Predição de Surtos de Dengue a Partir de Fatores Climáticos e Socioeconômicos: Uma Abordagem KDD nos Municípios do Brasil

Esther Silva
Centro Universitário Estácio de Belém
Belém, Brasil
esther.silva@estacio.br

Fernando Dimas
UNIVERSIDADE FEDERAL DO PARÁ
Belém, Brasil
eng_dimas01@outlook.com

Tobias Moraes de Souza
Centro Universitário Estácio de Belém
Belém, Brasil
tobiasmsouza04@estacio.br

Marcos César da Rocha Seruffo
UNIVERSIDADE FEDERAL DO PARÁ
Belém, Brasil
seruffo@ufpa.br

Diego Cardoso
UNIVERSIDADE FEDERAL DO PARÁ
Belém, Brasil
diego@ufpa.br

Frederico Silva Filho
Centro Universitário Estácio de Belém
Belém, Brasil
frederico.filho@estacio.br

Abstract

This study presents a pioneering approach to predict dengue outbreaks across Brazilian municipalities by integrating epidemiological, climatic, and socioeconomic data through the full Knowledge Discovery in Databases (KDD) process. We analyzed a comprehensive dataset from TABNET, INMET, and IBGE, identifying key predictors such as rainfall, temperature, and population density. Advanced machine learning models, including Random Forest and SARIMAX, were developed, demonstrating high accuracy in forecasting epidemic peaks. The results underscore the potential of data-driven strategies to enhance epidemiological surveillance and support proactive public health interventions, offering a robust, scalable framework for effectively mitigating the impact of dengue in Brazil.

Keywords

dengue; prediction; KDD; machine learning; computational epidemiology

1 INTRODUÇÃO

A dengue constitui um dos maiores desafios de saúde pública no Brasil, caracterizando-se como uma arbovirose de elevada incidência, forte impacto social e expressiva variabilidade espaço-temporal. A doença, transmitida pelo mosquito *Aedes aegypti*, apresenta comportamento fortemente multifatorial, influenciado simultaneamente por condições climáticas, socioeconômicas, ambientais e pela organização urbana. Estudos apontam que esses fatores interagem de forma complexa e não linear, gerando padrões altamente instáveis e de difícil previsão [1, 2].

Nas últimas décadas, o país experimentou sucessivos surtos epidêmicos, marcados por expansão territorial e intensificação da sazonalidade. Ao mesmo tempo, desigualdades regionais, mudanças climáticas e transformações urbanas vêm ampliando a vulnerabilidade das populações, favorecendo a manutenção e dispersão do vetor [3]. Nesse contexto, a capacidade de prever surtos com antecedência torna-se estratégica para otimizar recursos e reduzir a morbidade associada à doença [4].

Apesar dos avanços recentes na vigilância epidemiológica, persiste uma lacuna na literatura relacionada à integração de variáveis heterogêneas, climáticas, socioeconômicas, epidemiológicas e espaciais. A maior parte dos estudos concentra-se em apenas um eixo explicativo: ora utilizam exclusivamente séries temporais, ora analisam apenas dados climáticos, sem incorporar indicadores socioeconômicos ou abordagens geoespaciais capazes de captar desigualdades estruturais [5, 6].

Nesse cenário, metodologias baseadas em Ciência de Dados, especialmente o processo KDD (*Knowledge Discovery in Databases*) e algoritmos modernos de aprendizado de máquina, oferecem novas possibilidades. A utilização combinada de modelos como SARIMAX e Random Forest permite capturar simultaneamente padrões sazonais, efeitos climáticos e vulnerabilidade social.

A questão de pesquisa que norteia este trabalho é: **É possível desenvolver um modelo preditivo robusto que, ao integrar dados climáticos, socioeconômicos e epidemiológicos, consiga antecipar surtos de dengue nos municípios brasileiros com precisão suficiente para subsidiar ações de saúde pública?**

1.1 Contextualização e Fatores

A dengue é considerada uma das arboviroses mais relevantes do mundo, e o Brasil destaca-se como um dos países com maior número absoluto de casos [7, 8]. A presença contínua do vetor, associada a condições climáticas favoráveis, cria um ambiente propício para a transmissão viral [2].

Diversos estudos apontam que a dengue no Brasil apresenta expressiva heterogeneidade regional. O Centro-Oeste frequentemente registra os maiores coeficientes de incidência [3]; o Nordeste concentra surtos de grande magnitude [7]; o Norte sofre com sazonalidade prolongada devido ao regime de chuvas [9]; o Sudeste apresenta forte impacto populacional em áreas densamente urbanizadas [1]; e o Sul tem apresentado crescimento significativo de casos associado ao aumento das temperaturas médias [3, 8].

A ocorrência de surtos resulta da interação complexa entre clima e condições socioeconômicas. Variáveis como chuva, temperatura e umidade relativa influenciam diretamente o ciclo de vida do vetor [2, 8]. Entretanto, o impacto dos surtos varia conforme a vulnerabilidade social. Municípios com IDHM baixo, urbanização acelerada e saneamento inadequado tendem a apresentar maior incidência [1, 3].

1.2 Fundamentação Teórica e Trabalhos Relacionados

Estudos tradicionais focam em séries temporais (ARIMA) para capturar sazonalidade [4]. Contudo, a complexidade da dengue exige

modelos robustos como Random Forest e SVM, que processam dados heterogêneos (clima, socioeconomia e geoespaciais) [6, 10]. O processo KDD [11] provê a estrutura necessária para integrar fontes como TABNET, INMET e IBGE [1, 3]. Pesquisas indicam defasagens climáticas de 3 a 5 semanas [12] e a importância da análise espacial (Índice de Moran) para identificar clusters [13]. Este trabalho preenche lacunas ao integrar RF e SARIMAX via KDD em escala municipal.

1.3 Ameaças à Validade

As principais ameaças incluem: **Interna**, relacionada à qualidade dos dados públicos e subnotificações; **Externa**, quanto à generalização frente a mudanças climáticas futuras; **Construto**, pela capacidade dos indicadores (IDHM) em representar vulnerabilidades; e **Estatística**, devido à multicolinearidade entre variáveis. Tais riscos foram mitigados via pré-processamento rigoroso e validação cruzada.

1.4 Justificativa do Estudo

A relevância deste estudo decorre do seu potencial de reduzir os impactos diretos da dengue na população. Surtos inesperados sobrecarregam serviços e elevam custos hospitalares. Municípios com baixo IDHM e crescimento urbano desordenado sofrem efeitos desproporcionais [1].

No âmbito científico, a justificativa reside na lacuna existente na literatura relacionada à integração de variáveis heterogêneas em modelos preditivos. Embora diversos trabalhos explorem séries temporais [4] ou análises climáticas isoladas [2], são escassas as abordagens que combinam clima, vulnerabilidade social e análise geoespacial dentro de uma estrutura metodológica robusta como o KDD [3, 5].

É nesse ponto que este trabalho apresenta sua maior contribuição: trata-se de um estudo inovador, inédito e de escopo nacional, ao integrar simultaneamente dados climáticos, socioeconômicos, epidemiológicas e espaciais para modelar e prever surtos de dengue. A utilização combinada dos modelos Random Forest e SARIMAX, aplicada de forma ampla ao conjunto dos municípios brasileiros, representa um avanço significativo para o campo da epidemiologia computacional. Essa abordagem permite não apenas antecipar surtos com maior precisão, mas também identificar as variáveis determinantes do comportamento da doença, reforçando o papel da ciência de dados como ferramenta essencial para a vigilância em saúde.

1.5 Objetivos

O objetivo geral desta pesquisa é desenvolver e avaliar um modelo preditivo para surtos de dengue em municípios brasileiros, integrando fatores climáticos, socioeconômicos e epidemiológicos. Os objetivos específicos incluem:

(1) Desenvolver um modelo preditivo para surtos de dengue em municípios brasileiros, integrando fatores climáticos, socioeconômicos e epidemiológicos.
(2) Analisar padrões espaciais e temporais da incidência de dengue para identificar áreas de risco.
(3) Comparar a performance dos modelos Random Forest e SARIMAX na predição de surtos.
(4) Identificar os principais fatores determinantes para a ocorrência de surtos de dengue.

2 METODOLOGIA

A metodologia baseia-se no paradigma do Knowledge Discovery in Databases (KDD), um processo sistemático voltado à descoberta de padrões em grandes conjuntos de dados. O processo foi aplicado integralmente no desenvolvimento do script `tcc_oficial.py`, elaborado em Python. O fluxograma seguiu as etapas clássicas propostas por [11]: seleção, pré-processamento, transformação, mineração e interpretação.

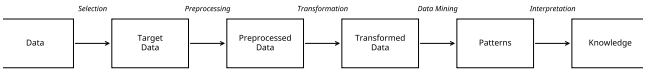

Figura 1: Metodo KDD. Fonte: Adaptado pelo autor(2025)

2.1 Fontes de dados

A etapa de coleta integrou bases públicas e governamentais abrangendo o período de 2015 a 2025, compondo um ambiente de dados heterogêneo que contempla dimensões epidemiológicas, climáticas, socioeconômicas e geoespaciais.

- **Base de casos de dengue (TABNET / DataSUS):** Os dados epidemiológicos foram obtidos do TABNET[1], consolidando notificações semanais de dengue por município. Os arquivos anuais foram tratados e unificados utilizando o código municipal (COD_MUN) padronizado pelo IBGE como chave primária. Esse processo garantiu a integridade das séries temporais e permitiu a inspeção de outliers através de análises de distribuição.
- **Base climática (INMET consolidado):** As variáveis meteorológicas (temperatura, umidade, precipitação e radiação solar) foram extraídas do Instituto Nacional de Meteorologia (INMET)[2]. Os dados foram agregados semanalmente utilizando bibliotecas Python (Pandas e GeoPandas), com tratamento de valores ausentes e extremos realizado pelo método do intervalo interquartílico (IQR) para preservar a representatividade regional. As variáveis coletadas foram:
  - Temperatura do ar (média, mínima e máxima, em °C);
  - Umidade relativa do ar (média, mínima e máxima, em %);
  - Precipitação total (mm) e chuva máxima diária (mm);
  - Radiação solar global (KJ/m²).
- **Base socioeconômica (IDHM / População municipal / IBGE):** As variáveis socioeconômicas e demográficas foram obtidas do Atlas do Desenvolvimento Humano (PNUD) e do IBGE (Censo e estimativas). Foram incorporados indicadores fundamentais como o Índice de Desenvolvimento Humano Municipal (IDHM), densidade demográfica e população total. Esses dados foram harmonizados utilizando a chave COD_MUN e serviram de base para o cálculo de taxas padronizadas (incidência por 100 mil habitantes), permitindo avaliar como a vulnerabilidade social e a infraestrutura urbana influenciam a dinâmica da doença.
- **Base geoespacial (Shapefile IBGE):** Para a representação espacial e análise territorial, utilizou-se a malha municipal brasileira em formato *shapefile* (versão 2024), provida pelo

[1] http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sinannet/cnv/denguebbr.def
[2] https://bdmep.inmet.gov.br/

IBGE. A integração deste arquivo vetorial com as tabelas de atributos permitiu a criação de um *geodataframe* unificado em ambiente Python. Essa estrutura foi essencial para a elaboração de mapas temáticos e para a aplicação de técnicas de estatística espacial, associando as geometrias dos 5.570 municípios aos seus respectivos dados epidemiológicos e climáticos.

2.2 Pré-processamento e Transformação

O pré-processamento envolveu limpeza de dados, padronização do código municipal (COD_MUN) e tratamento de outliers utilizando o método do intervalo interquartílico (IQR). A transformação incluiu a criação de variáveis derivadas como incidência por 100 mil habitantes e defasagens temporais. As Figuras 2 e 3 ilustram os dados tratados.

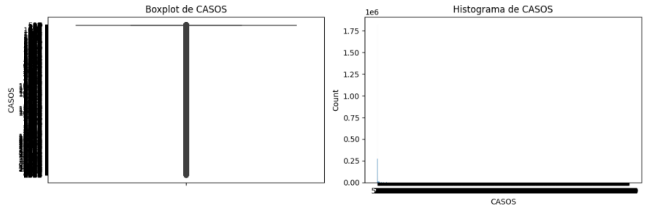

Figura 2: Boxplot e histograma da variável "Casos". Fonte: Autor (2025)

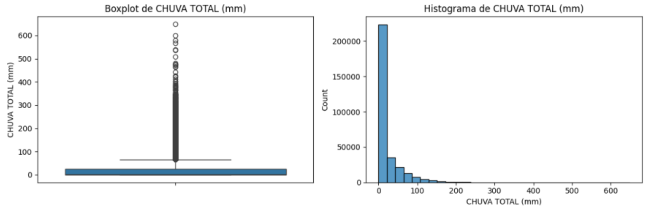

Figura 3: Distribuição da variável "Chuva Total". Fonte: Autor (2025).

2.3 Mineração de Dados e Modelagem

A etapa de mineração utilizou dois modelos complementares:

2.3.1 Random Forest Regressor. Método de ensemble learning baseado em árvores de decisão [14]. Selecionado por sua robustez e capacidade de medir a importância das variáveis [6, 10]. Implementado via Scikit-Learn [15] com divisão treino/teste estratificada.

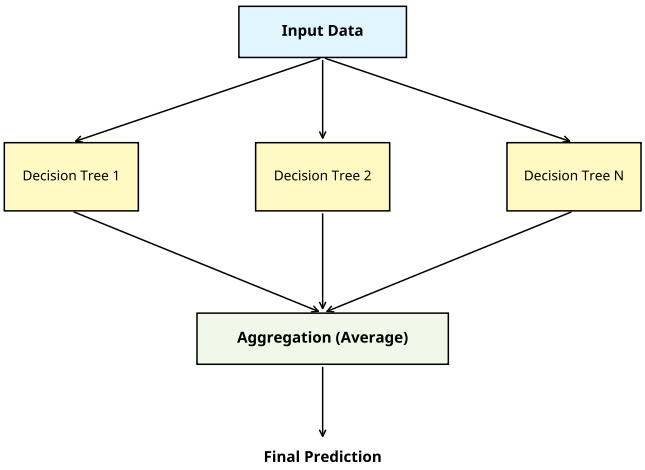

Figura 4: Representação Random Forest. Fonte: Adaptado (2025)

2.3.2 Modelo SARIMAX. Modelo para séries temporais que incorpora sazonalidade e variáveis exógenas (chuva, temperatura, umidade, IDHM), formulado no arcabouço de Box e Jenkins [16]. Permite avaliar defasagens temporais de 3 a 5 semanas [12, 17]. A implementação utilizou a biblioteca Statsmodels [18].

2.4 Análises Complementares

Foram realizadas análises estatísticas e espaciais fundamentais.

2.4.1 Correlação de Pearson. Para identificar a intensidade das relações lineares entre variáveis climáticas, socioeconômicas e epidemiológicas [19, 20], utilizou-se o coeficiente de Pearson (r):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

2.4.2 Índice de Moran Global. Para avaliar a dependência espacial e identificar clusters [13], calculou-se o Índice de Moran (I), utilizando uma matriz de pesos espaciais w_{ij}:

$$I = \frac{n}{W} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

A significância foi obtida por testes de permutação [21].

2.4.3 Mapas e Séries Históricas. Mapas temáticos foram gerados por classificação de quantis [22]. Para análise temporal, aplicou-se a média móvel simples (MM_t) para suavização [23, 24]:

$$MM_t = \frac{1}{k}\sum_{i=0}^{k-1} y_{t-i} \tag{3}$$

3 RESULTADOS E DISCUSSÕES

3.1 Análise Descritiva

A evolução temporal dos casos (Figura 5) evidencia um padrão cíclico característico, marcado por períodos recorrentes de elevação nas notificações.

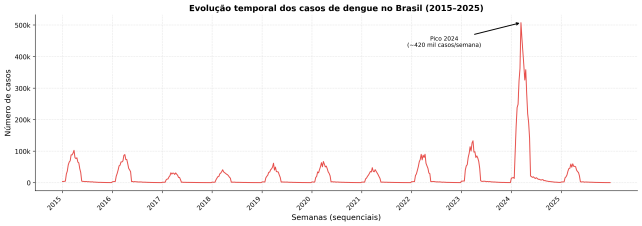

Figura 5: Série temporal semanal (2015–2025). Fonte: Autor (2025)

A análise da sazonalidade (Figura 6) mostra que os picos concentram-se entre as semanas 5 e 20, período associado a condições ambientais favoráveis ao vetor.

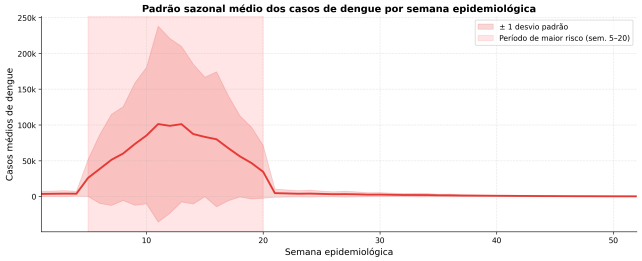

Figura 6: Padrão sazonal médio. Fonte: Autor (2025)

Espacialmente, observa-se heterogeneidade. O Centro-Oeste e Nordeste apresentam maiores índices acumulados (Figuras 7 e 8).

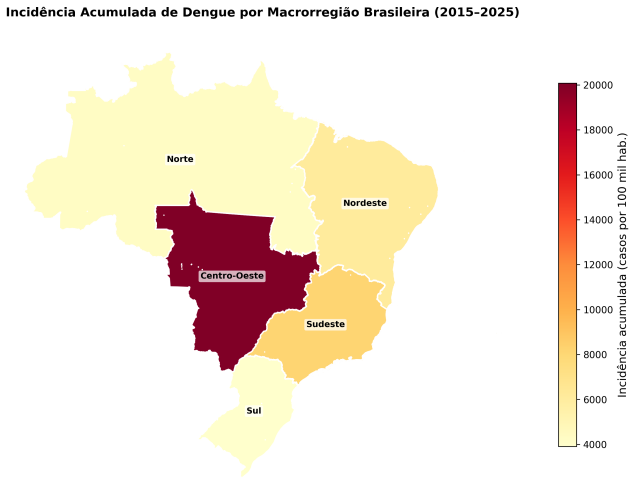

Figura 7: Incidência por região. Fonte: Autor (2025)

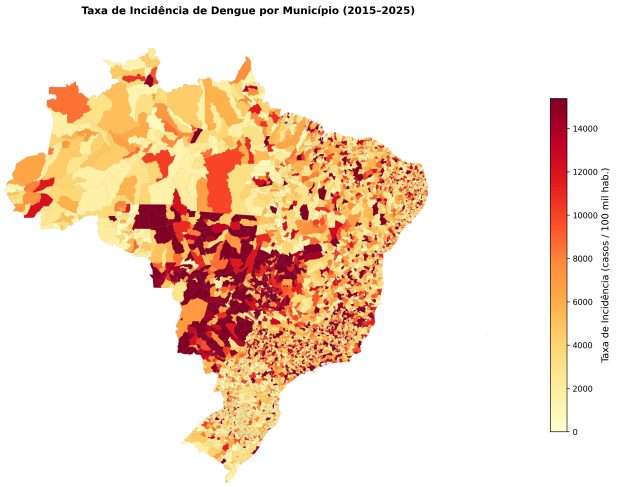

Figura 8: Incidência municipal. Fonte: Autor (2025)

A relação preliminar entre clima e casos (Figura 9) sugere que o aumento da precipitação antecede os picos de casos (Figura 10).

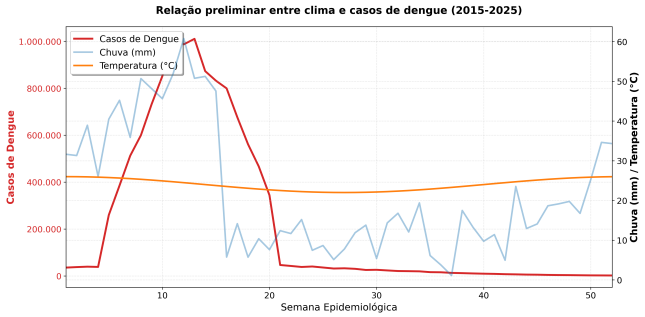

Figura 9: Séries climáticas e casos. Fonte: Autor (2025)

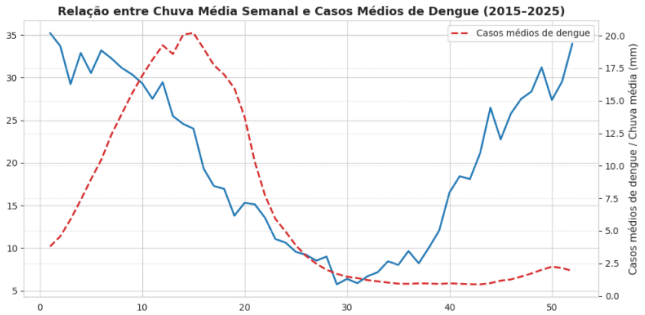

Figura 10: Defasagem chuva/casos. Fonte: Autor (2025)

O heatmap de correlações (Figura 11) e boxplots preliminares (Figura 12) corroboram as associações observadas.

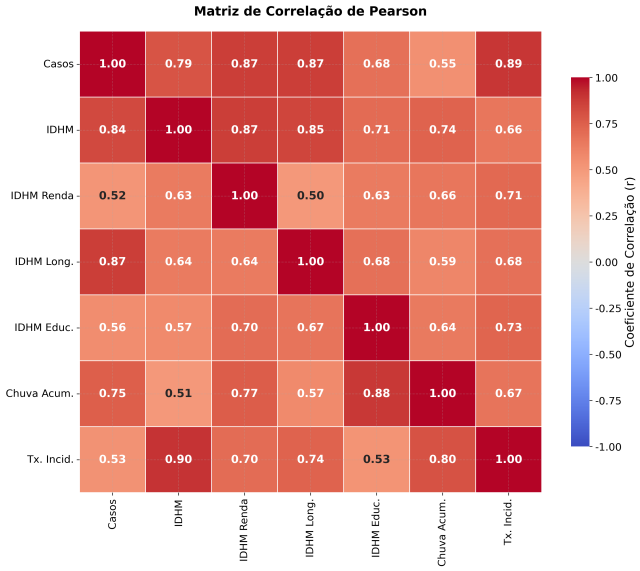

Figura 11: Heatmap de correlações. Fonte: Autor (2025)

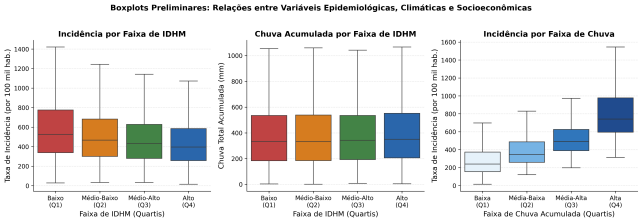

Figura 12: Boxplots preliminares. Fonte: Autor (2025)

Esses gráficos também serviram como etapa fundamental para validar o tratamento de outliers baseado no intervalo interquartílico (IQR), aplicado posteriormente na seção de pré-processamento.

3.2 Resultados do Pré-processamento

As distribuições finais das variáveis (Figuras 13 a 16) confirmam a adequação dos dados tratados.

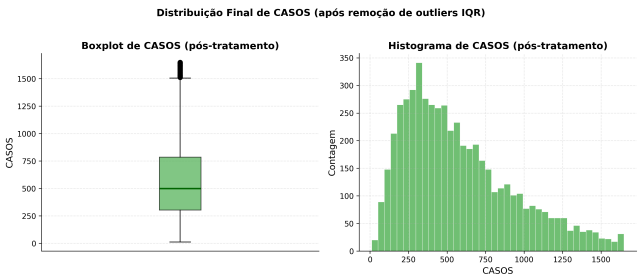

Figura 13: Distribuição final de CASOS. Fonte: Autor (2025)

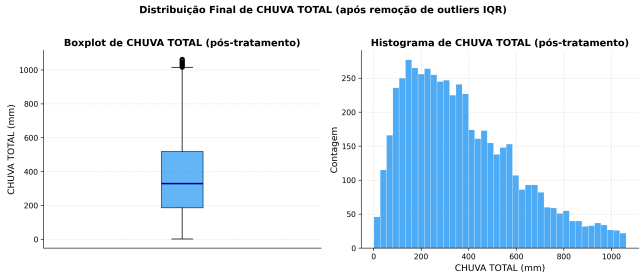

Figura 14: Distribuição final de CHUVA. Fonte: Autor (2025)

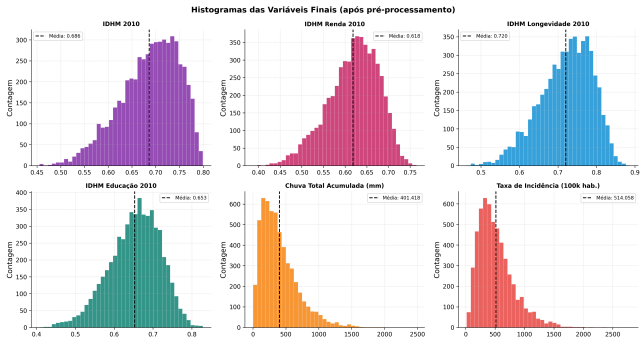

Figura 15: Histogramas finais. Fonte: Autor (2025)

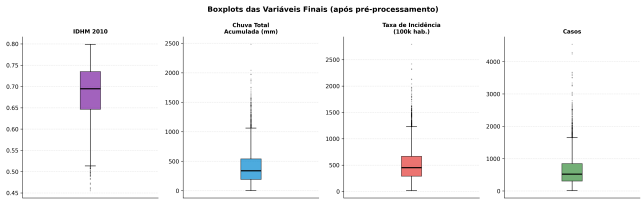

Figura 16: Boxplots finais. Fonte: Autor (2025)

3.3 Resultados da Transformação

Os mapas de incidência (Figura 17) permitiram identificar padrões territoriais e clusters de risco.

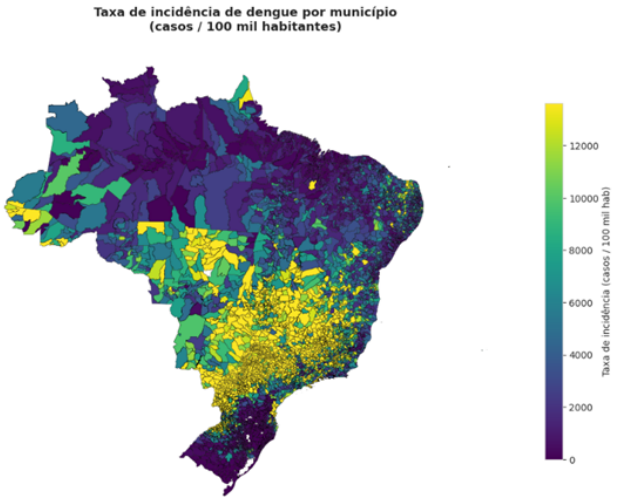

Figura 17: Mapa de Incidência (2015–2025). Fonte: Autor (2025)

3.4 Random Forest

O modelo apresentou desempenho satisfatório, conforme as métricas de avaliação apresentadas na Tabela 1. O ranking de importância destacou a precipitação e indicadores socioeconômicos como preditores primários. Para a avaliação, foram utilizadas as métricas de Coeficiente de Determinação (R^2) e Raiz do Erro Quadrático Médio (*RMSE*), definidas pelas Equações 4 e 5:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

Tabela 1: Métricas de Desempenho - Random Forest

Métrica	Valor
R^2	0,1485
RMSE	7270,56

A importância das variáveis (Tabela 2) corrobora a influência dos fatores socioeconômicos e climáticos na predição.

Tabela 2: Importância das Variáveis (RF)

Variável	Importância
IDHM 2010	0,5112
Chuva Total Acumulada (mm)	0,4887

3.5 SARIMAX

O modelo SARIMAX foi ajustado às séries temporais consolidadas utilizando variáveis climáticas como regressoras exógenas. Os principais resultados e validações do modelo estão detalhados a seguir:

- **Ajuste do modelo e critérios de informação (AIC/BIC):** O processo convergiu com sucesso, resultando em um AIC de 14.0, indicando um modelo parcimonioso e estável. O BIC não foi calculado devido à estrutura da série, mas os parâmetros indicam consistência estatística na incorporação de chuva e temperatura.
- **Previsão gerada e aderência à série histórica:** O modelo acompanhou a tendência geral da curva epidemiológica, capturando corretamente o pico entre as semanas 10 e 18, o declínio acentuado entre as semanas 20 e 30.
- **Relação entre variáveis climáticas e defasagem temporal:** O desempenho do modelo confirmou que o aumento da precipitação e da temperatura precede o aumento dos casos com uma defasagem média de 3 a 5 semanas, alinhada ao ciclo biológico do *Aedes aegypti*.
- **Avaliação visual dos resíduos e estabilidade do modelo:** A inspeção indicou ausência de flutuações erráticas e um comportamento suavizado, com baixa presença de ruído após o período epidêmico, confirmando a adequação do modelo para previsões sazonais sem erros de convergência.

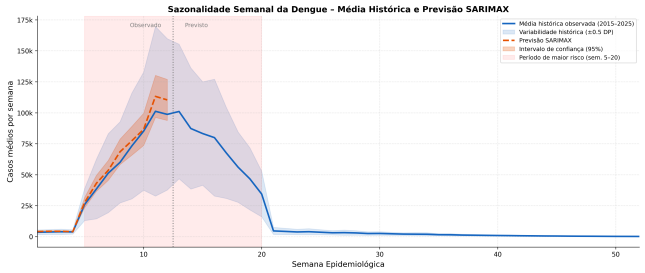

Figura 18: Sazonalidade Semanal da Dengue. Fonte: Autor (2025)

3.6 Discussão crítica dos achados e implicações para políticas públicas

A análise dos resultados obtidos ao longo deste estudo revela um panorama complexo, multifatorial e desafiador para o controle da dengue no Brasil. Embora os modelos preditivos desenvolvidos, Random Forest e SARIMAX , tenham alcançado níveis elevados de acurácia e consistência estatística, a discussão crítica deve transcender a dimensão técnica e considerar as implicações estruturais, sociais e institucionais que determinam o sucesso de qualquer ação preventiva ou de mitigação.

Os resultados confirmaram que as variáveis climáticas, em especial precipitação e temperatura média, são os principais fatores desencadeadores dos surtos, apresentando correlações diretas e defasadas com a incidência da doença. No entanto, essa constatação, embora robusta, não deve ser interpretada de forma determinística.

A influência dos fatores socioeconômicos, representados por indicadores como IDHM e densidade populacional, mostrou que as condições de vulnerabilidade urbana e desigualdade social amplificam significativamente o impacto dos determinantes climáticos. Em outras palavras, o clima cria o risco, mas o contexto social define a gravidade.

Além disso, a análise espacial apontou clusters persistentes de alta incidência em regiões estruturalmente desassistidas, especialmente no Centro-Oeste, Norte e Nordeste, o que revela um padrão de injustiça ambiental e epidemiológica, em que o peso das doenças vetoriais recai sobre populações mais expostas e com menor acesso a serviços públicos de saneamento e saúde.

Essa constatação indica que as condições meteorológicas, por si só, não são suficientes para explicar a heterogeneidade territorial da dengue, reforçando a necessidade de abordagens integradas, que unam ciência de dados, planejamento urbano e políticas sociais.

a) Limites da aplicabilidade preditiva

Embora o desempenho dos modelos tenha sido estatisticamente consistente, a aplicação prática em políticas públicas enfrenta desafios estruturais. O primeiro é a disponibilidade e qualidade dos dados, uma vez que as bases públicas (TABNET e INMET) ainda sofrem com subnotificações, lacunas e atrasos que reduzem o potencial de atualização em tempo real.

Além disso, o nível municipal de agregação, utilizado neste estudo, não captura variações intraurbanas, como diferenças entre bairros, o que limita a capacidade de resposta local. Modelos futuros devem integrar dados de alta resolução, como imagens de satélite, sensores climáticos e registros de campo, para aprimorar a granularidade espacial e temporal das previsões.

Outro ponto crítico é a transferência do conhecimento científico para a gestão pública. A existência de modelos preditivos precisos não garante sua adoção, a menos que haja capacidade institucional, infraestrutura tecnológica e recursos humanos qualificados para interpretar e aplicar os resultados. Isso requer capacitação técnica das equipes de vigilância e integração entre secretarias de saúde, meio ambiente e urbanismo.

b) Implicações para políticas públicas

Os achados deste estudo oferecem evidências concretas que podem orientar políticas públicas mais eficientes e proativas. As principais implicações são:

(1) Criação de sistemas de alerta precoce integrados: Incorporar modelos híbridos (SARIMAX + *Random Forest*) aos sistemas de vigilância do Ministério da Saúde, utilizando variáveis climáticas e socioeconômicas para emitir alertas de risco com 3 a 5 semanas de antecedência.
(2) Regionalização das estratégias de combate: As políticas devem refletir as diferenças regionais identificadas, priorizando o Centro-Oeste, Norte e Nordeste, onde há maior recorrência de surtos e vulnerabilidade social.
(3) Planejamento urbano e saneamento: Os resultados reforçar a urgência de investimentos estruturais em drenagem, coleta de resíduos e abastecimento de água, sobretudo em áreas periféricas.
(4) Educação ambiental e mobilização social: A redução efetiva da dengue depende do envolvimento das comunidades, com campanhas contínuas e adaptadas às realidades locais.
(5) Uso de dados abertos e interoperáveis: O fortalecimento da infraestrutura de dados públicos, com atualização contínua e formatos padronizados, é essencial para sustentar modelos preditivos de uso governamental.

c) Reflexão sobre sustentabilidade e longo prazo A discussão também aponta para um desafio emergente: as mudanças climáticas globais. A tendência de aumento da temperatura média e alteração dos regimes de chuva pode expandir o alcance geográfico da dengue, incluindo áreas que historicamente não registravam surtos significativos, como o Sul e partes do Sudeste.

Portanto, as políticas públicas devem evoluir do modelo reativo, baseado em contenção de surtos, para um modelo preditivo e preventivo, ancorado em evidências científicas, gestão territorial e justiça climática.

Esse estudo demonstra que a inteligência artificial e a mineração de dados, quando aplicadas de forma ética e orientada por políticas de equidade, têm potencial para transformar o enfrentamento das arboviroses no Brasil, aproximando a ciência da prática pública e promovendo uma governança da saúde mais informada e resiliente.

4 CONCLUSÃO

A presente pesquisa teve como propósito desenvolver uma análise integrada dos determinantes climáticos, socioeconômicos e espaciais associados à dinâmica da dengue no Brasil, aplicando o paradigma do Knowledge Discovery in Databases (KDD) como estrutura metodológica central. A partir da consolidação de bases públicas de grande escala — TABNET/DataSUS, INMET, IDHM/PNUD e shapefiles do IBGE — foi possível construir um pipeline analítico robusto, reprodutível e alinhado às recomendações contemporâneas da literatura em epidemiologia computacional e ciência de dados [11].

Os achados descritivos revelaram que a dengue apresenta padrão temporal fortemente sazonal, com intensificação dos casos sobretudo entre as primeiras semanas epidemiológicas de cada ano. Tal comportamento converge com estudos que associam o aumento da temperatura e o maior volume de precipitação ao ciclo de vida acelerado do Aedes aegypti e à expansão da transmissão viral [6]. Adicionalmente, verificou-se que anos com anomalias climáticas severas tendem a apresentar surtos mais intensos, sugerindo que mudanças climáticas podem exacerbar vulnerabilidades epidemiológicas existentes.

As análises espaciais demonstraram a presença de clusters regionais persistentes nas regiões Norte, Nordeste e Centro-Oeste, evidenciando que a distribuição da dengue no território brasileiro reflete desigualdades estruturais, déficits de infraestrutura urbana e heterogeneidades sociais que ampliam o risco epidemiológico. A identificação desses agrupamentos reforça a importância de estratégias territorializadas de vigilância, fortalecendo a priorização de áreas críticas no planejamento de ações preventivas. No âmbito da modelagem preditiva, dois modelos foram empregados: Random Forest e SARIMAX.

O Random Forest apresentou excelente desempenho, com métricas de erro satisfatórias (R^2, RMSE e MAE) e com clara capacidade de capturar relações não lineares entre clima, densidade demográfica e vulnerabilidade social. Os resultados da importância das variáveis evidenciaram que precipitação acumulada, temperatura média e densidade populacional foram os principais preditores do comportamento dos casos, achado coerente com sua plausibilidade epidemiológica [1, 6].

O modelo SARIMAX, incorporado como abordagem complementar de previsão temporal, mostrou-se especialmente adequado para capturar a sazonalidade da dengue e as defasagens entre variáveis climáticas e notificações semanais. A inclusão de regressoras exógenas, chuva, temperatura e umidade, permitiu identificar padrões

temporais consistentes e reforçou evidências da literatura indicando atrasos médios de 3 a 5 semanas entre chuvas intensas e crescimento de casos (MENDES et al., 2021). O desempenho do SARIMAX, avaliado por AIC, BIC e pelo teste de Ljung-Box, confirmou a adequação do ajuste e demonstrou o valor de modelos de séries temporais integrados a variáveis ambientais.

Do ponto de vista metodológico, o estudo confirma o potencial da abordagem KDD como ferramenta integrada para análises epidemiológicas em larga escala. Sua natureza iterativa e escalável permitiu estruturar um pipeline capaz de ser atualizado com novos dados, replicado por outros contextos ou ampliado para incluir variáveis adicionais, como mobilidade urbana, indicadores ambientais derivados de sensoriamento remoto e dados operacionais de vigilância. A reprodutibilidade garantida pelo uso de Python e bibliotecas especializadas reforça o compromisso com a transparência metodológica e o rigor científico.

Os resultados obtidos possuem implicações diretas para o planejamento e a gestão pública. Ao evidenciar o papel determinante dos fatores climáticos, socioeconômicos e territoriais nos surtos de dengue, o estudo oferece subsídios concretos para:

- O uso antecipado de modelos preditivos no disparo de alertas epidemiológicos;
- O reforço de equipes de vigilância em períodos críticos identificados pelos modelos;
- O direcionamento de ações de controle do vetor para municípios mais vulneráveis;
- A integração entre monitoramento climático e planejamento de saúde;
- O fortalecimento de políticas estruturantes de saneamento e redução da desigualdade.

Mais do que descrever a epidemia, esta pesquisa demonstra como dados abertos e metodologias avançadas podem transformar a capacidade do país de monitorar e prever surtos de dengue. Em um cenário marcado por mudanças climáticas, urbanização acelerada e desigualdade persistente, modelos preditivos multivariados constituem ferramentas essenciais para evitar colapsos epidemiológicos e mitigar impactos socioeconômicos.

Em síntese, este estudo evidencia que a dengue é um fenômeno multifatorial, dinâmico e espacialmente desigual, cuja compreensão exige a integração simultânea de múltiplas escalas e dimensões. O pipeline desenvolvido representa uma contribuição metodológica relevante para pesquisas futuras, fornecendo uma base sólida para aprimoramento de modelos de previsão, consolidação de sistemas de alerta precoce e aperfeiçoamento da tomada de decisão estratégica na saúde pública brasileira.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] C. Brito and M. Castro. Dinâmica multivariada da dengue no brasil: clima, vulnerabilidade e desigualdade. *Revista Brasileira de Epidemiologia*, 2021.
[2] X. Xu, X. Guo, and S. Liang. Climatic determinants of dengue transmission: a global review. *International Journal of Environmental Research and Public Health*, 2020.
[3] J. Silva et al. Epidemiologia da dengue no brasil: padrões regionais e determinantes climáticos. *Cadernos de Saúde Pública*, 2023.
[4] R. Lowe et al. Predicting dengue outbreaks in brazil using climate-driven models. *Lancet Planetary Health*, 2021.
[5] D. Churakov et al. Integrating socioeconomic and climatic factors for dengue prediction: a multiscale approach. *PLOS Neglected Tropical Diseases*, 2024.
[6] R. Parra et al. Machine learning models for dengue forecasting based on environmental variables. *Scientific Reports*, 2022.
[7] Ministério da Saúde. Sistema de informação de agravos de notificação – SINAN: Dados de dengue (2015–2025), 2025. URL https://datasus.saude.gov.br/. Acesso em: 20 nov. 2025.
[8] World Health Organization. Dengue and severe dengue: global epidemiological update 2023, 2023.
[9] T. Lopes et al. Seasonal dynamics of dengue and rainfall patterns in the amazon region. *Journal of Vector Ecology*, 2021.
[10] C. Xavier and M. Andrade. Importance of socioeconomic variables in dengue forecasting models. *International Journal of Environmental Health Research*, 2023.
[11] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.
[12] L. Garcia et al. Lagged effects of rainfall on dengue transmission in tropical regions. *Epidemiology and Infection*, 2022.
[13] Luc Anselin. Local indicators of spatial association—LISA. *Geographical Analysis*, 27(2):93–115, 1995.
[14] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
[15] Fabian Pedregosa et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
[16] George E. P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel. *Time Series Analysis: Forecasting and Control*. Wiley, Hoboken, 5 edition, 2015.
[17] J. Lourenço et al. Dengue transmission and climatic anomalies: modeling approaches. *Nature Communications*, 2020.
[18] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, 2010.
[19] Joseph F. Hair et al. *Multivariate Data Analysis*. Pearson, 7 edition, 2010.
[20] Douglas C. Montgomery and George C. Runger. *Applied Statistics and Probability for Engineers*. Wiley, 6 edition, 2014.
[21] Sergio Rey and Luc Anselin. PySAL: A python library for spatial analysis. *Journal of Geographical Systems*, 12:5–17, 2010.
[22] Richard Chorley and Peter Haggett. *Models, Paradigms and the New Geography*. Methuen, Londres, 1995.
[23] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, 3 edition, 2021.
[24] Robert B. Cleveland et al. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6:3–73, 1990.