Mitigação de Disparidades de Gênero em Modelos de Regressão de Redes Neurais Multitarefas Aplicado a Doença de Parkinson

Bruno Pires M. Silva, Lilian Berton, Luiz Leduino de Salles Neto

[1]Instituto de Ciência e Tecnologia (ICT)
Universidade Federal de São Paulo (UNIFESP).
São José dos Campos – SP – Brasil

bruno.pires22@unifesp.br, lberton@unifesp.br, luiz.leduino@unifesp.br

Abstract. *Fairness in predictive models is crucial, especially in sensitive contexts such as healthcare, where disparities can result in severe consequences. This work proposes a solution to mitigate such disparities in regression problems. The approach employs a modified multitask neural network, in which modifications are made to the loss function to reduce the disparity in errors between genders. The model was applied to a dataset of patients with Parkinson's disease to predict a score on disease progression. Preliminary results show the ability to predict gender and satisfactory performance in the regression task.*

Resumo. *Fairness em modelos preditivos é um aspecto crucial, especialmente em contextos sensíveis como a saúde, onde disparidades podem resultar em consequências graves. Este trabalho propõe uma solução para mitigar tais disparidades em problemas de regressão. A abordagem emprega uma rede neural multitarefa modificada, na qual são realizadas modificações na função de perda para redução da disparidade dos erros entre os sexos. O modelo foi aplicado num conjunto de dados de pacientes com Parkinson para prever um escore sobre a progressão da doença. Os resultados preliminares mostram a capacidade de predição do sexo e um desempenho satisfatório na tarefa de regressão.*

1. Introdução

O uso de inteligência artificial (IA) na área da saúde tem impulsionado os avanços no diagnóstico e monitoramento de doenças, permitindo que modelos preditivos auxiliem médicos na tomada de decisões e na personalização de tratamentos [Sun et al. 2025, Gulshan et al. 2016]. Redes neurais profundas, por exemplo, têm demonstrado desempenho semelhante ao de especialistas humanos em tarefas como a detecção de doenças em imagens médicas [Gulshan et al. 2016]. Além disso, técnicas baseadas em aprendizado de máquina são cada vez mais empregadas para prever a progressão de doenças crônicas, como Parkinson e Alzheimer, permitindo um acompanhamento mais próximo e preciso dos pacientes [Yu and Luo 2024].

Entretanto, estes modelos podem apresentar viés, resultando em previsões que desfavorecem certos grupos demográficos [Cozman and Kaufman 2022]. Estudos mostram que modelos treinados em dados clínicos podem exibir disparidades em métricas de erro quando avaliados em populações de diferentes sexos, idades ou grupos raciais [Obermeyer et al. 2019, Pfohl et al. 2021]. Em particular, há evidências

de que algoritmos de predição de risco em hospitais frequentemente subestimam a gravidade da condição de pacientes negros, devido a vieses históricos nos dados [Obermeyer et al. 2019]. Essas desigualdades são amplamente estudadas em tarefas de classificação, onde diversas abordagens foram propostas para mitigar viés, como por exemplo, técnicas de otimização e aprendizado adversarial [Zemel et al. 2013, Zhang et al. 2018].

Fairness refere-se ao desenvolvimento de sistemas de IA que tratam todos os grupos de maneira equitativa, sem favorecer ou prejudicar indevidamente qualquer grupo específico baseado em características sensíveis como raça, gênero, idade, etnia, entre outras [Barocas et al. 2018]. No entanto, *fairness* em problemas de regressão ainda é uma área menos explorada [Gursoy and Kakadiaris 2022]. Enquanto *fairness* em classificação geralmente se baseia em métricas como *disparate impact* e *equalized odds* [Pessach and Shmueli 2022] que podem ser associadas de forma mais fácil a conceitos de justiça como equidade individual, equidade coletiva, entre outros. Porém, um dos desafios da regressão é estabelecer esses paralelos e introduzir mecanismos em modelos já existentes que promovam uniformidade entre grupos sensíveis durante as regressões.

Neste trabalho, investigamos a aplicação de fairness em modelos de regressão voltados à predição da condição clínica de pacientes com Parkinson, utilizando medições biomédicas da voz. Consideramos o sexo como atributo sensível e propomos uma rede neural multitarefa, na qual o modelo realiza simultaneamente a regressão do escore clínico e a classificação do sexo. Inspirada em métodos adversariais de classificação [Zhang et al. 2018], essa abordagem busca reduzir disparidades nos erros entre grupos sensíveis ao incorporar a equidade como parte da função de perda em problemas de regressão. Além de propor um novo método, analisamos como a introdução de fairness afeta o desempenho preditivo do modelo, explorando o trade-off entre precisão e equidade por meio da variação de um hiperparâmetro na função de perda. Nosso objetivo é contribuir para o desenvolvimento de soluções mais justas em aprendizado de máquina na saúde.

2. Materiais e métodos

2.1. Conjunto de dados e tratamentos

O estudo utiliza o conjunto de dados *Parkinsons Telemonitoring*, disponibilizado pelo *UCI Machine Learning Repository* [Tsanas and Little 2009]. O conjunto de dados contém medições de pacientes com Parkinson obtidas por sensores de voz, com o atributo alvo sendo o grau de severidade da doença, descrito pelos UPDRS (*Unified Parkinson's Disease Rating Scale*), incluindo o motor UPDRS, que avalia especificamente os sintomas motores da doença, e o total UPDRS, que engloba também funções não motoras, como capacidade cognitiva, alterações no humor e atividades diárias. As principais variáveis consideradas para o treinamento incluem medições biomédicas e o sexo dos pacientes, sendo este último utilizado para a análise de *fairness*. O conjunto de dados contém 5.875 exames de 42 pessoas, dos quais 28 são mulheres e 14 são homens, resultando em 4.008 exames de indivíduos do sexo feminino e 1.867 do masculino. Informações sobre a variável alvo pode ser vista na tabela 1.

Os dados foram divididos em conjuntos de treinamento (80%) e teste (20%), conforme práticas padrão em modelos de aprendizado de máquina. A amostragem estrati-

Tabela 1. Resumo estatístico das variáveis motor_UPDRS e total_UPDRS

Estatística	motor_UPDRS	total_UPDRS
mean	21.296229	29.018942
std	8.129282	10.700283
min	5.037700	7.000000
25%	15.000000	21.371000
50%	20.871000	27.576000
75%	27.596500	36.399000
max	39.511000	54.992000

ficada foi utilizada para preservar a proporção de indivíduos de cada grupo de sexo em ambos os conjuntos, garantindo que a distribuição da variável categórica de interesse fosse semelhante. Essa abordagem é amplamente recomendada para reduzir viés na avaliação do modelo e assegurar que os resultados sejam generalizáveis. Os dados de treino foram submetidos a um processo de normalização z-score, no qual cada variável x é transformada segundo a equação:

$$x' = \frac{x - \mu}{\sigma},$$

onde μ e σ representam a média e o desvio padrão da variável, respectivamente. Esse procedimento garante que todas as *features* tenham média zero e variância unitária, facilitando o treinamento do modelo.

2.2. Arquitetura e modificações do modelo

A arquitetura proposta é baseada em uma rede neural do tipo *Multilayer Perceptron* (MLP), adaptada para executar simultaneamente duas tarefas: regressão e classificação. A tarefa de regressão visa estimar o escore motor da doença de Parkinson (*motor_UPDRS*), enquanto a tarefa de classificação tem como objetivo determinar o sexo do paciente. Essa abordagem multitarefa permite que o modelo compartilhe representações aprendidas entre as duas tarefas, o que pode melhorar a capacidade de generalização e reduzir o viés em relação a subgrupos específicos.

A rede é composta por duas camadas densas, com 64 e 32 neurônios, seguidas por duas saídas independentes: uma para regressão e outra para classificação. As camadas ocultas utilizam a função de ativação *ReLU* para introduzir não linearidade. A saída da regressão é uma única variável contínua, enquanto a saída da classificação consiste em uma camada densa com duas unidades e a função *softmax*, que estima a probabilidade de cada classe (feminino ou masculino). Como forma de promover *fairness* em relação ao atributo sensível, propomos uma função de perda combinada que pondera as perdas de regressão e classificação, definida da seguinte maneira:

$$Loss = \alpha \cdot Loss_{regressão} + (1 - \alpha) \cdot Loss_{classificação},$$

onde α é um hiperparâmetro que controla a contribuição relativa de cada tarefa. A perda da regressão é calculada utilizando o erro quadrático médio (MSE), enquanto a perda de classificação é calculada utilizando a entropia cruzada (*cross-entropy*). A escolha do

α permite ajustar o equilíbrio entre a precisão da regressão e a uma classificação mais equânime, explorando diferentes valores para minimizar disparidades entre grupos. Inicialmente, foram testados valores de α no intervalo de 0.1 a 0.9, com o objetivo de explorar o impacto da ponderação entre as tarefas de regressão e classificação. Essa abordagem manual forneceu uma visão preliminar sobre como α afeta o desempenho do modelo e a disparidade entre grupos. No entanto, para determinar o valor ótimo de α de maneira mais precisa e eficiente, pretende-se utilizar técnicas especializadas, como *grid search*, métodos de otimização como *simulated annealing* ou abordagens baseadas em gradiente.

O modelo foi treinado utilizando o otimizador *Adam*, com uma taxa de aprendizado fixa de 0.01 e um total de 50 épocas. Durante o treinamento, a função de perda combinada foi minimizada, integrando as perdas das tarefas de regressão e classificação. A avaliação do desempenho foi realizada no conjunto de teste, com métricas específicas para cada tarefa. Para a regressão, foram calculados o erro quadrático médio (MSE) e o erro absoluto médio (MAE) de cada sexo, permitindo a análise de disparidades entre os grupos. Já para a classificação, a acurácia foi utilizada como métrica principal, uma vez que o foco não era otimizar diretamente essa tarefa, mas sim utilizá-la como um mecanismo de interação com a tarefa principal, seguindo a lógica de algoritmos que implementam aprendizado adversário para promover justiça algorítmica em classificações.

3. Resultados parciais

Os resultados obtidos até o momento evidenciam a influência do parâmetro α na relação entre precisão preditiva e equidade do modelo de regressão. Para contextualizar os resultados, comparamos o modelo multitarefa com um modelo simples de regressão, que possui a mesma estrutura, mas é treinado apenas para prever a pontuação motora de Parkinson. No modelo simples, o MSE foi de 9.3082 para o grupo feminino e 10.0149 para o grupo masculino, com uma disparidade de -0.7066. Já o MAE foi de 2.2147 para o grupo feminino e 2.2914 para o grupo masculino, resultando em uma disparidade de -0.0766. Esses valores fornecem uma linha de base para avaliar o impacto da abordagem multitarefa na equidade e no desempenho do modelo.

De acordo com a Figura 1, observa-se que, para valores menores de α (0.1 a 0.3), a disparidade no erro quadrático médio (MSE) entre os grupos apresenta valores negativos, indicando um favorecimento ao grupo feminino. No entanto, conforme aumenta, a disparidade oscila, atingindo valores próximos de zero em $\alpha = 0.5$, antes de diminuir novamente para valores negativos em $\alpha = 0.7$ e retornar a zero em $\alpha = 0.9$.

Essa variação sugere que não há uma tendência clara de convergência para um ponto de equilíbrio estável, mas sim um comportamento oscilatório. Já a disparidade no erro absoluto médio (MAE) permanece próxima de zero ao longo de todos os valores de α, indicando que essa métrica é menos sensível à variação do hiperparâmetro em comparação com o MSE.

Quanto aos resultados da tarefa de classificação, sua função principal é auxiliar na incorporação de aspectos de equidade (fairness) à tarefa de regressão. Embora faça parte da arquitetura multitarefa, essa tarefa não é um objetivo final do modelo, sobretudo porque atributos sensíveis como o sexo são geralmente conhecidos em contextos práticos. Por esse motivo, seus resultados não foram analisados neste trabalho. No entanto, reconhecemos que essa análise pode representar uma linha interessante para investigações

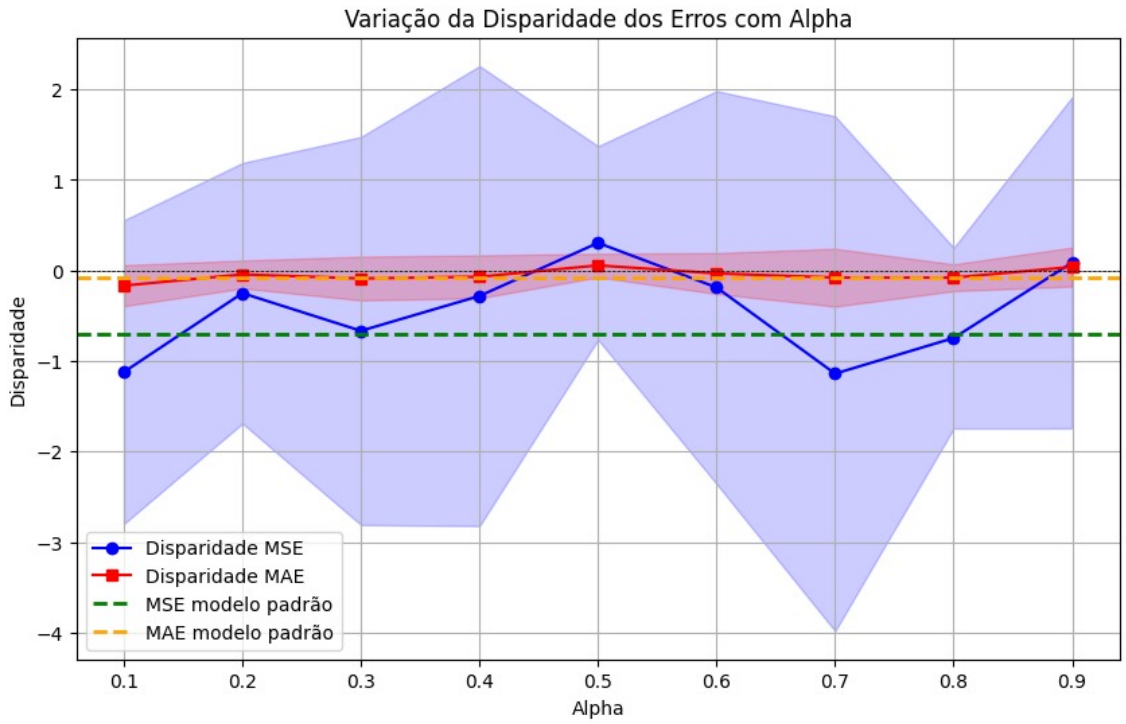

Figura 1. Variação da disparidade dos erros médios (MSE e MAE) entre os grupos sensíveis (sexo masculino e feminino) em função de α.

futuras.

4. Discussão e perspectivas futuras

Os resultados parciais obtidos neste trabalho demonstram que a escolha do hiperparâmetro α tem um impacto significativo no desempenho e na equidade do modelo multitarefa. No entanto, a abordagem atual de testar diferentes valores manualmente serve apenas como uma análise exploratória para ilustrar os efeitos dessa variação, não sendo uma estratégia escalável ou eficiente para encontrar o valor ótimo. Como próxima etapa, pretende-se implementar métodos de otimização, como o *simulated annealing*, *grid search* ou métodos baseados em gradiente, para automatizar a busca pelo α ideal.

Além disso, está em andamento a investigação de uma abordagem baseada em Minimax para a função de perda. Essa estratégia visa minimizar o pior caso de disparidade entre os grupos, garantindo que o modelo não favoreça excessivamente nenhum subgrupo em detrimento de outro. A aplicação dessa abordagem pode ser particularmente útil em contextos onde se tem como objetivo melhorar tipos específicos de justiça, como por exemplo equidade individual. Em contextos clínicos, este método pode ser justificado pelas consequências que podem ser geradas a partir de uma única classificação incorreta. Outra direção futura é validar o modelo em conjuntos de dados com características distintas, como distribuições desbalanceadas ou variáveis sensíveis adicionais, a fim de avaliar sua capacidade de generalização e adaptabilidade a diferentes contextos.

A aplicação de inteligência artificial na saúde tem potencial para transformar diagnósticos, tratamentos e a alocação de recursos. No entanto, há uma lacuna na literatura quanto à garantia de fairness em modelos de regressão, especialmente em contextos

clínicos, onde a equidade é essencial. Este trabalho busca contribuir nesse sentido, investigando estratégias para promover justiça algorítmica e assegurar que pacientes de diferentes grupos recebam previsões precisas e imparciais. A adoção de técnicas como a otimização de hiperparâmetros e a validação em múltiplos conjuntos de dados visa fortalecer a construção de sistemas de IA mais confiáveis, éticos e inclusivos.

5. Agradecimentos

Agradecemos o apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP (Grant 2021/14725-3) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Referências

Barocas, S., Hardt, M., and Narayanan, A. (2018). Fairness and machine learning. fairmlbook. org, 2019.

Cozman, F. G. and Kaufman, D. (2022). Viés no aprendizado de máquina em sistemas de inteligência artificial: a diversidade de origens e os caminhos de mitigação. *Revista USP*, (135):195–210.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410.

Gursoy, F. and Kakadiaris, I. A. (2022). Error parity fairness: Testing for group fairness in regression tasks. *arXiv preprint arXiv:2208.08279*.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Pessach, D. and Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44.

Pfohl, S. R., Foryciarz, A., and Shah, N. H. (2021). An empirical characterization of fair machine learning for clinical risk prediction. *Journal of biomedical informatics*, 113:103621.

Sun, Q., Akman, A., and Schuller, B. W. (2025). Explainable artificial intelligence for medical applications: A review. *ACM Trans. Comput. Healthcare*, 6(2).

Tsanas, A. and Little, M. (2009). Parkinsons Telemonitoring. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5ZS3N.

Yu, K. and Luo, X. (2024). Intelligent diagnosis and progression analysis of alzheimer's disease using machine learning. In *Proceedings of the 2024 5th International Symposium on Artificial Intelligence for Medicine Science*, pages 66–72.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.