Interação Humano-Dados na Predição de Desfechos Clínicos em Tuberculose: Um Protótipo Web com Streamlit para Explicabilidade e Visualização

Ronilson W. S. Pereira[1], **Igor Falcão** [2], **Saul Carneiro** [3], **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

[3]Hospital Universitário João de Barros Barreto – Universidade Federal do Pará
Belém – PA – Brasil

ronilsonengenharia@gmail.com, igorufpa2013.4@gmail.com

saul@ufpa.br, seruffo@ufpa.br, karlafigueiredo@ime.uerj.br

Abstract. Introduction*: Machine learning models, such as Random Forest, have proven effective in predicting clinical outcomes, although the interpretability of these models still poses a challenge.* ***Objective:*** *This paper presents an interactive web prototype, developed with Streamlit, aimed at explaining and visualizing clinical predictions in patients with tuberculosis.* ***Methodology or Steps****: The solution uses the Random Forest algorithm to generate predictions and incorporates interpretability techniques, such as SHAP, allowing interaction with the data and providing visual explanations about the contribution of variables to the results.* ***Expected Results:*** *The system is expected to promote an accessible and understandable interface, bringing artificial intelligence closer to clinical practice through human-data interaction.*
Keywords *Machine learning, Random Forest, Interpretability, SHAP, Tuberculosis.*

Resumo. Introdução: *Modelos de aprendizado de máquina, como o Random Forest, têm se mostrado eficazes na predição de desfechos clínicos, embora a interpretabilidade desses modelos ainda represente um desafio.* ***Objetivo:*** *Este trabalho apresenta um protótipo web interativo, desenvolvido com Streamlit, voltado à explicabilidade e visualização de predições clínicas em pacientes com tuberculose.* ***Metodologia ou Etapas****: A solução utiliza o algoritmo Random Forest para gerar predições e incorpora técnicas de interpretabilidade, como SHAP, permitindo a interação com os dados e fornecendo explicações visuais sobre a contribuição das variáveis nos resultados.* ***Resultados Esperados:*** *Espera-se que o sistema promova uma interface acessível e compreensível, aproximando a inteligência artificial da prática clínica por meio da interação humano-dados.*
Palavras-Chave *Aprendizado de máquina, Random Forest, Interpretabilidade, SHAP, Tuberculose.*

1. Introdução

A tuberculose (TB) permanece como um desafio relevante para a saúde pública, especialmente em regiões com alta incidência e recursos limitados [World Health Organization 2024]. Apesar dos avanços no tratamento, a predição de desfechos clínicos, como cura, abandono ou óbito, ainda apresenta limitações que afetam a eficácia das intervenções. Algoritmos de aprendizado de máquina (*Machine Learning*- ML) têm demonstrado potencial na análise de dados clínicos para apoiar a tomada de decisão [Yasin et al. 2024], mas sua adoção, na prática é restrita pela baixa interpretabilidade e pela falta de transparência [Paixão et al. 2022].

A área de Interação Humano-Dados (*Human-Data Interaction* – HDI) surge como uma abordagem relevante para enfrentar tais desafios, promovendo maior transparência, controle e compreensão dos sistemas baseados em dados. Trabalhos anteriores têm explorado técnicas de explicabilidade aliadas à visualização de dados como estratégias para tornar os modelos mais acessíveis e compreensíveis aos usuários finais [Tendedez et al. 2022], mas ainda são escassas as aplicações focadas no domínio da tuberculose. Este trabalho propõe um protótipo web que integra essas técnicas, visando facilitar o uso de modelos preditivos por profissionais de saúde e promover uma interação mais transparente e centrada no usuário.

2. Cuidados Éticos

Conforme a Resolução nº 466/12 do Conselho Nacional de Saúde, que estabelece diretrizes para pesquisas envolvendo seres humanos, este estudo não utilizou dados que permitissem a identificação direta ou indireta dos participantes. Por esse motivo, não houve exigência de submissão ao Comitê de Ética em Pesquisa. Todos os dados empregados foram previamente anonimizados e estão disponíveis em domínio público, assegurando a proteção da privacidade e a integridade dos indivíduos envolvidos.

3. Metodologia Aplicada

Nesta seção, descreve-se o percurso metodológico adotado no estudo, conforme ilustrado na Figura 1, que contempla desde a obtenção dos dados até a construção do protótipo *web*, passando pelas etapas de pré-processamento, modelagem com *Random Forest*, avaliação de desempenho e interpretação dos resultados por meio da técnica SHAP.

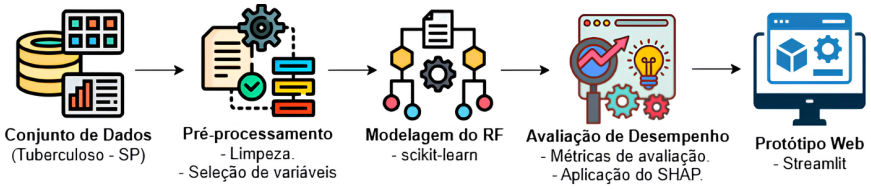

Figura 1. Fluxo de trabalho proposto.

3.1. Conjunto de Dados

Utilizou-se uma base pública de casos de tuberculose do estado de São Paulo,[1] com 103.846 prontuários de pacientes com tuberculose, contendo variáveis clínicas, demográficas e socioeconômicas.

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

3.2. Pré-processamento e Modelagem Preditiva

O pré-processamento teve como objetivo melhorar a qualidade dos dados, reduzindo ruídos e inconsistências. Foram aplicadas técnicas de imputação [2] para lidar com valores ausentes e, em casos mais críticos, optou-se pela exclusão de variáveis. Também foram ajustadas as estruturas dos atributos para garantir a relevância das informações.

Em seguida, realizou-se a seleção de variáveis com base em estudos anteriores [Orjuela-Cañón et al. 2022, Kanesamoorthy e Dissanayake 2021] e em análises estatísticas. Foram escolhidas 30 variáveis para compor o conjunto final de dados. A modelagem foi feita com o algoritmo *Random Forest*, utilizando a biblioteca *scikit-learn*, para classificar os desfechos clínicos como cura ou abandono. O modelo foi escolhido por sua robustez e bom desempenho em dados estruturados [Hartshorn 2016]. Durante o processo, foram aplicadas técnicas como normalização, divisão dos dados em treino e teste, e regularização para evitar *overfitting*.

3.3. Avaliação de Desempenho e Explicabilidade

O desempenho do modelo foi avaliado por meio das métricas*Accuracy*, *Precision*, *Recall*, *F1-score* macro e AUC. Devido ao forte desbalanceamento das classes, optou-se por priorizar o *F1-score macro*, que atribui peso igual a todas as classes e é mais adequado para cenários desproporcionais.

Para garantir transparência, foram utilizados valores SHAP (*SHapley Additive exPlanations*), que identificam a influência de cada atributo nas predições [Ma et al. 2023]. Baseados na teoria dos jogos cooperativos, esses valores mensuram a contribuição individual das variáveis e suas interações, oferecendo uma visão clara dos fatores mais determinantes nos desfechos clínicos.

3.4. Protótipo Web com Streamlit

O desenvolvimento de um protótipo web interativo constitui o principal objetivo deste estudo, visando aproximar os modelos preditivos dos profissionais da saúde por meio de uma interface acessível, explicável e funcional. Para isso, foi utilizada a biblioteca *Streamlit*[3], escolhida por sua facilidade de integração com modelos desenvolvidos em *Python* e por oferecer suporte nativo à construção de dashboards interativos com visualização de dados em tempo real.

O protótipo foi projetado para permitir a exploração prática dos resultados do modelo de classificação e a interpretação das predições geradas. A interface inclui recursos como o *upload* de novos dados para predição, a visualização das predições por paciente, a apresentação de gráficos SHAP para explicabilidade local e global e a utilização de filtros interativos para explorar padrões nos dados.

4. Resultados preliminares e discursões

A Figura 2 apresenta a interface inicial do protótipo **web** interativo desenvolvido com *Streamlit*. O sistema permite selecionar um paciente (por ID) e visualizar a predição do desfecho clínico com base em um modelo *Random Forest* (RF), treinado a partir de dados

[2]https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html
[3]`https://streamlit.io/`

sobre o tratamento da tuberculose. O modelo apresentou desempenho satisfatório, com *Accuracy* = 0,8851, *Precision* = 0,9146, *Recall* = 0,9596 e *F1-score* = 0,9366, indicando boa capacidade de generalização e equilíbrio entre precisão e abrangência das previsões.

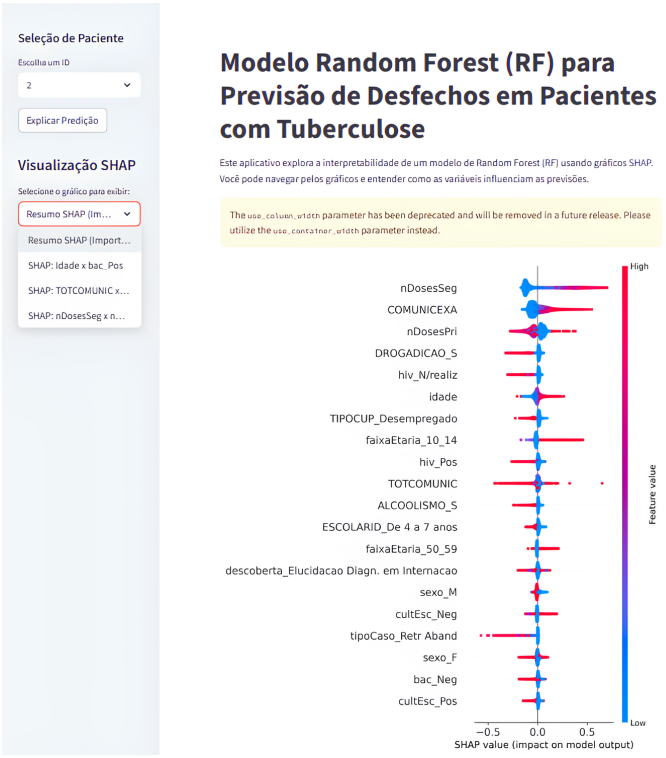

Figura 2. Interface inicial do protótipo web interativo desenvolvido.

Na área principal da interface, observa-se a exibição de um gráfico de importância global SHAP, que apresenta as variáveis mais relevantes para o modelo *Random Forest*. Cada linha representa um atributo, ordenado pela sua contribuição média para as predições. Os pontos indicam instâncias individuais dos pacientes: em vermelho, valores altos da variável; em azul, valores baixos. A posição horizontal dos pontos mostra o impacto no resultado previsto (positivo ou negativo). Esse tipo de visualização permite identificar, de forma clara, quais fatores exercem maior influência nas previsões de desfechos clínicos.

5. Conclusão

Este trabalho propõe um protótipo *web* interativo voltado à explicabilidade e visualização de predições clínicas em casos de tuberculose, com ênfase na interação humano-dados, isto é, na criação de interfaces que permitam aos profissionais de saúde compreender, explorar e aplicar informações geradas por modelos de inteligência artificial em seu contexto de atuação. A proposta visa contribuir para a adoção consciente de soluções baseadas em inteligência artificial na saúde pública, tornando os modelos preditivos mais compreensíveis, transparentes e acessíveis aos profissionais da área.

Como trabalhos futuros, propõe-se realizar testes com profissionais da saúde para avaliar a usabilidade e a utilidade da ferramenta na prática clínica, incorporar recursos de acessibilidade, ampliar a base de dados para abrangência nacional (como os dados do SINAN), além de explorar novas abordagens de explicabilidade e visualização interativa.

Referências

Hartshorn, S. (2016). Machine learning with random forests and decision trees: A visual guide for beginners. *Kindle edition*.

Kanesamoorthy, K. e 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.

Ma, F.-q., He, C., Yang, H.-r., Hu, Z.-w., Mao, H.-r., Fan, C.-y., Qi, Y., Zhang, J.-x., e Xu, B. (2023). Interpretable machine-learning model for predicting the convalescent covid-19 patients with pulmonary diffusing capacity impairment. *BMC Medical Informatics and Decision Making*, 23(1):169.

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

Paixão, G. M. d. M., Santos, B. C., Araujo, R. M. d., Ribeiro, M. H., Moraes, J. L. d., e Ribeiro, A. L. (2022). Machine learning na medicina: revisão e aplicabilidade. *Arquivos Brasileiros de Cardiologia*, 118(1):95–102.

Tendedez, H., Ferrario, M.-A., McNaney, R., e Gradinar, A. (2022). Exploring human-data interaction in clinical decision-making using scenarios: co-design study. *JMIR human factors*, 9(2):e32456.

World Health Organization (2024). *Global Tuberculosis Report*. Number September.

Yasin, P., Yimit, Y., Cai, X., Aimaiti, A., Sheng, W., Mamat, M., e Nijiati, M. (2024). Machine learning-enabled prediction of prolonged length of stay in hospital after surgery for tuberculosis spondylitis patients with unbalanced data: a novel approach using explainable artificial intelligence (xai). *European Journal of Medical Research*, 29(1):383.