

Aprendizado de Máquina para Predição de Chikungunya no Brasil: Estudo de Caso em Manaus (AM)

Lorena F. Souza¹, Giovanni E. Zanardo¹,
Fabíola G. Nakamura¹, Eduardo F. Nakamura¹

¹Instituto de Computação (IComp) – Universidade Federal do Amazonas (UFAM)
Av. Gen. Rodrigo Octávio, 6200, Coroado I, Setor Norte
69080-900 – Manaus – AM – Brasil

{lorena.flavia,gio.zanardo,fabiola,nakamura}@icomp.ufam.edu.br

Abstract. *Neglected Tropical Diseases represent a major public health challenge in tropical regions. In Brazil, chikungunya stands out due to its recurring epidemiological impact in the Northern Region, where climatic and socio-environmental factors favor vector proliferation. This paper presents partial results of a study applying machine learning models to the weekly forecasting of chikungunya cases using InfoDengue data (2010–2025). The Manaus time series comprises 810 epidemiological weeks (mean 3.2 cases/week, peak 62 cases). The optimized LightGBM achieved RMSE of 3.84 and MAE of 3.01, while the multivariate configuration without optimization obtained RMSE of 3.78 and MAE of 2.99. Results validate the proposed pipeline and support expansion toward multi-step models, deep learning, ensembles, and all 27 Brazilian federative units.*

Keywords: *Neglected tropical diseases; Chikungunya; Machine learning; Time series forecasting; Gradient boosting.*

Resumo. *As Doenças Tropicais Negligenciadas representam um desafio à saúde pública em regiões tropicais. No Brasil, a chikungunya destaca-se pelo impacto epidemiológico recorrente na Região Norte, onde fatores climáticos e socioambientais favorecem a proliferação do vetor. Este artigo apresenta resultados parciais de um estudo que aplica modelos de aprendizado de máquina à previsão semanal de casos de chikungunya usando dados da plataforma InfoDengue (2010–2025). A série temporal de Manaus compreende 810 semanas epidemiológicas (média de 3,2 casos/semana, pico de 62 casos). O LightGBM otimizado alcançou RMSE de 3,84 e MAE de 3,01; a configuração multivariada sem otimização obteve RMSE de 3,78 e MAE de 2,99. Os resultados validam o pipeline proposto e fundamentam a expansão para modelos multi-step, aprendizado profundo, ensembles e as 27 unidades federativas brasileiras.*

Palavras-chave: *Doenças tropicais negligenciadas; Chikungunya; Aprendizado de máquina; Previsão de séries temporais; Gradient boosting.*

1. Introdução

As Doenças Tropicais Negligenciadas (DTNs) incidem predominantemente sobre populações fragilizadas em zonas tropicais e subtropicais [Rosário et al. 2017], respondendo por estimativa de 500 mil a 1 milhão de mortes anuais na América do Sul

[Organização Pan-Americana da Saúde (OPAS) 2024]. Entre elas, destaca-se a chikungunya, arbovirose causada pelo vírus CHIKV (*Togaviridae*), transmitida pelos mosquitos *Aedes aegypti* e *Aedes albopictus*, com quadro clínico que evolui em fase aguda (febre e artralgia intensa) e fase crônica (dor articular persistente por semanas ou anos) [Castro et al. 2016]. Na Região Norte do Brasil, fatores climáticos e socioeconômicos favorecem surtos periódicos e dificultam respostas rápidas da vigilância epidemiológica.

Pesquisas recentes demonstram o potencial do Aprendizado de Máquina (AM) na estimação da incidência de arboviroses a partir de variáveis climáticas e histórico de notificações [Cabrera et al. 2022, Salim et al. 2021]. Este trabalho aplica modelos de AM à predição semanal de casos de chikungunya em Manaus (AM), como estudo de caso inicial. As contribuições incluem: (i) um *pipeline* reproduzível de coleta e modelagem de séries temporais epidemiológicas; (ii) caracterização quantitativa da base de dados; (iii) avaliação de LightGBM e XGBoost nas configurações *one-step* univariada e multivariada com reporte dos intervalos de busca em grade; e (iv) discussão sobre o impacto das variáveis climáticas e padrões observados na otimização.

2. Trabalhos Relacionados

Cabrera et al. [Cabrera et al. 2022] revisaram sistematicamente modelos preditivos de dengue na América Latina, mostrando que variáveis climáticas combinadas ao histórico de notificações produzem previsões acuradas. Salim et al. [Salim et al. 2021] demonstraram que métodos de *Gradient Boosting* superam abordagens estatísticas tradicionais em cenários de não linearidade. Sebastianelli et al. [Sebastianelli et al. 2024] evidenciaram que a combinação de múltiplos preditores reduz variância e melhora a generalização na previsão de surtos de dengue. Chowdhury et al. [Chowdhury et al. 2022] aplicaram SHAP para identificar fatores de maior influência em modelos preditivos de arboviroses. Diferentemente desses trabalhos, o presente estudo concentra-se na chikungunya no Brasil, com foco em Manaus (AM) como caso inicial e expansão prevista para as demais capitais da Região Norte.

3. Metodologia

3.1. Base de Dados e Pré-processamento

Os dados foram obtidos via API pública da plataforma InfoDengue [InfoDengue], cobrindo registros semanais de casos confirmados de chikungunya de 2010 a 2025 nas sete capitais da Região Norte (Acre, Amapá, Amazonas, Pará, Rondônia, Roraima e Tocantins). Para os experimentos iniciais, adotou-se Manaus (AM) como estudo de caso.

Caracterização da série (Manaus): A série compreende 810 semanas epidemiológicas (janeiro/2010 a março/2025), divididas em 729 semanas de treino (90%) e 81 de teste (10%). A média é de 3,2 casos/semana (desvio-padrão: 5,8), com mediana de 1 e pico de 62 casos (2016). Aproximadamente 38% das semanas apresentam zero ou um caso, configurando série esparsa com picos bem definidos em 2016 e 2022–2023. A taxa de missings nas variáveis meteorológicas foi inferior a 2%, tratada por interpolação temporal.

As *features* utilizadas foram: número semanal de casos (*casos*), temperatura média (*tempmed*) e umidade relativa média (*umidmed*), escolhidas por sua relação com

a dinâmica de transmissão do CHIKV, dado que fatores ambientais influenciam o ciclo de vida do vetor e a disseminação viral [Salvi et al. 2021]. O pré-processamento incluiu: reorganização cronológica; remoção de colunas com 100% de valores ausentes ou apenas zeros; tratamento de nulos por interpolação e imputação (média, mediana e moda); e conversão de tipos de dados. O *pipeline* completo é ilustrado na Figura 1.

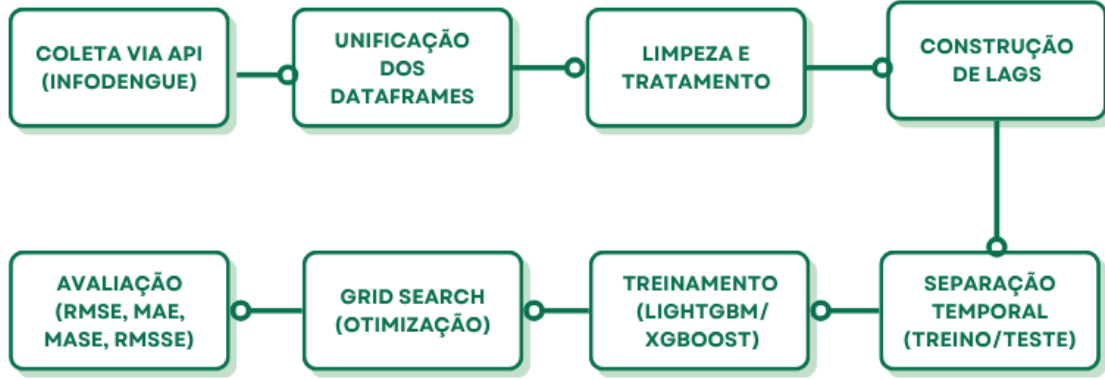


Figura 1. Pipeline metodológico do estudo. Fonte: Autores (2026).

3.2. Modelagem

Foram construídas variáveis defasadas (*lags*) por janelas temporais deslizantes de $T=10$ semanas — janela suficiente para capturar tendências recentes, sazonalidade de curto prazo e atrasos da influência climática, sem excesso de ruído. Do ponto de vista epidemiológico, o ciclo de vida do vetor e o período de incubação do vírus sugerem impacto climático escalonado ao longo de semanas subsequentes, adequadamente coberto por $T=10$. Nos modelos *one-step*, prevê-se o número de casos na semana $t+1$ a partir das informações disponíveis até t .

A divisão treino/teste (90%/10%) respeitou a ordem cronológica, sem embaralhamento. A otimização foi conduzida por busca em grade com validação cruzada temporal (`TimeSeriesSplit`, 5 subdivisões), garantindo que cada partição de validação contenha apenas observações posteriores ao conjunto de treinamento correspondente.

Foram avaliados dois algoritmos de *gradient boosting*: LightGBM e XGBoost [Chen and Guestrin 2016], ambos nas configurações univariada (`casos`) e multivariada (`casos`, `tempmed`, `umidmed`).

Intervalos de busca em grade (LightGBM): `num_leaves` $\in \{5, 15, 31\}$; `n_estimators` $\in \{100, 200, 300\}$; `learning_rate` $\in \{0,01, 0,05, 0,1\}$; `max_depth` $\in \{-1, 5, 10\}$; `feature_fraction` $\in \{0,6, 0,8, 1,0\}$; `lambda_l1` $\in \{0, 0,1, 1,0\}$.

3.3. Métricas de Avaliação

O desempenho foi avaliado por MAE e RMSE [Hyndman and Athanasopoulos 2018]:

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t|, \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2} \quad (1)$$

onde y_t é o valor observado e $\hat{y}_t$ a predição na semana t . O RMSE penaliza mais severamente erros de maior magnitude, relevante em contexto epidemiológico pois grandes desvios em surtos representam falhas críticas. Complementarmente, foram calculados MASE e RMSSE [Hyndman and Athanasopoulos 2018], em que valores inferiores a 1 indicam desempenho superior ao *baseline* ingênuo.

4. Resultados Parciais

4.1. LightGBM

O modelo *one-step* univariado obteve, na configuração inicial, RMSE de 4,01 e MAE de 3,19. Após otimização, os melhores hiperparâmetros foram: `feature_fraction: 0,8`; `lambda_l1: 0,1`; `learning_rate: 0,01`; `max_depth: -1`; `n_estimators: 100`; `num_leaves: 5` — reduzindo os erros para RMSE de 3,84 e MAE de 3,01.

A configuração multivariada sem otimização obteve RMSE de 3,78 e MAE de 2,99, melhor desempenho no conjunto de teste. Com busca em grade, os valores foram RMSE de 4,24 e MAE de 3,35. Esse resultado contraintuitivo tem explicação metodológica: a busca em grade seleciona hiperparâmetros pelo desempenho médio em 5 partições de validação cruzada temporal, favorecendo configurações mais conservadoras e generalizáveis. O conjunto de teste, com apenas 81 semanas (10% da série), introduz variância de avaliação não negligenciável em séries com picos episódicos — o modelo sem otimização pode ter se beneficiado de ajuste fortuito a esse recorte específico. O modelo otimizado, por sua vez, apresenta maior estabilidade entre partições temporais, característica mais relevante para uso prospectivo em vigilância epidemiológica. A incorporação de variáveis climáticas foi benéfica, sugerindo que temperatura e umidade carregam sinal preditivo relevante, possivelmente por capturar relações com o ciclo do vetor ausentes no histórico de casos isolado. A Figura 2 apresenta a comparação entre valores reais e previstos no período de teste para o modelo multivariado sem otimização (melhor RMSE).

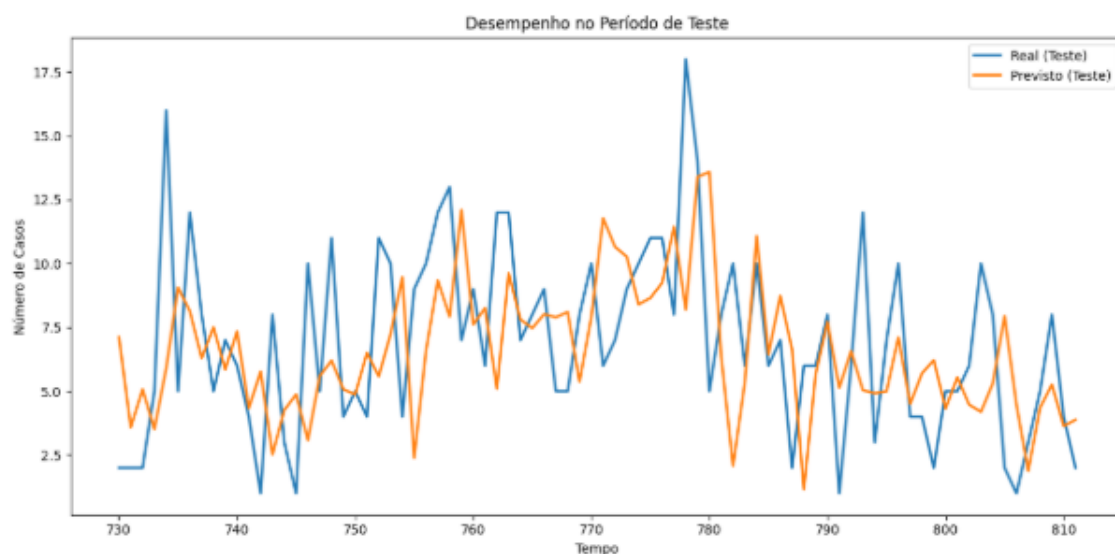


Figura 2. Valores reais vs. previstos no período de teste – LightGBM multivariado (RMSE = 3,78). Semanas 730–810 (2023–2025). Fonte: Autores (2026).

4.2. XGBoost

O XGBoost *one-step* univariado obteve RMSE de 4,13 e MAE de 3,11, desempenho comparável ao LightGBM na configuração inicial. A análise comparativa sistemática entre os dois algoritmos — considerando erro, generalização e estabilidade — será conduzida na etapa subsequente, com busca em grade idêntica à realizada para o LightGBM.

4.3. Síntese Comparativa

A Tabela 1 sumariza todos os resultados. Em todos os cenários, os erros permaneceram entre 3 e 5 casos semanais, coerente com a média de 3,2 casos/semana e o padrão de baixa incidência predominante em Manaus. Valores de RMSE ≈ 4 representam, contudo, erro relativo considerável durante picos (até 62 casos), indicando que os modelos tendem a subestimar surtos — limitação esperada em configurações *one-step*.

Tabela 1. Resultados no conjunto de teste (Manaus, AM; 81 semanas).

Modelo	Configuração	RMSE	MAE
LightGBM	One-step univariado	4,01	3,19
LightGBM	One-step univariado + <i>grid search</i>	3,84	3,01
LightGBM	One-step multivariado	3,78	2,99
LightGBM	One-step multivariado + <i>grid search</i>	4,24	3,35
XGBoost	One-step univariado	4,13	3,11

Os resultados confirmam a aplicabilidade do *pipeline* e evidenciam que variáveis climáticas podem melhorar o desempenho preditivo, em consonância com a literatura [Salvi et al. 2021]. O comportamento do modelo otimizado multivariado aponta para a necessidade de refinamento do espaço de busca, possivelmente com faixas maiores de regularização ou maior número de estimadores.

5. Trabalhos Futuros

Com base nos resultados, o projeto avançará em quatro eixos: (i) Modelos *multi-step*: arquiteturas que estimam múltiplos horizontes temporais para subsidiar planejamento preventivo; (ii) Aprendizado Profundo: modelos LSTM e GRU para capturar dependências de longo prazo entre variáveis epidemiológicas, climáticas e demográficas; (iii) *Ensembles* Clássicos: experimentação com AR, MA, ARIMA, ETS e SARIMAX em combinações via média simples, ponderada e estratégias avançadas [Sebastianelli et al. 2024]; e (iv) Expansão e Explicabilidade: extensão às demais capitais da Região Norte e às 27 UFs brasileiras, aplicação de SHAP [Chowdhury et al. 2022] para identificar covariáveis relevantes por região, e integração a um *dashboard* interativo para apoio à decisão em saúde pública.

6. Conclusão

Este trabalho apresentou resultados parciais de um estudo voltado à predição semanal de casos de chikungunya em Manaus (AM), como estudo de caso inicial. A série analisada compreende 810 semanas (2010–2025), com média de 3,2 casos/semana e pico de 62 casos. O *pipeline* proposto — coleta via API InfoDengue, pré-processamento, janelas de $T=10$ semanas e validação cruzada temporal — demonstrou robustez e reprodutibilidade.

LightGBM e XGBoost produziram RMSE entre 3,78 e 4,24, validando a abordagem para séries temporais epidemiológicas. A incorporação de variáveis climáticas evidenciou potencial de melhora preditiva, enquanto o comportamento do modelo otimizado multivariado aponta para a necessidade de refinamento do espaço de busca. Os próximos passos incluem modelos *multi-step*, aprendizado profundo, *ensembles* e expansão nacional, visando contribuir ao avanço científico na interface entre ciência de dados e epidemiologia e ao aprimoramento da vigilância epidemiológica no Brasil.

Referências

- Cabrera, M. et al. (2022). Dengue prediction in latin america using machine learning and the one health perspective: a literature review. *Tropical Medicine and Infectious Disease*, 7(10):322.
- Castro, A. P. C. R. d., Lima, R. A., and Nascimento, J. d. S. (2016). Chikungunya: a visão do clínico de dor. *Revista Dor*, 17:299–302.
- Chen, T. and Guestrin, C. (2016). Xgboost: a scalable tree boosting system. In *ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 785–794, New York. ACM.
- Chowdhury, S. U. et al. (2022). Shapley-additive-explanations-based factor analysis for dengue severity prediction using machine learning. *Journal of Imaging*, 8(9):229.
- Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts, Melbourne, 2 edition.
- InfoDengue. Infodengue – serviços de api. Disponível em: <https://info.dengue.mat.br/services/api>. Acesso em: 15 fev. 2026.
- Organização Pan-Americana da Saúde (OPAS) (2024). Doenças tropicais negligenciadas: Dia mundial chama atenção para fortalecimento da resposta na região das américas. Disponível em: <https://www.paho.org/pt/noticias/>. Acesso em: 29 mar. 2025.
- Rosário, M. S. et al. (2017). Doenças tropicais negligenciadas: caracterização dos indivíduos afetados e sua distribuição espacial. *Revista Brasileira de Pesquisa em Saúde*, 19(3):118–127.
- Salim, N. A. M. et al. (2021). Prediction of dengue outbreak in selangor malaysia using machine learning techniques. *Scientific Reports*, 11(1):939.
- Salvi, F. I. et al. (2021). Environmental and climatic factors associated with the *Aedes aegypti* occurrence. *Research, Society and Development*, 10(9):e56410918544.
- Sebastianelli, A. et al. (2024). A reproducible ensemble machine learning approach to forecast dengue outbreaks. *Scientific Reports*, 14:3807.