XXXI Congresso de Iniciação Científica Unicamp 2023

Fairness em Aprendizado de Máquina: Fundamentos e Estudo de Caso em Viés Racial

Palavras-Chave: fairness, classificação, viés racial, aprendizado de máquina

Autores:

Ana Claudia Messias [FEEC/Unicamp]

Prof. Dr. Romis Ribeiro de Faissol Attux (orientador) [FEEC/Unicamp]

INTRODUÇÃO:

Diante do crescente interesse na área de Inteligência Artificial, que passou a ser utilizada em diversos setores para automatização de processos na vida humana, surge a necessidade de questionar eticamente seus efeitos diante da sociedade, e ter uma visão clara destes aspectos durante o uso de algoritmos de aprendizado de máquina, tanto de uma perspectiva filosófica quanto de uma perspectiva técnica. O conceito de aprendizado de máquina dialoga com o fato de que os computadores podem aprender através de dados, sem que seja preciso programá-los explicitamente [1]. Sendo assim, os dados, que descrevem escolhas feitas na sociedade, contribuem para o aumento da desigualdade, visto que o preconceito racial na sociedade é reproduzido nos dados e, indiretamente, nos produtos que envolvem aprendizado de máquina. Nestas condições, é possível comparar este ciclo de aumento de viés ao realimentar dados enviesados com a existência do racismo estrutural, visto que a definição em [2] relata como a própria estrutura social colabora para a manutenção do racismo. Embora o tema em questão tenha grande destaque na abordagem de reconhecimento de imagens devido aos problemas notados nas ferramentas de grandes empresas de tecnologia (Google, Facebook, Amazon e outras), o viés racial encontra-se presente também em outros cenários, como por exemplo inteligências artificiais geradoras de texto. Este cenário destaca a importância de medir e avaliar os produtos do aprendizado de máquina: por esse motivo, este projeto realiza um estudo sobre o uso de fairness em aprendizado de máquina no contexto de viés racial.

Para o desenvolvimento deste projeto, inicialmente, foi feito um estudo teórico acerca dos fundamentos de fairness em Aprendizado de Máquina [3] e revisão bibliográfica sobre o problema de viés racial no contexto de inteligência artificial e seu impacto na sociedade. Por fim, o último processo realizado foi a seleção de uma base de dados, em que serão utilizadas métricas de fairness para avaliação e mitigação do viés racial presente na classificação realizada. Tais esforços foram complementados por reuniões periódicas com o orientador.

METODOLOGIA:

A metodologia adotada no trabalho em questão se baseou em [4], e foi dividida nas seguintes etapas: seleção e tratamento dos dados, classificação binária e avaliação do modelo, que tem como objetivo identificar a renda de um indivíduo com base em alguns

atributos e verificar equidade no modelo de classificação diante da presença de uma variável sensível (raça).

1. Base de Dados

Para abordar o problema deste trabalho, citado anteriormente, os dados utilizados foram retirados de um dataset público que não possui viés, PNAD - 2014 [5] (Pesquisa Nacional por Amostra de Domicílios). Trata-se de um levantamento estatístico realizado com cidadãos brasileiros, e, de acordo com [6], tem como objetivo suprir a falta de informações sobre a população brasileira durante o período intercensitário e estudar temas insuficientemente investigados ou não contemplados nos censos demográficos decenais realizados por aquela instituição.

A manipulação dos dados ocorreu através da criação de um diretório no google drive, onde foi realizada a importação do dataset. O acesso a este *dataframe* foi efetuado pelo Google Colab através da biblioteca *pandas*. Cada coluna do dataset representava uma variável (pergunta do questionário), ao passo que cada linha correspondia a um indivíduo. Por se tratar de um questionário, havia muitas perguntas que não foram respondidas ou que não eram relevantes para o problema abordado. Sendo assim, os dados foram tratados da seguinte forma: criou-se um novo *DataFrame* apenas com os atributos selecionados como variáveis para a classificação, as variáveis 'renda' e 'raça' foram transformadas em variáveis binárias, e, em seguida, foram removidas as linhas que apresentaram campos com respostas nulas. Por fim, foi feita a normalização dos dados.

Transformação das variáveis 'renda' e 'raça' em binárias: O ajuste nos dados tem bastante relevância, visto que facilita a representação, simplifica o modelo e auxilia a interpretação dos resultados. Esse processo foi realizado da maneira apresentada a seguir, seguindo o modelo em [4]. O valor limitante de renda escolhido equivale a 1,5 salário mínimo no ano de 2023, e a categorização da raça no grupo desprivilegiado teve como critério de seleção raças que sofrem de racismo estrutural.

Tabela 1. Transformação da variável raça em binária

Raça Entrada	Grupo Atribuído	Valor Atribuído
branca amarela	Privilegiado	1
preta parda indigena	Desprivilegiado	0
sem declaração	-	removida

Tabela 2. Transformação da variável renda em binária

Renda	Valor Atribuído
maior que 1980	1
menor que 1980	0

2. Classificação por Renda

Para o desenvolvimento dos algoritmos de classificação, a linguagem de programação utilizada foi Python, em razão das diversas bibliotecas no contexto de aprendizagem de máquina, scikit-learn, por exemplo. Como citado anteriormente, o código foi desenvolvido na plataforma google Colab, pela facilidade em trabalhar com o dataset armazenado no próprio google drive, além de oferecer um boa visualização dos resultados e ajudar na organização e clareza do algoritmo.

O seguinte processo foi realizado com o intuito de investigar a presença de viés racial uso de machine learning ao classificar dados que possuem uma variável sensível. Com base nos atributos (raça, posição no trabalho, faixa de horas de trabalho semanal, maior nível educacional atingido, idade, grupamento de atividade principal de empreendimento do trabalho, grupamento ocupacionais do trabalho), os indivíduos da base de dados

foram classificadas de acordo com sua renda (entre aqueles que recebem um valor menor que R$1980,00 e os que recebem um valor maior). Para investigar a presença de viés no modelo, inverte-se apenas o valor da variável sensível (raça) de cada indivíduo, e é realizada uma nova classificação para investigar se o modelo se comporta de maneira semelhante mantendo o mesmo número de pessoas classificadas, ou se haverá discrepância, indicando viés, uma vez uma pessoa com as mesmas características mudando apenas a sua cor passa a ser classificada de maneira diferente pelo mesmo modelo.

Os atributos utilizados foram obtidos através da referência fornecida em [4] e com algumas modificações adotadas por critérios empíricos.

Os modelos de classificação abordados foram regressão logística, vantajoso devido sua eficiência computacional, simplicidade na implementação/interpretação, e gradiente boosting, vantajoso por utilizar árvores de decisão, que são fáceis de interpretar e lidam bem com dados categóricos. Os dados foram separados em conjuntos de treino e teste com proporção de 80% e 20%. Foi possível realizar o treinamento destes modelos, através do comando direto de regressão logística e gradiente boosting da biblioteca scikit-learn.

Até o presente momento, o modelo foi avaliado apenas através de comparações visuais de dados fornecidos pela matriz de confusão e acurácia. Como encaminhamento, o modelo ainda será avaliado através de medidas de justiça, como paridade estatística e igualdade de oportunidade

RESULTADOS E DISCUSSÃO:

Os modelos de regressão logística e gradiente boosting foram treinados com aqueles 80% dos dados, que foram separados para treinamento. Para testá-los, foi utilizado apenas os dados de teste rotulados como grupo privilegiado, em seguida, foi realizado outro teste do mesmo modelo com os mesmos dados, substituindo manualmente apenas a variável sensível por grupo desprivilegiado.

1. Regressão Logística

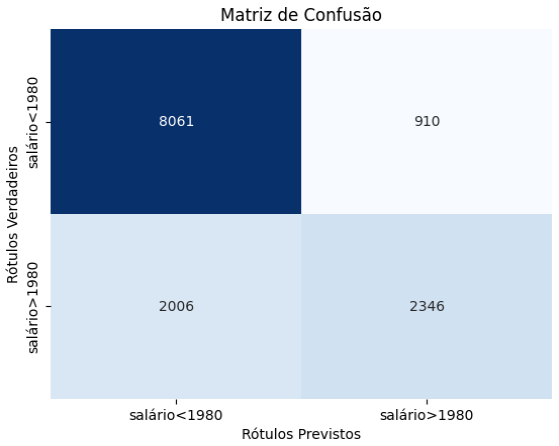

Figura 2 - Matriz de confusão - Regressão logística raça = 1

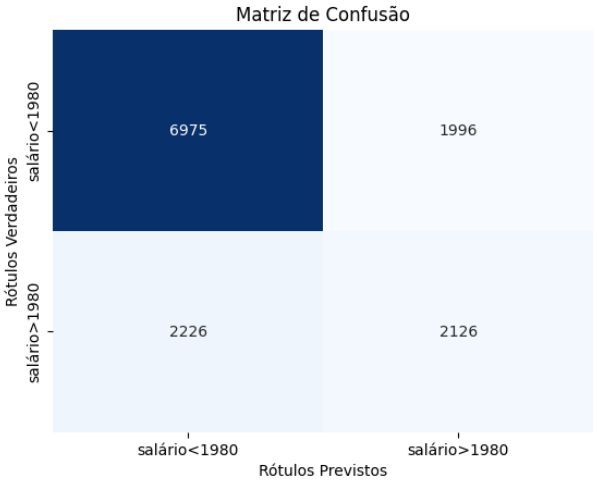

Figura 2 - Matriz de confusão - Regressão logística inverso de grupo raça = 0

2. Gradient Boosting

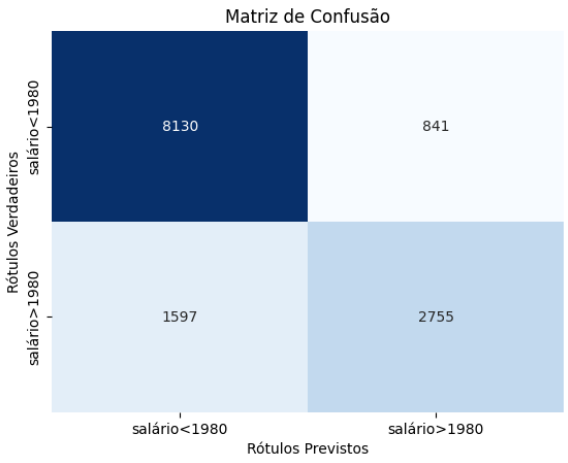

Figura 2 - Matriz de confusão - Gradiente Boosting

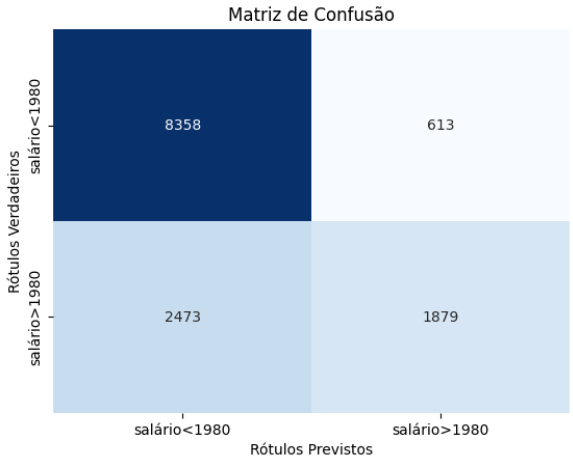

Figura 2 - Matriz de confusão - Gradiente Boosting

Tabela de comparação dos modelos de classificação

Dados Teste	Modelo	Acurácia
grupo privilegiado	Regressão Logística	0.78
grupo privilegiado com Inversão do atributo raça	Regressão Logística	0.68
grupo privilegiado	Gradiente Boosting	0.82
grupo privilegiado com Inversão do atributo raça	Gradiente Boosting	0.77

Analisando apenas a acurácia, é possível identificar uma queda em 10% no modelo de regressão logística apenas mudando a cor de cada indivíduo, o valor foi menor com relação ao gradiente boosting, ainda assim 5% é uma diferença bastante significativa. Com relação às matrizes de confusão, é possível observar que alterar a raça no modelo de regressão logística fez com que os rótulos previstos em renda maior que 1980 aumentassem de 3256 para 4122 pessoas, no entanto, para o caso do classificador com gradiente boosting que foi consideravelmente mais eficaz esse valor caiu de 3596 para 2492 pessoas classificadas com renda acima da média estabelecida, indicando viés racial e um peso considerável da variável sensível.

CONCLUSÃO:

O trabalho em questão visava estudar fairness e a presença de viés racial em modelos de aprendizado de máquina tanto no âmbito ético quanto técnico, analisando-os através do uso das métricas de avaliação estudadas.

Com relação à proposta inicial do trabalho, os testes conduzidos possibilitaram identificar a presença de viés racial no modelo mesmo sem o uso de métricas de justiça, e reforça a necessidade de avaliar e mitigar este viés, visto que os algoritmos têm impactado cada vez mais na tomada de decisão dos seres humanos, esse comportamento implica em um grande perigo quando essas decisões são enviesadas [7].

Como encaminhamento para o término do trabalho, serão realizados testes de *feature selection,* que realiza combinações de atributos para selecionar os mais importantes na classificação, com a finalidade de melhorar o modelo de regressão logística. Além disso, será alterado o parâmetro para definir a renda que separa os grupos entre privilegiados e não privilegiados, por fim, serão utilizadas técnicas já citadas de fairness para avaliar o viés e serão feitas intervenções no modelo para que mantenha uma boa acurácia e minimize a presença do viés. Por fim, é importante ressaltar que embora as variáveis sensíveis evidenciam os problemas na estrutura da sociedade e reforçam a presença de viés na classificação, eliminá-las no algoritmo pode não ser uma boa prática pela possibilidade de perder uma informação relevante.

AGRADECIMENTOS

Agradecemos à Unicamp e ao PIBIC, o suporte e apoio financeiro para a realização desta pesquisa.

BIBLIOGRAFIA

[1] [Géron, A. (2019)] A. Géron, Mãos à Obra: Aprendizado de Máquina com Scikit-Learn TensorFlow. Alta Books, 2019.

[2] [Almeida, 2019] S. Almeida, Racismo estrutural. Pólen Produção Editorial LTDA, 2019.

[3] [Barocas et al., 2021] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, acessado em https://fairmlbook.org/.

[4] [Cesaro, 2021] J. Cesaro, Avaliação de Discriminação em Aprendizagem de Máquina usando Técnicas de Interpretabilidade. Tese de Doutorado. Universidade de São Paulo, 2021.

[5] Base de microdados da PNAD. 2014. Disponível em: <https://centrodametropole.fflch.usp.br/>

[6] Informações gerais sobre a PNAD. Disponível em: <http://portal.mec.gov.br>

[7] [Silva, 2020] T. Silva, "Racismo Algorítmico em Plataformas Digitais: microagressões e discriminação em código. Comunidades, algoritmos e ativismos digitais: olhares afrodiaspóricos", p. 121-135, 2020.