

Um Estudo Comparativo de Modelos de Regressão Linear e Aprendizado de Máquina para Predição de Casos de Dengue nos Biomas Brasileiros

Vinícius Souza Pereira de Lima¹, Andréza Leite de Alencar¹

¹Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)
Recife, PE, Brasil

{vinicius.plima, andreza.leite}@ufrpe.br

Abstract. *Dengue had a record epidemic in the Americas in 2024. This study compares five ML algorithms and two baselines (linear regression and seasonal naive) for weekly dengue prediction in four Brazilian biomes. Using climatic and autoregressive features, validation was performed for 2024 (epidemic) and 2025 (post-epidemic). MLR performed best in 2024 (RMSE=26.57, $R^2=0.954$), while in 2025 Random Forest had the lowest median RMSE (1.43). Ablation of autoregressive features severely degraded performance. Outbreak detection sensitivity exceeded 91% for all supervised models. These preliminary results suggest linear regression was more robust to the extreme 2024 epidemic. Further work will include more municipalities.*

Resumo. *A dengue teve epidemia recorde nas Américas em 2024. Este estudo compara cinco algoritmos de ML e dois baselines (regressão linear e modelo ingênuo) na predição semanal de dengue em quatro biomas brasileiros. Usando dados climáticos e autorregressivos, a validação foi feita para 2024 (epidemia) e 2025 (pós-epidemia). A RLM teve melhor desempenho em 2024 (RMSE=26,57; $R^2=0,954$), enquanto em 2025 Random Forest obteve a menor mediana de RMSE (1,43). A ablação das variáveis autorregressivas degradou severamente o desempenho. A sensibilidade na detecção de surtos foi superior a 91%. Estes resultados preliminares sugerem que a regressão linear foi mais robusta à epidemia extrema de 2024. O estudo será expandido.*

1. Introdução

A dengue é a arbovirose de maior impacto global, com cerca de metade da população mundial em risco [Bhatt et al. 2013]. Nas Américas, 2024 foi o ano de maior epidemia registrada, com mais de 8,2 milhões de casos, e o Brasil respondeu pela maior parcela [Rodriguez-Morales et al. 2024, Al Bannoud et al. 2026]. Apesar dos avanços em modelos preditivos baseados em aprendizado de máquina (ML) [Roster et al. 2022, Zanardo et al. 2025, Stolerman et al. 2019], a maioria dos estudos agrega resultados nacionalmente, sem explorar o desempenho em diferentes biomas ou em contextos de epidemia extrema versus pós-epidemia. Esta lacuna é crítica, pois modelos baseados em árvores possuem capacidade limitada de extrapolação, tendendo a produzir previsões dentro da faixa observada no treinamento. Este artigo apresenta resultados parciais de uma pesquisa em andamento, com o objetivo de avaliar a viabilidade da metodologia em uma

amostra inicial de quatro municípios, servindo como prova de conceito para estudos futuros com maior abrangência.

Diante do exposto, este estudo compara cinco algoritmos de ML (Random Forest, XGBoost, LightGBM, SVM-RBF e rede neural) e dois baselines (Regressão Linear Múltipla – RLM – e modelo ingênuo sazonal) na predição semanal de dengue em Belém (Amazônia), Brasília (Cerrado), São Paulo (Mata Atlântica) e Porto Alegre (Pampa), avaliando o desempenho em 2024 (epidemia histórica) e 2025 (pós-epidemia).

2. Materiais e Métodos

2.1. Dados

Utilizaram-se dados diários do INMET (temperatura, precipitação, umidade, vento) agregados em semanas epidemiológicas (domingo a sábado) para quatro municípios: Belém/PA (Amazônia, Af), Brasília/DF (Cerrado, Aw), São Paulo/SP (Mata Atlântica, Cfa) e Porto Alegre/RS (Pampa, Cfa).

A incidência semanal foi calculada como o número de casos notificados por 100.000 habitantes. Os casos foram obtidos do Sistema de Informação de Agravos de Notificação (SINAN) para o período de 2014 a 2025. As estimativas populacionais foram obtidas do Instituto Brasileiro de Geografia e Estatística e do Tribunal de Contas da União (IBGE/TCU), com interpolação linear para os anos sem estimativa disponível (2022 e 2023). A variável resposta foi transformada por $\ln(1 + \text{incidência})$ para estabilizar a variância [Al Bannoud et al. 2026].

Municípios com mais de 5% de valores ausentes após a remoção das quatro semanas iniciais foram excluídos. Os códigos municipais do INMET (7 dígitos) foram convertidos para o formato IBGE de 6 dígitos, e os nomes foram normalizados por meio de expressões regulares e correções manuais para permitir o merge com a base do SINAN.

2.2. Engenharia de Variáveis (features)

A partir dos dados coletados, foram criadas as variáveis: lags climáticos (1–8 semanas) e médias móveis (3–4 semanas) para temperatura, precipitação, umidade e vento; lags da incidência transformada (1–4 semanas) e média móvel de 4 semanas; codificação sazonal cíclica (seno e cosseno da semana) e encoding ordinal do município. O modelo completo utilizou 56 features; o modelo de ablação (apenas clima) removeu as 5 features autorregressivas.

2.3. Modelos e validação

Cinco modelos de ML foram treinados (Random Forest – RF, XGBoost, LightGBM, SVM com kernel RBF, rede neural – ANN) e dois baselines: Regressão Linear Múltipla (RLM) e modelo ingênuo sazonal (que prevê para a semana t do ano y o valor observado na mesma semana t do ano $y - 1$; por exemplo, a semana 10 de 2025 recebe o valor da semana 10 de 2024). A busca de hiperparâmetros para RF, XGBoost e LightGBM usou RandomizedSearchCV com TimeSeriesSplit (5 folds); os valores finais estão no código-fonte. SVM e ANN usaram configurações fixas da literatura [Stolerman et al. 2019, Al Bannoud et al. 2026].

A validação temporal consistiu em: treino (2014–2022, 1.817 obs), validação (2023, 208), teste primário (2024, 208) e teste secundário (2025, 201). As métricas foram RMSE, MAE e R^2 (escala original). Intervalos de confiança de 95% para RMSE foram calculados por block bootstrap circular (blocos de 4 semanas, 1000 iterações) [Politis and Romano 1992]. Para detecção de surtos, definiu-se o percentil 90 da série histórica de cada município (calculado no treino). As previsões contínuas foram binarizadas: semanas com incidência observada ou prevista acima do limiar foram classificadas como epidêmicas. A significância das diferenças foi avaliada pelo teste de Wilcoxon pareado (ANN vs. demais) com correção de Bonferroni para 5 comparações ($\alpha = 0,05/5 = 0,01$) [Wilcoxon 1945].

3. Resultados

3.1. Desempenho global – 2024

A Tabela 1 apresenta o desempenho global em 2024. A RLM obteve o menor RMSE (26,57) e maior R^2 (0,954), superando os demais modelos neste estudo. Entre os modelos de ML, a ANN apresentou o melhor resultado (RMSE = 43,50; $R^2 = 0,875$). Os modelos de árvore (RF, XGBoost, LightGBM) e SVM tiveram RMSE entre 102 e 110 e R^2 entre 0,21 e 0,32. O modelo ingênuo sazonal teve RMSE de 132,91 e R^2 negativo. Os intervalos de confiança indicam menor incerteza para RLM e ANN.

Tabela 1. Desempenho global – Teste A (2024). RMSE e MAE em casos/100k hab.

Modelo	RMSE	IC95% RMSE	MAE	R^2
Reg. Linear	26,57	[16,47–35,04]	11,92	0,954
ANN	43,50	[22,11–60,82]	17,92	0,875
SVM	101,81	[56,63–139,34]	40,46	0,318
LightGBM	107,98	[59,95–149,62]	41,90	0,232
Random Forest	108,60	[60,17–150,18]	42,18	0,224
XGBoost	109,70	[61,14–151,80]	42,67	0,208
Naive Sazonal	132,91	[79,46–178,03]	59,33	-0,163

3.2. Desempenho por Bioma – 2024

A Tabela 2 mostra o RMSE estratificado. Belém teve erros muito baixos (1,6–4,0). Em Brasília e São Paulo, ANN e RLM tiveram erros bem menores (32–78) que os demais (130–172). Em Porto Alegre, todos os erros foram moderados (15–19).

Tabela 2. RMSE por município e bioma – Modelo completo – Teste A (2024). Valores em negrito indicam o menor RMSE por município.

Bioma	Município	ANN	LGBM	RF	SVM	XGB	RLM
Amazônia	Belém	2,71	2,97	2,51	4,01	2,33	1,59
Cerrado	Brasília	78,65	169,70	171,11	155,30	171,59	37,60
Mata Atlântica	São Paulo	32,65	132,54	132,80	130,27	135,86	33,87
Pampa	Porto Alegre	17,65	16,22	15,94	18,83	15,15	16,17

3.3. Inversão de ranking – 2025

A Tabela 3 mostra a mediana do RMSE por modelo nos dois anos. Em 2024, ANN e RLM tiveram as menores medianas de RMSE. No entanto, no RMSE global (Tabela 1), a RLM foi melhor (26,57 vs. 43,50 da ANN). Em 2025, o ranking inverteu-se: Random Forest obteve a menor mediana (1,43), seguido por XGBoost (1,66), RLM (1,67), LightGBM (1,86), ANN (2,02) e SVM (2,09).

Tabela 3. Mediana do RMSE por modelo – Teste A (2024) e Teste B (2025).

Modelo	Mediana RMSE 2024	Mediana RMSE 2025
Reg. Linear	26,57	1,67
ANN	25,15	2,02
Random Forest	74,37	1,43
XGBoost	75,50	1,66
LightGBM	74,38	1,86
SVM	74,55	2,09

3.4. Detecção de surtos

A Tabela 4 apresenta as métricas de classificação de surtos (percentil 90). Todos os modelos supervisionados apresentaram sensibilidade $\geq 91,4\%$, com destaque para Random Forest (96,4%) e Regressão Linear (95,7%). A especificidade variou entre 76,5% (ANN) e 89,7% (Random Forest). O modelo ingênuo sazonal teve sensibilidade muito baixa (29,3%).

Tabela 4. Detecção de surto – Percentil 90 – Teste A (2024).

Modelo	Sensibilidade	Especificidade	PPV	NPV	Acurácia
Random Forest	0,964	0,897	0,951	0,924	0,942
Reg. Linear	0,957	0,897	0,950	0,910	0,938
XGBoost	0,957	0,882	0,944	0,909	0,933
LightGBM	0,950	0,897	0,950	0,897	0,933
ANN	0,936	0,765	0,891	0,852	0,880
SVM	0,914	0,897	0,948	0,836	0,909
Naive Sazonal	0,293	0,912	0,872	0,385	0,495

3.5. Modelo de ablação

Ablação consiste na remoção de um conjunto de variáveis com o objetivo de quantificar sua contribuição preditiva. Neste estudo, removemos as 5 features autorregressivas (lags 1, 2, 3 e 4 da incidência transformada e a média móvel de 4 semanas) para isolar a contribuição exclusiva das variáveis climáticas. Com apenas as variáveis climáticas, acrescentando a codificação sazonal (seno e cosseno da semana) e o identificador numérico do município, todos os algoritmos apresentaram $RMSE > 133$ e R^2 negativo (ex.: LightGBM $RMSE = 133,80$, $R^2 = -0,179$). Esse resultado confirma que as variáveis climáticas isoladamente são insuficientes para predição semanal e que a transmissão passada (os valores de incidência de semanas anteriores) domina o sinal preditivo.

3.6. Análise de resíduos

A estatística de Durbin-Watson foi muito baixa para os modelos de árvore (0,056–0,083), indicando previsões aproximadamente constantes em 2024 – comportamento típico de extrapolação fora do domínio de treinamento. A ANN apresentou $DW = 0,73$, sugerindo maior capacidade de acompanhar a série temporal. O teste de Ljung-Box rejeitou a hipótese de não autocorrelação para todos os modelos ($p < 0,001$).

4. Discussão

O achado mais notável foi a superioridade da RLM sobre todos os modelos de ML em 2024. Possivelmente devido à limitação de extrapolação dos modelos baseados em árvores (propriedade intrínseca desses métodos, demonstrada empiricamente pela análise de resíduos): o Durbin-Watson próximo de zero confirma que esses modelos produziram previsões quase constantes, equivalentes à média do treino, comportamento esperado quando o valor a ser previsto está fora do domínio de treinamento. Em 2025, com a incidência dentro da faixa histórica, os modelos de árvore recuperaram o desempenho esperado, invertendo o ranking. Este resultado tem implicações práticas: para sistemas de vigilância que exigem predições de curto prazo (1 semana), neste conjunto de quatro municípios, os modelos lineares simples mostraram-se mais robustos à mudança extrema de regime observada em 2024, enquanto os modelos baseados em árvores obtiveram melhor desempenho em 2025, quando a incidência na maioria dos municípios retornou à faixa histórica (exceção de Porto Alegre).

5. Limitações e trabalhos futuros

A seleção de apenas um município por bioma foi determinada pela disponibilidade de séries climáticas completas na base do INMET. Os biomas Caatinga e Pantanal não estão representados porque nenhum município dessas regiões apresentou série temporal contínua no período analisado. Os resultados, portanto, não devem ser generalizados para todos os municípios desses biomas.

Apenas quatro municípios (um por bioma) foram analisados nesta versão do trabalho, impossibilitando a validação cruzada espacial e reduzindo a validade estatística. A mediana do RMSE foi calculada sobre apenas quatro observações, não sendo estatisticamente robusta. O modelo ingênuo sazonal em 2025 utiliza como referência o ano epidêmico de 2024, o que pode levar a superestimação sistemática.

Os modelos SVM e ANN não passaram por otimização de hiperparâmetros específica para este conjunto de dados, adotando-se configurações previamente reportadas na literatura [Stolerman et al. 2019, Al Bannoud et al. 2026]. Isso pode ter afetado o desempenho relativo desses algoritmos.

Estudos futuros devem ampliar o número de municípios, incluir novas bases climáticas, variáveis de mobilidade e sorotipos, e explorar horizontes mais longos (2–4 semanas).

6. Conclusão

Este estudo preliminar, baseado em uma amostra de quatro municípios, indicou que a Regressão Linear Múltipla apresentou melhor desempenho que os algoritmos de ML avaliados no ano epidêmico de 2024, enquanto o ranking se inverteu em 2025 com Random

Forest obtendo a menor mediana de RMSE. Recomenda-se que baselines lineares (modelos de referência simples, como regressão linear ou o modelo ingênuo sazonal) sejam sempre incluídos como referência em estudos de predição epidemiológica, pois muitos trabalhos comparam apenas modelos de ML entre si, omitindo esses baselines e potencialmente superestimando o ganho preditivo. Para predição em regimes epidemiológicos dentro da distribuição histórica, Random Forest demonstrou desempenho superior nesta amostra; para epidemias históricas extremas, modelos com capacidade de extrapolação linear (RLM, ANN) mostraram-se mais adequados.

Declaração de Uso da IA

O *ChatGPT* (GPT-5.3) e *DeepSeek* (DeepSeek-V3) foram utilizados para revisão textual e refinamento gramatical. Todas as sugestões produzidas pelas ferramentas foram analisadas de forma crítica e revisadas integralmente pelos autores, que assumem plena responsabilidade pelo conteúdo final apresentado.

Referências

- Al Bannoud, M., Mendes, J. L., Sacay, R. S., Martins, T. D., Bresolin, I. R. A. P., and Lourenço, J. B. (2026). Machine learning forecasting of dengue in são paulo using virtual data augmentation and urban incident predictors: Addressing the exceptional surge of cases in 2024. *Acta Tropica*, 275:107992.
- Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., Drake, J. M., Brownstein, J. S., Hoen, A. G., Sankoh, O., Myers, M. F., George, D. B., Jaenisch, T., Wint, G. R. W., Simmons, C. P., Scott, T. W., Farrar, J. J., and Hay, S. I. (2013). The global distribution and burden of dengue. *Nature*, 496(7446):504–507.
- Politis, D. N. and Romano, J. P. (1992). A circular block resampling procedure for stationary data. In LePage, R. and Billard, L., editors, *Exploring the Limits of Bootstrap*, pages 263–270. John Wiley & Sons.
- Rodriguez-Morales, A. J., Montenegro-Idrogo, J. J., Celis-Salinas, J.-C., Angerami, R., Villamil-Gómez, W. E., Sarute, N., Orduna, T., Perret, C., Rodriguez-Enciso, H. D., Savio-Larriera, E., Riskey, A., Forero-Peña, D. A., Ulloa-Gutierrez, R., Cabrera, M., Cimerman, S., Sah, R., Acosta-España, J. D., Navarro, J.-C., Sandoval, N., and Suarez, J. A. (2024). Unraveling the unparalleled 2024 epidemic of dengue in the americas. *Revista Chilena de Infectología*, 41(3):421–428.
- Roster, K., Connaughton, C., and Rodrigues, F. A. (2022). Machine-learning-based forecasting of dengue fever in brazilian cities using epidemiologic and meteorological variables. *American Journal of Epidemiology*, 191(10):1803–1812.
- Stolerman, L. M., Maia, P. D., and Kutz, J. N. (2019). Forecasting dengue fever in brazil: An assessment of climate conditions. *PLOS ONE*, 14(8):e0220106.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.
- Zanardo, G. E., Souza, É. L., Colonna, J. G., Nakamura, F. G., and Nakamura, E. F. (2025). Forecasting weekly dengue cases in brazilian federative units. In *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2025)*, pages 353–364, Porto Alegre, RS, Brasil. SBC.