Identification of important predictors, using V-Cramer technique and Permutation feature importance, in predicting Tuberculosis treatment status

Jacó M. Santos[1], Elson P. Vasques[2], Adriane F. Valentin[2], Eliane C. Nogueira[2], Ericle L. Costa [2], Juan C. Souza[2], John K. S. Silva[2], Elias E. R. Marques[2], July E. S. Silva[2], João G. S. Gomes[2]

[1]Universidade Federal do Amazonas - Instituto de Computação, Programa de Pós Graduação em Informática PPGI (ICOMP) – Manaus – AM – Brasil

[2]Diretoria de Inteligência de Dados (DID) – Secretaria de Saúde Municipal – SEMSA – Manaus, AM – Brasil.

{jaco.santos,elson}@icomp.ufam.edu.br,{adrianefariascosta.afc, eliacampos,juan.choque,ericle.costa,eliasemanuel12ribeiro,july081503, guilhermejgsg01}@gmail.com, john.silva@pmm.am.gov.br

Abstract. *In this study, we investigate the use of a fully connected neural network implemented with Keras to create prediction models for the treatment outcomes of tuberculosis patients. The data used were from the Sistema de Informação de Agravos de Notificação (SINAN) of the year 2023, integrated by iTB, a software developed by the Health Department of Manaus-AM. This work is still in progress. So far, we have achieved an accuracy of 92.80% and F1-Score values of 0.90 for cure, 0.89 for abandonment, 0.98 for TB-related death, and 0.99 for drug resistance. Additionally, applying the permutation-importance technique, we obtained an accuracy of 93.31%. Experiments with patient consultation data will still be added to the test data, along with a Cramer's V analysis.*

Resumo. *Neste estudo, investigamos o uso de uma rede neural totalmente conectada implementada com Keras para criar modelos de previsão do desfecho do tratamento de pacientes com tuberculose. Os dados utilizados foram do Sistema de Informação de Agravos de Notificação (SINAN) do ano de 2023, integrados pelo iTB, um software desenvolvido pela Secretaria de Saúde de Manaus-AM. Este trabalho ainda está em andamento. Até o momento, alcançamos uma acurácia de 92,80% e valores de F1-Score de 0.90 para cura, 0.89 para abandono, 0.98 para morte por TB e 0.99 para droga resistente. Além disso, ao aplicar a técnica de permutation-importance, obtivemos uma acurácia de 93,31%. Experimentos com dados de consultas dos pacientes ainda serão agregados aos dados de teste, juntamente com uma análise V de Cramer.*

1. Introdução

A tuberculose (TB) é uma doença infecciosa transmitida pelo ar, causada pelo *Mycobacterium tuberculosis*, que continua sendo uma das principais causas de morbidade e mortalidade em escala global. Se não for detectada e tratada precocemente, a TB pode levar a complicações graves e até mesmo ao óbito. A importância da detecção precoce é enfatizada pelo fato de que, se não tratada, há uma probabilidade significativa de 70% de um paciente falecer em um período de 10 anos após o diagnóstico. Modelos treinados alcançaram altas taxas de precisão, com 96.91% de acurácia, 99.38% de área sob a curva

(AUC), 91.81% de sensibilidade e 98.42% de especificidade para classificação multiclasse de imagens conforme observado em [Acharya et al., 2022].

Acharya et al. (2022) também afirmam que nos últimos anos a necessidade de ferramentas complementares para o diagnóstico e tratamento eficaz da tuberculose tem crescido, especialmente em países de renda média a baixa. Vale citar o *software* iTB, lançada em 2023 pela Secretaria Municipal de Saúde de Manaus, como uma dessas ferramentas que auxilia na inserção de dados no SINAN através de formulários semiautomatizados, permitindo a seleção de opções pré-determinadas pelo sistema, minimizando erros de digitação. Outros estudos como de Inacio B et al. (2023), Perlaza et al. (2023), Tabosa de Oliveira et al. (2020), Singh et al. (2022), Zhang et al. (2021) enfatizam a necessidade de recursos de computação para descobertas, otimização, mitigação de erros no campo da saúde. A estimativa da Organização Mundial da Saúde (OMS) é de que aproximadamente um quarto da população global esteja infectada pelo bacilo causador da tuberculose. Esse cenário apresenta um risco significativo de 5 a 10% de desenvolver a forma ativa da doença ao longo da vida, sendo essa probabilidade maior nos primeiros anos após a infecção inicial [Murrugara et al., 2023].

2. Métodos

Utilizou-se a base de dados do iTB, especificamente a tabela *tb_tuber_sinan*, composta por 67 colunas e 58.137 registros de casos de tuberculose. Para criação de uma base de dados válida, selecionaram-se 4.385 registros de notificação do período de 2023, porém com alguns filtros com o intuito de se ter dados mais qualificados.

Utilizou-se pacotes do *Scikit-learn* e do *TensorFlow* para treinamento de uma rede neural totalmente conectada. Alguns detalhes são mencionados a seguir, ressaltando que este trabalho está em andamento, onde pretende-se agregar novas variáveis aos dados.

2.1. Exploração dos dados

Algumas colunas apresentaram dados inválidos durante as consultas a anos anteriores, como pontos, vírgulas e tremas. Para evitar esta situação, aplicou-se filtro durante a consulta à *tb_tuber_sinan*. Alguns filtros são mostrados na Tabela 2.

Tabela 2. Filtros realizados em *PostgreSQL* para remoção de dados inválidos.

Filtros	Condição SQL	Importância
Data de Notificação	dt_notificacao >'01/01/2023' and dt_notificacao < '31/12/2023'	Seleciona notificações de 2023.
Data de Início do Tratamento	dt_inicio_tratamento is not null	Inclui apenas pacientes que iniciaram o tratamento.
Situação do Encerramento	tp_situacao_encerramento in ('1','2','3','7')	Seleciona resultados específicos de tratamento (curado, abandono, morte por TB, Droga Resistente).
Data de Nascimento	dt_nascimento is not null	Garante a inclusão de pacientes com data de nascimento válida para calcular a idade.
Baciloscopia após 6 Meses	st_bacil_apos_6_mes not in (',','.')	Remove valores inválidos da variável st_bacil_apos_6_mes.
Sensibilidade	tp_sensibilidade not in('′′)	Remove valores inválidos da variável tp_sensibilidade.
Tipo de Entrada	tp_entrada in ('1','2','3','4','6')	Filtra por tipos específicos de entrada.

Além da remoção dos dados inválidos, foi atribuído "-1" aos dados em branco, pois dados em branco podem significar, resultados ainda não informados. Como a tabela

funciona de forma dinâmica, isto é, a tabela pode ser atualizada com dados novos, algumas colunas ainda aguardam resultados de etapas em andamento.

A coluna alvo, “Situação de Encerramento” (SE), assim como a maiorias das colunas, é do tipo categórica. Isso apontou para um problema de classificação. As classes que compõem a coluna alvo são: 1 – Cura, 2- Abandono, 3 - Óbito por TB, 4 - Óbito por outras causas, 5 – Transferência, 6 - Mudança de Diagnóstico, 7- TB-DR, 8 - Mudança de esquema, 9 – Falência, 10 - Abandono Primário.

2.2. Pré-processamento

Um filtro ('1','2','3','7') foi aplicado na coluna SE para compor o conjunto de dados apenas com registro de “cura”, “abandono”, “morte por TB” e “TB_DR”, que resultou em 1.488 instâncias. Com isso, foi possível aplicar a técnica *SMOTE* para auxiliar no balanceamento das classes minoritárias [Bahaweres et al., 2020].

Foi utilizado os pacotes *preprocessing*, *imblearn.over_sampling*, *metrics*, *inspection*, *scipy*, e*nsemble*, *model_selection* do *sklearn* para a normalização, codificação dos dados entres outras bibliotecas como *pandas*, *numpy*, *matplotlib*, *seaborn*.

2.3. Separação dos dados para Treino e Teste

Os dados de treino e teste foram separados por meio do *train_test_split* do *sklearn.* A Figura 1 mostra o balanceamento de 70% dos dados de treino, Figura 1.a, e o resultado em 1.b, aplicando a técnica *SMOTE* por meio do pacote *imblearn.over_sampling.*

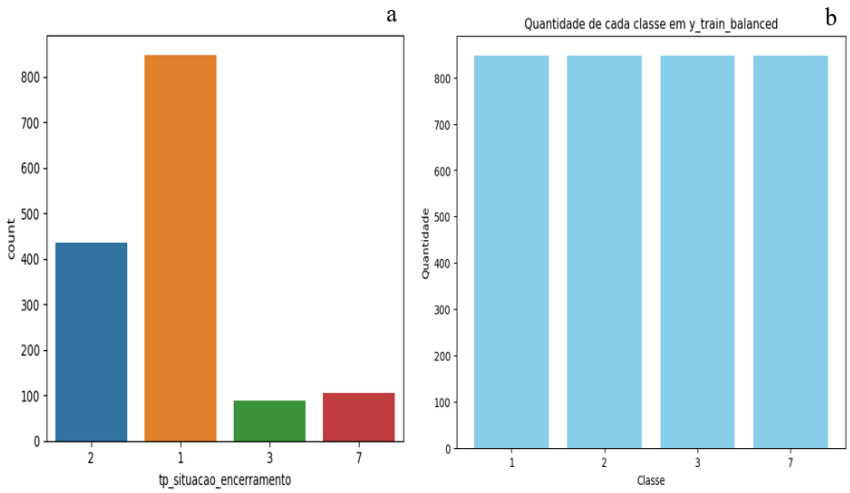

Figura 1. Dados de treino balanceado por *SMOTE*

O processo resultou em quantidades de 800 amostras para cada variável alvo.

2.4. Configuração da rede neural.

A rede é criada usando o modelo *Sequential*, o que significa que as camadas são empilhadas uma sobre a outra de forma sequencial. Entre cada camada densa, há uma camada de *Dropout* com uma taxa de 0.5. O *Dropout* é uma técnica de regularização que ajuda a prevenir *overfitting*, desativando aleatoriamente um número especificado de neurônios durante o treinamento. A última camada é uma camada densa com ativação softmax. A *softmax* é comumente usada na camada de saída de uma rede neural para problemas de classificação multiclasse, pois produz probabilidades normalizadas para cada classe.

Tabela 2. Rede neural com 03 camadas densas intercaladas com camadas de regulação.

Camada	Tipo	Unidades	Ativação	*Dropout*
1	*Dense*	256	*ReLU*	-
2	*Dropout*	-	-	0.5
3	*Dense*	256	*ReLU*	-
4	*Dropout*	-	-	0.5
5	*Dense*	128	*ReLU*	-
6	*Dropout*	-	-	0.5
7	*Dense*	4	*Softmax*	-

3. Resultados

A Figura 2 mostra o gráfico da acurácia e da matriz de confusão resultante.

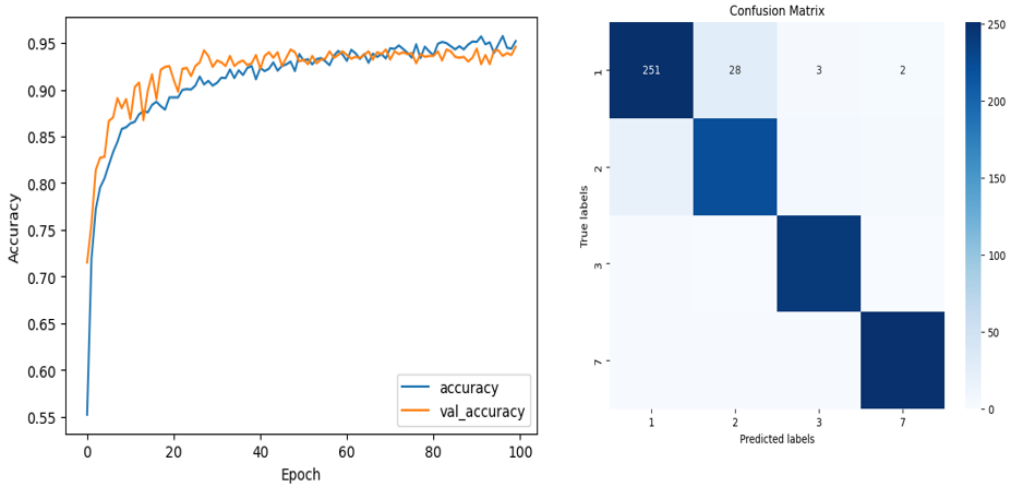

Figura 2. Acurácia e matriz de confusão.

3.1. Aplicação de Coeficiente de V de *Cramer*.

Para verificação da associação das preditoras X variável alvo utilizou-se a função *V* de C*ramer* implementada em *python* por meio da função *chi2_contingency* do pacote *scipy.stats*, para entendimento relacional das variáveis e posterior aplicação de seleção.

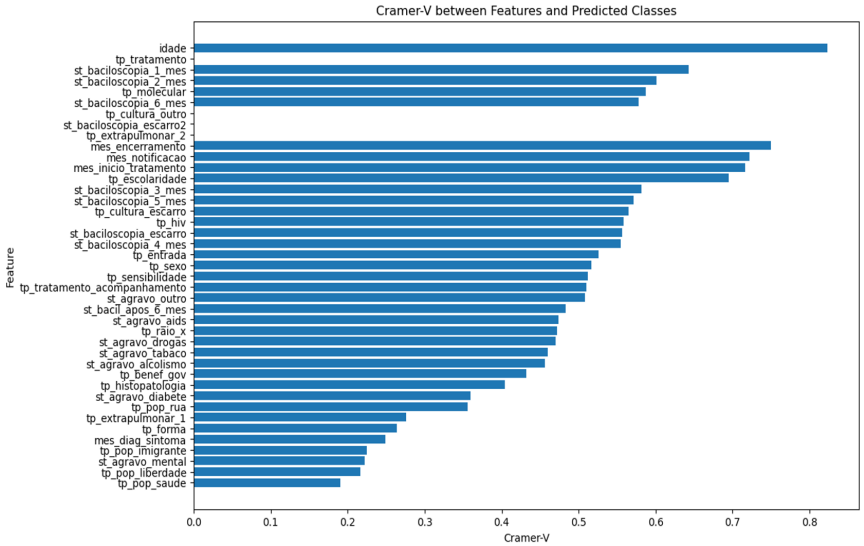

Figura 3. Gráfico resultante da associação das preditoras e variável alvo.

É possível observar que algumas preditoras não possuem associação alguma com a variável alvo.

3.2. Aplicação da biblioteca *permutation importance*.

Esta biblioteca pertence ao pacote *sklearn.inspection*, por meio dela foi possível aplicar o embaralhamento nos valores de cada variável preditora para avaliar o impacto no desempenho do modelo. Dessa forma, o valor da importância varia de 0 a 1, e que, quanto maior for a alteração causada no desempenho do modelo, maior é a importância da variável.

Na Figura 4, são apresentados os gráficos da acurácia e da importância atribuída às variáveis preditoras. Já na Figura 5, é ilustrada a aplicação da técnica *Permutation Feature Importance*, na qual, ao treinar o modelo com um limiar de importância das características estabelecido em 0,0011, observou-se uma melhoria na acurácia, que aumentou de 92,8% para 93,31%. O limiar de 0,0011 implica na seleção das características cuja importância seja igual ou superior a esse valor.

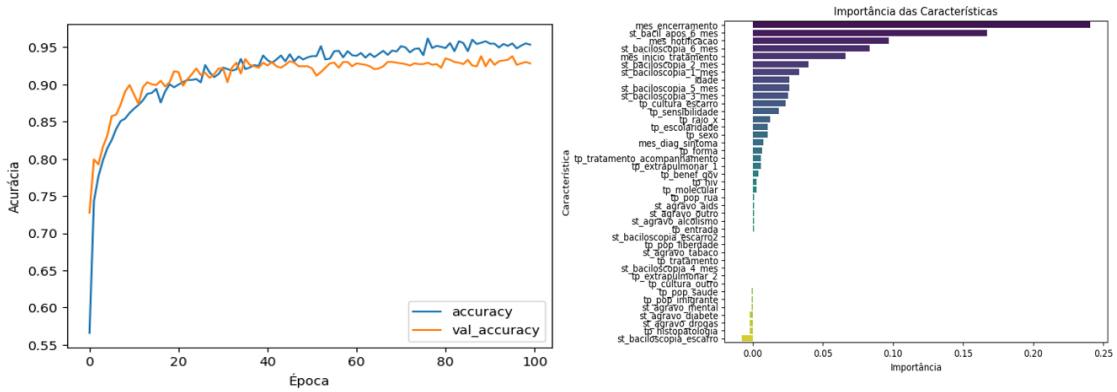

Figura 4. Gráfico da curácia e importância das preditoras na escala de 0 a 0,25.

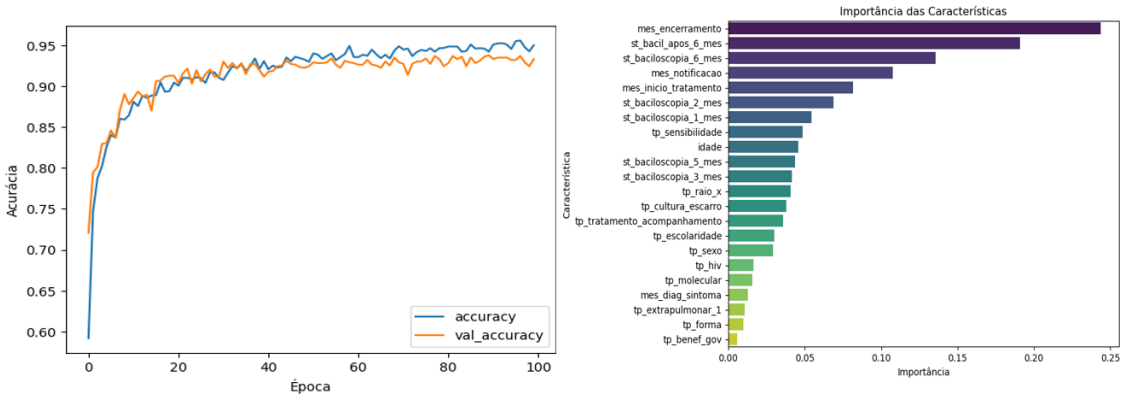

Figura 5. Seleção das características com importâcia maior ou igual a 0,0011.

4. Conclusão

O experimento com dados do SINAN procedentes do iTB, resultou em um modelo que prevê *status* do encerramento do tratamento de tuberculose. Com acurácia de aproximadamente 94% para os desfechos de cura, abandono, morte por TB e droga resistente, elaborou-se um dashboard com as previsões do mês de janeiro a 20 de março de 2024.

Pretende-se utilizar do iTB, dados de acompanhamento onde constam o quantitativo de consultas de cada paciente, números e acompanhamentos de contatos e dados de geolocalização que podem convergir para certas características relacionadas a região e condições socioeconômicas.

Pretende-se também estudar a relação dinâmica de reincidência de tuberculose utilizando a inteligência artificial para descobertas de padrões com dados relevante do contanto e do caso índice.

Referências

Acharya, V., Dhiman, G., Prakasha, K., Bahadur, P., Choraria, A., Prabhu, S., Chadaga, K., Viriyasitavat, W., & Kautish, S. (2022). AI-Assisted Tuberculosis Detection and Classification from Chest X-Rays Using a Deep Learning Normalization-Free Network Model. *Hindawi Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/2399428

Bahaweres, R. B., Agustian, F., Hermadi, I., Suroso, A. I., & Arkeman, Y. (2020). Software defect prediction using neural network based smote. *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, *2020-October*. https://doi.org/10.23919/EECSI50503.2020.9251874

Inacio B, A., Henrique Q, L., Arruda, I., Carminati S, P., Caroline C V, A., & Maia M S, C. (2023). *View of Predição do desfecho de pacientes com Tuberculose*. REVISTA OBSERVATORIO DE LA ECONOMIA LATINOAMERICANA. https://doi.org/10.55905/oelv21n9-094

Murrugara, L. K. S., Lima, L. V., Pavinati, G., Silva, I. G. P., Monteiro, L. R. S., Gil, N. L. M., & Magnabosco, G. T. (2023). Desfecho do tratamento da tuberculose latente em pessoas vivendo com vírus da imunodeficiência humana. *Revista de Saúde Pública do Paraná*, *6*(1), 1–15. https://doi.org/10.32811/25954482-2023v6n1.709

Perlaza, C. L., Mosquera, F. E. C., Murillo, L. M. R., Sepulveda, V. B., & Arenas, C. D. C. (2023). Factors of abandonment of tuberculosis treatment in the public health network. *Revista de Saude Publica*, *57*. https://doi.org/10.11606/S1518-8787.2023057004454

Singh, M., Pujar, G. V., Kumar, S. A., Bhagyalalitha, M., Akshatha, H. S., Abuhaija, B., Alsoud, A. R., Abualigah, L., Beeraka, N. M., & Gandomi, A. H. (2022). Evolution of Machine Learning in Tuberculosis Diagnosis: A Review of Deep Learning-Based Medical Applications. *Electronics 2022, Vol. 11, Page 2634*, *11*(17), 2634. https://doi.org/10.3390/ELECTRONICS11172634

Tabosa de Oliveira, T., Sampaio, V., & Endo, P. T. (2020). Configuração de hyper-parâmetros de modelos deep learning para auxílio no pós-diagnóstico de Tuberculose. *Anais Estendidos do Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, 87–92. https://doi.org/10.5753/SBCAS.2020.11563

Zhang, J., Qiang, Y., & Xu, X. (2021). Detection of Tuberculosis based on Deep Learning based methods You may also like An Ultrasensitive Electrochemical Aptasensor for Thrombin Detection Using MoS 2 Nanoparticles Loaded Iron-Porphyrinic Metal-Organic Framework as Signal Amplifier. *J. Phys.: Conf. Ser*, *1767*, 12004. https://doi.org/10.1088/1742-6596/1767/1/012004