Racismo Algorítmico: Uma discussão fundamentada

Pedro M. O. Menezes[1], **Lucas A. Barbosa**[1]

[1]Centro Universitário SENAI CIMATEC – Salvador – BA – Brasil

pedromenezes01@hotmail.com, lucas.barbosa@fieb.org.br

Abstract. *Machine Learning Models, a subfield of Artificial Intelligence, are becoming increasingly common in everyday life. However, there is a need to analyze these models' impact on socially vulnerable minorities and the biases that can be presented, regardless of their cause and mitigating their consequences. For this, it is vital to understand incidents that have already occurred, noting their potential technical failures and related legal consequences. Based on this analysis, the work proposes solutions to minimize the impact of failures.*

Resumo. *Modelos de Aprendizado de Máquina, subárea da Inteligência Artificial, vem se tornando cada vez mais comuns no cotidiano. Contudo, há a necessidade de analisar o impacto que esses modelos podem ter em minorias socialmente vulneráveis e os vieses que podem ser apresentados, independente de sua causa e mitigando sua consequência. Para isso, é importante entender incidentes que já ocorreram, observando suas potenciais falhas técnicas e consequências jurídicas relacionadas. A partir dessa análise, o trabalho propõe soluções que podem minimizar o impacto das falhas.*

1. Introdução

A Inteligência Artificial (IA) vem se tornando uma ferramenta relevante à vida em sociedade nos últimos anos. Grande parte dos aplicativos, sites e serviços que são utilizados atualmente contam com algum tipo de suporte inteligente, podendo qualquer desenvolvedor contratar esses recursos [T. Silva, 2020]. Nesse contexto, há o destaque para o aprendizado de máquina (*machine learning* - ML), que trata do reconhecimento de padrões através de uma base de dados e posterior aplicação do aprendizado no reconhecimento das variáveis em outras unidades ou conjuntos de dados [Oliveira, 2018]. Em primeira instância, isso não parece ter nada de problemático ou perigoso, porém, ao perceber que o resultado de um processo de predição é função dos dados que lhe são fornecidos, surgem questionamentos: o que acontece se esses dados forem, propositalmente ou não, distorcidos? Esses modelos são passíveis de manipulação?

O presente artigo tem como propósito analisar a literatura acadêmica sobre racismo e discriminação algorítmica, trazendo à tona as discussões levantadas, relacioná-las e interpretar os problemas gerados por algoritmos que impactam minorias raciais e étnicas. Também é feita uma análise de datasets para relacioná-los às discussões e validá-las, elencando possíveis soluções, mensurando sua relevância e tangibilidade.

2. Fundamentação Teórica

T. Silva (2020) usa o conceito de branquitude para discutir a falsa objetividade de sistemas automatizados, a predominância de pessoas brancas como desenvolvedoras de tecnologias e o conceito de "caixa preta": dispositivos que funcionam em termos de entrada

de informações e decisões e saída de resultados e operações que ocultam os modos pelos quais ciência e tecnologia são construídas. O autor também levanta os vieses de dados de treinamento como agravante da falsa objetividade. M. L. Silva & Araújo (2020) dialogam diretamente com essas ideias ao fazer uma análise exploratória de casos de racismo algorítmico [T. Silva, 2022] para argumentar que a presença de algoritmos na vida cotidiana se tornou uma ferramenta no chamado racismo estrutural-algorítmico.

do Amaral et al. (2021) faz uma pesquisa investigativa sobre a neutralidade de algoritmos de bancos de imagem, comprovando a hipótese de que o "neutro" seria interpretado como "branco" pelos sistemas. Carrera (2020), em pesquisa similar, questiona a falta de interesse dos donos de grandes sistemas de modificá-los quando são apontados como problemáticos, também questionando se é necessário existir um estopim para fazer mudanças significativas em seus produtos em vez de fazê-las voluntariamente. A autora explica, também, o funcionamento de bancos de imagens digitais para exemplificar como a ideia de objetividade é equivocada, retomando o discutido por T. Silva (2020).

Fidalgo (2022) volta seu olhar para o uso de algoritmos nas relações trabalhistas, mostrando que além de preconceito racial, eles também são capazes de agir de forma machista, discriminando até mesmo mulheres grávidas. A autora também retoma as questões de diversidade dos meios de produção e treinamento das tecnologias.

O uso de sistemas imbutidos por algoritmos racistas por entidades estatais e na aplicação da lei já é reconhecido nas discussões legais e tecnológicas. de Lucena (2019) disserta sobre o Policiamento Preditivo — o uso de algoritmos para auxiliar a polícia a prever crimes —, explicando seu funcionamento, apontando suas inconsistências e mostrando como se trata de mais um mecanismo de perpetuação de vieses discriminatórios usando a bandeira da neutralidade. Possa (2022) argumenta como algoritmos de reconhecimento facial servem para agravar a segregação racial do sistema carcerário brasileiro ao afetar pessoas negras desproporcionalmente, e como seu uso é "consentido" pela sociedade sob a justificativa de auxiliar na segurança pública. Pereira (2020), também sobre o reconhecimento facial, traz o conceito de inimigo no direito penal, aqueles que devem ser vigiados e perseguidos pelo Estado e sociedade, para expandir como essa tecnologia e outras utilizadas em automatização de decisões judiciais afetam mais as pessoas negras. Costa & Kremer (2022) dizem que essa tecnologia, também, é uma ferramenta de controle estatal e que, em seu estado atual de desenvolvimento e treinamento, ela serve para continuar atingindo aqueles que já são considerados inimigos.

Diante do cenário exposto de dados de treinamento enviesados e falsa objetividade de sistemas, buscaram-se *datasets* de disponibilidade pública para avaliar sua distribuição e pensar quanto aos problemas levantados. As coleções análisadas são desbalanceadas quanto à raça, conforme pode ser visto na Figura 1. Para a elaboração da figura, foram selecionadas os 4 grupos étnicos mais representados entre os 6 conjuntos avaliados, com os demais agrupados na categoria "Outros". Os *datasets* de prisões em Chicago e Nova York diferenciavam "Brancos Hispânicos"de "Negros Hispânicos", mas, para esta análise, esses subconjuntos foram agrupados num único, "Hispânicos". No conjunto de dados faciais, diferenciou-se "Asiáticos"de "Indianos", mas essas categorias foram agrupadas por questões geográficas e de coerência com os demais conjuntos.

Os conjuntos referentes a prisões são numerosos em registros de "Negros" [OP,

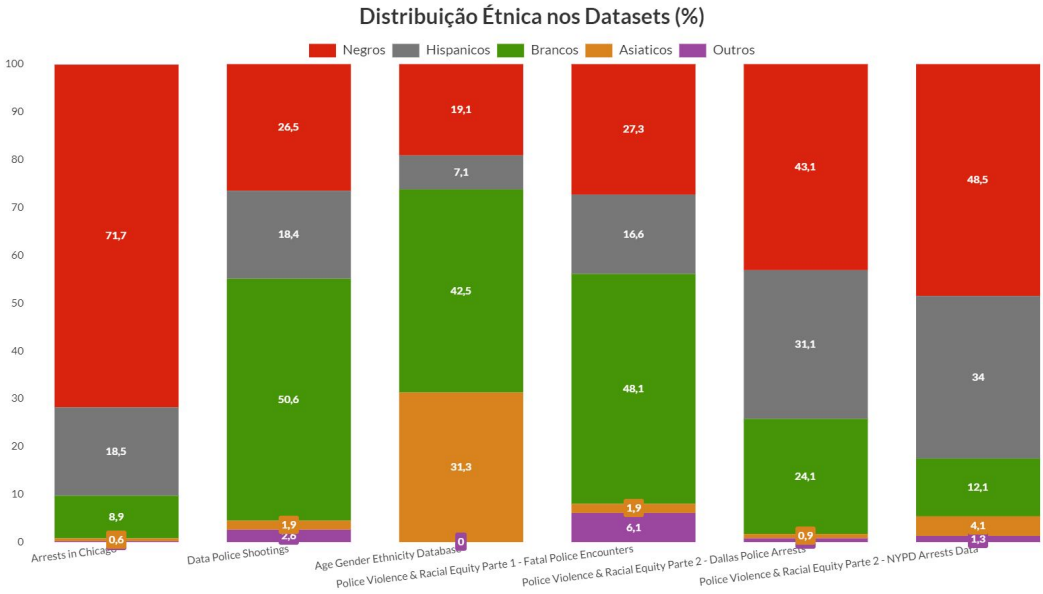

Figura 1. Comparativo de distribuição étnico-racial das bases de dados

2014] [JohnM, 2014a] [JohnM, 2014b]. Os referentes a fatalidades cometidas por policiais tem mais cadastros de "Brancos", tendo, também, números elevados de "Negros" e "Hispânicos" atingidos [Samoshyn, 2015] [JohnM, 2014c]. No conjunto de dados faciais, "Brancos" são a raça individual mais representada, porém, diferente de todos os outros *datasets*, "Negros" constituem menos de 20% de sua distribuição [Arora, 2020]. Com a falta de equilíbrio apontada, o problema levantado por T. Silva (2020) de vieses de dados de treinamento e, considerando a origem policial da maioria dos dados, os levantamentos de Possa (2022) e Pereira (2020) sobre a desproporcionalidade de alvos negros são ratificados.

Com base nos dados apresentados, é possível afirmar que as disparidades encontradas na análise dos *datasets* representam um universo constantemente utilizados em aplicações e sistemas diversos, apesar do trabalho ter sido executado em uma amostra pequena. Essas disparidades podem induzir a construção de soluções tecnológicas que reproduzem mazelas, viéses e preconceitos da sociedade.

3. Considerações finais

O trabalho apresentado busca trazer luz ao tema dos vieses que podem ocorrer no processo de construção, treinamento e utilização de modelos inteligentes. De forma muito homogênea, a base de trabalhos anteriores que discutiram o tema sempre converge no ponto da ideia vendida de que a tecnologia em si é neutra, incapaz de carregar nenhum tipo de preconceito. Conforme discutido, isso não é verdade por diversos fatores, como seu desenvolvimento e dados fornecidos.

A mais promissora solução apontada pelos autores está relacionada diretamente à diversidade: Diversidade nos dados utilizados para construir e treinar os modelos; Diversidade nas equipes de programadores; Diversidade na difusão desses modelos e nos locais de utilização dos mesmos. Quanto mais diversos forem os ambientes, menor é a possibilidade de um viés problemático passar sem ser percebido.

Como trabalhos futuros relacionados a essa pesquisa, um processo de validação de potenciais soluções e estratégias de mitigação levantadas na literatura é extremamente pertinente. É possível criar um ambiente controlado e forçar os erros que já ocorreram antes e, assim, validar se a solução apontada de fato minimiza o impacto do viés naquele contexto. É coerente, também, a criação de diretrizes de desenvolvimento que visem diminuir a criação de vieses nos modelos matemáticos, podendo fazer parte das respectivas legislações.

Referências

Arora, N. (2020). *AGE, GENDER AND ETHNICITY (FACE DATA) CSV*.

Carrera, F. (2020). Comunidades, algoritmos e ativismos digitais: olhares afrodiaspóricos. In *Racismo e sexismo em bancos de imagens digitais: análise de resultados de busca e atribuição de relevância na dimensão financeira/profissional* (p. 149-165). São Paulo: Literarua.

Costa, R. S., & Kremer, B. (2022). Inteligência artificial e discriminação: desafios e perspectivas para a proteção de grupos vulneráveis frente às tecnologias de reconhecimento facial. *Revista Brasileira de Direitos Fundamentais & Justiça*, *16*(1).

de Lucena, P. A. C. (2019). Policiamento preditivo, discriminação algorítmica e racismo: potencialidades e reflexos no brasil. In *Anais do vi simpósio internacional lavits: "assimetrias e (in)visibilidades: Vigilância, gênero e raça"*. Salvador: Lavits.

do Amaral, A. J., Martins, F., & Elesbão, A. C. (2021). Racismo algorítmico: uma análise da branquitude nos bancos de imagens digitais. *Pensar : Revista de Ciências Jurídicas*, *26*(4).

Fidalgo, L. B. B. (2022). Discriminações algorítmicas: racismo e sexismo nas relações laborais: Algorithmic discrimination: racism and sexism in labor relations. *Brazilian Journal of Development*, *8*(10).

JohnM. (2014a). *Dallas Police Arrests. In: Police Violence & Racial Equity - Part 2 of 2*.

JohnM. (2014b). *NYPD_Arrests_Data__Historic_. In: Police Violence & Racial Equity - Part 2 of 2*.

JohnM. (2014c). *Police Violence & Racial Equity - Part 1 of 2*.

Oliveira, C. (2018). Aprendizado de máquina e modulação do comportamento humano. In *A sociedade de controle: manipulação e modulação nas redes digitais.* São Paulo: Hedra.

OP, M. (2014). *Arrests in the City of Chicago (2014 - 2023).*

Pereira, D. F. M. (2020). O uso de câmeras de reconhecimento facial em contexto de pós democracia – uma ferramenta contra o inimigo no direito penal? *ADPEB*.

Possa, A. (2022). O reconhecimento facial como instrumento de reforço do estado de coisas inconstitucionais no brasil. *IDP Law Review*, *1*(2).

Samoshyn, A. (2015). *Data Police shootings.*

Silva, M. L., & Araújo, W. F. (2020). Biopolítica, racismo estrutural-algorítmico e subjetividade. *Revista Educação Unisinos*, *24*.

Silva, T. (2020). Visão computacional e racismo algorítmico: Branquitude e opacidade no aprendizado de máquina. *Revista da Associação Brasileira de Pesquisadores/as Negros/as (ABPN)*, *12*(31).

Silva, T. (2022). *Linha do tempo do racismo algorítmico.* Blog do Tarcizio Silva.