Comparação entre aplicação de redes neurais e regressão logística na predição de diabetes utilizando gradiente descendente

Ester Miriã de Sousa Freire[1], Paulo Anderson Gonçalves de Lima[1], Ângelo Carlos Avelino Feitosa[1], Lucas Fontineli Alves Sousa[1], Simone Azevedo Bandeira de Melo Aquino[2], Daniel Duarte Costa[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Maranhão – Campus Imperatriz (IFMA)

CEP 65.906-335 – Imperatriz – MA – Brasil

[2]Departamento de Ensino Superior e Tecnologia - Instituto Federal do Maranhão, IFMA

[2]UFMA: Centro de Ciências de Imperatriz - Universidade Federal do Maranhão.

{emiria, paulo.anderson, angelo.a, sousafontineli}@acad.ifma.edu.br, simonebandeira@ifma.edu.br, daniel.dc@ufma.br

Abstract. *This article presents a manual implementation of a neural network optimized with gradient descent, applied to predicting the occurrence of diabetes in Pima women using the Pima Indians Diabetes Database. The study covers preprocessing, training, and evaluation with accuracy, confusion matrix, and ROC curve. The results show that the neural network achieved slightly better performance compared to logistic regression, making it ideal for more complex situations.*

Resumo. *Este artigo apresenta uma implementação manual de uma rede neural com otimização por gradiente descendente, aplicado à previsão da ocorrência de diabetes em mulheres da etnia Pima, a partir do conjunto de dados Pima Indians Diabetes Database. O estudo abrange pré-processamento, treinamento e avaliação com acurácia, matriz de confusão e curva ROC. Os resultados evidenciam que a rede neural obteve um desempenho levemente superior em relação à regressão logística, sendo ideal para situações de maior complexidade.*

1. Introdução

A prevenção e o diagnóstico precoce do diabetes, especialmente do tipo 2, têm ganhado destaque pelo impacto na saúde pública, permitindo reduzir complicações e possibilitar tratamento inicial [Antunes et al, 2021]. Nesse cenário, algoritmos de aprendizado de máquina baseados em variáveis clínicas e biomédicas, como redes neurais, despontam como alternativa promissora.

Com o avanço computacional, pesquisas têm explorado novas aplicações, inclusive industriais, em que as redes neurais artificiais (RNA) se mostram altamente atrativas [Fleck et al, 2016]. Este trabalho tem como objetivo implementar uma RNA para prever diabetes com

fins educacionais, demonstrando o fluxo de *machine learning* desde a implementação até a avaliação, sem uso de bibliotecas prontas.

A base de dados utilizada foi o *Pima Indians Diabetes Database* (Kaggle), com informações médicas de 768 mulheres da etnia Pima, incluindo variáveis como glicose, pressão arterial, gestações e idade [Mouza et al, 2023]. O artigo está organizado da seguinte forma: a seção 2 descreve a metodologia, a seção 3 apresenta os resultados e a seção 4 detalha considerações finais.

2. Metodologia

A etapa de preparação da base de dados é fundamental para a análise, pois envolve a exploração e a sumarização das informações, possibilitando a compreensão de suas características, propriedades e estrutura [Aithal, 2023]. Assim, o conteúdo *Pima Indians Diabetes Database*, contido no repositório da base de dados *Kaggle* foi utilizado para a modelagem desta rede neural. O *dataset* contido possui 768 amostras com oito atributos: número de gestações, dosagens de glicose sanguínea, pressão arterial, espessura da pele, valores séricos de insulina, índice de massa corporal (IMC), função hereditária (Diabetes Pedigree Function) e idade. Já o alvo (label) pode ser binário: 0 (não diabético) ou 1 (diabético).

O desenvolvimento do código envolveu cinco etapas: análise exploratória, tratamento de valores ausentes, normalização com z-score, divisão em treino e teste, e implementação e avaliação.

Valores nulos foram substituídos pela média de cada atributo, mantendo a coerência da base. Após a normalização (média 0 e desvio padrão 1), o conjunto foi dividido em 70% para treino e 30% para teste, garantindo equilíbrio entre amostras de aprendizado e validação.

3. Resultados

A acurácia obtida foi de aproximadamente 78,3%, indicando um bom desempenho geral do modelo e superando a regressão logística, que alcançou cerca de 70%. Esse resultado demonstra que, dentre as previsões realizadas pela rede neural, a maioria correspondeu corretamente aos casos reais de diabetes, evidenciando uma capacidade satisfatória de generalização.

A matriz de confusão (Figura 1) mostrou que o modelo com rede neural apresentou uma taxa menor de erros em comparação à regressão logística, com cinco falsos negativos a menos e dois falsos positivos a menos. Isso indica que o modelo foi mais eficaz tanto em identificar corretamente os casos positivos quanto em reduzir as classificações incorretas.

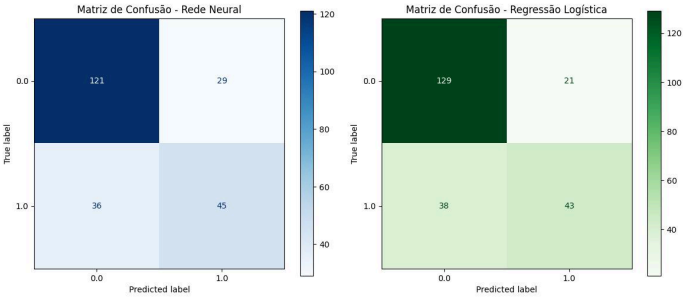

Figura 1. Comparativo entre as matrizes de confusão da regressão logística e da rede neural.

O F1-score obtido foi de aproximadamente 58%, demonstrando um equilíbrio razoável entre precisão (proporção de casos previstos como positivos que realmente eram positivos) e revocação (proporção de casos positivos corretamente identificados). Esse resultado reflete uma melhora na detecção de casos reais de diabetes, atingindo valores de 63% de precisão e 71,7% de revocação, respectivamente.

A curva ROC (Figura 2) apresentou uma área sob a curva (AUC) de 0,80, o que indica uma boa capacidade discriminativa do modelo, ainda que levemente inferior à da regressão logística. Em outras palavras, a rede neural consegue distinguir de forma satisfatória entre indivíduos com e sem diabetes, reforçando a robustez de sua generalização.

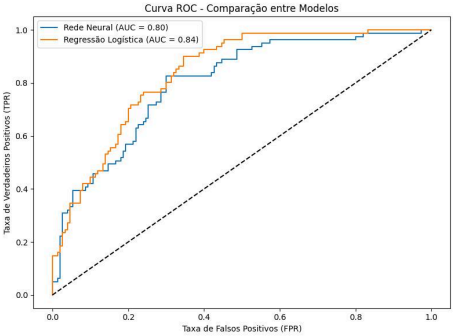

Figura 2. Curva ROC da rede neural.

A curva de aprendizado (Figura 3) evidenciou estabilidade na acurácia de treino e pequenas variações na validação, o que sugere que o modelo não sofreu com sobreajuste significativo. Já a curva log-loss (Figura 4), que representa a convergência do gradiente descendente, demonstrou uma boa redução da perda ao longo das iterações, confirmando o sucesso do processo de otimização.

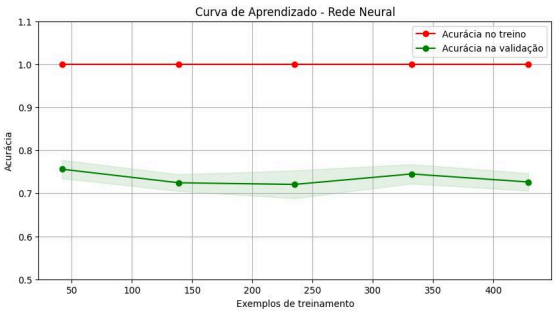

Figura 3. Curva de aprendizado da rede neural.

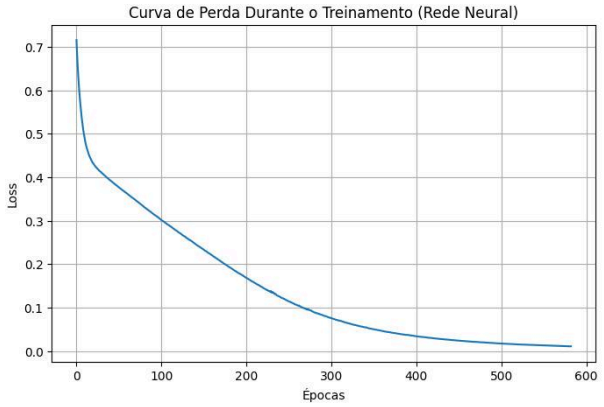

Figura 4. Curva log-loss.

De modo geral, a rede neural mostrou-se mais sensível à detecção de casos positivos, o que é desejável em cenários clínicos. Embora a diferença em relação à regressão logística seja pequena, o modelo neural oferece maior flexibilidade e tende a se destacar em problemas de maior complexidade, como os que envolvem imagens ou séries temporais.

4. Considerações finais

Este trabalho teve como objetivo apresentar a implementação de um modelo de rede neural com gradiente descendente para predição de diabetes. Os resultados demonstraram uma eficácia da abordagem levemente superior à regressão logística, servindo também como exercício prático de compreensão de conceitos fundamentais, como o aprendizado supervisionado, otimização e avaliação do modelo.

Referências

Antunes, Y. R., de Oliveira, E. M., Pereira, L. A., & Picanço, M. F. P. (2021). Diabetes Mellitus Tipo 2: A importância do diagnóstico precoce da diabetes Type 2 Diabetes Mellitus: The importance of early diabetes diagnosis. *Brazilian Journal of Development*, *7*(12), 116526-116551.

Fleck, L., Tavares, M. H. F., Eyng, E., Helmann, A. C., & Andrade, M. A. D. M. (2016). Redes neurais artificiais: Princípios básicos. *Revista Eletrônica Científica Inovação e Tecnologia*, *1*(13), 47-57.

MOUSA, A., MUSTAFA, W., & MARQAS, R. B. (2023). A comparative study of diabetes detection using the Pima Indian diabetes database. *methods*, *7*, 8.

Tariq, D. T. H. S., & Aithal, P. S. (2023). Visualization and explorative data analysis. *International Journal of Enhanced Research in Science, Technology & Engineering*, *12*(3), 11-21.