Atividade 1 MO810/MC959

[*] O link contendo o jupyter notebook e outros conjuntos de dados e também o datacar estão no drive:

https://drive.google.com/drive/folders/1XbJe1t7qfeV_G1yKNZamV8UuZNwLuPZP?usp=sharing

1st Betania E R da Silva
IC-Unicamp

2nd Decio Miranda Filho
IC-Unicamp

3rd Felipe Scalabrin Dosso
IMECC-Unicamp

I. Introdução

Problemas sociais sempre estiveram presentes em todo o mundo, frequentemente afetando grupos vulneráveis de maneira desproporcional. As políticas sociais têm como objetivo mitigar essas desigualdades, oferecendo suporte a esses grupos. E na era da inteligência artificial (IA) e das extensas bases de dados, torna-se possível aplicar sistemas de IA para o bem social (*AI for Social Good*) [1], visando ampliar o impacto dessas políticas.

Os sistemas de IA têm o potencial de contribuir com questões em diversas áreas, como por exemplo, para os 17 Objetivos de Desenvolvimento Sustentável da ONU, já que essas tecnologias são capazes de gerar relatórios, análises, e predições, trazendo um impacto positivo em larga escala, facilitando a tomada de decisões e a distribuição de recursos de forma mais eficaz.

No entanto, assim como preconceitos e vieses estão arraigados na sociedade, eles também se refletem nos dados utilizados para treinar sistemas de IA. A representatividade nos dados é um exemplo claro disso. No campo da engenharia, por exemplo, homens superam as mulheres em uma proporção de quase 9:1 [2], o que demonstra que, a maioria das decisões tomada a partir desses dados será enviesada, penalizando fortemente o grupo de menor representatividade, nesse caso, as mulheres. Outro exemplo disso foi exposto em 2018, o sistema de recrutamento da Amazon, que apresentou um viés contra mulheres, isto é, as mulheres nunca eram selecionadas para as vagas, dado que os currículos usados para treinamento eram predominantemente de homens [3].

Assim, é importante que, ao desenvolver modelos de IA, haja um cuidado com a representatividade dos dados e com os vieses embutidos tanto nos dados quanto nos próprios desenvolvedores, pois o modelo tende a aprender o que é mais fácil, ou seja, o padrão, e isso pode resultar em comportamentos discriminativos e antiéticos.

A devida preocupação, monitoramento contínuo e utilização de técnica pode ser capaz de mitigar e criar sistemas mais justos e equitativos, que realmente contribuem para o bem-estar social, em vez de perpetuar as desigualdades já existentes.

Neste trabalho, utilizaremos o conjunto de dados *Adult* [4], também conhecido como *Census Income*. Essa base de dados foi extraída do Censo dos Estados Unidos de 1994 por Barry Becker e contém informações demográficas e socioeconômicas de indivíduos. A lista de atributos vão de idade, nível de educação, ocupação até sexo, país de origem e raça, proporcionando uma visão abrangente do perfil socioeconômico da população.

O objetivo principal deste *dataset* é prever se a renda de um indivíduo excede ou não US$50.000 anuais com base em seus atributos. Embora inicialmente voltado para o estudo da renda, o conjunto de dados pode ser aplicado em diferentes cenários práticos, como a concessão de crédito ou a análise de elegibilidade para benefícios sociais.

Neste trabalho, abordaremos a análise de elegibilidade para benefícios sociais, com foco em determinar se um indivíduo está apto a receber auxílio financeiro. Utilizando o conjunto de dados *Adult*, será possível automatizar o processo de pré-seleção de indivíduos, acelerando a distribuição de auxílios para que as pessoas que realmente necessitam recebam o suporte de forma mais rápida e eficiente.

Vale ressaltar que, neste *dataset* o primeiro desbalanceamento observado foi associado as mulheres, que estão em minoria, são 20380 homens para 9782 mulheres, o qual será melhor explicado na Seção III. Esse desbalanceamento é um exemplo de como questões de representatividade podem afetar o desempenho e a justiça de modelos preditivos.

O restante do trabalho está organizado da seguinte forma: Seção II discute os trabalhos relacionados, Seção III aborda a metodologia utilizada para a coleta e tratamento de dados.

II. Trabalhos Relacionados

A utilização de dados demográficos, como os do *Adult*, é comum em estudos que exploram a desigualdade de renda e buscam identificar vieses em modelos preditivos. O *Adult* tornou-se uma referência na área, apesar de suas limitações.

Barocas et al. [5], no livro *FairMLBook*, destacam um comportamento peculiar do *Adult* que o torna menos ideal para o estudo de vieses em aprendizado de máquina. Eles argumentam que os dados são relativamente antigos e pouco refletem a sociedade atual. A fixação de um corte de renda em US$50.000 anuais, por exemplo, posiciona quase todas as pessoas negras e grande parte das mulheres abaixo desse ponto de corte, dificultando a análise justa de vieses.

No estudo de Li-Pang Chen [6], o conjuntos de dados *Adult* foi explorado para identificar os principais fatores que afetam a renda dos adultos. O autor descobriu que os atributos

relationship, *native.country* e *capital.gain* têm maior impacto na predição de salários. No entanto, foi constatado que a distribuição de renda no *dataset* é desbalanceada: 76,1% dos indivíduos ganham menos de US$50.000 por ano, enquanto apenas 23,9% superam esse valor. Esse desbalanceamento pode introduzir vieses no modelo preditivo.

Recentemente, Girhepuje [7], realizou um estudo para identificar os vieses no conjunto de dados *Adult* a fim de mitigá-los. O autor identificou um viés significativo contra mulheres na predição de salários. Além disso, o estudo verificou que o viés contra cidadãos nascidos nos EUA era insignificante, dado que a maioria da amostra (27.504 indivíduos) era composta por pessoas nascidas nos Estados Unidos, em comparação com 2.658 indivíduos nascidos fora do país.

Outro estudo relevante é o de Ding et al [8], que discute as limitações do *Adult*, apesar de seu uso frequente no desenvolvimento de ferramentas, como o framework AI Fariness 360 [9] e critérios de *fairness*. Entre as limitações apontadas, está o limiar de renda binário fixado em US$50.000, o que pode levar a distorções nos resultados preditivos. Os autores propõem novos conjuntos de dados, também baseados em pesquisas do Censo dos EUA, mas que abrangem outras áreas, como saúde, transporte e habitação, além de ampliar o limiar de renda para valores entre US$60.000 e US$70.000. No entanto, os autores alertam que o aumento da quantidade de dados não elimina necessariamente as disparidades algorítmicas, ressaltando a complexidade do problema.

Já no contexto da análise de elegibilidade para benefícios sociais, vê-se a necessidade de análise de dados similares aos do *Adult*. O caso do Instituto Nacional do Seguro Social (INSS) no Brasil exemplifica a urgência de se implementar sistemas de inteligência artificial para automatizar processos e agilizar a concessão de auxílios. De acordo com Barchilon e Escovedo [10], em 2019, o INSS enfrentou um acúmulo alarmante de 833 mil novos requerimentos por mês, sendo processados por aproximadamente 6 mil servidores. A alta demanda e a baixa capacidade de processamento resulta em longas filas de espera, afetando especialmente os grupos mais vulneráveis, como idosos e pessoas com doenças graves, que muitas vezes aguardam por anos para ter seus pedidos analisados.

A aplicação de IA nesse cenário pode contribuir de forma a agilizar a análise dos requerimentos, otimizando o fluxo de concessão de auxílios.

III. Metodologia

A. Coleta e Tratamento de Dados

Inicialmente a coleta dos dados envolveu uma procura vasta e extensiva de datasets que contivessem variáveis sensíveis e ao mesmo tempo a atribuição pelo fator **renda**. Desta forma o *Dataset Adult* [4], retirado diretamente de sua fonte primária, presente na base de dados da *UCI-Machine Learning Repository* [11] foi o escolhido para realizarmos o projeto.

Averiguou-se os dados brutos e foi feita a limpeza primária, verificando valores *missing* ou *null*, foram encontradas variaveis cujos nomes possuiam espaços em branco, os quais foram retirados. Realizou-se também a atribuição dos nomes às colunas, as quais estavam num arquivo separado. A partir disso iniciou-se com dados de brutos de 32561 linhas e 15 colunas para o arquivo data_adult enquanto o arquivo de teste- que já veio separado e dividido, o qual contém 33% das amostras totais (16281 amostras).

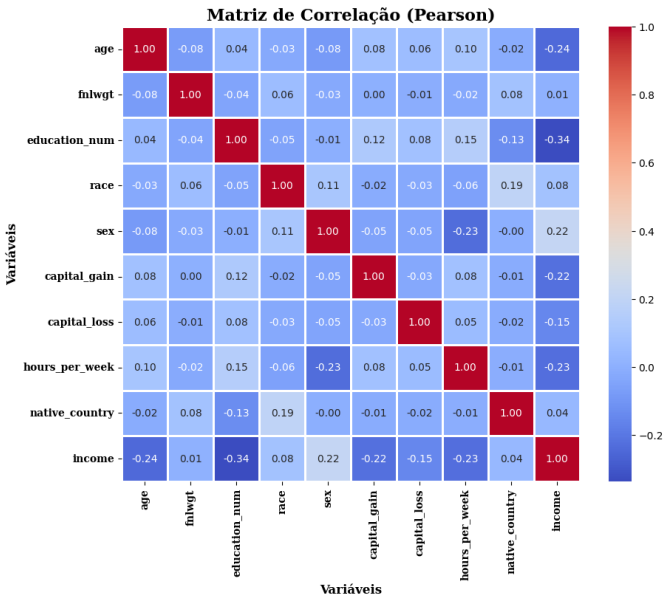

Figura 1. Gráfico de correlação.

Em seguida, mantendo os dados de teste separados do processo para etapas posteriores iniciou-se a veficação dos dados. Visualizou-se a distribuição das features e as estatísticas sumárias para verificar se havia algo a considerar a priori. Também utilizou-se a matriz de correlação (de Pearson) conforme a Figura 1, não foram identificadas variáveis altamente correlacionadas com ***income*** *(Target)* e nem com as variáveis preditorias entre si, não demonstrando problemas de multicolineariade.

Outra consideração é que para algumas features categóricas foram exculídas amostras que continham o caractere ”?”, pois não ajudaria nas análises.

O processo completo da etapa de processamento, passando pelas observações feitas sobre as features sensíveis e até a exportação dos dados prontos para a modelagem futura pode ser verificado melhor no fluxograma 2 que simplifica as etapas seguidas nesta fase.

B. Reconhecendo Atributos Sensíveis e Tratamento de Dados

Assim como outros trabalhos na área de IA ético e fairness observa-se que ocorre a codificação binária para variáveis sensíveis em muitos trabalhos [12] [13]. Pensando no fato rotulatação das variáveis sensíveis, o atributo protegido é rotulado como ”1”para facilitar a detecção de desigualdade ou viés decidiu-se atribuir para os grupos protegidos o rótulo 1 (Feminino, no caso de Sexo;’Não Branco ’ no caso de raça;

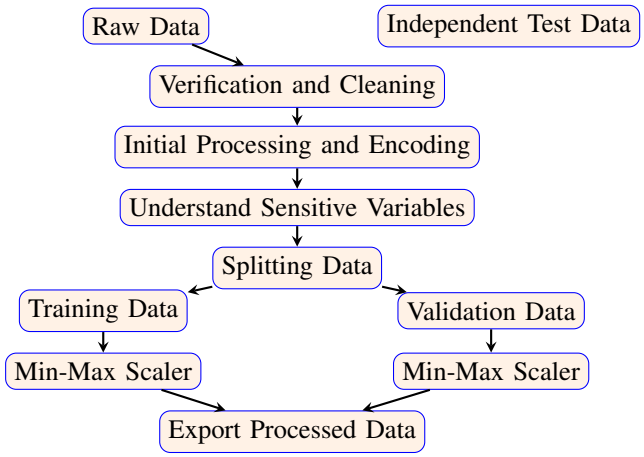

Figura 2. Fluxograma do processamento de dados.

Imigrante, no caso País de Origem)- assim como em trabalhos semelhantes [14] [15], enquanto para o grupo não protegido atribui-se 0 ((Masculino, no caso de Sexo;'Branco'no caso de raça;Nascido nos EUA, no caso de País de origem), mais detalhes descritos na tabela I. Após a categorização de binários para as features citadas realizou-se uma codificação ***One-Hot*** para as features categóricas restantes, tornando o dataset agora com 55 colunas (em vez das 15 inicialmente).

Em seguida, realizou-se um *split* dos dados brutos *(raw-data)* entre treino e validação. Com isso, as proporção de treino,validação e teste seguiu a proporção de 1/3 do total para teste e restando 50% do total para treino e cerca de 10% para validação.

Em conjunto com a fase da análise descritiva observou-se os dados relacionados aos aitributos sensíveis. A tabela I mostra a proporção de diferentes categorias dos atributos sensíveis e da variável ***income***. As colunas estão organizadas por variáveis (Renda, Raça, Gênero/Sexo e País de Origem), cada uma com seus respectivos percentuais.

Ao analisá-la percebemos uma diferença na proporção para todas as variáveis explicitadas. renda apresenta 24,89% das pessoas têm uma renda maior que 50 mil dólares, enquanto 75,11% têm renda menor ou igual a 50 mil dólares. Raça mantém-se com 14,02% das pessoas não são brancas, e 85,98% são brancas. Para sexo/gênero, 32,43% são mulheres e 67,57% são homens.E também, para país de origem somente 8,81% das pessoas não são nativas e 91,19% são nativas (naturais dos EUA).

Portanto, a partir da descrição fica evidente que há desproporção grande intra-grupos, o que deve ser levado em consideração para mitigação dos vieses algorítimicos

Renda (%)	Raça (%)	Sexo (%)	País de Origem (%)
24.89 (>50K)	14.02 (Não Branco)	32.43 (Feminino)	8.81 (Imigrante)
75.11 (≤50K)	85.98 (Branco)	67.57 (Masculino)	91.19 (Nativo-EUA)

Tabela I
PROPORTION OF SENSITIVE CATEGORIES

A figura apresentada 3 explicita a distribuição de renda por país de origem (Estados Unidos e Imigrantes), dividida entre renda alta e baixa. Os dados indicam que a maioria dos indivíduos com baixa renda é originária dos Estados Unidos (20.509), enquanto um número menor é imigrante (2.145). Para indivíduos com renda alta, o padrão é semelhante: 6.995 pessoas são dos Estados Unidos e apenas 513 são imigrantes. Isso revela uma disparidade significativa entre nativos e imigrantes tanto em termos de quantidade quanto de distribuição de renda, com imigrantes representando uma minoria nas duas categorias de renda.

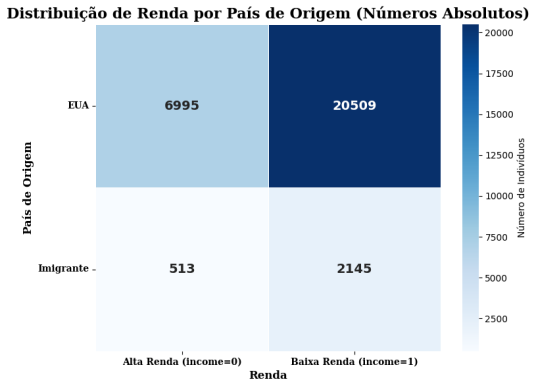

Figura 3. Distribuição de renda por país de origem.

A figura 4 mostra a distribuição das variáveis raça pela variável sexo, segmentada em duas categorias de renda: alta e baixa com seus devidos atributos. No gráfico à esquerda (renda alta), observa-se que, tanto para mulheres quanto para homens, a maior parte dos indivíduos é branca, com uma pequena proporção sendo não branca.

No entanto, a quantidade total de homens é consideravelmente maior do que a de mulheres. No gráfico à direita (renda baixa), há um número muito maior de indivíduos em geral, com uma predominância de homens brancos e um número menor de mulheres. Em ambos os gráficos, a proporção de indivíduos não brancos é menor, mas há mais diversidade racial entre os indivíduos de baixa renda, especialmente entre os homens.

IV. COMO OS DADOS SERÃO UTILIZADOS

Para que seja possível entender como os dados serão utilizados, é preciso definir um procedimento rigoroso para trabalhar com eles. Esse processo é chamado de ciclo de vida de aprendizado de máquina e é constituído por seis etapas: especificação do problema, entendimento dos dados, processamento dos dados, modelagem, validação do modelo e implementação e monitoramento. Cada etapa é fundamental para garantir que o sistema seja preciso, justo e ético, e envolve a colaboração de diversos profissionais e partes interessadas e podendo envolver constantes iterações entre as fases [16] e detalhamentos maiores a depender do caso [17]. A seguir, detalha-se como cada fase é organizada e o papel de cada ator envolvido no ciclo.

Na fase inicial, os proprietários do problema e cientistas de

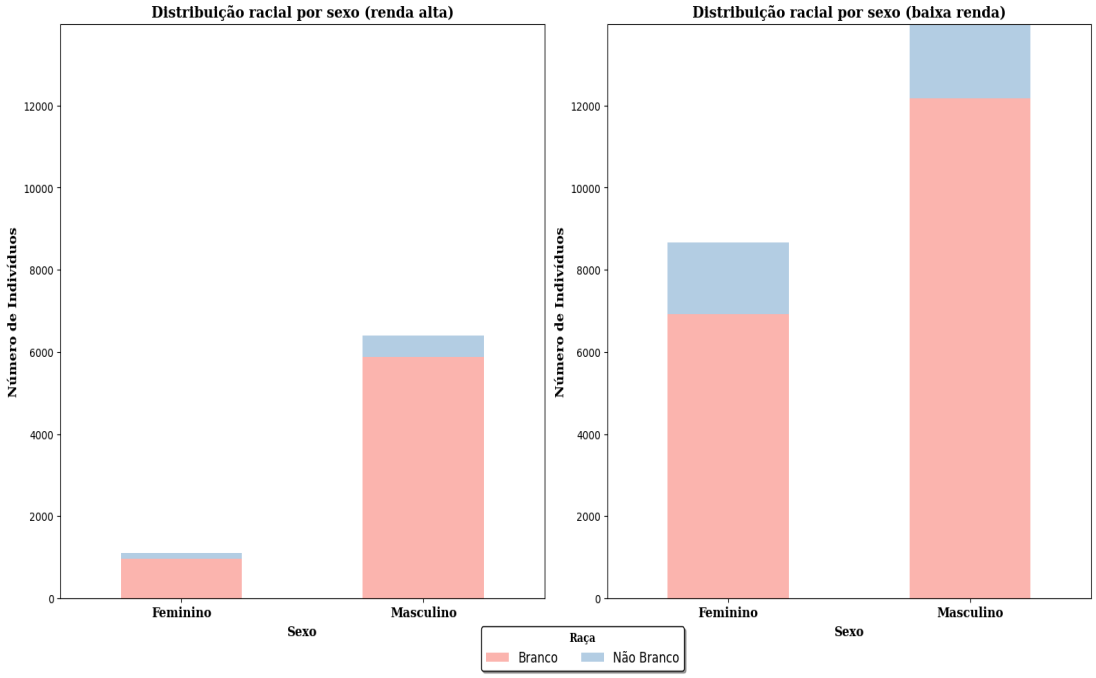

Figura 4. Distribuição de sexo e raça.

dados trabalham juntos para definir claramente o problema a ser resolvido — neste caso, a predição de elegibilidade para benefícios sociais com base em características demográficas e socioeconômicas. O objetivo aqui é garantir que o modelo contribua para decisões mais justas e ágeis, beneficiando quem realmente necessita do auxílio. *Stakeholders* afetados, como beneficiários potenciais, podem fornecer feedback sobre como o sistema impacta suas vidas, ajudando a ajustar o foco das previsões e métricas de sucesso.

No passo seguinte, engenheiros de dados e cientistas de dados exploram o conjunto de dados *Adult*, analisando a estrutura dos dados, identificando valores ausentes e avaliando a distribuição das variáveis sensíveis, como sexo, raça e país de origem. Nessa etapa, eles trabalham em estreita colaboração com especialistas em ética para detectar atributos que gerem possíveis vieses e o que deve ser feito para mitigar o impacto desses vieses no modelo.

Após o entendimento dos dados, engenheiros de dados realizam a limpeza e transformação dos dados, o que inclui a codificação de variáveis categóricas, normalização de atributos numéricos, remoção de valores ausentes e separação dos dados em treinamento, validação e teste. Aqui, são tomadas decisões sobre como lidar com atributos sensíveis para garantir que o modelo não aprenda padrões discriminatórios. Especialistas em *fairness* podem ser chamados para revisar a preparação dos dados, garantindo que as técnicas utilizadas mitiguem adequadamente os vieses observados.

Na fase de modelagem, os cientistas de dados selecionam e treinam modelos de classificação, como árvores de decisão, para prever a elegibilidade. Além das métricas tradicionais de performance, como a acurácia, são aplicadas métricas de *fairness* para garantir que o modelo não esteja favorecendo certos grupos em detrimento de outros. *Stakeholders* de políticas públicas podem ajudar a definir critérios justos e aplicáveis para garantir que o modelo funcione de acordo com os objetivos sociais pretendidos.

Model validators realizam testes rigorosos para avaliar o desempenho e a robustez do modelo. Essa avaliação inclui a análise da equidade, verificando se o modelo trata igualmente todos os grupos demográficos. Reguladores e auditores externos podem ser envolvidos para garantir que o modelo esteja em conformidade com leis e regulamentações de proteção de dados e ética.

Uma vez implementado, o modelo será monitorado continuamente por engenheiros de operações de aprendizado de máquina. A performance do modelo será acompanhada ao longo do tempo para detectar novos vieses introduzidos. Usuários finais, como órgãos governamentais ou instituições sociais, fornecerão feedback sobre o impacto do sistema, e os tomadores de decisões garantirão que o modelo continue alinhado com as políticas sociais em vigor. Novos ajustes poderão ser feitos com base nesse monitoramento e feedback.

V. Conclusão

Portanto, mediante o apresentado nota-se que os dados podem possuir disparidades envolvendo atributos sensíveis, os quais podem impactar negativamente o *fairness* e reforçar vieses que já são trazidos diretamente pelos dados. Com essa primeira atividade da discipina, crê-se que nas próximas etapas conseguiremos modelar e melhorar essa prioridade, além de tornar os modelos explicáveis.

REFERÊNCIAS

[1] McKinsey Company, "Applying artificial intelligence for social good," https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good, 2018, accessed: 2024-09-08.

[2] A. Patrick, C. Riegle-Crumb, and M. Borrego, "Examining the gender gap in engineering professional identification," *J Women Minor Sci Eng*, vol. 27, no. 1, pp. 31–55, 2021, author manuscript; available in PMC 2024 Jan 19. Published in final edited form as J Women Minor Sci Eng.

[3] I. A. Hamilton, "Why it's totally unsurprising that amazon's recruitment ai was biased against women," https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10, 2018, accessed: 2024-09-08.

[4] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20. [Online]. Available: https://archive.ics.uci.edu/dataset/2/adult

[5] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[6] L.-P. Chen, "Supervised learning for binary classification on us adult income," *Journal of Modeling and Optimization 2021;13(2):80-91*, 2021, received: 25 April 2021; Accepted: 30 July 2021; Available online: 10 November 2021.

[7] S. Girhepuje, "Identifying and examining machine learning biases on adult dataset," *arXiv preprint arXiv:2310.09373*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.09373

[8] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring adult: New datasets for fair machine learning," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 6478–6490.

[9] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[10] N. Barchilon and T. Escovedo, "Machine learning applied to the inss benefit request," in *Proceedings of the Department of Informatics, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)*. Rio de Janeiro, RJ, Brazil: PUC-Rio, 2024.

[11] M. Kelly, R. Longjohn, and K. Nottingham, "The uci machine learning repository," 2023, accessed: 2023-09-09. [Online]. Available: https://archive.ics.uci.edu

[12] K. Kobayashi and Y. Nakao, "One-vs.-one mitigation of intersectional bias: A general method to extend fairness-aware binary classification," *ArXiv*, vol. abs/2010.13494, 2020.

[13] A. Shah, M. Shen, J. Ryu, S. Das, P. Sattigeri, Y. Bu, and G. Wornell, "Group fairness with uncertainty in sensitive attributes," *ArXiv*, vol. abs/2302.08077, 2023.

[14] C. Mougan, J. M. Alvarez, S. Ruggieri, and S. Staab, "Fairness implications of encoding protected categorical attributes," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 454–465. [Online]. Available: https://doi.org/10.1145/3600211.3604657

[15] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *ArXiv*, vol. abs/1808.00023, 2018.

[16] C. Yang, W. Wang, Y. Zhang, Z. Zhang, L. Shen, Y. Li, and J. See, "Mlife: A lite framework for machine learning lifecycle initialization," *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, 2021.

[17] C. Weber, P. Hirmer, P. Reimann, and H. Schwarz, "A new process model for the comprehensive management of machine learning models," pp. 415–422, 2019.