Imparcialidade

Bruna Almeida Osti* e Fabio Fogliarini Brolesi†
Instituto de Computação, Universidade Estadual de Campinas - UNICAMP
Campinas, SP
Email: *b231024@dac.unicamp.br, †brolesi@gmail.com

***Resumo*—Este trabalho busca abordar a identificação e correção de possíveis imparcialidade ligadas aos dados sensíveis para o projeto da disciplina, mostrando diferentes métricas de avaliação e abordagens de mitigação das imparcialidades sendo abordagens de pré-processamento, in-processamento e pós-processamento.**

***Index Terms*—Trustworthy, AI, Ethical**

I. INTRODUÇÃO

Os modelos prognósticos usam vários fatores em combinação para prever o risco de resultados clínicos futuros em pacientes. Um bom modelo deve (i) fornecer previsões precisas que informam os pacientes e seus cuidadores, (ii) apoiar a pesquisa clínica e (iii) permitir decisões para melhorar os resultados dos tratamentos aos pacientes, segundo [1]. Um modelo prognóstico tem três fases principais: desenvolvimento do modelo (incluindo validação interna), validação externa e investigações de impacto na prática clínica. Embora muitos modelos prognósticos sejam propostos, poucos são atualmente usados na prática, conforme mostra [1].

Mas há sérios riscos em aprender padrões com exemplos, não é um processo de simplesmente memorização. Envolve generalizar e aprimorar os detalhes que são característicos de uma classe em geral, não apenas indivíduos específicos que aparecem nos exemplos. Este é o processo de indução: extrair regras gerais de exemplos específicos — regras que efetivamente respondem por casos passados, mas também se aplicam a casos futuros, ainda não vistos [2].

O fato de ser baseado em evidências não garante que os modelos levaram a decisões precisas, confiáveis ou justas. Nossos exemplos históricos dos resultados relevantes quase sempre refletirão preconceitos históricos contra certos grupos sociais, esteriótipos culturais predominantes e desigualdades demográficas existentes, e aprender padrões sobre esses dados significa geralmente replicar esses mesmos preconceitos [2].

Neste trabalho pretende-se estruturar uma automação que seja capaz de predizer quais pacientes sairão da unidade de terapia intensiva dadas características presentes nos datasets fornecidos que tem relação com cirurgias ortopédicas e dados de estadia em unidades de terapia intensiva, entendendo as features sensíveis e os possíveis viéses relacionados a elas, além de aplicar técnicas de mitigação de possíveis viéses.

II. TRABALHOS RELACIONADOS

Em 2018, o estudo de prognóstico [3] incluiu admissões de adultos num centro médico acadêmico em vários locais entre 2015 e 2017. Foi desenvolvido um modelo de previsão da mortalidade por todas as causas (incluindo a iniciação de cuidados hospitalares) no prazo de 60 dias após a admissão. A generalizabilidade do modelo é avaliada na validação temporal no contexto de um potencial enviesamento demográfico. Um estudo de coorte prospectivo subsequente foi realizado nos mesmos locais entre Outubro de 2018 e Junho de 2019, além disso, foi treinado um modelo em todo o conjunto de treino, e aplicado o teste para sub-coortes por combinações de demografia sensível (por exemplo, mulheres negras admitidas em Brooklyn) e são relatadas várias medidas de desempenho do modelo. Este procedimento é repetido para um segundo modelo onde todos os dados demográficos sensíveis são excluídos ou "mascarados"durante a formação. A validação temporal demonstra uma boa discriminação modelo para a mortalidade em 60 dias. Pequenas variações de desempenho são observadas em subpopulações demográficas. O modelo foi implementado prospectivamente e produziu com sucesso estimativas significativas de risco dentro de minutos após a admissão.

Ainda em 2018, o estudo [4] utilizou a abordagem de aprendizado adversarial para desenvolver um modelo "equitativo"de previsão de risco para a doença cardiovascular aterosclerótica (ASCVD) com EHR. Utilizaram o gerador para construir o preditor de risco e discriminador para impor probabilidades equalizadas para os riscos previstos em diferentes grupos protegidos.

Em 2021, o estudo [5] comparou diferentes métodos para reduzir viéses de modelos de machine learning utilizados no cenário clínico real para predição de depressão pós-parto. Os modelos de machine learning (regressão logística [LR], random forest, e extreme gradient boosting) foram treinados para 2 resultados binários: depressão pós-parto (PPD) e utilização do serviço de saúde mental pós-parto. Para cada resultado foram utilizados modelos lineares generalizados ajustados ao risco para avaliar a potencial disparidade na coorte associada à raça binarizada (Preto ou Branco).

Os métodos para reduzir o enviesamento, incluindo a ponderação de instâncias, o removedor de preconceitos, e a remoção da raça dos modelos, foram examinados através da análise das mudanças nas métricas de justiça em comparação com os modelos de base. Foram comparadas as características de base dos indivíduos do sexo feminino no decil de maior risco previsto para diferenças sistemáticas. Métricas de equidade de impacto díspar (DI, 1 indica equidade) e diferença de oportunidades iguais (EOD, 0 indica equidade).

Os resultados deste estudo sugerem que o desempenho variou em função do modelo, etiqueta de resultados, e método

para reduzir o enviesamento. Modelos de previsão clínica treinados em dados potencialmente tendenciosos podem produzir resultados injustos com base nas métricas escolhidas. Portanto, seria necessário comparar diferentes abordagens para o contexto em que são aplicadas.

III. METODOLOGIA

As noções de igualdade propostas na literatura geralmente são classificadas em grandes áreas, como: (1) definições baseadas na semelhança das métricas estatísticas entre grupos identificados por diferentes valores em atributos protegidos (ex. masculino e feminino, pessoas em diferentes grupos etários); (2) definições centradas na prevenção de tratamentos diferentes para indivíduos considerados semelhantes à respeito de uma tarefa específica; (3) definições que defendam a necessidade de encontrar a causalidade entre as variáveis para evitar os impactos injustos nas decisões [6].

De modo geral, se concentram em três tipos de metodologias: métodos de pré-processamento, que tentam remover o viés diretamente dos dados; métodos de processamento, impondo justiça como uma restrição ou como uma perda adicional durante a otimização; pós-processamento, trabalhando diretamente nas saídas resultantes do modelo.

A. Features sensíveis

Os dados que serão utilizados foram disponibilizados pelo Professor e são extraídos dos registros hospitalares do Instituto Nacional de Traumatologia e Ortopedia Jamil Haddad (INTO)[1]. Eles referem-se a pacientes que foram submetidos a cirurgias ortopédicas e registros de internações em unidades de terapia intensiva de alguns pacientes. Após tratamento dos dados para colocá-los em um único dataset, identificamos como features sensíveis:

- `AGE`: idade do paciente
- `SEX`: sexo do paciente (masculino ou feminino)

Neste trabalho trataremos o grupo de indivíduos como privilegiados pertencendo ao sexo feminino. Entretanto, quando analisamos o contexto do problema verificamos que a label positiva é na verdade de mais óbitos, ou seja, na verdade não privilegiados.

B. Métricas de imparcialidade

As métricas de imparcialidade são medidas que detectam a presença de viéses nos dados ou nos modelos, portanto, detectam a presença de injustiça em relação a um ou outro grupo. Estando ciente que há interferência dos viéses nas análises, podemos mitigá-los utilizando alguns métodos de pré-processamento, in-processamento e pós-processamento, que serão discutidos nos próximos tópicos.

[1]https://www.into.saude.gov.br/

1) Disparate Impact: Disparate Impact é uma métrica para avaliar a imparcialidade, no qual compara a proporção de indivíduos que recebem uma saída positiva para dois grupos: um grupo não privilegiado e um grupo privilegiado. Essa métrica precisa ter como resultado 1 para ser considerada justa [7].

$$DI = \frac{P(Y = 1|A = minoria)}{P(Y = 1|A = privilegiado)} \tag{1}$$

No qual, Y são as predições do modelo e A é o grupo do qual o atributo sensível pertence.

O cálculo é a proporção do grupo não privilegiado que recebeu o resultado positivo dividida pela proporção do grupo privilegiado que recebeu o resultado positivo.

O padrão da indústria é uma regra de quatro quintos: se o grupo não privilegiado receber um resultado positivo inferior a 80% de sua proporção do grupo privilegiado, isso é uma violação de impacto díspar. No entanto, isso é definido de acordo com o problema estudado.

2) Average Absolute Odds Difference: Esta é a média da diferença em falsos positivos e verdadeiros positivos entre grupos não privilegiados e privilegiados. Um valor de 0 implica que ambos os grupos têm o mesmo benefício, ou seja, é justa [7].

$$\begin{aligned} AAOD = \frac{1}{2}[&|FPR_{A=minoria} - FPR_{A=privilegiado}| \\ &+|TPR_{A=minoria} - TPR_{A=privilegiado}|] \end{aligned} \tag{2}$$

No qual, Y são as predições do modelo e A é o grupo do qual o atributo sensível pertence.

C. Pré-processamento

Na etapa de pré-processamento do pipeline de modelagem, segundo [7], ainda não existe um modelo treinado. Sendo assim, os métodos de pré-processamento não podem incluir explicitamente métricas de imparcialidade que envolvam previsões de modelo. Portanto, a maioria dos métodos de pré-processamento está focada na visão de mundo "somos todos iguais", mas não exclusivamente. Existem várias maneiras de pré-processar um conjunto de dados de treinamento: (1) aumentar o conjunto de dados com dados adicionais, (2) aplicar pesos de instância aos dados e (3) alterar os rótulos.

1) Data Augmentation: Um dos algoritmos mais simples para pré-processamento do conjunto de dados de treinamento é inserir registros inicialmente inexistentes. Esses registros são construídos a partir dos registros já existentes, e trocando valores de atributos protegidos (como imparcialidade contrafactual) [8]. Optamos pela utilização desse método pois os dados da base são desbalanceados, como podemos observar

pela Figura 1. No entanto, se não houver dados redundantes, é mais comum a sobreamostragem de grupos minoritários. Algoritmos populares, como a técnica de sobreamostragem de minorias sintéticas (SMOTE) ou as suas variações, tais como SMOTE-ENC, Borderline-SMOTE conforme encontramos em [9].

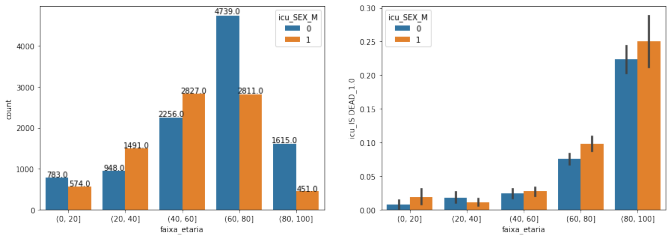

Figura 1. Distribuição do dataset levando em conta (1) quantidade de pessoas por sexo e faixa etária, (2) quantidade de mortes por sexo e faixa etária

2) Ponderação de instâncias: Outra maneira de pré-processar o conjunto de dados de treinamento é por meio de ponderações de amostra, semelhantes à ponderação de probabilidade inversa e ponderação de importância. O método de repesagem é voltado para melhorar a paridade estatística (visão de mundo "somos todos iguais"), que pode ser avaliada antes que o modelo de gerenciamento de cuidados seja treinado e é uma métrica de imparcialidade do conjunto de dados conforme detalha [10].

Os atributos detectados são necessários nos dados de treinamento para o modelo, mas não precisam fazer parte do modelo ou dos dados no deploy. O aumento e a repesagem dos dados não alteram os dados de treinamento que você tem das decisões históricas de gerenciamento de cuidados.

D. Processamento

Os algoritmos de mitigação de viés no processamento utiliza uma restrição ou um termo de regularização ao objetivo de otimização existente usando uma métrica de imparcialidade [7].

1) Prejudice Remover: O preconceito refere-se ao fato de que existe dependência estatística entre o atributo protegido e o resultado previsto ou outras variáveis independentes. Visa a aprendizagem de um preditor cujas previsões são independentes do atributo protegido [11] .

2) Adversarial Debiasing: A aprendizagem adversarial é um paradigma de aprendizagem originalmente concebido para gerar amostras falsas para confundir o modelo. Geralmente, existe um gerador que garante as amostras falsas geradas que estão próximas das amostras reais, e um discriminador para discriminar as amostras falsas das amostras reais. O objetivo da aprendizagem adversarial é aprender um gerador para gerar amostras que o discriminador não pode realmente dizer que falsificaram ou não [12].

E. Pós-processamento

Os métodos de mitigação de viés pós-processamento tentam realizar modificações posteriores de forma a satisfazer as restrições de imparcialidade. Só é possível alterar as previsões de saída para atender às métricas de equidade do grupo desejado com base em uma visão de mundo (ou seja, inverter os rótulos previstos de positivos para negativos ou vice-versa). A ideia básica é encontrar um limite adequado usando a função de pontuação original para cada grupo. Nenhum retreinamento/mudança é necessário para o classificador, segundo [7].

1) Equalized Odds Postprocessing - Threshold Optimizer: O Threshold Optimizer é um algoritmo de pós-processamento baseado no artigo "Equality of Opportunity in Supervised Learning" [13]. Essa técnica é implementada pela biblioteca Fairlearn e usa como entrada um classificador existente e a feature sensível e deriva uma transformação monótona da previsão do classificador para impor as restrições de paridade especificadas.

O classificador é obtido pela aplicação de limiares específicos do grupo ao estimador fornecido. Os limites são escolhidos para otimizar o objetivo de desempenho fornecido, sujeito às restrições de imparcialidade fornecidas.

2) Reject Option Classification: A classificação de Reject Option Classification (opção de rejeição de classificação) é uma técnica de pós-processamento que fornece resultados favoráveis a grupos não privilegiados e resultados desfavoráveis a grupos privilegiados em uma faixa de confiança em torno do limite de decisão com a maior incerteza. Assim, explorando a região de baixa confiança de um classificador para redução de discriminação e rejeitando suas previsões, podemos reduzir o viés nas previsões do modelo [14].

F. Privacidade diferencial

Conforme [15], privacidade diferencial é uma definição de privacidade recente que garante que o resultado de um modelo seja insensível a qualquer registro específico no conjunto de dados ou mesmo atributos sensíveis.

Um algoritmo é considerado diferencialmente privado se, ao olhar para a saída, não se pode dizer se os dados de qualquer indivíduo foram incluídos no conjunto de dados original ou não. Em outras palavras, a garantia de um algoritmo diferencialmente privado é que seu comportamento dificilmente muda quando um único indivíduo entra ou sai do conjunto de dados. Qualquer coisa que o algoritmo possa produzir em um banco de dados contendo informações de algum indivíduo tem quase a mesma probabilidade de ter vindo de um banco de dados sem as informações desse indivíduo e esta garantia vale para qualquer indivíduo e qualquer conjunto de dados. Desta forma, independentemente de quão diferentes sejam os detalhes de qualquer indivíduo e independentemente dos detalhes de qualquer outra pessoa no banco de dados, a garantia de privacidade diferencial ainda é válida. Isso dá uma garantia formal de que as informações de nível individual sobre os participantes no banco de dados não vazam.

Modelo	Etapa	disparate_impact	average_abs_odds_difference	Acurácia	Recall	Matriz de Confusão
Regressão Logística	baseline sem mitigação	0.7522	0.0401	0.93	0.32	[3668, 57] [213, 102]
Regressão Logística + RandomUnderSample	pré-processamento	0.7522	0.1013	0.83	0.86	[266, 63] [44, 271]
Regressão Logística + SMOTE	pré-processamento	0.6293	0.0421	0.93	0.91	[3526, 215] [323, 3372]
Regressão Logística + Ponderação de instâncias	pré-processamento	0.9440	0.0098	0.93	0.29	[3674, 51] [223, 92]
Prejudice Remover	in-processamento	0.9549	0.0117	0.93	0.25	[4414, 65] [275, 94]
AdversarialDebiasing	in-processamento	0.6327	0.0062	0.93	0.08	[4471, 8] [341, 28]
EqOddsPostprocessing	pós-processamento	0.9955	0.0052	0.93	0.28	[3666, 59] [228, 87]
RejectOptionClassification	pós-processamento	0.9437	0.0305	0.84	0.84	[3111, 614] [50, 265]

Tabela I
RESULTADOS OBTIDOS UTILIZANDO ABORDAGENS DE MITIGAÇÃO NO PRÉ-PROCESSAMENTO, IN-PROCESSAMENTO E PÓS-PROCESSAMENTO

IV. RESULTADOS E DISCUSSÕES

Neste trabalho buscamos aplicar métodos de mitigação de viéses sendo eles aplicados as etapas de pré-processamento, in-processamento e pós-processamento.

Na tabela I são apresentados os resultados levando em conta tanto métricas de avaliação dos modelos, quanto as métricas de fairness. Portanto, avaliaremos tanto a performance quanto a justiça presente na classificação.

Todos os métodos foram treinados com 70% da base, e testados 30%. Além disso, o modelo de baseline e todos os modelos com mitigação de viéses na etapa de pré-processamento utilizaram cross-validation para a escolha do melhor modelo, no qual a métrica responsável pela escolha foi o fbeta_score com o beta=2, para melhorar o recall. Por outro lado, para a implementação dos outros modelos foi utilizado a biblioteca AIF360.

O melhor modelo encontrado é o Reject Option Classification, no qual utiliza o método de pós-processamento. Além de manter a acurácia em 84%, o que é consideravelmente alta, manteve o recall em 84%. Isso é bem interessante pois nesse problema os falsos negativos tem muito mais impactos do que o oposto.

De mesmo modo, manteve o average abs odds difference próximo de 0, e o disparate impact próximo de 1, o que é considerável justo nos dois casos.

Nos casos dos modelos: baseline, Poderação de instâncias, Prejudice Remover, Adversarial Debiasing e EqOddsPostProcessing. Apesar de o modelo apresentar acurácia alta, obtiveram o Recall muito baixo, abaixo de 50%. E portanto, podem ser prejudiciais à classificação.

Para os modelos com mudança de amostras como a aplicação de RandomUnderSample e SMOTE, podemos notar que apesar de o SMOTE apresentar uma melhora significativa nas métricas de desempenho do modelo (83% contra 93%), obteve a métrica dispate_impact um pouco pior. O viés possivelmente é causado pela repetição das classes não privilegiadas.

Entretanto, não é possível afirmar que haja um problema de justiça nesses dados devido a natureza do problema estudado, as complicações resultantes que levaram pacientes à óbito podem estar relacionadas com estilo de vida do paciente, comorbidades, idade, resistência física do paciente entre outros.

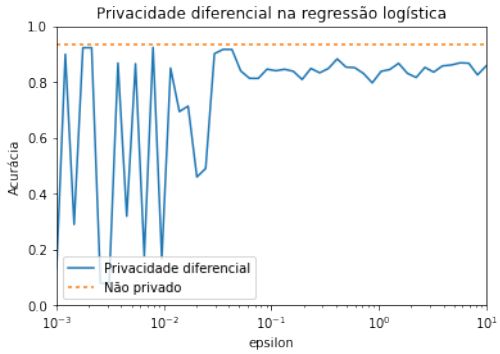

Figura 2. Distribuição do dataset levando em conta (1) quantidade de pessoas por sexo e faixa etária, (2) quantidade de mortes por sexo e faixa etária

Com relação à privacidade diferencial, realizando o treinamento com a regressão linear sem privacidade diferencial versus com privacidade diferencial, podemos visualizar o trade-off entre acurácia e o ϵ para privacidade diferencial conforme a Figura 2 considerando [16].

REFERÊNCIAS

[1] C. M. Patino and J. C. Ferreira, "Prognostic studies for health care decision making," *Jornal Brasileiro de Pneumologia*, vol. 43, pp. 252–252, Aug. 2017.

[2] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[3] "Prediction models for 30-day mortality and complications after total knee and hip arthroplasties for veteran health administration patients with osteoarthritis," *The Journal of Arthroplasty*, vol. 33, no. 5, pp. 1539–1545, 2018.

[4] S. Pfohl, B. J. Marafino, A. Coulet, F. Rodriguez, L. Palaniappan, and N. H. Shah, "Creating fair models of atherosclerotic cardiovascular disease risk," *CoRR*, vol. abs/1809.04663, 2018.

[5] Y. Park, J. Hu, M. Singh, I. Sylla, I. Dankwa-Mullan, E. Koski, and A. K. Das, "Comparison of Methods to Reduce Bias From Clinical Prediction Models of Postpartum Depression," *JAMA Network Open*, vol. 4, pp. e213909–e213909, 04 2021.

[6] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini, "A clarification of the nuances in the fairness metrics landscape," *Scientific Reports*, vol. 12, Mar. 2022.

[7] K. R. Varshney, *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published, 2022.

[8] S. Sharma, Y. Zhang, J. M. Ríos Aliaga, D. Bouneffouf, V. Muthusamy, and K. R. Varshney, "Data augmentation for discrimination prevention and bias disambiguation," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, (New York, NY, USA), p. 358–364, Association for Computing Machinery, 2020.

[9] "SMOTE: Synthetic data augmentation for tabular data." https://towardsdatascience.com/smote-synthetic-data-augmentation-for-tabular-data-1ce28090debc. Accessed: 2022-11-27.

[10] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[11] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Machine Learning and Knowledge Discovery in Databases* (P. A. Flach, T. De Bie, and N. Cristianini, eds.), (Berlin, Heidelberg), pp. 35–50, Springer Berlin Heidelberg, 2012.

[12] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," *CoRR*, vol. abs/1801.07593, 2018.

[13] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *CoRR*, vol. abs/1610.02413, 2016.

[14] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929, 2012.

[15] A. Friedman and A. Schuster, "Data mining with differential privacy," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, (New York, NY, USA), p. 493–502, Association for Computing Machinery, 2010.

[16] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization.," *Journal of Machine Learning Research*, vol. 12, no. 3, 2011.