A PERPETUAÇÃO DO PRECONCEITO EM MODELOS DE MACHINE LEARNING

Luísa Andrade CORREA (Unileste); Gabriel Drumond Costa MAGALHAES (Unileste); Beatriz Maria Fernandes DE JESUS (Unileste); Marcelo Vieira CORREA (Unileste)

Introdução: A perpetuação do preconceito em modelos de Inteligência Artificial (IA) e machine learning é uma questão importante para o desenvolvimento de algoritmos de treinamento de máquinas. Não reproduzir discriminações nesses modelos é essencial para a construção de uma sociedade inclusiva e igualitária. Em geral, a literatura apresenta diferentes aspectos considerados como influenciadores do viés nas decisões de modelos de machine learning, tais como o tamanho da base de dados e os atributos selecionados. Na abordagem do problema, identificou-se algoritmos que identificam e removem características enviesadas (sexo, raça), bem como realizam balanceamento de classes.
Objetivo: Neste contexto, teve-se o objetivo de realizar uma revisão bibliográfica para entender como a ciência vem desenvolvendo técnicas para mitigar esse problema. Uma das mais promissores investigações das causas do desbalanceamento com base na seleção dos dados e dos atributos referentes a cada ocorrência.
Metodologia: As palavras chaves “bias in machine learning”, “prejudice in machine learning models”, entre outros, foram usadas para a seleção de artigos em bases científicas. De maneira geral, pode-se dizer que a seleção de atributos é mais significativa na atenuação do problema do que o tamanho do conjunto de dados. O desafio então é o de reduzir a propagação do preconceito sem afetar a acurácia do modelo. Assim, foi desenvolvido um método (Fair-SMOTE) que remove propriedades enviesadas bem como realiza o balanceamento de distribuições internas, de modo que exemplos de classes positivas e negativas apresentam a mesma quantidade.
Resultados: Apesar de o Fair-SMOTE ter apresentado um resultado semelhante em termos de redução da intolerância como outros algoritmos clássicos, ele apresentou uma melhor performance. Além disso, em outro artigo, é feita uma análise mais profunda, em que tenta-se entender qual aspecto do sistema de machine learning influencia o viés (como atributos e dados de treinamento). Chegou-se à conclusão de que aumentando a quantidade de características analisadas, tem-se uma melhora de 38% na questão do preconceito. Outrossim, contrariando o senso comum de que menor tendência discriminatória, menor acurácia, observou-se que o aumento dos atributos ajuda tanto na diminuição do viés quanto na melhora da precisão. Ademais, comprovou-se que um aumento no conjunto de dados com atributos insuficientes não ajuda na justiça do algoritmo, trazendo, inclusive, aumento no viés. Isso se torna um tema muito importante quando falamos de cidadania, uma vez que esses modelos estão cada vez mais sendo usados para tomar decisões, como escolha de currículos, concessão de crédito, sentença criminal, entre outros. Desse modo, é importante garantir que eles diminuam a discriminação e não que a propaguem ou aumentem.
Conclusão: Conclui-se que há abordagens que tratam sobre o preconceito em modelos de machine learning e como mitigá-lo. Assim, o grupo pretende testar os algoritmos e

os aspectos estudados para melhor entender como eles afetam o viés, bem como testar novas possibilidades de abordagem do problema.

Palavras-chave: Preconceito. Machine learning. Algoritmos.