EVENTO ON-LINE
ConBRepro
XVI CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO
02 a 04 de dezembro de 2026
Inovação e Transformação Digital como Estratégia Competitiva nas Engenharias

Previsão de Casos de Dengue Correlacionada a Fatores Climáticos Utilizando Modelos de *Machine Learning*

Luíza Dalla Costa Perusso
Universidade Tecnológica Federal do Paraná – Pato Branco
Laryssa Rayane Zortea
Universidade Tecnológica Federal do Paraná – Pato Branco
Ramon Gomes da Silva
Pontifícia Universidade Católica do Paraná – Curitiba
Matheus Henrique Dal Molin Ribeiro
Universidade Tecnológica Federal do Paraná – Pato Branco

Resumo: A dengue, por ser uma doença epidemiológica, possui surtos de casos e rápida disseminação em diferentes regiões do mundo, sendo fortemente influenciada por fatores externos como condições climáticas e saneamento básico. O desafio da saúde pública é fornecer assistência de recursos, profissionais e alas hospitalares nos momentos de maior ascensão da doença. Para isso, o estudo de previsão de séries temporais auxilia o poder público na tomada decisão, uma vez que ao interpretar os dados históricos da dengue é possível modelar seus padrões para prever seus comportamentos. Esse trabalho utiliza dados do DATASUS, de janeiro de 2015 a dezembro de 2025 a respeito do número de pessoas suspeitas de dengue da região sudoeste do Paraná; e dados do Instituto Nacional de Meteorologia (INMET), acerca da temperatura máxima e mínima e umidade relativa do ar média da região, para analisar a influência que a variável exógena clima exerce sobre o número de casos. Neste trabalho, foram testados modelos estatísticos e de *machine learning*. Os resultados destacam que para modelos estatísticos a variável exógena contribui de forma positiva para menores erros, reduzindo em até 37,7% o valor do RMSE e 41,2% o valor do MAE, contudo o mesmo não ocorre para modelos de *machine learning*.
Palavras-chave: dengue, séries temporais, clima.

Dengue Case Forecasting Correlated with Climatic Factors Using Machine Learning Models

Abstract: As an epidemiological disease, dengue is characterized by case outbreaks and rapid spread across different regions of the world, heavily influenced by external factors such as climatic conditions and basic sanitation. The key of public health challenge is ensuring the availability of resources, professionals, and hospital wards during periods of peak disease activity. Time-series forecasting aids public authorities in decision-making; by interpreting historical dengue data, it is possible to model patterns and predict future behavior. This study utilizes data from DATASUS, covering January 2015 to December 2025, regarding suspected dengue cases in the southwestern region of Paraná, alongside data from the National Institute of Meteorology (INMET), specifically maximum and minimum temperatures and average relative humidity for the area, to analyze the influence of climate on case numbers. Both statistical and machine learning models were tested. The results indicate that, for statistical models, the exogenous variable contributes to lower error rates, reducing RMSE by up to 37.7% and MAE by 41.2%—whereas this is not the case for machine learning models.

Keywords: dengue, time series, climate.

1. Introdução

A dengue é considerada uma das principais doenças transmitidas por mosquitos, especialmente em regiões tropicais e subtropicais (CHEONG et al., 2025). Trata-se de uma doença viral transmitida principalmente pelo *Aedes aegypti*, pequeno mosquito do gênero *Aedes* caracterizado pela presença de listras brancas no corpo. Amplamente presente em regiões tropicais e subtropicais, esse vetor representa um grave problema de saúde pública, contribuindo para a ocorrência de surtos recorrentes e para o aumento da carga epidemiológica da dengue em diversos países (SILVA; PALA; NAKAMURA, 2026). Embora seja uma doença que afeta principalmente países subdesenvolvidos e em desenvolvimento, fenômenos como as mudanças climáticas e a globalização têm ampliado sua disseminação para regiões anteriormente livres da doença (PANJA et al., 2023).

A crescente presença do *Aedes aegypti* em áreas urbanas tem contribuído para o aumento da transmissão, especialmente nas Américas, onde os casos cresceram de 1,5 milhão na década de 1980 para 16 milhões no período de 2010 a 2019 (GURGEL-GONÇALVES; OLIVEIRA; CRODA, 2024). Em 2024, o Brasil enfrentou sua maior epidemia de dengue já registrada, contabilizando 6,5 milhões de casos prováveis e 6.321 mortes relacionadas à doença (MINISTÉRIO DA SAÚDE, 2026).

O vírus da dengue apresenta quatro sorotipos distintos, todos capazes de causar infecção em humanos. Após um período de incubação de aproximadamente 4 a 7 dias, a dengue pode apresentar desde quadros febris leves até formas graves, como febre hemorrágica e síndrome do choque por dengue (CHEONG et al., 2025; PANJA et al., 2023). Embora a maioria dos pacientes se recupere em poucas semanas, parte dos casos evolui para quadros graves. Além disso, ainda não existem antivirais específicos eficazes disponíveis para o tratamento da dengue (CHEONG et al., 2025).

A transmissão urbana da dengue é fortemente influenciada por fatores climáticos e condições de saneamento, que afetam a distribuição do vetor e a dinâmica de transmissão viral (FIGUEREDO et al., 2023; VILLELA, 2025). Variáveis como temperatura, precipitação e umidade relativa do ar favorecem a reprodução e a sobrevivência do *Aedes aegypti*, aumentando a disseminação da doença, especialmente em períodos de chuvas intensas (PANJA et al., 2023). Além disso, as mudanças nas condições climáticas globais e a circulação simultânea de múltiplos sorotipos contribuem para a intensificação dos surtos e dificultam o controle epidemiológico da dengue (FIGUEREDO et al., 2023; SILVA; PALA; NAKAMURA, 2026). Diante desse cenário, a preparação para surtos de doenças torna-se essencial, pois possibilita a alocação eficaz de esforços e recursos para as áreas que mais necessitam, ajudando a prevenir casos, particularmente os graves (VILLELA, 2025).

O controle da dengue baseia-se principalmente em medidas preventivas, como a eliminação de criadouros do mosquito, o uso de repelentes e ações de conscientização da população. Vacinas também vêm sendo incorporadas como estratégia complementar para reduzir a incidência e a gravidade da doença. Entretanto, métodos de controle vetorial ainda demandam elevado esforço operacional e econômico quando não associadas a estratégias eficientes de planejamento (CHEONG et al., 2025). A interação entre fatores epidemiológicos, ambientais e climáticos também dificulta a compreensão das tendências da doença e limita a capacidade de prever surtos com precisão (ZHAO et al., 2023).

Para atender a essa demanda, diferentes modelos vêm sendo utilizados na previsão de casos de dengue, incluindo abordagens mecanicistas, estatísticas e técnicas de *machine learning* (ML). Entre essas abordagens, modelos estatísticos e de *machine learning* baseados em dados históricos de incidência e fatores ambientais têm apresentado

resultados promissores, uma vez que esses modelos epidemiológicos estão sendo adotados para resolver problemas dessa natureza, incorporando variáveis climáticas (DA SILVA et al. 2020).

A modelagem de séries temporais, seja clássica ou moderna, permite que o matemático e pesquisador consiga identificar padrões em dados históricos coletados dentro de uma sequência de pontos temporais previamente ocorridos. Feito isso, a partir de uma análise detalhada de comportamentos como tendência, sazonalidade e ciclicidade se torna possível prever as procedências futuras dessa variável estudada, considerando um horizonte temporal antecedentemente selecionado, sendo de curto, médio ou longo prazo. Cabe lembrar que os resultados de previsões estão relacionados a um valor probabilístico, podendo ou não acontecer. Uma previsão é acompanhada por um intervalo de predição, que fornece uma *faixa* de valores que a variável aleatória pode assumir com probabilidade relativamente alta (HYNDMAN, RJ; ATHANASOPOULOS, G. (2021).

Para assegurar a assertividade das previsões existem medidas de desempenho, como por exemplo o Erro Absoluto Médio (MAE), Raiz do Erro Quadrático Médio (RSME) ou Erro Percentual Absoluto Médio Simétrico (Smape), cujo principal objetivo é quantificar a incerteza da previsão por meio de erros observados, os quais quando corrigidos, garantem um maior alinhamento das expectativas dos valores resultantes.

Considerando a influência das variáveis climáticas sobre a proliferação do vetor e a necessidade de ferramentas que auxiliem no planejamento da saúde, este trabalho tem como objetivo prever o cenário epidemiológico da dengue na Região Sudoeste do Paraná. Para isso, serão aplicados modelos de previsão estatísticos e de *machine learning* utilizando dados climáticos como fatores externos à variável resposta (casos de dengue).

Desse modo, o presente estudo visa contribuir para a literatura, tanto em cunho matemático, por meio das análises comparativas utilizando diferentes ferramentas de previsão como citadas anteriormente, quanto em cunho social, compreendendo a realidade do estado do Paraná e proporcionando um instrumento relevante para a formulação de políticas públicas na área da saúde, possibilitando ações governamentais antecipadas, como investimento e alocação de recursos para as regiões mais necessitadas.

Por fim, este artigo está divido em 5 seções. A primeira, contextualiza o tema geral do trabalho, ou seja, a predição dos casos de dengue e seus desafios levando em consideração as variáveis exógenas. A segunda seção, traz a metodologia utilizada para realizar essa pesquisa, indicando o DATASUS e o INMET como bases de dados. A terceira aponta os trabalhos correlatos encontrados na literatura cujo objetivo de trabalho condiz com esse presente. A quarta aborda os resultados obtidos por meio da análise entre modelos preditivos das séries temporais. Por fim, a quinta apresenta as conclusões a respeito da análise comparativa entre os modelos de previsão.

2. Metodologia

Essa seção está dividida em duas partes, sendo a primeira referente a organização dos dados da pesquisa e a segunda em relação a análise prévia dos modelos implementados para a previsão dos casos de dengue.

2.1 Organização dos dados

Os dados dos casos de dengue referentes à região Sudoeste do Paraná foram obtidos a partir dos dados nacionais disponibilizados publicamente pelo sistema DATASUS. Foram selecionados casos suspeitos de dengue entre janeiro de 2015 e dezembro de 2025,

independente da confirmação laboratorial, uma vez que a investigação de um caso já caracteriza a utilização de recursos de saúde pública. Os registros foram agregados em nível mensal, resultando em uma série temporal de 132 observações.

Os dados meteorológicos foram obtidos no site oficial do INMET, utilizando os registros de três estações localizadas em municípios da região Sudoeste do Paraná. As variáveis coletadas foram: temperatura máxima, temperatura mínima e umidade relativa do ar. Os registros horários foram agregados em nível diário e, em seguida, para nível mensal. Para compor um único conjunto de dados representativo da região, foi calculada a média aritmética entre as três estações para cada mês.

A partir das variáveis climáticas originais, foram calculadas duas variáveis derivadas com o objetivo de enriquecer o conjunto exógeno. A primeira corresponde à temperatura média mensal, obtida pela média entre as temperaturas máxima e mínima. A segunda consiste em um termo de interação entre temperatura e umidade, calculado pelo produto entre a temperatura média e a umidade média. Essa variável foi incluída para representar o efeito conjunto desses dois fatores, cuja interação é biologicamente relevante para o ciclo de vida do vetor *Aedes aegypti*.

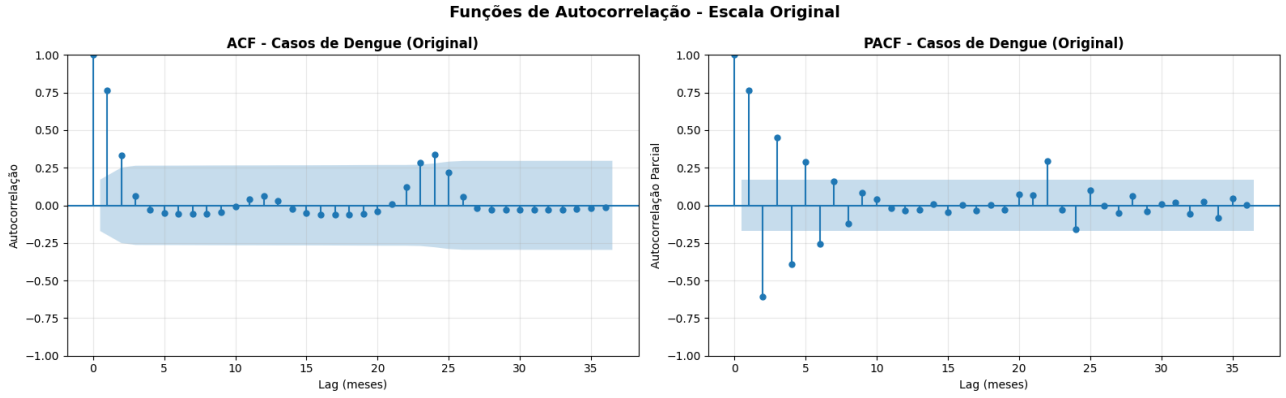
2.2 Análise prévia e modelos implementados

Uma análise exploratória da série alvo foi realizada antes da implementação dos modelos, buscando compreender seu comportamento temporal e orientar as decisões metodológicas.

A sazonalidade foi investigada utilizando diagramas de caixa (*boxplots*) construídos a partir da série transformada em escala logarítmica, empregada para facilitar a visualização dos dados. O gráfico resultante indicou um padrão sazonal bem definido, com maior concentração de casos entre os meses de março e maio, posteriormente confirmado pelo teste de Kruskal-Wallis, o qual não pressupõe normalidade. No entanto, assume que as observações em cada grupo provêm de populações com a mesma forma de distribuição e que as amostras são aleatórias e independentes (OSTERTAGOVÁ et. al, 2014).

As funções de autocorrelação (ACF) e autocorrelação parcial (PACF) foram utilizadas para avaliar a dependência temporal da série. A função ACF indicou forte dependência do mês imediatamente anterior, além de um pico no *lag* 24, sugerindo a presença de sazonalidade bienal além da anual. A função PACF indicou que a dependência temporal se estende por aproximadamente seis meses.

Figura 1: Funções de Autocorrelação – ACF e PACF



Fonte: Autores (2026)

Os resultados da análise exploratória orientaram a escolha dos modelos utilizados neste estudo. O Modelo Autorregressivo Integrado de Médias Móveis Sazonal (SARIMA) foi empregado como modelo linear de referência, enquanto o Modelo Autorregressivo Integrado de Médias Móveis Sazonal com Regressores Exógenos (SARIMAX) foi empregado para incorporar as variáveis climáticas exógenas. O modelo de aprendizado de máquina *Random Forest* (RF) foi utilizado como abordagem não linear tanto na modelagem univariada quanto na multivariada, adotando a estratégia de previsão recursiva.

Para os modelos lineares, os parâmetros foram selecionados automaticamente por funções de minimização do Critério de Informação de Akaike (AIC), que buscam o modelo com melhor equilíbrio entre ajuste aos dados e complexidade.

O modelo SARIMA selecionado apresentou uma ordem não sazonal (0,1,1) e sazonal (2,0,1)[12], indicando a aplicação de uma diferenciação de primeira ordem (d = 1). Os termos de médias móveis não sazonal (q = 1) e sazonal (Q = 1) apontam para a presença de dependência de curto prazo e de um ciclo sazonal, respectivamente. Os dois termos autorregressivos sazonais (P = 2) sugerem a influência dos mesmos meses de anos anteriores, abrangendo tanto o ciclo anual quanto o bienal, padrões que também haviam sido identificados nos gráficos de autocorrelação.

O modelo SARIMAX resultou em uma ordem não sazonal (0,0,2) e sazonal (2,0,1)[12]. A ausência de diferenciação indica que a série foi tratada como estacionária, enquanto os dois termos de médias móveis não sazonais capturam a dependência de curto prazo. Os componentes sazonais, por sua vez, refletem os ciclos anual e bienal observados na análise exploratória. Como variáveis exógenas, foram incluídas as defasagens climáticas de um a seis meses.

Para a implementação da abordagem não linear utilizando o *Random Forest*, foram consideradas duas configurações. A configuração univariada utilizou apenas os atrasos da série de dengue, definidos em seis meses com base na análise da PACF, enquanto a configuração multivariada incluiu também as defasagens climáticas de um a seis meses.

Os hiperparâmetros foram otimizados por meio de uma busca em grade com validação cruzada temporal, utilizando o RMSE como métrica de minimização. O modelo final que apresentou o menor erro médio foi configurado com 500 árvores, profundidade máxima sem restrição e um mínimo de uma amostra por folha. Essa configuração foi adotada por apresentar melhor equilíbrio entre capturar os picos epidêmicos e não se ajustar excessivamente a ruídos. O uso da validação cruzada temporal garantiu que os parâmetros fossem testados em diferentes períodos da série, permitindo verificar se o modelo mantinha um desempenho consistente ao longo do tempo.

3. Trabalhos Relacionados

Utilizando a base de dados ScienceDirect, com um recorte temporal de 2020 a 2026, foram selecionados 10 artigos cujo objetivo de trabalho condiz com esse presente. Todos os trabalhos selecionados abordam a previsão de casos de dengue, utilizando o clima como variável exógena.

Em síntese, nota-se que no âmbito geral são aplicados quatro tipos de modelos matemáticos para realizar a previsão dos casos de dengue utilizando outras covariáveis. Modelos clássicos, modelos de aprendizado por árvores, modelos de redes neurais e modelos de desempenho, os quais são normalmente combinados com outros modelos matemáticos.

Em relação às variáveis exógenas relacionadas às pesquisas, a maior parte delas trouxe o fator climático como agente determinante para a proliferação dos casos de dengue. Outros

fatores como geográficos, socioeconômicos e de infraestrutura também se mostraram influentes em alguns dos trabalhos.

Segundo K.H. Cheong et al. (2025), o uso do Google Trends para a previsão de casos de dengue se demonstrou uma ferramenta eficaz e de baixo custo. Além disso, quando combinada com modelos matemáticos de *Deep Learning,* como o GRU (*Gated Recurrent Unit*) e BiLSTM *(Bidirectional Long-Term Memory Neural Network)* ou LSTM (*Long-Short Term Memory*), apresentam resultados concisos de previsão para um horizonte de uma a quatro semanas e desempenho melhor que quando comparadas a métodos tradicionais de *maching learning*.

De acordo com W. Juraphanthong e K. Kesorn (2025), o modelo ARIMAS (extensão do ARIMA) demonstra-se eficaz nas previsões, uma vez que inova no que diz respeito à análise ontológica para compreender os vetores, sintomas e demais covariáveis relacionadas à dengue. Neste trabalho foi possível perceber o impacto que o modelo trouxe para a tomada de decisões na área de saúde pública, principalmente no que diz respeito à alocação de recursos, em diferentes regiões geográficas do planeta, provando assim sua capacidade de adaptabilidade para diversos casos.

Para M. S. Islam et al. (2024), o modelo matemático Prophet foi o que obteve o melhor desempenho nos resultados de previsão quando comparados com ARIMA e LSTM na análise de dados realizada, pois apresentou uma melhor performance para compreender a sazonalidade e os pontos de mudança da série. Nessa pesquisa os autores enfatizam a significância não apenas de fatores climáticos, como também fatores estruturais/sanitários para o aumento de casos de dengue.

Conforme X. Zhao et al. (2023), um modelo híbrido de previsão, combinando redes neurais e *deep learning* – CNN (Convolutional Neural Network) e BiLSTM - fornece uma melhoria significativa na previsão semanal de casos de dengue. Neste trabalho, os autores concluíram que o modelo aprende o padrão da série somente com os dados históricos já registrados, não necessitando de variáveis exógenas, como os fatores climáticos.

Na ótica de W. Anggraeni et al. (2024), determinar a previsão de casos de dengue é um problema de modelagem complexa. Desse modo, para uma melhor eficácia, o modelo matemático deve ser capaz de mapear a não linearidade da série. Sendo assim, foi utilizado o GRNN *(Generalized Regression Neural Network),* em conjunto com o EMD *(Empirical Mode Decomposition),* extraindo os padrões de não estacionariedade e o PSO *(Particle Swarm Optimization),* algoritmo utilizado para encontrar parâmetros ideais, os quais o modelo GRNN necessariamente utilizaria um processo exaustivo para determiná-los. Essa abordagem híbrida demonstrou-se melhor que modelos clássicos como ARIMA, especificamente para séries com dados escassos ou com características intermitentes.

Sob a perspectiva de M. Panja et al. (2023), o modelo XEWNet *(Ensemble Wavelet Neural Network with Exogenous Factors)* trouxe resultados positivos para previsões de seis meses e um ano, correlacionando em específico a variável de precipitação nos locais onde a pesquisa foi realizada, superando modelos clássicos e de *deep learning*. Seu principal diferencial é prever surtos de dengue ao integrar a decomposição de sinais por *wavelets*.

Segundo aponta A. Ceia-Hasse et al. (2023), o uso de modelos de *deep learning,* como o xgbTree *(extreme gradient boosting tree*), RF, NNET (*neural network*) e DNN (*deep neural network*) para a previsão de casos semanais de dengue oferece uma precisão significativamente maior que os métodos tradicionais. Os autores destacaram a importância da modelagem para sistemas de alerta precoce, auxiliando os órgãos públicos na tomada de decisão para a implementação de medidas de controle do mosquito.

Citando L. Lober et al. (2024), os modelos matemáticos para previsões demonstram diferentes resultados de desempenho conforme os dados das séries e suas características

socioeconômicas e geográficas. Dessa forma, evidencia-se a necessidade e relevância de testes prévios que apontam o menor erro possível. Neste estudo, o modelo *Random Forest* trouxe melhores resultados de previsão quando comparado com o XGBoost (*Extreme Gradient Boosting*) para casos de dengue, contudo, o mesmo não acontece para outras doenças como COVID-19.

Para J. Bracher et at. (2022), a previsão dos casos de dengue se enquadra em uma abordagem cujo modelo de previsão mais ideal é o EE (*Endemic-Epidemic*), uma vez que ele decompõe a dinâmica da transmissão em dois componentes, a taxa de base e a propagação. Sendo assim, neste estudo, o EE superou modelos sazonais naïve e apresentou desempenho superior ou equivalente aos modelos GLARMA *(Generalized Linear Autoregressive Moving Average)*, especialmente para horizontes de longo prazo.

Em consonância com os trabalhos registrados anteriormente, K. Roster et al. (2022), enfatiza a importância de escolher um modelo que forneça os melhores resultados para cada série a ser analisada. Aqui, os autores realçam a relevância de observar as semelhanças entre diferentes doenças, pois assim, utilizando modelos de transferência, é possível treinar a máquina para reconhecer padrões básicos de propagação de doenças em séries que possuem poucos dados disponíveis para serem observados.

A tabela a seguir sintetiza as informações essenciais encontradas na literatura.

Tabela 1: Trabalhos Correlatos com Modelos de Previsão de Casos de Dengue

Referências	Ano	Modelos	Medidas de desempenho	Variáveis exógenas
K. H. Cheong et al.	2025	LSTM, BiLSTM, GRUs, RNNs[1], CNNs, RF, SVR[2], GBM[3], KNN[4]	$R2$, MAPE, MAE, RMSE	Características climáticas e geográficas.
W. Juraphanthong e K. Kesorn	2025	ARIMAS	MSE, MAPE, MAE, RMSE	Características climáticas (temperatura, chuva, umidade) e geográficas.
M. S. Islam et al.	2024	ARIMA, LSTM, Prophet	MAE, MSE, RMSE	Características climáticas e de infraestrutura.
X. Zhao et al.	2023	RNN, CNN, BiLSTM, GRU	MAPE, MAE, RMSE	Características climáticas e geográficas (nesse caso não impactam diretamente nas previsões).
W. Anggraeni	2024	EMD, GRNN, PSO	MSE, MAE, SMAPE	Características geográficas (planaltos, montanhas), climáticas (temperatura, velocidade do vento, umidade, precipitação) e densidade populacional.
M. Panja et al.	2023	XEWNet,	RMSE, MAE, SMAPE, MASE	Características climáticas (precipitação).
A. Ceia-Hasse et al.	2023	xgbTree RF, NNET, DNN	—	Características climáticas (temperatura, umidade do ar e precipitação).
L. Lober et al.	2024	XGBoost, RF	MASE, MAE	Características climáticas, geográficas e socioeconômicas.
J. Bracher e L. Held	2022	GLARMA, EE	—	—
K. Roster et al.	2022	RF, NNET	MAE, MSE	—

Fonte: Autores (2026)

[1] *Recurrent Neural Network* (RNN)
[2] *Support Vector Regression* (SVR)
[3] *Gradient Boosting Machine*(GBM)
[4] *K-Nearest Neighbors* (KNN)

4. Resultados

Nesta seção serão apresentados os resultados dos modelos clássico e de ML analisando a relevância da inclusão da variável exógena na previsão dos casos de dengue.

Os modelos selecionados para observar o comportamento da variável dos casos de dengue de forma isolada foram o SARIMA (modelo clássico) e o *Random Forest* univariado (modelo de ML); enquanto a análise da influência da variável exógena foi realizada pelo SARIMAX (modelo clássico) e com o *Random Forest* multivariado (modelo de ML).

As métricas de desempenho observadas são o RMSE, MAE e sMAPE. A escolha delas justifica-se respectivamente por permitir avaliar a magnitude do erro com maior penalidade aos desvios, a quantificar de maneira mais robusta e intuitiva o erro e para possibilitar a análise entre os modelos selecionados no contexto de saúde pública para o replanejamento de recursos em picos de casos de dengue.

Tabela 2: Desempenho dos modelos preditivos

Modelo	RMSE médio	MAE médio	sMAPE médio
SARIMA	3316,36	2883,19	81,22%
SARIMAX	2063,31	1694,32	72,70%
RF univariado	2120,32	1680,94	61,83%
RF multivariado	2272,98	1863,14	59,39%

Fonte: Autores (2026)

Como mostra a tabela 2, é possível concluir que o modelo SARIMAX superou o modelo SARIMA na realização das previsões de casos de dengue nas três métricas utilizadas, ou seja, a inclusão dessa variável exógena diminui o erro esperado.

O RMSE reduziu aproximadamente em 37,7% dos casos e MAE reduziu em torno de 41,2% dos casos. Apesar de o modelo SARIMAX apresentar melhores resultados em relação ao RMSE, o percentual de erro visto no sMAPE ainda é alto, isso prova a alta variabilidade da doença e indica também a presença de períodos com baixa incidência de casos de dengue, bem como períodos com ascensões abruptas.

Analisando as medidas de desempenho utilizando os modelos de machine learning, os resultados não apontam uma melhora uniforme para as três métricas analisadas. O RMSE teve um aumento aproximado de 7,2% dos casos utilizando a variável exógena como componente na previsão da série temporal; o MAE também apresentou aumento em cerca de 10,8% dos casos, ambos os resultados destoam do sMAPE, cujo percentual do RF multivariado decaiu em relação ao RF univariado.

Esses resultados sugerem que a inclusão da variável exógena não contribui de forma consistente para a previsão de casos de dengue utilizando o modelo RF, isto é, as informações agregadas não foram suficientes para reduzir os erros médios dessa previsão e não complementam a leitura que o modelo realizou em relação às informações já fornecidas da série original.

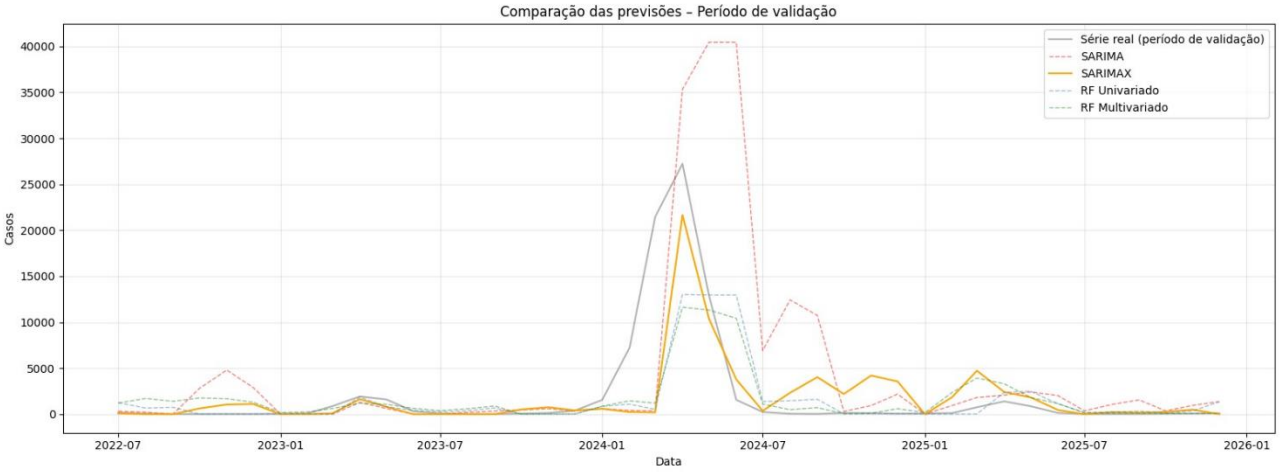
Em suma, a inclusão da variável exógena beneficiou os modelos estatísticos, pois sua combinação com a variável resposta considera a autocorrelação, a sazonalidade e a dependência temporal da série, aumentando a precisão dos resultados. O mesmo não ocorreu nos modelos de *machine learning*, visto que os resultados das métricas analisadas foram inconsistentes. Esse comportamento sugere que o *Random Forest* já capturava os padrões históricos da série, fazendo com que a variável exógena atuasse como fonte de ruído ou variabilidade na previsão.

A partir da imagem gráfica, é possível analisar a correção que o uso do modelo SARIMAX realiza em relação ao SARIMA, posto que é visível a diferença de amplitude entre ambos. Isso ocorre devido à falta das variáveis climáticas no SARIMA, na medida que ao incorporá-las no SARIMAX explica-se a variabilidade de caos, ou seja, os momentos de pico.

A figura 1 apresenta a comparação da previsão dos casos de dengue entre os quatro modelos testados, SARIMA, SARIMAX, RF univariado e RF multivariado em relação a série real. Esse gráfico destaca o modelo SARIMAX, devido a sua maior capacidade de preditividade em relação aos demais modelos.

A análise gráfica também valida os resultados obtidos nas medidas de desempenho em relação aos modelos de *machine learning*, mostrando como as previsões se distanciam dos valores observados na série em estudo.

Figura 2: Comparação das previsões dos casos de dengue utilizando os modelos SARIMA, SARIMAX, RF univariado e RF multivariado a partir do período de validação



Fonte: Autores (2026)

Para compreender as razões do desempenho inferior do *Random Forest* multivariado, foram conduzidos testes complementares. Uma das hipóteses consideradas foi que o melhor desempenho dos modelos lineares estivesse relacionado a uma estrutura predominantemente linear da série de dengue. Para investigá-la, foi aplicado o teste BDS (*Brock-Dechert-Scheinkman*) aos resíduos do modelo SARIMA ajustado à série. O teste rejeitou a hipótese nula, com p-valor de 0,0000, indicando a presença de componentes não lineares não capturados pelo modelo.

A relação entre as variáveis climáticas e os casos de dengue também foi analisada, evidenciou-se que a inclusão de termos não lineares proporcionou melhor ajuste aos dados. Assim, os resultados não sustentam a hipótese de que a predominância de linearidade tenha favorecido os modelos SARIMA e SARIMAX em relação ao *Random Forest*, indicando que o desempenho inferior do modelo de ML deve estar associado a outros fatores

Uma segunda hipótese considerada foi que o elevado número de variáveis climáticas exógenas comprometesse o desempenho do *Random Forest*. Para investigá-la, a correlação de Pearson foi calculada entre cada preditor climático e a série de casos de dengue, selecionando aqueles que apresentaram correlação significativa ($p < 0,02$). Foram selecionados sete preditores climáticos, apresentados na Tabela 3, correspondentes a defasagens de até três meses, indicando que as informações climáticas relevantes para a previsão se concentram em defasagens curtas.

Tabela 3: Preditores climáticos selecionados por correlação de *Pearson*

Variável climática	Defasagens selecionadas
Interação Temperatura–Umidade	lags 1, 2 e 3
Temperatura média	lags 2 e 3
Temperatura mínima	lags 2 e 3

Fonte: Autores (2026)

O conjunto reduzido de preditores foi então utilizado para reavaliar os modelos *Random Forest* multivariado e SARIMAX. Os resultados são apresentados na Tabela 4.

Tabela 4. Desempenho dos modelos com o conjunto completo e o conjunto reduzido de variáveis exógenas.

Modelo	Variáveis	RMSE médio
Random Forest univariado	-	2120,32
Random Forest multivariado	30	2272,98
Random Forest multivariado	7	2166,30
SARIMAX	30	2063,31
SARIMAX	7	1699,00

Fonte: Autores (2026)

A seleção de variáveis (indicadas na tabela 3) melhorou o desempenho de ambos os modelos, evidenciando a influência direta do conjunto de preditores na qualidade das previsões. O ganho foi mais expressivo no SARIMAX, cujo RMSE médio diminuiu de 2063,31 para 1699,00. No *Random Forest*, o RMSE também foi reduzido, de 2272,98 para 2166,30, embora o modelo tenha permanecido inferior à versão univariada. Esses resultados indicam que o excesso de variáveis exógenas contribuiu para a perda de desempenho do *Random Forest*.

A terceira hipótese levantada foi a de que a alta flexibilidade da configuração original do *Random Forest* poderia estar causando *overfitting*. Para testar essa possibilidade, o modelo foi re-treinado com valores mais restritivos de *min_samples_leaf*, utilizando 2 e 4, enquanto os demais hiperparâmetros foram mantidos fixos. Os resultados apresentados na Tabela 4 mostram que a restrição da flexibilidade piorou o desempenho do modelo. Esse resultado indica que a redução da flexibilidade não trouxe benefícios ao modelo, sugerindo que a configuração original conseguiu capturar melhor os padrões presentes na série.

Tabela 5. Impacto do ajuste de flexibilidade do *Random Forest* no desempenho preditivo.

min_samples_leaf	RMSE
1	2166,30
2	2256,84
4	2707,64

Fonte: Autores (2026)

5. Conclusões

A dengue, por ser considerada uma doença epidemiológica, traz consigo inúmeros desafios para a saúde pública. Determinar de forma assertiva a previsão de casos dessa doença auxilia os órgãos responsáveis para a tomada de decisão quando ela se encontra em picos de proliferação.

Para isso, leva-se em consideração a influência de variáveis climáticas como fator direto no aumento de casos de dengue especialmente em regiões de clima subtropicais. Desse

modo, esse trabalho propôs observar como modelos matemáticos clássicos e de machine learning reagem à inserção dessas variáveis exógenas na previsão de casos de dengue.

De modo geral, os resultados evidenciam que a incorporação da variável exógena torna a capacidade preditiva dos modelos estatísticos mais aprimorada, sendo o SARIMAX superior ao SARIMA. Essa inclusão permitiu representar de maneira mais fiel a dinâmica da incidência de dengue, especialmente durante o pico epidêmico anômalo, reduzindo significativamente os erros de previsão. Informações complementares, nesse contexto, auxiliam na interpretação dos dados históricos. Ademais, dentre os quatro modelos avaliados, o SARIMAX apresentou o melhor desempenho e consistência preditiva.

Em contrapartida, no âmbito do *machine learning*, o *Random Forest* multivariado não garantiu melhorias em relação à versão univariada. Os modelos de aprendizado de máquina demonstraram maior sensibilidade à seleção e organização das variáveis de entrada, fazendo com que o excesso de informações complementares introduzisse ruído e comprometesse a generalização. Contudo, o modelo não deve ser descartado, mas sim submetido a uma engenharia de atributos mais criteriosa, com melhor exploração da sazonalidade e de defasagens temporais.

Cabe ressaltar que a escolha adequada das variáveis preditoras é tão determinante quanto a seleção do próprio algoritmo, pois informações representativas da dinâmica epidemiológica aumentam a robustez das estimativas. Por fim, embora um horizonte de previsão menor possa gerar estimativas pontualmente mais precisas, o horizonte adotado neste estudo mostrou-se ideal para antecipar a evolução da dengue na região, favorecendo o planejamento preventivo, a alocação de recursos e a organização dos serviços de saúde para evitar a sobrecarga do sistema em picos epidêmicos.

Espera-se para trabalhos futuros uma análise mais aprofundada em relação a seleção de atributos relacionados as variáveis exógenas, mudanças de horizontes temporais e incorporação de algoritmos de aprendizagem profunda, visando combater as limitações encontradas até então.

Referências

ANGGRAENI, Wiwik; YUNIARNO, Eko Mulyanto; RACHMADI, Reza Fuad; SUMPENO, Surya; PUJIADI, Pujiadi; SUGIYANTO, Sugiyanto; SANTOSO, Joan; PURNOMO, Mauridhi Hery. A hybrid EMD-GRNN-PSO in intermittent time-series data for dengue fever forecasting. **Expert Systems with Applications**, *[S. l.]*, v. 237, p. 121438, 2024. DOI: 10.1016/j.eswa.2023.121438.

BRACHER, Johannes; HELD, Leonhard. Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. **International Journal of Forecasting**, *[S. l.]*, v. 38, n. 3, p. 1221–1233, 2022. DOI: 10.1016/j.ijforecast.2020.07.002.

CEIA-HASSE, Ana; SOUSA, Carla A.; GOUVEIA, Bruna R.; CAPINHA, César. Forecasting the abundance of disease vectors with deep learning. **Ecological Informatics**, *[S. l.]*, v. 78, p. 102272, 2023. DOI: 10.1016/j.ecoinf.2023.102272.

CHEONG, Kang Hao; LI, Kainan; YU, Dengxiu; ZHAO, Xinxing. Forecasting dengue cases through time-series modeling with Google Trends and deep neural networks. **Chaos, Solitons & Fractals**, *[S. l.]*, v. 201, p. 117290, 2025. DOI: 10.1016/j.chaos.2025.117290.

DA SILVA, Ramon Gomes; RIBEIRO, Matheus Henrique Dal Molin; MARIANI, Viviana Cocco; COELHO, Leandro dos Santos. **Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables**. *Chaos, Solitons & Fractals*, v. 139, p. 110027, 2020. DOI: 10.1016/j.chaos.2020.110027.

FIGUEREDO, Marcos Batista; MONTEIRO, Roberto Luiz Souza; DO NASCIMENTO SILVA, Alexandre; DE ARAÚJO FONTOURA, José Roberto; DA SILVA, Andreia Rita; ALVES, Carolina Aparecida Pereira. Analysis of the correlation between climatic variables and Dengue cases in the city of Alagoinhas/BA. **Scientific Reports**, *[S. l.]*, v. 13, n. 1, p. 7512, 2023. DOI: 10.1038/s41598-023-34349-8.

GURGEL-GONÇALVES, Rodrigo; OLIVEIRA, Wanderson Kleber De; CRODA, Julio. The greatest Dengue epidemic in Brazil: Surveillance, Prevention, and Control. **Revista da Sociedade Brasileira de Medicina Tropical**, *[S. l.]*, v. 57, p. e00203, 2024. DOI: 10.1590/0037-8682-0113-2024.

HYNDMAN, RJ, & ATHANASOPOULOS, G. (2021) ***Previsão: princípios e prática***, 3ª edição, OTexts: Melbourne, Austrália. OTexts.com/fpp3. Acessado em: 24 de setembro de 2026.

ISLAM, Md Shahidul; SHAHREAR, Pabel; SAHA, Goutam; ATAULLHA, Md; RAHMAN, M. Shahidur. Mathematical analysis and prediction of future outbreak of dengue on time-varying contact rate using machine learning approach. **Computers in Biology and Medicine**, *[S. l.]*, v. 178, p. 108707, 2024. DOI: 10.1016/j.compbiomed.2024.108707.

JURAPHANTHONG, Wanarat; KESORN, Kraisak. Autoregressive integrated moving average with semantic information: An efficient technique for intelligent prediction of dengue cases. **Engineering Applications of Artificial Intelligence**, *[S. l.]*, v. 143, p. 109985, 2025. DOI: 10.1016/j.engappai.2024.109985.

LOBER, Luiza; ROSTER, Kirstin O.; RODRIGUES, Francisco A. Forecasting infectious diseases in Brazilian cities: Integrating socio-economic and geographic data from related cities through a machine learning approach. **Chaos, Solitons & Fractals**, *[S. l.]*, v. 187, p. 115417, 2024. DOI: 10.1016/j.chaos.2024.115417.

MINISTÉRIO DA SAÚDE. **Monitoramento das arboviroses**. 2026. Disponível em: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/a/aedes-aegypti/monitoramento-das-arboviroses. Acessado em: 13 de maio de 2026.

OSTERTAGOVÁ, E; OSTERTAG, O; KOVÁC, J. (2014) **Metodologia e aplicação do teste de Kruskal-Wallis.** *Mecânica Aplicada e Materiais*. ISSN: 1662-7482, Vol. 611, pp 115-120. doi:10.4028/www.scientific.net/AMM.611.115.

PANJA, Madhurima; CHAKRABORTY, Tanujit; NADIM, Sk Shahid; GHOSH, Indrajit; KUMAR, Uttam; LIU, Nan. An ensemble neural network approach to forecast. Dengue outbreak based on climatic condition. **Chaos, Solitons & Fractals**, *[S. l.]*, v. 167, p. 113124, 2023. DOI: 10.1016/j.chaos.2023.113124.

ROSTER, Kirstin; CONNAUGHTON, Colm; RODRIGUES, Francisco A. Forecasting new diseases in low-data settings using transfer learning. **Chaos, Solitons & Fractals**, *[S. l.]*, v. 161, p. 112306, 2022. DOI: 10.1016/j.chaos.2022.112306.

SILVA, Viviane C.; PALA, Luiz O. O.; NAKAMURA, Luiz R. Interacting climatic drivers of dengue incidence: a GARMA-based time series analysis in a Brazilian health region. **Acta Tropica**, *[S. l.]*, v. 278, p. 108089, 2026. DOI: 10.1016/j.actatropica.2026.108089.

VILLELA, Daniel A. M. Predicting high dengue incidence in municipalities of Brazil using path signatures. **Scientific Reports**, *[S. l.]*, v. 15, n. 1, p. 26733, 2025. DOI: 10.1038/s41598-025-12356-1.

ZHAO, Xinxing; LI, Kainan; ANG, Candice Ke En; CHEONG, Kang Hao. A deep learning based hybrid architecture for weekly dengue incidences forecasting. **Chaos, Solitons & Fractals**, *[S. l.]*, v. 168, p. 113170, 2023. DOI: 10.1016/j.chaos.2023.113170.