DP), Paridade de Taxa Positiva Verdadeiro/Falso (TPR/FPR) e Probabilidades Equalizadas (EO).

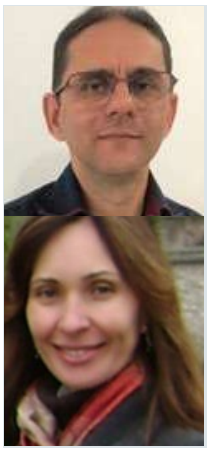
**Correção do desbalanceamento dos dados**: técnicas de balanceamento de dados podem prover a inclusão de dados com exemplos sub-representados e eliminar exemplos ruidosos da classe majoritária, que estão localizados no lado errado da borda de decisão. Pode-se também produzir dados sinteticamente para este fim.

**Ponderação de dados**: ponderar dados atribuindo pesos diferentes pode ser uma estratégia a ser adotada, dependendo dos dados sensíveis.

É importante sensibilizar organizações públicas e/ou privadas a respeito da justiça de dados e algorítmica e sua importância para a sociedade. Explicar modelos preditivos tornou-se, da mesma forma, uma importante área de pesquisa. Além disso, é imprescindível a adoção de políticas, o envolvimento da sociedade e esforços para uma educação realmente cidadã.


**Referências**

1. ALPAYDIN, E. Introduction to Machine Learning. 3rd Edition. Massachusetts: MIT Press, 2010.
2. CASTANEDA, J.; JOVER, A.; CALVET, L.; YANES, S.; JUAN, A.A.; SAINZ, M. Dealing with Gender Bias Issues in Data-Algorithmic Processes: A Social-Statistical Perspective. Algorithms 2022, 16 303.
3. DASTIN, J. Amazon scraps secret AI recruiting tool that showed bias against women. In Ethics of Data and Analytics; Auerbach Publications: B.R., FL, USA, 2018; pp. 296–299.
4. FIRMANI, D.; TANCA, L.; TORLONE, R. Ethical dimensions for data quality. Journal Data and Information Quality, Association for Computing Machinery, New York, NY, USA, v. 12, n. 1, 2019. ISSN 1936-1955.
5. PAGANO, T.P. et al. Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. Big Data Cogn. Comput. 2023, 7, 15.
6. SAHOO, Nihar et al. Detecting Unintended Social Bias in Toxic Language Datasets. ArXiv abs/2210.11762 (2022).
7. STOYANOVICH, J; LEWIS, A. Teaching Responsible Data Science: Charting New Pedagogical Territory. Int. Journal of Artificial Intelligence in Education (IJAIED), 2021.
8. VARONA, D.; SUÁREZ, J.L. Discrimination, Bias, Fairness, and Trustworthy AI. Appl. Sci. 2022, 12, 5826.


**ALEX SANDRO DA CUNHA RÊGO** é Professor Titular do Instituto Federal da Paraíba (IFPB), Doutor em Ciência da Computação pela Universidade Federal de Campina Grande. Atua em pesquisas associadas a Ciência de Dados, Aprendizado de Máquina e Processamento de Linguagem Natural.

**DAMIRES YLUSKA DE SOUZA FERNANDES** é Professora Titular e pesquisadora do PPGTI/IFPB, com mestrado e doutorado em Ciência da Computação pela UFPE. Atua em pesquisas associadas a Gerenciamento de Dados, Ciência de Dados, Aprendizado de Máquina e Análise de Sentimentos..