IX Jornada Científica do PRODER
II Conferência Internacional de Saúde e Desenvolvimento Sustentável da UFCA
17 a 19 de Novembro de 2025

# Aplicação de Técnicas de *Ensemble Learning* na Previsão da Incidência de Tuberculose

*José Wivo Gomes (UFCA – wivo.gomes@aluno.ufca.edu.br)*
*Jair Paulino de Sales (UFCA – jair.paulino@ufca.edu.br)*

**RESUMO:** A tuberculose (TB), causada pela bactéria *Mycobacterium tuberculosis*, permanece como um dos principais desafios da saúde pública mundial, devido às elevadas taxas de incidência e mortalidade, especialmente em países em desenvolvimento. Nesse contexto, a aplicação de técnicas de *machine learning* (ML) em séries temporais representa uma abordagem importante para modelar e prever a evolução da doença a fim de subsidiar o planejamento de ações preventivas mais eficazes. O presente estudo teve como objetivo avaliar diferentes abordagens de ML, individuais e *ensembles*, para a previsão da incidência de tuberculose na região Nordeste do Brasil, identificando combinações de modelos que proporcionem maior acurácia e menor erro de previsão em séries temporais mensais de casos confirmados. A pesquisa, de natureza aplicada e abordagem quantitativa, utilizou dados do Sistema de Informação de Agravos de Notificação (SINAN) e do Departamento de Informática do Sistema Único de Saúde (DATASUS), obtidos por meio da ferramenta TABNET. Os registros dos nove estados nordestinos foram consolidados em uma série temporal mensal representando o total de casos confirmados. O processamento e a modelagem foram realizados em Python (versão 3.11.13), utilizando janelas de 12 defasagens (lags) para capturar as dependências temporais da série. Foram avaliados quatro modelos individuais de ML: *Support Vector Regression* (SVR), *Multilayer Perceptron* (MLP), *Random Forest Regressor* (RFR) e *Extreme Gradient Boosting* (XGB). Além dos modelos individuais, foram testados diferentes *ensembles*. As combinações foram realizadas considerando o desempenho dos respectivos modelos individuais no conjunto de treinamento. Entre as combinações testadas, o *ensemble* XGB + RFR (dois melhores modelos no treinamento) apresentou MAPE de 7,39%, o ensemble XGB + RFR + SVR (três melhores modelos no treinamento) obteve MAPE de 7,26%, e o ensemble XGB + RFR + SVR + MLP (quatro melhores modelos no treinamento) alcançou o melhor desempenho, com MAPE de 7,22%. Assim, verifica-se que o aumento no número de modelos individuais que compõem o *ensemble* resultou na redução do MAPE. Assim, a inclusão de novos algoritmos de previsão pode melhorar, ainda mais, a performance geral do sistema. Como trabalhos futuros, objetiva-se a construção de *ensembles* baseados em outras técnicas de combinação discutidas na literatura, bem como o desenvolvimento de sistemas de múltiplos preditores mais robustos, considerando o *trade-off* entre diversidade e acurácia do pool de modelos gerado.

**Palavras-chave:** tuberculose; series temporais; modelos preditivos; ensemble.

## Application of Ensemble Learning Techniques in Forecasting Tuberculosis Incidence

**ABSTRACT:** Tuberculosis (TB), caused by the bacterium Mycobacterium tuberculosis, remains one

of the main global public health challenges due to its high incidence and mortality rates, especially in developing countries. In this context, the application of machine learning (ML) techniques to time series represents an important approach for modeling and predicting the evolution of the disease, supporting the planning of more effective preventive actions. This study aimed to evaluate different ML approaches, both individual and ensemble, for forecasting tuberculosis incidence in the Northeast region of Brazil, identifying model combinations that provide higher accuracy and lower forecasting errors in monthly time series of confirmed cases. The applied, quantitative research used data from the Sistema de Informação de Agravos de Notificação (SINAN) and the Departamento de Informática do Sistema Único de Saúde (DATASUS), obtained via the TABNET tool, consolidating the records from the nine northeastern states into a single monthly time series representing the total number of confirmed cases. Processing and modeling were performed in Python (version 3.11.13), using 12-lag windows to capture temporal dependencies. Four individual ML models were evaluated: Support Vector Regression (SVR), Multilayer Perceptron (MLP), Random Forest Regressor (RFR), and Extreme Gradient Boosting (XGB). In addition, different ensemble combinations were tested based on the performance of the individual models in training. Among the tested combinations, the XGB + RFR ensemble (the two best-performing models in training) achieved a MAPE of 7.39%, the XGB + RFR + SVR ensemble (the three best models) obtained a MAPE of 7.26%, and the XGB + RFR + SVR + MLP ensemble (the four best models) achieved the best performance, with a MAPE of 7.22%. It was observed that increasing the number of individual models composing the ensemble resulted in a reduction of MAPE, indicating that the inclusion of additional forecasting algorithms can further improve overall system performance. Future work will focus on constructing ensembles based on other combination techniques discussed in the literature and developing more robust multiple-predictor systems that consider the trade-off between diversity and accuracy within the generated model pool.

**Keywords:** tuberculosis; time series; predictive models; ensemble.