Forecasting Dengue Outbreaks with Machine Learning: A Case Study in Porto Alegre

Franklin Sales de Oliveira[1], Dulcinéia Esteves Santos[2], Fábio Raphael Pascoti Bruhn[2], Bianca Conrad Bohm[2], Brenda Salenave Santana[1]

[1]Technology Development Center – Federal University of Pelotas
Gomes Carneiro Street, 01 – Pelotas – RS – Brazil

[2]Faculty of Veterinary Medicine – Federal University of Pelotas
Post Office Box 354 – 96160-000 – Capão do Leão – RS - Brazil

{fsoliveira,bssalenave}@inf.ufpel.edu.br, dulcineiaestevessantos@gmail.com

biankabohm@hotmail.com, fabio_rpb@yahoo.com.br

Abstract. *Dengue remains a critical global health concern intensified by climatic changes. Despite advances, few studies validate machine learning models in subtropical regions or handle data imbalance. This work compares Decision Tree, Random Forest, SVM, and XGBoost to predict dengue case acceleration in Porto Alegre, Brazil (2020–2024). Using time-lagged climatic features and Balanced Accuracy optimization, XGBoost achieved the best performance (0.67), identifying 10–15-day averages of dew point and humidity as key predictors. The results highlight the value of imbalance-aware modeling for reliable early warning systems in public health.*

1. Introduction

Dengue represents a significant global health challenge, imposing severe socioeconomic burdens through high hospitalization costs, reduced productivity, and overloaded public health systems [Bertoldo et al. 2025]. With nearly half of the world's population at risk, the World Health Organization estimates 100 to 400 million new infections annually, underscoring the urgent need for effective predictive strategies for disease control. This scenario highlights the magnitude of the problem and reinforces the need for innovative and effective strategies for monitoring, preventing, and controlling the disease.

Furthermore, there is an intrinsic relationship between the number of dengue cases and climatic factors. Anthropogenic environmental changes, such as deforestation, climate change resulting from pollution, and disorderly urbanization, act as predisposing factors that favor vector mosquito distribution, diversity, and abundance. This, in turn, increases the risk of diseases like dengue [De Sousa Azevedo and Araújo 2023].

Machine learning has emerged as a promising approach for predicting dengue outbreaks. Studies in diverse climatic regions, such as Bangladesh, have demonstrated the potential of models like the Support Vector Machine (SVM) [Sarwar and Al Mamun 2022]. However, comparative analyses often indicate that ensemble methods offer superior performance; for instance, a study by [Ong et al. 2023] found that XGBoost outperformed several other algorithms. Despite these advances, two critical gaps persist in the literature: the specific validation of these models in subtropical regions

with distinct seasonality, and the rigorous methodological treatment of the inherent class imbalance found in outbreak prediction data.

In an effort to address existing gaps, we conducted a systematic comparative analysis of four machine learning models (Decision Tree (DT), Random Forest (RF), SVM, and XGBoost (XGB) to predict periods of dengue case acceleration in the subtropical city of Porto Alegre, Brazil. A key aspect of our methodology is the explicit handling of the imbalanced dataset, optimizing the models for Balanced Accuracy to ensure clinical relevance. Our results show that the XGBoost model provided the most effective predictions, achieving the highest recall and balanced accuracy. Furthermore, the analysis revealed that time-lagged climatic variables, specifically the 10- to 15-day moving averages of dew point and humidity, are the most critical predictors, validating the model's ability to capture the underlying biological drivers of the disease. To ensure reproducibility, all code and data are publicly available[1].

2. Methodology

The study is based on a unified dataset constructed from two official sources. Climatic variables (Table 1) were obtained from the National Institute of Meteorology (*INMET*), using records from a meteorological station in Porto Alegre. Concurrently, epidemiological records were extracted from the Notifiable Diseases Information System (*SINAN*). A critical preprocessing step was applied to the raw *SINAN* data to filter and consolidate only confirmed dengue cases, which were then aggregated into daily counts. The two datasets were integrated based on the notification date using the ArbovirusFramework [de Oliveira et al. 2025], a framework for manipulating and processing CSV data, resulting in a final dataset with 1,814 daily records for Porto Alegre, covering the period from 2020-01-01 to 2024-12-18.

Table 1. Description of the base features used in the model.

Variable (Feature)	Description	Unit
`precipitation`	Daily accumulated rainfall.	mm
`dew_point`	Air saturation temperature with water vapor.	°C
`mean_temperature`	Daily average air temperature.	°C
`humidity`	Average relative humidity of the air.	%
`season`	Season of the year corresponding to the date.	Categorical

Feature engineering was essential for preparing the data for modeling. To capture the biological and temporal dynamics of dengue transmission, two main techniques were applied. First, to incorporate climatic history, *time-lagged variables* were created. These variables, computed as cumulative sums and moving averages over 5-, 10-, and 15-day windows, reflect the biological delays inherent in the process, including the mosquito vector's life cycle and the virus incubation period in humans. According to the World Health Organization (WHO), dengue symptoms typically appear 4 to 10 days after the bite of an infected mosquito. Second, the categorical variable *season* underwent *one-hot encoding* to convert it into a numerical format.

[1]Available at: `https://github.com/LEPIVET`

The target variable, *cases_increase*, was defined as a binary label (1 for increase, 0 for no increase) by comparing the number of cases on a given day with that of 15 days prior. This approach guides the models to focus on predicting periods of disease acceleration rather than absolute case counts, which are more relevant for early warning systems. The dataset was split into training (70%) and testing (30%) sets using stratified sampling to preserve the class proportion of the target variable, a crucial step for imbalanced data. Model optimization was performed via *five-fold stratified cross-validation* on the training set only. Grid Search was used to identify the optimal hyperparameter combination for each algorithm (DT, RF, SVM, and XGB).

All modeling steps were implemented using the *scikit-learn*[2] and *XGBoost*[3] libraries. The primary metric for hyperparameter optimization and model selection was *Balanced Accuracy*, chosen for its effectiveness in providing a fair evaluation for imbalanced data. For a more detailed diagnosis, especially regarding the minority class (days with increasing cases), *F1-score, Precision, and Recall* was also computed and analyzed on the test set.

3. Results

The performance evaluation of each model was carried out on the test dataset, which remained untouched during the training and hyperparameter optimization stages. Table 2 summarizes their performance on the test set. The evaluation metrics *F1-score*, *Precision*, and *Recall* refer specifically to the minority class (class 1), i.e., days with increasing cases.

Table 2. Performance of the Models on the Test Set (2020-2024)

Model	Balanced Accuracy	F1-score	Precision	Recall
DT	0.6513	0.48	0.56	0.42
RF	0.6485	0.47	0.60	0.39
SVM	0.6685	0.51	0.49	0.54
XGB	0.6738	0.52	0.46	0.60

The feature importance analysis for the tree-based models (DT, RF, and XGB) consistently highlighted a common set of primary predictors, with the 10- and 15-day moving averages of dew point and humidity dominating the top of the ranking across all algorithms. Despite this agreement in ranking, the distribution of importance scores reflects the unique architecture of each algorithm. The Decision Tree, which computes importance through impurity reduction (Gini Importance), concentrated most of its relevance on a few key variables. In contrast, the Random Forest, which averages importance scores across hundreds of trees, exhibited a more distributed scoring pattern. XGBoost, in turn, showed the widest distribution, where importance is measured by the sequential Gain each feature contributes to the model, resulting in more moderate contributions from a broader set of variables. This convergence in identifying the same key variables across algorithms with distinct calculation mechanisms provides strong evidence that recent humidity and dew point histories are the primary drivers of dengue case acceleration in the studied region.

[2]Available at: `https://scikit-learn.org/stable/`

[3]Available at: `https://xgboost.readthedocs.io/en/stable/`

4. Conclusion

This study developed and systematically compared four machine learning models to predict periods of dengue case acceleration in Porto Alegre, Brazil. The results demonstrate that the XGBoost model exhibited superior performance, achieving the highest Balanced Accuracy (0.6738) and Recall (0.60), highlighting its effectiveness in minimizing false negatives. Furthermore, a feature importance analysis consistently identified time-lagged climatic variables, specifically the 10- to 15-day moving averages of dew point and humidity, as the most dominant predictors across all tree-based models. This finding aligns with the strong influence of weather on the dengue transmission cycle.

The primary contribution of this work lies in its methodological rigor applied to a specific geographical context. By explicitly addressing the challenge of class imbalance and optimizing for relevant metrics like Balanced Accuracy, this study provides a reliable work for epidemiological forecasting in Porto Alegre. The findings have direct implications for public health, validating that a concise set of meteorological indicators can drive an effective and automated early warning system, enabling proactive resource allocation and vector control measures.

Future work should aim to incorporate additional data streams, including socioeconomic indicators, human mobility patterns, and information from multiple cities to enhance generalization across the subtropical climate of Southern Brazil. Additionally, exploring sequential models such as Long Short-Term Memory (LSTM) networks and developing ensemble methods that integrate the strengths of the top-performing algorithms could further improve predictive accuracy and robustness, enabling more comprehensive and reliable dengue early warning systems.

References

Bertoldo, S. O. L., Moreira Jorcelino, T., Filho, L. A. d. S., Viana, H. T. d. O., and Almeida, F. C. S. d. (2025). Modelo de plano integrado de vigilância em saúde aplicado à dengue. *Gestão & Cuidado em Saúde*, 3(1):e13893. Acesso em: 26 ago. 2025.

de Oliveira, F. S., Santos, D. E., Bohm, B. C., and Santana, B. S. (2025). Framework integrado para análise de dados climáticos e epidemiológicos: Potencial para monitoramento. In *Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais (WCAMA)*, pages 340–343. SBC.

De Sousa Azevedo, M. L. and Araújo, M. A. P. (2023). Consequências de impactos ambientais na saúde humana: uma análise estatística dos casos de dengue no estado do rio de janeiro. *CONTRIBUCIONES A LAS CIENCIAS SOCIALES*, 16(10):18835–18846.

Ong, S. Q., Isawasan, P., Ngesom, A. M. M., Shahar, H., Lasim, A. m. M., and Nair, G. (2023). Predicting dengue transmission rates by comparing different machine learning models with vector indices and meteorological data. *Scientific reports*, 13(1):19129.

Sarwar, M. T. and Al Mamun, M. (2022). Prediction of dengue using machine learning algorithms: Case study dhaka. In *2022 4th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*, pages 1–6. IEEE.