Previsão de Desfechos Clínicos em Pacientes com Tuberculose Usando Redes Neurais Perceptron Multicamadas: Análise Interativa de Dados e Visualizações

Ronilson W. S. Pereira[1], **Marcos Seruffo**[2], **Karla Figueiredo**[1]

[1]Instituto de Matemática e Estatística – Universidade do Estado do Rio de Janeiro
Rio de Janeiro – RJ – Brasil

[2]Instituto de Tecnologia – Universidade Federal do Pará
Belém – PA – Brasil

ronilsonengenharia@gmail.com, karlafigueiredo@ime.uerj.br,

seruffo@ufpa.br

Abstract. *Tuberculosis is one of the world's leading infectious diseases, and its treatment requires effective data analysis. This study uses Multilayer Perceptron (MLP) Neural Networks and Interactive Visualizations to predict clinical outcomes in tuberculosis patients. The aim is to improve understanding and clinical decision-making through predictive analysis and interactive visualizations. Normalization and balancing techniques were applied to train the MLP model. Interactive tools were used to display data distributions, performance metrics and confusion matrices. The results show that the combination of MLP with interactive visualizations is effective for interpreting clinical outcomes and assisting in treatment planning.*

Resumo. *A tuberculose é uma das principais doenças infecciosas do mundo, e seu tratamento exige análise eficaz de dados. Este estudo usa Redes Neurais Perceptron Multicamadas (MLP) e Visualizações Interativas para prever desfechos clínicos em pacientes com tuberculose. O objetivo é melhorar a compreensão e a tomada de decisões clínicas por meio de análises preditivas e visualizações interativas. Aplicaram-se técnicas de normalização e balanceamento para treinar o modelo MLP. Ferramentas interativas foram empregadas para exibir distribuições de dados, métricas de desempenho e matrizes de confusão. Os resultados mostram que a combinação de MLP com visualizações interativas é eficaz para interpretar desfechos clínicos e auxiliar no planejamento de tratamentos.*

1. Introdução

A tuberculose (TB) é uma das principais causas de doenças infecciosas globalmente, impactando significativamente a saúde pública e a qualidade de vida dos pacientes. Com o aumento da complexidade no tratamento e no acompanhamento dos casos, a análise eficaz de dados se torna crucial para otimizar estratégias terapêuticas e melhorar os desfechos clínicos [Obeagu and Obeagu 2024]. O sucesso do tratamento da TB depende de vários fatores, incluindo adesão à medicação, resistência aos medicamentos, comorbilidades como o HIV, condições socioeconômicas e outros dados clínicos [Motta et al. 2023].

A aplicação de Aprendizado de Máquina (ML) e técnicas de Inteligência Artificial (IA) tem sido vista como uma solução promissora nesse contexto. As arquiteturas de MLPs e outras Redes Neurais Artificiais (RNAs) foram mostradas como ferramentas úteis para prever os resultados do tratamento da TB [Simarmata et al. 2023]. A combinação de MLP com visualizações interativas representa uma inovação significativa na análise de dados clínicos. Enquanto as redes neurais oferecem poderosas capacidades preditivas, as ferramentas de visualização interativa permitem que os dados e os resultados do modelo sejam apresentados de maneira compreensível e acessível.

Os recentes avanços em métodos de ML têm gerado diversas aplicações em múltiplos campos da ciência e da indústria. Como parte integrante da IA, os métodos de ML são ferramentas capazes de aprender automaticamente a partir de dados de amostra (dados de treinamento), fornecendo *insights* valiosos. No entanto, embora existam diversos métodos de ML para resolver problemas do mundo real, esses métodos ainda não são suficientes em abordagens críticas de tomada de decisão, como nas aplicações médicas, onde a participação humana continua sendo essencial ao longo do ciclo de ML [Maadi et al. 2021]. No estudo, os autores fornecem uma revisão da interação humano-Inteligência Artificial para aplicações de ML para informar como combinar melhor a experiência do domínio humano e poder computacional dos métodos de ML.

À medida que a medicina digital avança, as aplicações de visualização em saúde da população oferecem cada vez mais meios para pesquisadores e profissionais para explorar e comunicar descobertas, apoiando a descoberta de conhecimento a partir de grandes volumes de dados. No estudo de [Chishtie et al. 2022], os autores realizaram uma Revisão Sistemática de Escopo para descrever e resumir as evidências das aplicações de visualização interativa, métodos e ferramentas usadas em pesquisa em saúde populacional e serviços de saúde e seus subdomínios.

Em [Jannah and Al Kindhi 2024], os autores desenvolveram e avaliaram MLP e *Extreme Learning Machine* (ELM) para detecção precoce de tuberculose com base em dados clínicos. Os testes de balanceamento de dados foram realizados utilizando Técnica de Sobreamostragem Minoritária (SMOTE). O melhor resultado foi de 95% de acerto para o modelo MLP sem aplicação da técnica SMOTE.

Este estudo tem como objetivo utilizar Redes Neurais Perceptron Multicamadas (MLP) para prever desfechos clínicos (Cura ou Abandono) em pacientes com tuberculose, além de empregar visualizações interativas para aprimorar a compreensão e a interpretação dos resultados. A pesquisa se alinha à área de Interação Humano-Dados (IHD) ao investigar como a combinação de modelos preditivos e representações visuais interativas pode facilitar a tomada de decisões clínicas e promover uma maior literacia de dados entre os profissionais de saúde.

A questão que orienta o estudo é: de que maneira a combinação de Redes Neurais Perceptron Multicamadas (MLP) e visualizações interativas pode melhorar a previsão dos desfechos clínicos em pacientes com tuberculose, facilitando a tomada de decisões clínicas? Essa abordagem não se limita a análises preditivas tradicionais, mas busca proporcionar uma interação significativa entre humanos e dados, potencializando a compreensão dos resultados e contribuindo para decisões mais eficientes no planejamento de tratamentos, aspecto essencial para o cuidado dos pacientes com tuberculose.

O artigo está estruturado da seguinte forma: a Seção 2 apresenta os Materiais e Métodos utilizados. Os resultados e discussões são expostos na Seção 3. Por fim, as principais conclusões alcançadas são resumidas e as perspectivas para trabalhos futuros são discutidas na Seção 4, seguida pelos Cuidados Éticos (Seção 5).

2. Materias e Métodos

Neste estudo, utilizou-se uma abordagem de Aprendizado de Máquina para prever resultados clínicos em pacientes com tuberculose. Redes Neurais Perceptron Multicamadas e técnicas de visualização interativa foram utilizadas para analisar e interpretar os resultados. A metodologia está dividida em 6 etapas principais, conforme ilustrado na Figura 1.

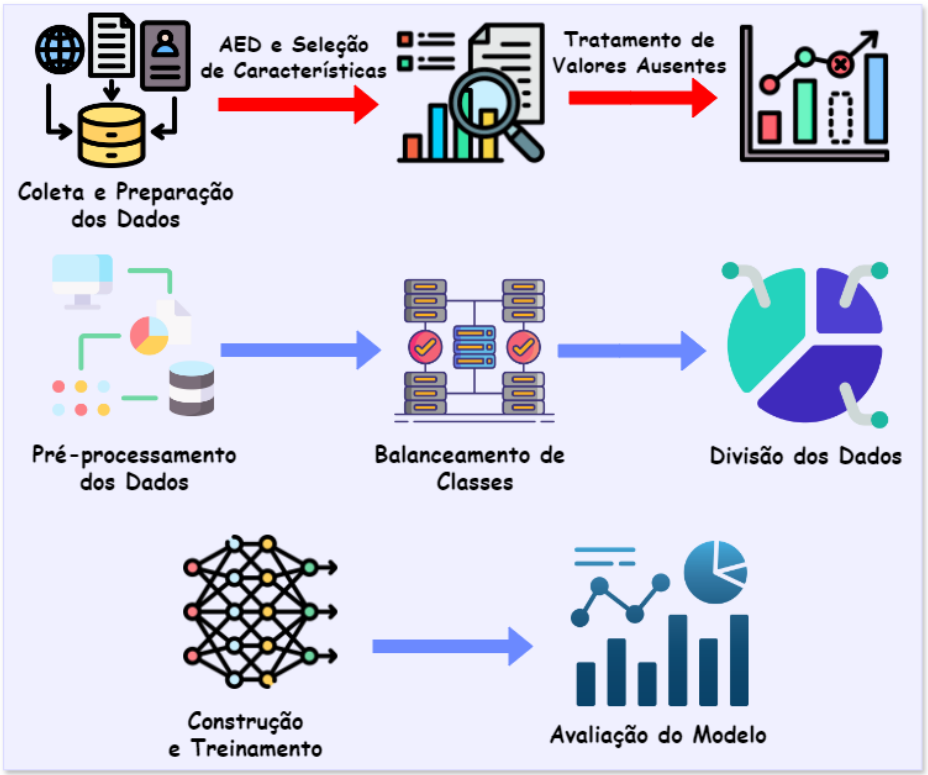

Figura 1. Fluxo de Trabalho da Metodologia Aplicada

2.1. Cuidados Éticos

As Normas e Diretrizes Regulamentadoras da Pesquisa Envolvendo Seres Humanos (Resolução n.º 466/12) estabelecem que a pesquisa apresentada neste estudo não utilizou dados identificáveis de indivíduos. Assim, o Comitê de Ética em Pesquisa com Seres Humanos não teve a obrigatoriedade de revisar o projeto. Para garantir que nenhum indivíduo fosse identificado ou exposto durante a realização do estudo, todos os dados utilizados foram anonimizados e são de domínio público.

Não houve interação direta com pacientes nem coleta de novos dados humanos, uma vez que o foco do estudo foi a aplicação de métodos computacionais para prever desfechos clínicos com base em dados já existentes. Como essa abordagem não comprometeu a integridade ou a privacidade dos indivíduos, não foi necessária a aprovação ética para a realização do trabalho.

2.2. Coleta e Preparação dos Dados

Os dados utilizados neste estudo foram extraídos de um conjunto de dados sobre pacientes com tuberculose, disponível no repositório público [1]. Este estudo analisou dados de 103.846 prontuários de pacientes com tuberculose no estado de São Paulo, obtidos de 2006 a 2016. A preparação dos dados envolveu as seguintes etapas:

2.2.1. Análise Exploratória dos Dados

Inicialmente, foi realizada a Análise Exploratória de Dados (AED) para compreender a distribuição das variáveis. Utilizou-se a biblioteca *plotly* em *Python* para criar visualizações interativas, incluindo histogramas para a distribuição das idades dos pacientes e gráficos de barras para a distribuição por sexo e situação atual. Faixas etárias foram definidas para analisar a distribuição dos pacientes em diferentes grupos etários.

Após a AED, foi realizada a seleção de características (*features*). As variáveis preditoras foram categorizadas em grupos baseados em aspectos sociais e clínicos para facilitar a análise. A seleção foi fundamentada em revisões da literatura, na disponibilidade de dados e na importância para a predição dos desfechos, conforme apresentado em [Orjuela-Cañón et al. 2022] e [Kanesamoorthy and Dissanayake 2021].

O conjunto final de dados foi composto por 23 características (*features*) distribuídas em duas categorias: Clínicas e Socioeconômicas. Quatro características socioeconômicas foram selecionadas: raça/cor, sexo, escolaridade e tipo de ocupação, uma vez que, como apontado em estudos anteriores [de Almeida Rodrigues et al.], fatores socioeconômicos e demográficos podem impactar significativamente a adesão dos pacientes ao tratamento e os desfechos clínicos. Além disso, 19 características clínicas foram selecionadas, incluindo dados sobre comorbidades, como diabete, HIV e outras condições. O número de doses primárias e secundárias recebidas serve como um indicador direto da adesão ao tratamento e da resposta clínica, enquanto idade e tipo de caso são fatores críticos no planejamento do tratamento [health Organisation 2023].

2.2.2. Tratamento de Valores Ausentes:

Valores ausentes foram tratados utilizando a imputação com a estratégia de moda, substituindo valores ausentes pelo valor mais frequente nas respectivas colunas. Isso garante que não haja mais valores ausentes no conjunto de dados.

2.3. Pré-processamento dos Dados

O pré-processamento de dados dos pacientes com TB envolve codificação e normalização dos dados. As etapas de pré-processamento incluíram: Codificação de Variáveis Categóricas, Normalização de Variáveis Numéricas e Transformação da Variável de Saída.

Procedeu-se à codificação de atributos categóricos e à normalização de atributos numéricos para padronizar o conjunto de dados. Os atributos categóricos foram codificados utilizando o *OneHotEncoder*, que converte cada categoria em uma nova coluna

[1] `https://figshare.com/articles/dataset/tuberculosis-data-06-16_csv/8066663?file=15032345`

binária[2]. Já os atributos numéricos foram normalizados com o *StandardScaler*[3], garantindo média 0 e desvio padrão 1 para esses valores [Viboonsang and Kosolsombat 2024]. A variável de saída foi transformada com o *LabelEncoder*, que converteu a variável categórica de saída ('sitAtual') em valores numéricos, criando uma variável binária onde '0' representa o desfecho de abandono e '1' representa o desfecho de cura. O conjunto de dados inclui 91.823 pacientes com situação de cura ('1') e 12.023 pacientes com situação de abandono ('0').

2.4. Balanceamento de Classes

Para abordar o desbalanceamento das classes, foi aplicado o método *SMOTETomek*, que combina o SMOTE (*Synthetic Minority Over-sampling Technique*) com *Tomek Links* para criar um *dataset* balanceado. Esse método visa melhorar a eficiência do treinamento do modelo e permitir uma classificação mais precisa das classes.

O método combina *Tomek Links*, que subamostra a classe majoritária para aproximá-la da classe minoritária, com o SMOTE, que realiza a sobreamostragem da classe minoritária, criando pontos de dados sintéticos para se aproximar da classe majoritária [Talukder et al. 2024]. Foram criados gráficos interativos para comparar a distribuição das classes antes e depois do balanceamento, fornecendo uma visão clara das alterações na distribuição dos dados.

2.5. Divisão dos Dados

O conjunto de dados balanceado foi dividido em conjuntos de treinamento, validação e teste, com 80% dos dados destinados ao treinamento, 20% ao teste e 20% dos dados de treinamento foram empregados na validação. Essa divisão permite uma avaliação adequada do modelo em dados não vistos durante o treinamento.

2.6. Construção e Treinamento do Modelo

O modelo MLP foi construído com duas camadas ocultas (128 e 64 neurônios) e regularizado com *dropout* de 50%. A função de ativação ReLU foi usada nas camadas ocultas e sigmoide na camada de saída para a tarefa de classificação binária. O modelo foi treinado com o otimizador Adam e a função de perda de entropia cruzada, por 50 épocas e com tamanho de lote de 32.

2.7. Avaliação do Modelo

O desempenho do modelo foi avaliado utilizando as métricas de acurácia, precisão, *recall* e *F1-score*, calculadas com base nas previsões feitas sobre o conjunto de teste. Foram empregadas as seguintes abordagens de avaliação:

- Curvas de Aprendizado e Acurácia: As curvas de aprendizado e acurácia foram visualizadas para avaliar o desempenho do modelo durante o treinamento, fornecendo *insights* sobre a convergência e a eficácia da regularização.
- Matriz de Confusão: A matriz de confusão foi gerada para comparar as previsões do modelo com os valores reais, facilitando a identificação de erros de classificação e a avaliação detalhada do desempenho por classe.

[2]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
[3]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

- Métricas de Avaliação: As métricas de avaliação, incluindo acurácia, precisão, *recall* e *F1-score*, foram calculadas e visualizadas em gráficos interativos para fornecer uma visão abrangente do desempenho do modelo.

3. Resultados e Discussões

Esta seção discute os resultados da Análise Exploratória de Dados (AED) e os resultados de avaliação de desempenho obtidos pelo modelo na previsão de desfechos clínicos (cura ou abandono) em pacientes com tuberculose, utilizando visualizações interativas.

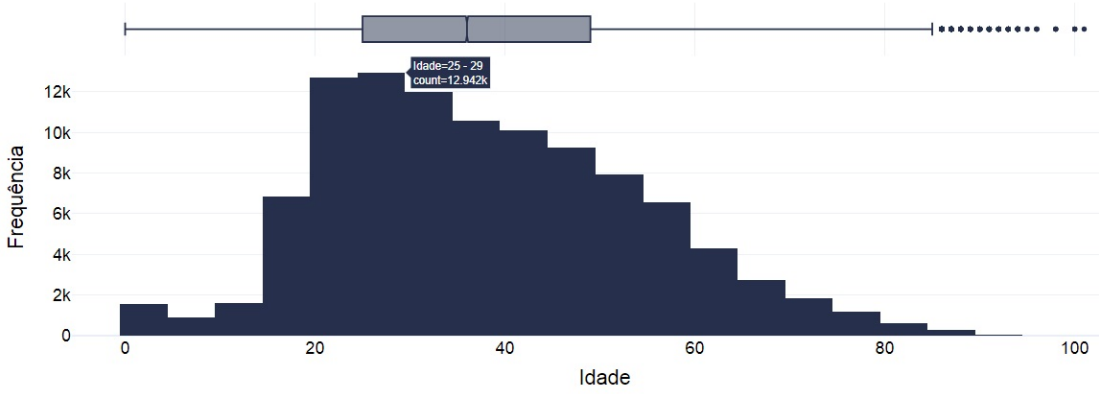

Figura 2. Distribuição das Idades dos Pacientes

A Figura 2, gerada com visualizações interativas, mostra a distribuição etária dos pacientes, destacando a maior concentração entre 25 e 29 anos, totalizando 12.942 casos de pacientes com TB. Através dessas visualizações, também se observa uma menor incidência de TB em crianças de 5 a 9 anos, com um total de 911 casos e uma redução nos casos a partir dos 60 anos.

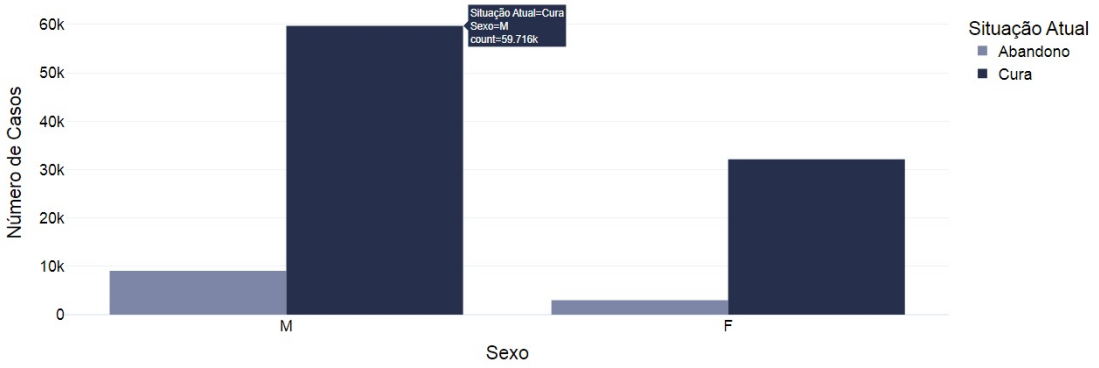

Figura 3. Distribuição por Sexo e Situação Atual

Na análise interativa por sexo, conforme ilustrado na Figura 3, foram identificados 68.755 pacientes do sexo masculino (M) e 35.091 do sexo feminino (F). Além disso, as visualizações revelam que pacientes do sexo masculino apresentam um índice de cura superior ao de abandono, totalizando 59.716 casos com situação de cura. Diferente do sexo feminino, que possui apenas um total de 32.107 casos de cura e 2.984 de situação de abandono do tratamento.

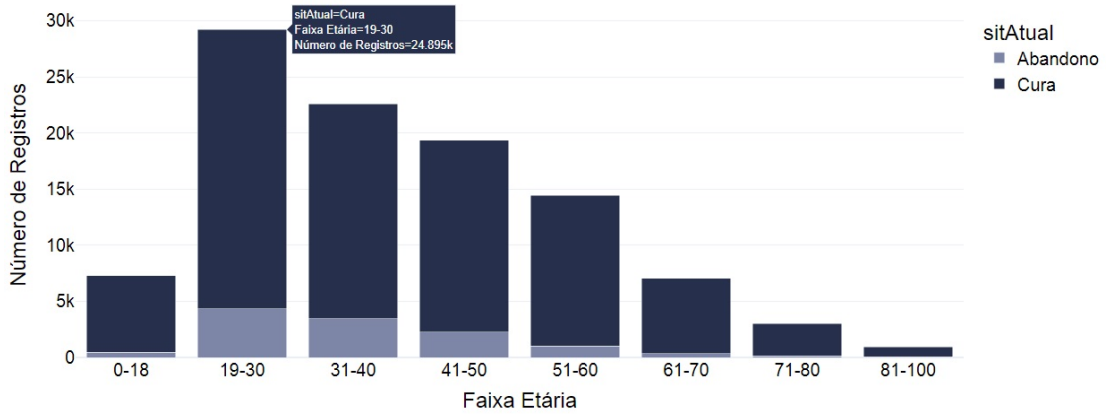

Figura 4. Número de Registros por Faixa Etária e Situação Atual.

Os registros foram agrupados e analisados por faixa etária e situação atual do paciente (abandono ou cura) utilizando visualizações interativas. A Figura 4 revela que a maioria dos pacientes curados está na faixa de 19 a 30 anos, somando 24.895 casos de pacientes curados e 4.320 pacientes que abandonaram o tratamento.

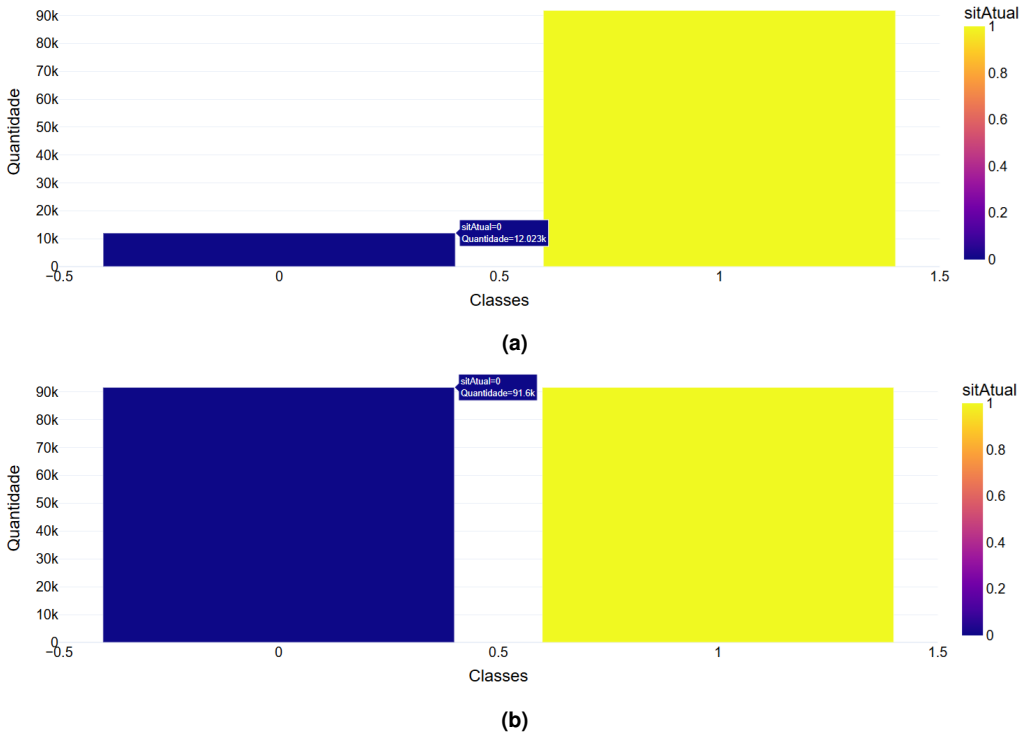

Figura 5. (a) Distribuição da classe de saída antes do balanceamento dos dados; (b) Distribuição da classe de saída após o balanceamento dos dados com *SMOTETomek*

As visualizações interativas também destacaram o desbalanceamento entre as classes de cura e abandono, conforme mostrado na Figura 5a, com uma razão de classes de 7,64. A Figura 5b mostra o resultado do balanceamento, aumentando o número de registros da classe ’0’ e promovendo uma distribuição mais equitativa entre as classes ’0’ e ’1’.

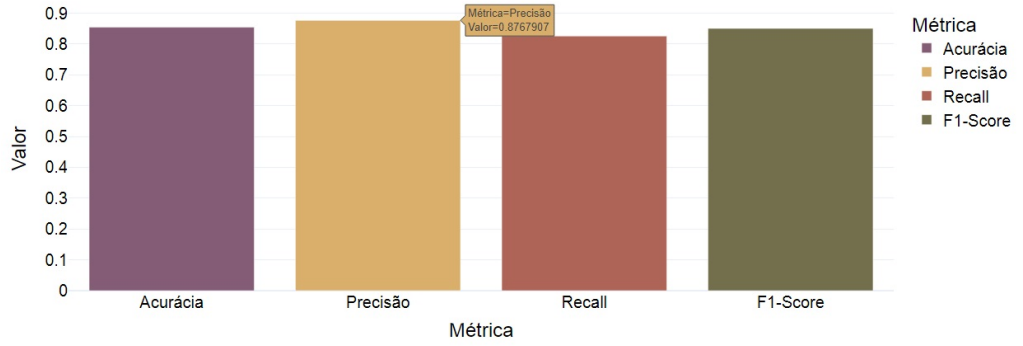

Figura 6. Resultados da Avaliação de Desempenho do Modelo.

A partir dos resultados da Figura 6, o modelo alcançou uma precisão de 87,67%, o *recall* foi de 83,45%, indicando que o modelo identificou a maioria dos casos positivos, essencial para evitar o sub-registro de casos de tuberculose. O F1-Score, que combina precisão e *recall*, foi de 85,49%. As visualizações interativas evidenciam o impacto do balanceamento de classes, uma técnica comum para corrigir conjuntos de dados desbalanceados, que no caso da doença de tuberculose deste estudo, apresentam mais casos de cura do que abandono.

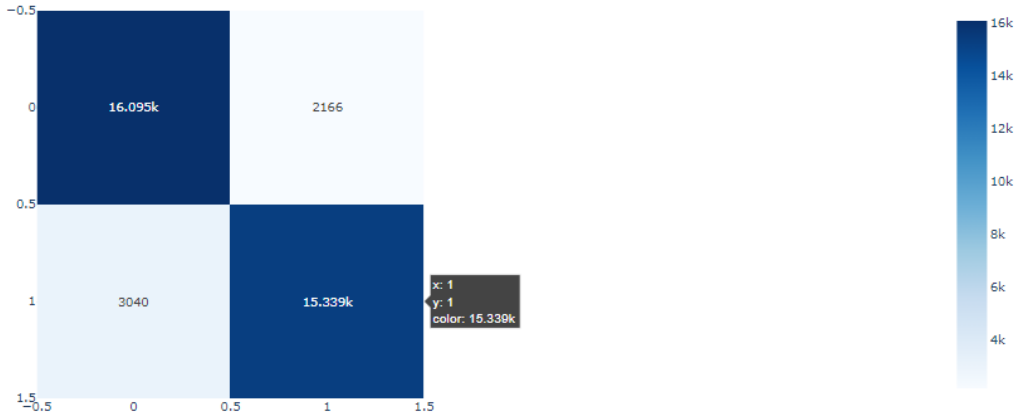

Figura 7. Matriz de Confusão.

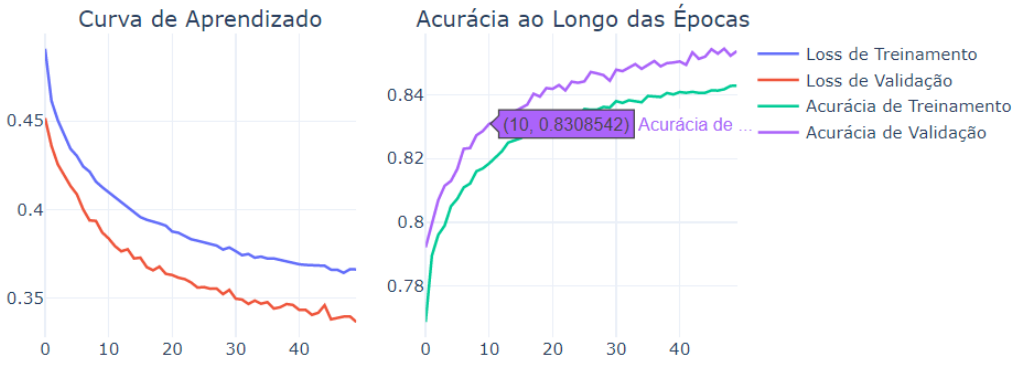

Figura 8. Curvas de Aprendizado e Acurácia ao Longo das Épocas.

A Figura 7 apresenta a matriz de confusão utilizada para avaliar o desempenho do modelo. Os valores nas células representam a quantidade de predições corretas e incorretas realizadas. O modelo obteve 16.095k verdadeiros negativos (TN) e 15.339k verdadeiros positivos (TP), indicando uma alta precisão nas classificações corretas. No entanto, também foram registrados 3.040 falsos negativos (FN) e 2.166 falsos positivos (FP), evidenciando os erros nas predições. Esses resultados utilizando visualizações interativas sugerem que o modelo tem um desempenho satisfatório, mas ainda apresenta espaço para melhorias na redução dos erros.

A Figura 8 exibe as curvas de aprendizado e acurácia ao longo das épocas, fundamentais para avaliar o comportamento do modelo durante o treinamento. No gráfico à esquerda, observa-se a evolução das funções de perda (*loss*) para os conjuntos de treinamento e validação, ambas apresentando uma tendência decrescente, o que indica uma melhoria contínua no ajuste do modelo. No gráfico à direita, as curvas de acurácia mostram um aumento progressivo até a estabilização, sugerindo que o modelo alcançou um nível adequado de generalização.

4. Conclusões

Este estudo explorou o uso de Redes Neurais Perceptron Multicamadas (MLP) combinadas com visualizações interativas para a previsão de desfechos clínicos em pacientes com tuberculose. Os resultados do modelo foram comparados com outros métodos, demonstrando um desempenho superior. O modelo alcançou uma acurácia de 85% e uma precisão de 87%. A eficácia da nossa abordagem é evidenciada ao compará-la com o trabalho de [Orjuela-Cañón et al. 2022], que utilizou um conjunto de dados diferente e apresentou resultados inferiores.

Com base nos resultados apresentados, foi possível responder satisfatoriamente à pergunta de pesquisa proposta, evidenciando que a combinação de modelos preditivos com visualizações interativas promove uma interação mais significativa entre os usuários e os dados, contribuindo para uma compreensão aprofundada dos desfechos clínicos. A pesquisa, portanto, reforça a importância da IHD como um campo interdisciplinar que potencializa a aplicabilidade prática das soluções de Machine Learning por meio de visualizações interativas. Como trabalho futuro, pretende-se aplicar o modelo em outras bases de dados de doenças infecciosas e a exploração de outras técnicas de aprendizado de máquina e visualização que possam aprimorar ainda mais a interpretação dos desfechos clínicos.

Referências

Chishtie, J., Bielska, I. A., Barrera, A., Marchand, J.-S., Imran, M., Tirmizi, S. F. A., Turcotte, L. A., Munce, S., Shepherd, J., Senthinathan, A., et al. (2022). Interactive visualization applications in population health and health services research: systematic scoping review. *Journal of medical Internet research*, 24(2):e27534.

de Almeida Rodrigues, M. G., Sampaio, V., Lynn, T., and Endo, P. T. A brazilian classified data set for prognosis of tuberculosis, between january 2001 and april 2020.

health Organisation, W. (2023). *Report 20-23*, volume t/malaria/.

Jannah, A. W. and Al Kindhi, B. (2024). Optimization of early detection of tuberculosis: Use of multilayer perceptron and extreme learning machine with clinical data. *Jurnal Indonesia Sosial Teknologi*, 5(5).

Kanesamoorthy, K. and Dissanayake, M. B. (2021). Prediction of treatment failure of tuberculosis using support vector machine with genetic algorithm. *The International Journal of Mycobacteriology*, 10(3):279–284.

Maadi, M., Akbarzadeh Khorshidi, H., and Aickelin, U. (2021). A review on human–ai interaction in machine learning and insights for medical applications. *International journal of environmental research and public health*, 18(4):2121.

Motta, I., Boeree, M., Chesov, D., Dheda, K., Günther, G., Horsburgh Jr, C. R., Kherabi, Y., Lange, C., Lienhardt, C., McIlleron, H. M., et al. (2023). Recent advances in the treatment of tuberculosis. *Clinical Microbiology and Infection*.

Obeagu, E. and Obeagu, G. (2024). Understanding immune cell trafficking in tuberculosis-hiv coinfection: The role of l-selectin pathways. *Elite Journal of Immunology*, 2(2):43–59.

Orjuela-Cañón, A. D., Jutinico, A. L., Awad, C., Vergara, E., and Palencia, A. (2022). Machine learning in the loop for tuberculosis diagnosis support. *Frontiers in Public Health*, 10:876949.

Simarmata, T. S., Isnanto, R. R., and Triwiyatno, A. (2023). Detection of pulmonary tuberculosis using neural network with feature extraction of gray level run-length matrix method on lung x-ray images. In *2023 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 570–574. IEEE.

Talukder, M. A., Sharmin, S., Uddin, M. A., Islam, M. M., and Aryal, S. (2024). Mlstl-wsn: machine learning-based intrusion detection using smotetomek in wsns. *International Journal of Information Security*, 23(3):2139–2158.

Viboonsang, P. and Kosolsombat, S. (2024). Network intrusion detection system using machine learning and deep learning. In *2024 IEEE International Conference on Cybernetics and Innovations (ICCI)*, pages 1–6. IEEE.