

UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE CIÊNCIAS HUMANAS, LETRAS E ARTES
DEPARTAMENTO DE CIÊNCIAS SOCIAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM POLÍTICAS PÚBLICAS

ADRIANO LEONARDO BENASSI

**ESTRATIFICAÇÃO DE ÁREAS DE RISCO E MONITORAMENTO DE SURTOS DE
DENGUE UTILIZANDO CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL.**

MARINGÁ

2025

ADRIANO LEONARDO BENASSI

**ESTRATIFICAÇÃO DE ÁREAS DE RISCO E MONITORAMENTO DE SURTOS DE
DENGUE UTILIZANDO CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL.**

Dissertação apresentada ao Programa de Pós Graduação em Políticas Públicas do Departamento de Ciências Sociais da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Políticas Públicas. Área de concentração: Políticas Públicas e Desenvolvimento.

Orientadora: Profa. Dra. Marguit Neumann

MARINGÁ

2025

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

B456e

Benassi, Adriano Leonardo

Estratificação de áreas de risco e monitoramento de surtos de dengue utilizando ciência de dados e inteligência artificial / Adriano Leonardo Benassi. -- Maringá, PR, 2026.
54 f. : il. color., figs., tabs., mapas

Orientadora: Profa. Mr. Marguit Neumann.

Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Ciências Sociais Aplicadas, Departamento de Administração, Programa de Pós-Graduação em Políticas Públicas - Mestrado Profissional, 2026.

1. Dengue. 2. Saúde pública. 3. Modelo preditivo. 4. Vigilância epidemiológica. 5. Inteligência artificial. I. Neumann, Marguit, orient. II. Universidade Estadual de Maringá. Centro de Ciências Sociais Aplicadas. Departamento de Administração. Programa de Pós-Graduação em Políticas Públicas - Mestrado Profissional. III. Título.

CDD 23.ed. 614.58852

Rosana de Souza Costa de Oliveira - 9/1366

ADRIANO LEONARDO BENASSI

**Estratificação das áreas de risco e monitoramento de surtos de dengue utilizando
ciência de dados e inteligência artificial**

Dissertação apresentada ao Programa de Pós-Graduação em Políticas Públicas do Centro de Ciências Humanas, Letras e Artes da Universidade Estadual de Maringá, como requisito parcial do título de Mestre em Políticas Públicas, avaliada pela Comissão Julgadora composta pelos membros:

Documento assinado digitalmente
 **MARGUIT NEUMANN**
Data: 26/02/2026 16:34:29-0300
Verifique em <https://validar.iti.gov.br>

Prof^ª. Dr^ª. Marguit Neumann
(Presidente/orientadora)

Documento assinado digitalmente
 **KERLA MATTIELLO**
Data: 26/02/2026 16:37:10-0300
Verifique em <https://validar.iti.gov.br>

Prof^ª. Dr^ª. Kerla Mattiello
(Membro – PPP/UEM)

Documento assinado digitalmente
 **JULIANE ANDRESSA PAVÃO**
Data: 27/02/2026 09:24:27-0300
Verifique em <https://validar.iti.gov.br>

Prof^ª. Dr^ª. Juliane Andressa Pavão
(Membro externo – UNESPAR)

Documento assinado digitalmente
 **ANDRE LUIS DE CASTRO**
Data: 27/02/2026 09:44:07-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. André Luís de Castro
(Membro – PPP/UEM)

Aprovada em: 26 de novembro de 2025

Local de defesa: Bloco B-12, sala 001 da Universidade Estadual de Maringá

AGRADECIMENTOS

No decorrer desta trajetória, gostaria de agradecer as pessoas que me apoiaram nessa caminhada, que contribuíram direta ou indiretamente para o resultado dessa dissertação.

Primeiramente agradeço à Deus, pela oportunidade e por abrir novos caminhos na minha vida: ele que me dá força e me sustenta.

Agradeço à minha esposa Marla Cristina Barros Drews, por todo o apoio e dedicação.

Agradeço à minha mãe, Marcia Aparecida Lopes e ao meu pai, Valdir Benassi, pelo apoio incondicional.

Agradeço aos meus amigos, que tanto me ajudaram, especialmente à Michelle Plomcoski Guerreiro Gongora, Icaro da Costa Francisco, Eduardo Alcantara Ribeiro e Joel Victor Ravaneli Gonzaga.

Agradeço especialmente à minha orientadora, Dra. Marguit Neumann, por toda atuação no processo de construção desta pesquisa.

Também, uma menção especial à Universidade Estadual de Maringá, pela oportunidade da realização deste sonho.

À nossa filha, Juliana, que Deus nos enviou para adoçar mais a nossa vida.

E a tantos outros que me ajudaram. Não consigo agradecer nominalmente a todo mundo, mas cada contribuição, por menor que pareça, foi fundamental para a realização desta pesquisa.

“Suba o primeiro degrau com fé. Não é necessário que você veja toda escada. Apenas dê o primeiro passo” (Martin Luther King).

RESUMO

A dengue constitui um desafio para a saúde pública brasileira, demandando respostas rápidas e baseadas em evidências por parte dos administradores. Este estudo visa ao desenvolvimento de um modelo preditivo, baseado em aprendizado de máquina, para prever surtos de dengue no município de Maringá, Estado do Paraná. O objetivo é subsidiar políticas públicas mais eficazes e direcionadas para ações de controle da dengue. Utilizaram-se dados epidemiológicos do Sistema de Informação de Agravos de Notificação (SINAN) e variáveis climáticas do Instituto Nacional de Meteorologia (INMET), abrangendo o período de 2008 a 2019 (excluindo anos epidêmicos). A metodologia empregou o algoritmo *Random Forest*, com validação cruzada do tipo *Leave-One-Year-Out* (LOYO), para prever a incidência de casos com antecedência no município. Os resultados indicaram que o modelo com horizonte de 8 semanas demonstrou um bom desempenho. Obteve-se um coeficiente de determinação (R^2) de 0,917, erro absoluto médio (MAE) de 18,98 casos e raiz do erro quadrático médio (RMSE) de 31,54 casos. Conclui-se que o modelo possui alta capacidade preditiva. Pode ser integrado ao Programa Municipal de Controle da Dengue (PMCD) como ferramenta de apoio à decisão. Isso permite uma alocação mais eficiente de recursos e ações preventivas oportunas. Recomenda-se também, o aprimoramento da qualidade dos dados de endereçamento no SINAN para futuras modelagens mais assertivas com dados mais robustos e informações mais corretas.

Palavras chaves: Dengue; Inteligência Artificial; Aprendizado de Máquina; Políticas Públicas; Previsão Epidemiológica.

ABSTRACT

Dengue fever represents a significant challenge for Brazilian public health, requiring rapid and evidence-based responses from administrators. This study aims to develop a predictive model based on machine learning to forecast dengue outbreaks in the municipality of Maringá, Paraná State. The goal is to support more effective and targeted public policies for dengue control actions. Epidemiological data from the Notifiable Diseases Information System (SINAN) and climatic variables from the National Institute of Meteorology (INMET) were used, covering the period from 2008 to 2019 (excluding epidemic years). The methodology employed the Random Forest algorithm, with Leave-One-Year-Out (LOYO) cross-validation, to predict the incidence of cases in advance within the municipality. The results indicated that the model with an 8-week horizon demonstrated superior performance. A coefficient of determination (R^2) of 0.917 was obtained, along with a mean absolute error (MAE) of 18.98 cases and a root mean square error (RMSE) of 31.54 cases. We conclude that the model has high predictive capacity. It can be integrated into the Municipal Dengue Control Program (PMCD) as a decision-support tool. This allows for more efficient resource allocation and timely preventive actions. We also recommend improving the quality of address data in SINAN for future more assertive modeling with more robust data and accurate information.

Keywords: Dengue; Artificial Intelligence; Machine Learning; Public Policy; Epidemiological Forecasting.

LISTA DE IMAGENS

Figura 01 - Dengue no Brasil	18
Figura 02 - Série histórica arboviroses de 1986 à 2024	19
Figura 03 - Fluxograma das etapas da pesquisa	31
Figura 04: Ramificação de uma árvore de decisão	33
Figura 05: Correlação entre variáveis epidemiológicas	37
Figura 06 Correlação entre variáveis meteorológicas	38
Figura 07: Resultado do modelo preditivo com 12 semanas de horizonte	41
Figura 08 - Resultado do modelo preditivo com 08 semanas de horizonte	42
Figura 09 - Validação do modelo Previsão vs Observados, com dados do ano de 2008	43
Figura 10 Validação do modelo Previsão vs Observados, com dados do ano de 2009	44
Figura 11- Validação do modelo Previsão vs Observados, com dados do ano de 2016	44
Figura 12 - Comparativo de Horizontes de Previsão de 08 e 12 semanas.	45

LISTA DE ABREVIATURAS E SIGLAS

ACE	Agente Comunitário de Endemias
AI/IA	Artificial Intelligence/Inteligência Artificial
CO ₂	Dióxido de Carbono
DBF	Data Base File
ETL	Extract, Transform and Load (Extrair. Transformar e Carregar)
FUNASA	Fundação Nacional de Saúde
GDPR	Geral Data Protection Regulation (Regulamento Geral sobre a Proteção de Dados)
INMET	Instituto Nacional de Meteorologia
LGPD	Lei Geral de Proteção de Dados
LOYO	Leave-One-Year-Out (Validação Cruzada “Deixe-Um-Ano-Fora”)
MAE	Mean Absolute Error (Erro Absoluto Médio)
MS	Ministério da Saúde
OMS	Organização Mundial de Saúde
PIACD	Planejamento de Intensificação das Ações de Controle da Dengue
PMCD	Programa Municipal de Controle da Dengue
R ²	Coefficiente de Determinação
RMSE	Root Mean Square Error (Raiz do Erro Quadrático Médio)
SINAN	Sistema de Informação de Agravos de Notificação
SUS	Sistema Único de Saúde
SVS	Secretaria de Vigilância em Saúde

SUMÁRIO

1. INTRODUÇÃO	11
2. REVISÃO DE LITERATURA	14
2.1 Políticas Públicas	14
2.2 Dengue, Epidemiologia e IA: como a Análise dos dados e a Inteligência Artificial podem contribuir para o controle da dengue.	16
3. PROCEDIMENTOS METODOLÓGICOS	29
3.1 Etapas Metodológicas	30
3.2 Árvore de decisão	33
3.3 Random Forest	33
3.4 Validação de Modelos	34
4. PREPARAÇÃO DOS DADOS E ELABORAÇÃO DO MODELO	36
4.1 Análise Exploratória dos dados	36
4.2 Configuração do Modelo	39
4.3 Execução do Modelo	40
4.4 Validação do Modelo	42
5. RESULTADOS	45
6. PRODUTO E PROPOSTA DE INTERVENÇÃO	47
6.1 Relatório Técnico Conclusivo	48
REFERÊNCIAS	50
APÊNDICE	54

1- INTRODUÇÃO

A dengue é uma doença tropical causada por um vírus, sendo um arbovírus da família Flaviviridae, transmitida por espécies de mosquito do gênero *Aedes*, principalmente o *Aedes aegypti*, que é originário do Egito e significa o “odioso do Egito”; também é um importante vetor da febre amarela. O vírus da dengue originou-se, provavelmente, de vírus que circulavam em primatas não humanos nas proximidades da península da Malásia. Com o crescimento da população e a aproximação das habitações da região de selva, os mosquitos transmitiam o vírus de primatas da região e, por meio de mutações, foi possível a transmissão para os seres humanos (Brasil, 2002). É considerada a arbovirose mais prevalente no mundo, na qual cerca de 40% da população mundial está em risco e com projeções para que em 2085 cerca de 5 a 6 bilhões de pessoas – 50 a 60% da população global – possam ser infectadas (Viana; Ignotti, 2013).

O aumento crescente da dengue tem se constituído em objeto de preocupação para a sociedade, em especial para as autoridades de saúde, em razão da dificuldade de combate ao vetor e a adequação do sistema de saúde a fim de ter uma resposta mais rápida e eficaz quando há transmissão sustentada da doença nas cidades. (Barreto; Teixeira, 2008). Mesmo com avanços quanto a atualização de inseticidas e larvicidas realizadas pelo Ministério da Saúde (MS), desde os anos de 1980 e a constante capacitação dos trabalhadores que atuam diretamente no combate ao mosquito, a dengue e demais arboviroses transmitidas pelo *Aedes aegypti* continuam a ser um problema de relevância na saúde pública mundial (Viana; Ignotti, 2013). A dengue, vêm evoluindo com ciclos mais complexos que outros vírus e, com sua constante adaptação, seu controle vêm tornando-se cada vez mais difícil (Lara 2022).

Com base nisso, a busca por alternativas e agilidade no combate ao vetor tornam-se essenciais para o gestor público tomar decisões a fim de prever surtos e epidemias segundo as informações já coletadas pelo sistema de saúde. Nesse contexto, a Ciência de Dados surge como o alicerce metodológico para transformar grandes volumes de registros em conhecimento estratégico. A Inteligência Artificial (IA) pode ser uma ferramenta de auxílio nessa deliberação, realizando a rápida análise das informações, com base em dados do Sistema de Informação de Agravos de Notificação (SINAN Online) e demais variáveis, a fim de otimizar o tempo de reação e garantir ao usuário um trabalho ágil e de qualidade. Entre outras tecnologias que podem ser utilizadas no combate ao vetor, a IA (Brasil, 2025).

Braga e Valle (2007), Brasil (2002, 2024) e Chagas et al. (2023) concordam no contexto da importância do combate ao vetor da dengue e da resposta rápida do sistema de saúde. Fundado em dados obtidos da Gerência Financeira da Secretaria de Saúde de Maringá, 15º Regional de Saúde e Ministério da Saúde, a epidemia de 2023/2024 custou cerca de 5 milhões de reais, com investimentos das três esferas (Municipal, Estadual e Federal) (Maringá, 2025). As estimativas dos custos com a epidemia giraram em torno de 20 bilhões de reais, entre internamentos e produtividade (Oglobo, 2025). Como servidor público há 17 anos, todos eles envolvidos com o Programa Municipal de Controle da Dengue tive a oportunidade de participar de toda a cadeia de trabalho deste setor; isto me deu base para entender todo o processo de trabalho e melhor compreender sua dinâmica; com uma visão mais assertiva, podemos identificar falhas no processo e tentar minimizá-los. Isso me possibilitou perceber que o combate à dengue enfrenta obstáculos que transcendem as estratégias convencionais. Como já mencionei, as falhas nos processos (falta de agilidade no envio de informações, bancos de dados incompletos, entre outros) geralmente resultam da resistência em adotar inovações e da falta de integração entre os dados disponíveis e as ações práticas. Por exemplo, a falta de instrumentos preditivos fundamentados em dados climáticos e epidemiológicos restringe a habilidade de prever surtos e direcionar recursos de maneira mais eficaz.

Portanto, o objetivo deste estudo é criar um modelo de previsão para a ocorrência de dengue na cidade de Maringá, possibilitando a detecção antecipada de regiões com maior probabilidade de surtos de dengue. Essa previsão fornece dados estratégicos à administração pública municipal, permitindo o direcionamento de ações de campo mais acuradas e a implementação de campanhas de conscientização da população de maneira preventiva. Para isso, a pesquisa foi composta em três partes: revisão de literatura, apontando os conceitos de relevância para a pesquisa, metodologia e modelagem e demonstrativo de eficácia do modelo. Na modelagem, trabalharemos os dados com uma matriz de seleção de variáveis, na qual demonstramos, por meio da sua correlação, as mais indicadas para a confecção do modelo, em conjunto com as variáveis meteorológicas, utilizando a técnica de *Random Forest* para a confecção do modelo preditivo. Logo após a criação do modelo preditivo, com um período esperado de cerca de 3 meses de previsibilidade, realizaremos o demonstrativo de eficácia. Um aspecto fundamental na modelagem preditiva de doenças infecciosas é a definição do horizonte temporal de previsão, uma vez que a precisão das estimativas geralmente diminui à medida que o número de semanas projetadas aumenta. As previsões epidemiológicas têm maior precisão em horizontes curtos como demonstrado por Shaman e Karspeck (2012) em

seu estudo sobre surtos de influenza. Contudo, ao estender esse horizonte para oito e doze semanas, a pesquisa atual mostrou que é viável alcançar previsões consistentes e estatisticamente satisfatórias, mesmo em períodos mais prolongados. Esse resultado demonstra que, apesar de modelos tradicionais enfrentarem limitações com o aumento do horizonte, métodos baseados em aprendizado de máquina, como o Random Forest, conseguem superar essas dificuldades e oferecer estimativas confiáveis para apoiar o planejamento das ações de vigilância e controle da dengue. Para embasar essa afirmação estudos como de Zhao *et al.* (2020) mostram que modelos de aprendizado de máquina, como o *Random Forest*, podem superar modelos tradicionais em previsões de longo prazo para surtos de dengue. Isso se deve, em grande parte, à habilidade do *Random Forest* de combinar variáveis heterogêneas, como dados históricos de incidência, fatores meteorológicos (como temperatura, precipitação e umidade) o que possibilita a identificação de padrões complexos e não lineares na propagação de doenças.

Com este estudo, espera-se contribuir para o Programa Municipal de Controle da Dengue (PMCD), ao sugerir um modelo robusto e replicável para a previsão de surtos de dengue, empregando dados públicos e locais. Os resultados visam auxiliar na tomada de decisões pela administração municipal de saúde, permitindo uma resposta mais ágil, eficaz e econômica frente à ameaça de epidemias. O uso de um modelo automatizado também pode maximizar a eficiência de recursos humanos e financeiros, proporcionando um sistema a mais de monitoramento, em consonância com as orientações do Ministério da Saúde (Brasil, 2025), para a aplicação de tecnologias digitais e analíticas no enfrentamento das arboviroses.

2 REVISÃO DE LITERATURA

2.1 Políticas Públicas

As políticas públicas podem ser entendidas como um conjunto de decisões, ações e omissões do Estado voltadas para a solução de problemas que afetam a coletividade. Segundo Saraiva e Ferrarezi (2006, p. 30), política pública é “um sistema de decisões públicas que visa a ações ou omissões, preventivas ou corretivas, destinadas a manter ou modificar a realidade de um ou vários setores da vida social”. Isso significa admitir que toda política pública é, simultaneamente, uma resposta do Estado à realidade e uma tentativa de mudá-la. Capella (2018) complementa que a elaboração de políticas públicas está intimamente relacionada à forma como os problemas são definidos e considerados importantes pelo governo. Afirma que “toda a produção de políticas públicas está relacionada com a definição de problemas”, o que ressalta o caráter político e estratégico da ação estatal (Capella, 2018). Entende-se que a forma como um problema é apresentado afeta diretamente as opções que serão levadas em consideração e as possíveis abordagens para sua resolução.

Nesse cenário, entender o ciclo de políticas públicas é fundamental. De acordo com Capella (2018), esse ciclo abrange as fases: formação da agenda, formulação, tomada de decisão, implementação e avaliação. Destaca que essas etapas não seguem uma ordem linear, mas interativa, e que a forma como um problema é identificado e colocado em destaque na agenda afeta todas as fases subsequentes. Saraiva e Ferrarezi (2006) também entendem que o ciclo é uma construção teórica que facilita a sistematização e a compreensão do processo de criação de políticas estatais. Apesar de reconhecerem que, na prática, essas etapas não acontecem de forma sequencial e ordenada, ressaltam que elas costumam estar presentes, mesmo que se sobreponham em alguns contextos. Assim, Capella (2018) e Saraiva e Ferrarezi (2006) entendem que a definição do problema, sua inclusão na agenda e a sequência de decisões e ações que constituem o ciclo da política pública são influenciadas por fatores políticos, institucionais e sociais. Essa compreensão é fundamental, particularmente na área da saúde pública, em que questões como a dengue demandam respostas integradas do Estado, com medidas planejadas e avaliadas em todas as etapas do ciclo.

No ciclo de políticas públicas, como citado por Capella (2018) e Saraiva e Ferrarezi (2006), talvez a etapa mais desafiadora para o gestor seja a sua implementação. Está é a fase em que as intenções estabelecidas nas etapas anteriores do ciclo, particularmente na formulação e na decisão, são convertidas em ações efetivas do Estado. Desta maneira, trata-se de um momento em que as políticas públicas se concretizam, momento em que os planos,

estratégias e diretrizes previamente elaborados devem ser implementados pelas estruturas administrativas e técnicas do governo. Capella (2018) enfatiza que a implementação não deve ser vista apenas como uma etapa técnica ou neutra do processo. Ao contrário, ela é marcada por conflitos e diferentes interpretações dos problemas e desafios operacionais, como ocorrem no Brasil. Saraiva e Ferrarezi (2006) também sustentam que muitos dos problemas na implementação decorrem de erros na formulação, como a falta de definição de objetivos, a pouca clareza nas estratégias ou a falta de coordenação entre os participantes. Para entender o processo de implementação, segundo Capella, é necessário levar em conta os diversos agentes institucionais e as habilidades estatais, bem como os elementos políticos e sociais que influenciam o processo (Capella, 2018, p. 145). No campo da saúde pública, as políticas públicas assumem papel ainda mais crucial, pois envolvem a proteção da vida e o bem-estar da população, de maneira geral.

Portanto, com base no entender de Capella (2018) e Saraiva & Ferrarezi (2006) o ciclo de políticas públicas é fundamental para identificar os pontos fracos e fortes das ações governamentais oferecidas à população. No âmbito da saúde pública, especialmente quando se aborda o combate à dengue, cada fase do ciclo traz seus próprios desafios. A definição da agenda, por exemplo, depende não só da gravidade do problema do ponto de vista epidemiológico, mas também da pressão social e política para que o assunto seja colocado em primeiro plano. Por outro lado, a elaboração de políticas requer a combinação de evidências científicas, dados epidemiológicos e fatores socioambientais, o que frequentemente encontra obstáculos devido a restrições técnicas e institucionais. Além disso, a aplicação da ciência de dados no setor público é considerada uma abordagem promissora para aprimorar a gestão governamental. De Toni e Dorneles (2022) defendem que a análise sistemática de dados administrativos aumenta a habilidade do governo em criar políticas mais eficientes, fundamentadas em comprovações empíricas, diminuindo a dependência apenas de decisões reativas.

A fase de implementação, conforme já destacado por Capella (2018), é especialmente crítica, pois expõe as diferenças entre o que foi planejado e a real capacidade administrativa de execução. No contexto da dengue, isso resulta em problemas como a subnotificação de casos, falta de consistência nos registros encaminhados ao SINAN e problemas na coordenação entre os diversos níveis de governo. Esses desafios afetam tanto a implementação das ações quanto a fase de avaliação, que requer dados confiáveis e consistentes para medir a eficácia das políticas públicas.

A fase de avaliação também se torna um ponto estratégico do ciclo, pois possibilita a revisão de estratégias, a correção de rumos e a fundamentação de novas decisões. Capella (2018) destaca que a avaliação deve ser encarada não apenas como uma etapa final e isolada, mas como um processo contínuo que retroalimenta todo o ciclo de políticas públicas. A autora enfatiza que a maneira como os resultados são avaliados têm um impacto direto na reorganização de prioridades e na criação de novas estratégias. Saraiva e Ferrarezi (2006) também destacam que a avaliação permite entender até que ponto os objetivos estabelecidos foram alcançados e quais ajustes são necessários para garantir maior eficácia.

Dessa forma entende-se que a avaliação fecha o ciclo de políticas públicas ao permitir que o Estado entenda os efeitos reais de suas ações e, com base nisso, ajuste suas estratégias futuras. Mais do que uma etapa conclusiva, é um processo de aprendizado constante, no qual os erros e acertos da implementação são avaliados e convertidos em instrumentos para futuras decisões. Nesse cenário, o estudo realizado se apresenta como uma contribuição prática, fornecendo recursos para o aprimoramento da vigilância em saúde e na habilidade de resposta à dengue.

2.2 Dengue, Epidemiologia e IA: como a Análise dos dados e a Inteligência Artificial podem contribuir para o controle da dengue

Diante da necessidade de ações preditivas no controle da dengue, é necessário a compreensão de conceitos sobre a doença, a epidemiologia, instrumentos de ciência de dados e os problemas locais do município de Maringá. Entender a dinâmica da transmissão, os elementos ambientais e administrativos envolvidos, juntamente com o emprego de métodos de aprendizado de máquina, possibilita a criação de modelos preditivos mais acurados e relevantes para a tomada de decisões. Neste cenário, a análise exploratória de dados auxilia na verificação da qualidade da informação e na detecção de padrões pertinentes, solidificando a fundação para um modelo confiável e aplicável na vigilância em saúde.

A dengue é uma doença febril aguda, de etiologia viral, transmitida por mosquitos do gênero *Aedes*, podendo ser o *Aedes aegypti* e/ou *Aedes albopictus*. Estes mosquitos também são transmissores das viroses chikungunya e febre amarela (Brasil, 2002). As fêmeas dos mosquitos precisam de sangue para a maturação dos ovos, tornando assim um vetor de transmissão para o ser humano, com um mosquito infectado pelo vírus da dengue. A incubação da dengue é de 3 a 15 dias, após a picada do vetor sendo, em média, de 5 a 6 dias

(Brasil, 2002). Brasil (2024) atualiza esses dados ao ressaltar que, após a chegada do vírus Zika ao Brasil em 2015, a situação das arboviroses se tornou ainda mais complexa. A coexistência de dengue, chikungunya e zika, todas transmitidas pelo *Aedes aegypti*, tornou necessária uma resposta mais unificada dos serviços de saúde, tanto na vigilância epidemiológica quanto no controle clínico e nas técnicas de controle de vetores.

Brasil (2025) aponta que os fatores climáticos e sociais têm impacto no aumento da população do vetor da dengue no Brasil e no mundo. O desenvolvimento do *Aedes aegypti* é influenciado por condições climáticas como temperatura, pluviosidade e umidade. Fatores sociais como gestão de resíduos sólidos, saneamento básico e aspectos sociais também influenciam (Chagas et al. 2023). De acordo com Chagas et al. (2023) e Brasil (2002, 2024, 2025), o crescimento da população do mosquito *Aedes aegypti* está diretamente ligado às ações humanas, especialmente aquelas que contribuem para o aquecimento global, como o aumento das emissões de dióxido de carbono (CO₂), um dos principais gases do efeito estufa. A Organização Mundial da Saúde (OMS) destaca que períodos mais longos de chuva e temperaturas elevadas ajudam na propagação do mosquito, pois aumentam os locais onde ele pode colocar ovos e aceleram o seu desenvolvimento, (Chagas et al 2023).

A principal explicação para a epidemia de 2024 está relacionada às mudanças climáticas. A OMS destacou, em setembro de 2023, os impactos atuais e potenciais do fenômeno El Niño na saúde das populações. Entre as ameaças mencionadas no documento, estão as doenças de transmissão vetorial (Organização Mundial da Saúde, 2023). O período compreendido entre maio de 2023 e maio de 2024 bateu recordes históricos de altas temperaturas globais, segundo o observatório climático Copernicus Climate Change Service (2024). No Brasil, houve expansão territorial das localidades com epidemia de dengue no Brasil, e, naquelas em que a transmissão já ocorria de forma endêmica, ocorreu aumento da magnitude da epidemia. Este aumento ocorrido em 2024 não se deu de forma isolada no Brasil, também foi vivenciado por outros países das Américas, Ásia e Europa, e tornou a dengue uma ameaça alta em todo o mundo, em razão do incremento de casos e óbitos (Organização Mundial da Saúde, 2024), (Brasil, 2025, p..16)

O Brasil vem adotando medidas de combate ao vetor desde as primeiras epidemias registradas no início do século XX. Desde os anos 1980, segundo Braga e Valle (2007) vêm enfrentando epidemias de dengue subsequentes, no qual têm se um panorama onde 70% dos municípios do Brasil encontram-se infestados. Dados mais atuais apresentados pelo Agência Câmara (2023), apontam que quase 90% dos municípios encontram-se infestados pelo mosquito *Aedes aegypti*, conforme apresentado na Figura 1:

DENGUE NO BRASIL

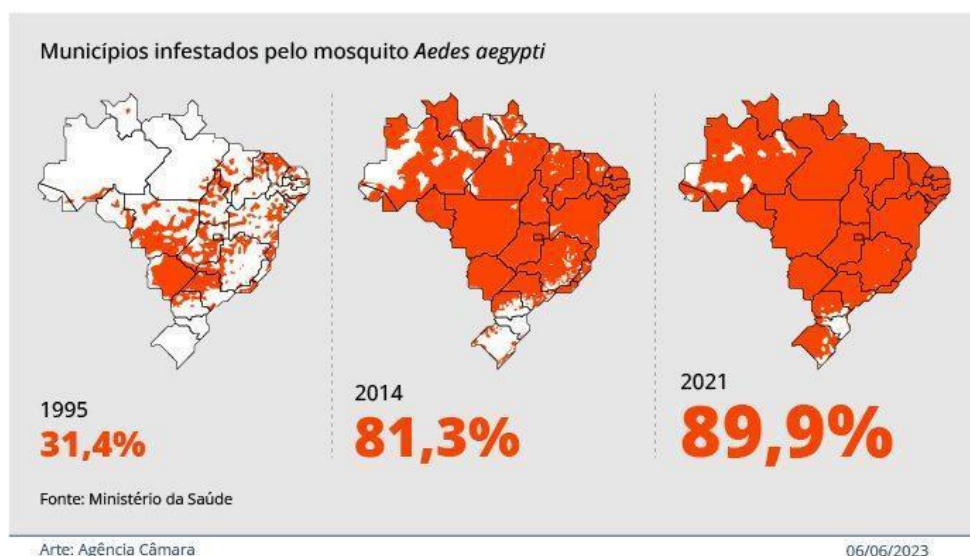
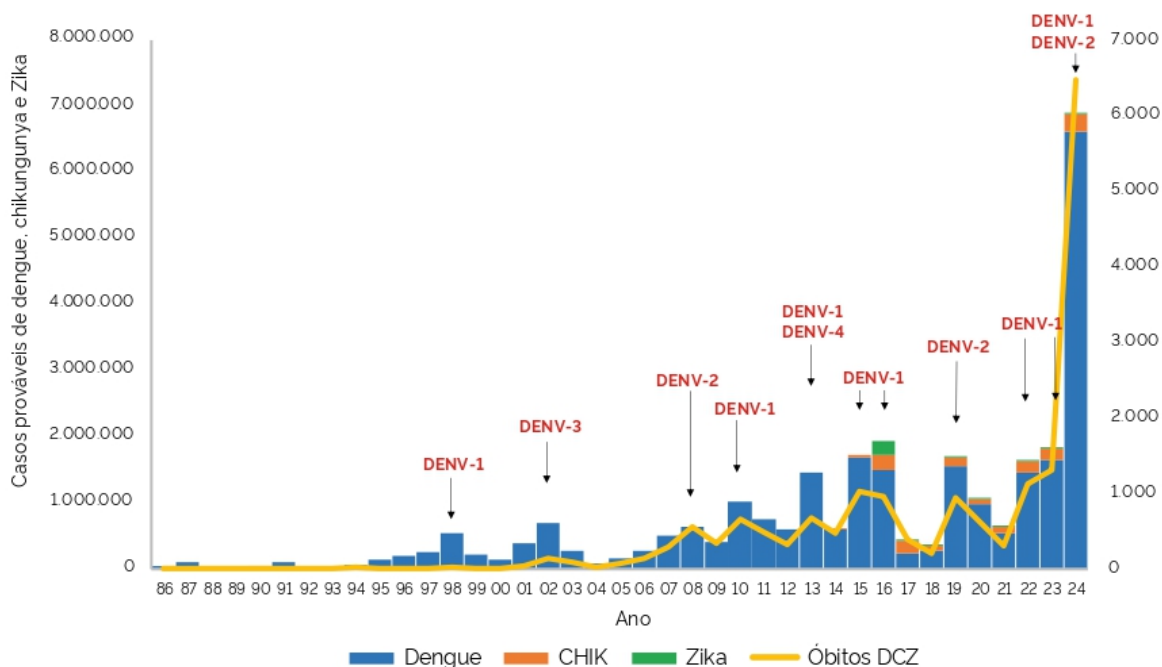


Figura 01 – Dengue no Brasil

Fonte: Ministério do Brasil/ Agência Câmara

A busca por novas tecnologias para combater o mosquito da dengue, como previsto no Decreto nº 5183, de 14 de março de 2024, do Estado do Paraná, é uma estratégia importante para tornar mais eficazes os esforços de controle da doença. Entre as medidas, o decreto incentiva a adoção de soluções inovadoras que possam melhorar a vigilância em saúde pública e ajudar a prevenir a propagação da dengue. Nesse sentido, o uso de ferramentas como Inteligência Artificial, georreferenciamento e análise preditiva pode ser muito útil para que os gestores tomem decisões mais rápidas e precisas. Essas tecnologias ajudam a fortalecer o combate às arboviroses, especialmente a dengue, alinhando-se às orientações do decreto e contribuindo para um controle mais eficiente. Brasil (2025) afirma que as novas tecnologias possibilitam a estratificação de risco intramunicipal, ou seja, a identificação de áreas prioritárias com maior risco de transmissão. Isso permite que as ações sejam mais direcionadas, eficientes e economicamente viáveis, em vez de dispersas e generalizadas, conforme demonstrado na Figura 2.

Série histórica dos casos prováveis, sorotipos e óbitos por dengue, chikungunya e Zika, Brasil, 1986 a 2024



Fonte: Sinan Windows/NET/On-line e e-SUS-VS.

Obs.: Dados atualizados em novembro de 2024.

Figura 02 – Série histórica arboviroses de 1986 à 2024.

Fonte: SINAN Windows/NET/On-line e e-SUS-VS

Quando se fala em epidemiologia, talvez a primeira situação seja o estudo de doenças em populações humanas. Bonita et al. (2010) afirma que a epidemiologia originou-se das observações de Hipócrates, na Grécia antiga, de fatores ambientais que influenciam na ocorrência de doenças. Foi somente no século XIX que a distribuição de doenças em grupos humanos específicos passou a ser medida em larga escala. Pereira e Veiga (2014) salientam que durante a revolução industrial, foram realizados censos para uma reforma sanitária onde as estatísticas vitais foram usadas para justificar e apoiar as novas políticas de saúde. Salienta que William Farr (1807-1883) e demais epidemiologistas ajudaram a desenvolver um sistema moderno de estatística que é usado até os dias de hoje. Farr alegou que algumas doenças, principalmente as crônicas, teriam uma causa multifatorial.

Deste modo, o estudo de Farr mostrou que existe uma correlação entre fatores que podem justificar alguns agravos e doenças. De acordo com Bonita et al. (2010), a dengue, assim como outras febres virais hemorrágicas, é um exemplo significativo da reemergência de

doenças infecciosas no final do século XX e no início do século XXI. A ocorrência dessas infecções, frequentemente intensificadas por alterações ecológicas e ambientais, tem aumentado o risco de transmissão entre indivíduos. Especificamente, a dengue faz parte de um grupo de arboviroses que passaram a refletir na saúde pública global, exigindo respostas coordenadas e eficazes da vigilância epidemiológica. A identificação e o controle desses surtos ressaltam a relevância da epidemiologia como uma ferramenta fundamental para entender os fatores que influenciam essas doenças, acompanhar sua propagação e direcionar intervenções oportunas e efetivas.

Com as contribuições de vários epidemiologistas, chega-se à epidemiologia atual. Bonita et al. (2010) nos mostra que o uso de métodos quantitativos para estudar a ocorrência de doenças como uma estratégia de prevenção e controle, é uma das principais abordagens da epidemiologia. Ainda nas afirmações de Bonita et al. (2010) e como descrito por John M. Last (1926-2019) a epidemiologia pode ser definida como “o estudo da distribuição e dos determinantes de estados ou eventos relacionados à saúde em populações específicas”. Deve-se ter como meio de partida que o alvo do estudo epidemiológico é sempre uma população humana. Pereira e Veiga (2014) completam afirmando que a epidemiologia é entendida como um estudo de determinada doença ou estado de saúde das populações e, também, por fatores que influenciam ou determinam essa contribuição.

Cabe a epidemiologia entender e estudar os fatores que influenciam algumas doenças para, com isso, ter uma abordagem mais assertiva e direta na prevenção e futuras medidas de controle. A epidemiologia pode ser definida como a ciência essencial da saúde pública, auxiliando diretamente na promoção da saúde ao gerar conhecimento sobre o perfil de doenças das populações. Através da avaliação de dados epidemiológicos, pode-se elaborar estratégias e medidas destinadas à pesquisa, acompanhamento e controle de enfermidades (Pereira e Veiga, 2014). Esta avaliação possibilita direcionar as decisões baseadas em evidências, aumentando a eficácia das ações em saúde pública. De acordo com Thomaz et al. (2015), as informações do Sistema de Informação de Agravos de Notificação (SINAN) permitem um diagnóstico rápido da ocorrência de eventos de saúde na população, fornecendo dados cruciais para o planejamento e implementação de medidas em variados níveis da administração pública.

A epidemiologia analisa os dados para a melhor tomada de decisões. Thomaz et al. (2015) afirma que os dados do SINAN permitem a realização do diagnóstico dinâmico da ocorrência de um evento na população, podendo fornecer informações cruciais para a tomada de decisão.

Fato importante para a tomada de decisão são dados específicos de determinadas comunidades. Thomaz et al. (2015) aponta que dentro de um panorama mais assertivo, as políticas de saúde descentralizadas precisam de estatísticas locais precisas para a identificação de fatores de risco e de populações vulneráveis. Com isso, entende-se que quanto mais dados e informações sobre uma comunidade específica, mais dinâmica propositivo será a abordagem e o direcionamento de uma determinada política pública.

A análise epidemiológica para o contexto da dengue é vital para o entendimento do comportamento da doença no Brasil. Conforme argumentado por Braga e Valle (2007) a estratificação epidemiológica da dengue ajudou os municípios a prepará-los para o enfrentamento de epidemias e surtos, levando em consideração que desde de 2001 a Fundação Nacional de Saúde, FUNASA, mudou sua meta em relação à dengue:

Em julho de 2001, a Funasa abandonou oficialmente a meta de erradicar o *Aedes aegypti* do País e passou a trabalhar com o objetivo de controlar o vetor. Foi implantado o Plano de Intensificação das Ações de Controle da Dengue (PIACD), que focalizou as ações em Municípios com maior transmissão da doença, considerados prioritários, escolhidos entre aqueles com infestação por *Ae. aegypti* e registro de transmissão de dengue nos anos de 2000-2001. (Braga; Valle, 2007 p. 116-117).

Diante do volume crescente de dados disponíveis e da necessidade de respostas rápidas e eficazes diante de surtos e epidemias, a Ciência de Dados apresenta-se como uma aliada estratégica da epidemiologia moderna. Como descrito por Bonita et al. (2010), o mundo vivenciou, nas últimas décadas, o ressurgimento de diversas doenças infecciosas, entre elas a dengue — uma arbovirose que representa importante ameaça à saúde pública global. Como mencionamos anteriormente, as mudanças ecológicas, ambientais e sociais contribuíram para o aumento da incidência e da gravidade dessas doenças, exigindo novos modelos de vigilância e resposta. Morettin e Singer (2019) afirmam que hoje em dia, os termos *Data Science* (Ciência de Dados) e *Big Data* (Megadados) são muito utilizados, contudo como se fossem conceitos mais recentes, diferentes do que os estatísticos lidavam há 2 séculos atrás. Talvez seja John Tukey (1915-2000) o primeiro matemático a dar importância à ciência de dados. Tukey sugeriu que fosse dado maior ênfase ao uso de tabelas, gráficos e outros dispositivos para uma análise preliminar de dados, previamente a uma análise confirmatória; a inferência estatística.

Como descrito por Moutinho (2024), observou-se um aumento exponencial de informações e mudanças quanto a maneira como os dados são coletados e armazenados. Essas mudanças favoreceram o desenvolvimento de uma nova maneira de uso de dados.

A ciência de dados é uma prática que busca resolver problemas complexos e encontrar insights acionáveis para negócios através da análise de dados avançada, a qual combina técnicas de matemática, estatística, programação, inteligência artificial (IA) e aprendizado de máquina. (ibm -site)

Os autores Morettin e Singer (2019) e Chollet (2018) concordam que com o volume de informações disponíveis na atualidade pode tornar a análise de dados uma ciência crucial como ferramenta estatística. A utilização da Inteligência Artificial (IA) e do Aprendizado de Máquina (*Machine Learning*) pode expandir as oportunidades de utilização da ciência de dados no campo da saúde pública. Segundo Morettin e Singer (2019), a IA tem se firmado como um instrumento apto a automatizar atividades intelectuais complexas, que antes eram executadas por humanos. Chollet (2018) reforça essa concepção ao caracterizar a IA como um esforço para automatizar atividades cognitivas humanas, englobando desde estratégias fundamentadas em regras (IA simbólica) até técnicas mais recentes como o aprendizado profundo.

IA é um tema muito comum nos dias de hoje. Morettin e Singer (2019) acreditam que o matemático Alan Turing (1912-1954) foi o primeiro a tratar do tema. Em 1950, o mesmo publicou um artigo com a primeira frase em destaque:

I propose to consider the question, “Can machines think?” (Alan Turing).

Com isso Morettin e Singer (2019) compreendem que a IA está cada vez mais imersa na nossa sociedade. Como mencionado, a IA é uma ferramenta para auxiliar a automatizar tarefas realizadas pelo ser humano. Chollet (2018) complementa esta afirmação, quando expõe que a IA é um esforço para automatizar tarefas realizadas por humanos:

Concisely, AI can be described as the effort to automate intellectual tasks normally performed by humans. As such, AI is a general field that encompasses machine learning and deep learning, but that also includes many more approaches that may not involve any learning. Consider that until the 1980s, most AI textbooks didn’t mention “learning” at all! Early chess programs, for instance, only involved hardcoded rules crafted by programmers, and didn’t qualify as machine learning. In fact, for a fairly long time, most experts believed that human-level artificial intelligence could be achieved by having programmers handcraft a sufficiently large set of explicit rules for manipulating knowledge stored in explicit databases. This approach is known as symbolic AI. It was the dominant paradigm in AI

from the 1950s to the late 1980s, and it reached its peak popularity during the expert systems boom of the 1980s. (Chollet, 2018, p.03).

Como pode-se observar, a principal função da IA é automatizar tarefas repetitivas e trabalhosas; fazer com que o computador possa pensar, executar e criar. Hoje a IA também é utilizada para criar imagens, textos, com *chatbots* inteligentes, algoritmos de aprendizado, etc. Segundo Chollet (2018) foi no período vitoriano que a primeira máquina analítica; a primeira máquina de uso geral conhecido, um computador mecânico foi desenvolvido. Foi criada para automatizar e realizar cálculos; a máquina criada por Charles Babbage era uma calculadora mecânica automática projetada para tabular funções polinômicas. Porém, foi nos anos 60 em que o aprendizado de máquina nasceu como um campo da inteligência artificial. Izbicki e Dos Santos (2020), argumentam sobre uma das funções do aprendizado de máquina seria aprender padrões com base em dados.

Métodos baseados em aprendizado de máquina, para resolver problemas das mais diversas áreas, são cada vez mais utilizados. Em alguns contextos, o aprendizado de máquina é apresentado como a própria ciência de dados:

O aprendizado de máquina conta com algoritmos para analisar conjuntos de dados enormes. Atualmente ela não fornece o tipo de IA apresentada nos filmes. Mesmo os melhores algoritmos não conseguem pensar, sentir, apresentar qualquer forma de autoconhecimento ou exercitar o livre-arbítrio. O que o aprendizado de máquina pode fazer é realizar análise preditiva bem mais rápido que qualquer ser humano. Como resultado, ela ajuda as pessoas a trabalhar mais eficientemente. Então o estado atual da IA é fazer análise, mas os seres humanos ainda precisam considerar as implicações dessa análise — tomando as decisões éticas e morais necessárias. (Mueller; Massaron, 2018, p.03).

Como mencionado por Mueller e Massaron (2019), a principal presunção do aprendizado de máquina (machine learning) é que, na teoria, o trabalho e a consciência (inteligência) da máquina é melhor do que uma pessoa. Talvez, pode-se pensar que o aprendizado de máquina seja uma maneira mais rápida e assertiva de aprender padrões estatísticos e criar modelos mais assertivos para a análise de dados.

Sob essa ótica, o aprendizado de máquina se apresenta como um recurso valioso quando combinado com métodos rigorosos de coleta e processamento de dados. Segundo Morettin e Singer (2019) e Moutinho (2024), a qualidade dos dados é fundamental para a criação de modelos sólidos e análises fiáveis. A recolha, estruturação e avaliação metódica das informações é o alicerce de qualquer pesquisa científica, sendo o primeiro passo para

assegurar a exatidão estatística e a pertinência dos achados. Portanto, mesmo que os algoritmos sejam capazes de detectar padrões com rapidez e exatidão superiores às habilidades humanas, conforme destacado por Mueller e Massaron (2018), sua efetividade está diretamente ligada à integridade e estrutura dos dados empregados.

A coleta, organização e análise dos dados são vitais para um bom desenvolvimento e a criação de modelo robusto. A coleta, como mencionada por Morettin e Singer (2019) e Moutinho (2024) concordam quando afirmam que é o passo crucial. A primeira etapa do desenvolvimento de uma pesquisa está na coleta das informações: é a base. Conclui-se que a pesquisa deve seguir uma abordagem estruturada e orientada pelos objetivos de qualquer estudo.

A coleta de dados deve ser meticulosamente organizada, levando em conta os propósitos do estudo e a natureza das variáveis. De acordo com Morettin e Singer (2019), essa fase é decisiva para assegurar a precisão das análises estatísticas, pois falhas na coleta podem prejudicar toda a pesquisa. Os autores ressaltam que a coleta de dados pode ser realizada através de pesquisas observacionais, amostrais ou experimentais. Além disso, a correta estruturação dos dados, através da elaboração de planilhas bem organizadas e definição precisa das variáveis, é indispensável para permitir análises consistentes. Ademais, realçam a relevância de tratar informações ausentes, prevenir falhas de digitação e seguir princípios éticos ligados à proteção da privacidade das informações. Godoy (1995), embora adote uma perspectiva diferente quanto ao enfoque metodológico e qualitativo da coleta de dados, enfatiza que a metodologia científica deve ser entendida como um processo organizado de busca pelo conhecimento, no qual cada fase — da coleta à análise — é fundamental para a legitimidade dos resultados alcançados. Portanto, é claro que a qualidade dos dados está intrinsecamente ligada à robustez metodológica. Nesta pesquisa, o uso de dados do SINAN e do INMET demandou atenção extra em relação à padronização, correção de inconsistências e preenchimento de lacunas. Esses procedimentos foram fundamentais para assegurar que o modelo preditivo criado pudesse fornecer resultados confiáveis, ao passo que enfatizam a importância de melhorar continuamente as práticas de coleta e registro de dados na saúde pública.

Os dados para esta pesquisa foram obtidos através de duas fontes principais: o SINAN, que fornece os dados semanais sobre os casos confirmados de dengue na cidade de Maringá e o INMET, de onde foram coletadas as variáveis climáticas, como a temperatura média, a umidade relativa, precipitação, entre outros dados. O período de coleta abrange uma série de dados de 10 anos, excluindo os anos com registros de surtos epidêmicos (2013 e

2014), conforme apontado pelo sistema InfoDengue. Estes anos foram considerados exceções devido ao seu comportamento incomum em comparação com o histórico habitual. As informações foram acessadas em formato digital, seguindo a Lei nº 12.527/2011, que garante o direito ao acesso à informação pública.

O acesso e a proteção dos dados é um aspecto fundamental nas práticas atuais da Ciência da Informação e da Ciência de Dados, desempenhando um papel importante ao longo de todo o ciclo de vida dos dados. De acordo com Moutinho (2024), com o aumento da complexidade dos fluxos de informações e a quantidade enorme de dados que são gerados todos os dias, não basta apenas usar tecnologias avançadas para armazenar e recuperar essas informações. É preciso também criar políticas éticas e legais que garantam a privacidade e os direitos das pessoas. A autora destaca que a proteção dos dados deve estar presente em todas as fases do ciclo — desde a coleta até o descarte — e que é importante seguir leis como a Lei Geral de Proteção de Dados (LGPD), aqui no Brasil, e o Regulamento Geral sobre a Proteção de Dados (GDPR). Além disso, a gestão adequada dos dados envolve estabelecer boas práticas de documentação, controle de acesso e manutenção da qualidade das informações, criando um ambiente seguro e confiável para o uso dos dados. Dessa forma, a proteção de dados, do ponto de vista da Ciência da Informação, vai além do aspecto técnico e se torna um compromisso ético com os usuários e a sociedade. Isso reforça a importância de os profissionais da área estarem sempre atualizados e treinados para lidar com as novas exigências do mundo digital.

Esta pesquisa irá primar pela segurança e confidencialidade das informações empregadas. Assim, os dados recolhidos serão submetidos a um processo minucioso de limpeza e filtragem do banco de dados, com o objetivo de fornecer apenas as informações indispensáveis para o progresso dos projetos. Informações confidenciais que possam identificar diretamente os pacientes, tais como nome, endereço, CPF ou qualquer outra informação pessoal, serão descartados na análise exploratória para assegurar a privacidade e o sigilo dos envolvidos. Assim, garante-se que a avaliação seja conduzida com fundamentos éticos e legais, protegendo a privacidade dos pacientes e cumprindo as leis em vigor, como a LGPD. Esta abordagem possibilita a utilização de dados pertinentes na elaboração dos modelos preditivos, mantendo a segurança e o anonimato dos participantes.

A limpeza dos dados é uma etapa fundamental para a ciência de dados. Ferreira (2020). Consiste na etapa de extração, transformação e carregamento dos dados (*Extract, Transform and Load - ETL*), assim como garantir o anonimato das informações. Nesta pesquisa, entende-se que a limpeza do banco é uma fase crucial para a elaboração de um

modelo preditivo mais assertivo. Como aborda Gaspar et al. (2023) a limpeza do banco antecede a análise e a modelagem:

Ao realizar a limpeza e o tratamento de dados, garantimos que eles estejam em um formato adequado para análise ou modelagem, reduzindo as chances de erro e aumentando as chances de conclusões mais confiáveis. O processo de limpeza de dados envolve tarefas como remover dados ausentes, substituir dados ausentes ou inconsistentes, aplicar transformações em dados numéricos, converter dados entre diferentes tipos de dados e até mesmo normalizar dados (16). O objetivo da limpeza de dados é tornar os dados mais confiáveis e consistentes para que possam ser usados para análise, ou seja com mais qualidade. Cabe destacar que este processo demanda um tempo importante em toda a análise de dados, recomenda-se destinar de 20% a 40% do tempo do projeto de análise para esta etapa. (Gaspar et al. 2023 p 25).

Chollet (2018) completa a afirmação, comentando que a limpeza do banco de dados é uma etapa fundamental no desenvolvimento de qualquer projeto de análise de dados. Morettin e Singer (2019), comentam que a análise exploratória se ocupa da organização e resumo dos dados, preparando-os para a inferência modelagem e a inferência estatística. Os autores ainda discorrem que a análise de dados é crucial para entender e organizar o banco de dados, para que seja possível sua interpretação mais correta. Reis (2008) complementa o pensamento afirmando que o papel da análise exploratória, antes chamada de Estatística Descritiva, consiste em resumir e organizar os dados coletados, através de tabelas, gráficos ou medidas numéricas, procurando alguma irregularidade que possa prejudicar a interpretação dos dados. Reis (2008) e Morettin e Singer (2019) concordam com a necessidade de uma análise exploratória bem realizada.

Morettin e Singer (2019) consideram a análise exploratória de dados como um passo essencial no processo estatístico, particularmente no âmbito da ciência de dados. Segundo eles, o propósito desta análise é estruturar, visualizar e sintetizar os dados, possibilitando ao pesquisador uma percepção inicial dos padrões e particularidades antes de utilizar métodos inferenciais ou modelos mais complexos. Além do disposto acima, destacam a importância dos *boxplots*, histogramas e gráficos como instrumentos para visualizar assimetrias, dispersão e a presença de valores extremos, *outliers* (valores atípicos na amostragem dos dados).

Para a elaboração do modelo preditivo desta pesquisa, foi utilizada a geolocalização dos casos confirmados e notificados de dengue registrados no banco de dados do SINAN, combinada com as variáveis climatológicas fornecidas pelo INMET. Essa abordagem visa explorar a relação espacial e ambiental dos casos de dengue, permitindo identificar padrões e prever possíveis surtos da doença. Conforme afirmado por Morettin e Singer (2019), a

modelagem preditiva se beneficia da integração de múltiplas fontes de dados para capturar as características do fenômeno estudado, o que pode resultar em modelos mais robustos e assertivos. A utilização de dados georreferenciados, aliada a informações climáticas, possibilita a construção de modelos que consideram tanto aspectos epidemiológicos quanto ambientais, essenciais para o monitoramento e controle da dengue.

Na análise exploratória, foi realizada a investigação da correlação entre as variáveis epidemiológicas, referentes aos casos confirmados e notificados de dengue, e as variáveis climáticas obtidas do INMET. Conforme destacado por Reis (2008), avaliar a correlação entre variáveis é essencial para identificar possíveis relações que possam influenciar o comportamento da doença, fornecendo uma base sólida para a construção do modelo preditivo. Como mencionado anteriormente, todo o processamento e análise dos dados seguirão rigorosamente as diretrizes da LGPD, garantindo a privacidade e o sigilo das informações dos pacientes, sem o uso de dados pessoais que possam identificar indivíduos.

Utilizou-se uma faixa de 10 anos de dados obtidos do SINAN com a incidência semanal de dengue e variáveis climatológicas extraídas do site INMET do mesmo período da cidade de Maringá, Estado do Paraná. Excluiu-se desta pesquisa os anos onde foram registrados ciclos epidêmicos no município, com informações obtidas no *site* Infodengue. Sendo assim, foram retirados os dados dos anos de 2010, 2013, 2014, 2016 e 2019. Entende-se que anos com dados fora da série histórica padrão, podem ser considerados com dados outliers. Gaspar et al. (2023) caracteriza *outlier* como uma observação que foge tanto do padrão que pode ser encarada como uma anomalia.

Os outliers são dados que estão muito distantes do padrão geral. Eles podem ser causados por erros de entrada de dados ou podem representar dados reais, por isso, é importante identificar e lidar com outliers para evitar problemas com a análise. O tratamento dos outliers podem seguir (ou não) as mesmas decisões tomadas para tratar os dados ausentes. (Gaspar et al, 2023, p.30).

Identificar *outliers* não é sempre uma tarefa fácil. Gaspar et al. (2023) comenta a importância de compreender e observar sempre os valores das variáveis, a fim de confrontá-las com os dados padrões. Reis (2008) completa a afirmação de Gaspar et al. (2023), quando comenta que valores *outliers* podem distorcer modelos. Por isso é importante interpretar corretamente o valor da média.

O Instituto Nacional de Meteorologia possui diversos bancos climatológicos, com dados abertos à população. O site “<https://bdmep.inmet.gov.br/>” abrange um grande banco de dados, que podem ser solicitados com, base na Lei nº12.527/2011, em que estabelece que

órgãos e entidades públicas devem divulgar, independente de solicitação, informações de interesse geral ou coletivo, exceto aqueles cuja confidencialidade esteja prevista no texto legal (Brasil, 2025).

3 PROCEDIMENTOS METODOLÓGICOS

Quanto à metodologia de pesquisa, a abordagem escolhida foi a quantitativa. Godoy (1995) entende que a pesquisa se caracteriza pela obtenção de dados sobre pessoas, lugares ou processos administrativos, procurando entender fenômenos ou situações. Pode-se entender com isso que a pesquisa quantitativa tem um viés de investigação, com o objetivo de entender um fato ou situação. Godoy (1995) completa quando argumenta que o processo de pesquisa quantitativa, com o passar dos tempos, gerou diversidade de métodos de investigação, análise e apresentação de resultados.

O objetivo deste trabalho é criar um modelo preditivo que possa antecipar o número de casos prováveis de dengue em Maringá, usando dados epidemiológicos e fatores meteorológicos. Esse objetivo requer uma abordagem quantitativa, uma vez que implica na medição de fenômenos observáveis, na análise estatística de grandes volumes de dados e na modelagem matemática de padrões encontrados em séries temporais.

A pesquisa quantitativa é apropriada para identificar relações entre variáveis, testar hipóteses e gerar estimativas objetivas com base em dados numéricos. Segundo Marconi e Lakatos (2003), essa abordagem possibilita ao pesquisador “quantificar os dados, aplicar técnicas estatísticas e generalizar resultados a partir de uma amostra representativa”. De forma semelhante, Morettin e Singer (2019) afirmam que análises estatísticas robustas requerem uma organização rigorosa dos dados, com a precisão das medições e a consistência das bases sendo fundamentais para a validação dos resultados.

Nesse cenário, a utilização de métodos de aprendizado de máquina, como o algoritmo Random Forest, é justificada pela habilidade de gerenciar diversas variáveis ao mesmo tempo, reconhecer padrões não lineares e gerar previsões fundamentadas em estruturas matemáticas robustas. Dessa forma, a natureza do problema em estudo, focada na previsão de casos de dengue com base em dados numéricos e séries temporais, requer a utilização de uma abordagem quantitativa, sustentada por métodos estatísticos e computacionais apropriados para lidar com a complexidade do fenômeno.

Esta pesquisa é classificada como quantitativa, uma vez que trata os dados de maneira sistemática e objetiva, empregando métodos estatísticos para examinar informações provenientes de sistemas de monitoramento de saúde, como o SINAN. De acordo com Thomaz et al. (2015), essas informações possibilitam a execução de diagnósticos dinâmicos sobre a incidência de eventos de saúde na população, auxiliando na tomada de decisões. Além disso, de acordo com Marconi e Lakatos (2003), a pesquisa quantitativa visa quantificar

variáveis e estabelecer conexões entre elas através de métodos estruturados e padronizados, o que está em consonância com as metas deste estudo.

3.1 Etapas Metodológicas

A metodologia desta pesquisa foi organizada em etapas sucessivas, alinhando-se aos princípios de Godoy (1995), que enfatiza a relevância de um processo sistemático e estruturado na realização de pesquisas quantitativas. De maneira semelhante, Marconi e Lakatos (2003) ressaltam que a transparência no delineamento metodológico aumenta a confiabilidade dos resultados, possibilitando sua replicação e validação por outros pesquisadores. Os autores destacam a necessidade de clareza e precisão na explicação dos procedimentos metodológicos para assegurar que outros pesquisadores possam reproduzir os estudos corretamente. Esse princípio é fundamental para a validação e a credibilidade dos resultados científicos. De Toni e Dornelles (2022) enriquecem quando falam da ciência de dados como uma ferramenta estratégica, que auxilia na criação e acompanhamento de políticas públicas: nesse contexto, os autores expandem essa visão ao destacar que, no âmbito das políticas públicas, a ciência de dados deve ser entendida como um instrumento estratégico, cuja utilização requer métodos claros, etapas bem definidas e alinhamento entre análise técnica e metas institucionais. Os autores afirmam que a organização metódica das etapas analíticas é essencial para que os resultados obtidos possam realmente apoiar a criação, a supervisão e a avaliação de políticas públicas.

A construção do modelo preditivo foi realizada usando o algoritmo Random Forest, conhecido por sua eficácia na modelagem de dados complexos, como mencionam Kuhn e Johnson (2013) e utilizando ferramentas gratuitas de software, como o Google Colab. A fase seguinte consistiu na avaliação do desempenho do modelo utilizando métricas estatísticas, como R^2 , MAE e RMSE. Os resultados foram analisados, levando em conta suas consequências para a criação de políticas públicas de saúde, a fim de apoiar a administração municipal no combate à dengue. A Figura 3 apresenta as etapas seguidas pela pesquisa modelo das etapas:

Fluxograma da Metodologia

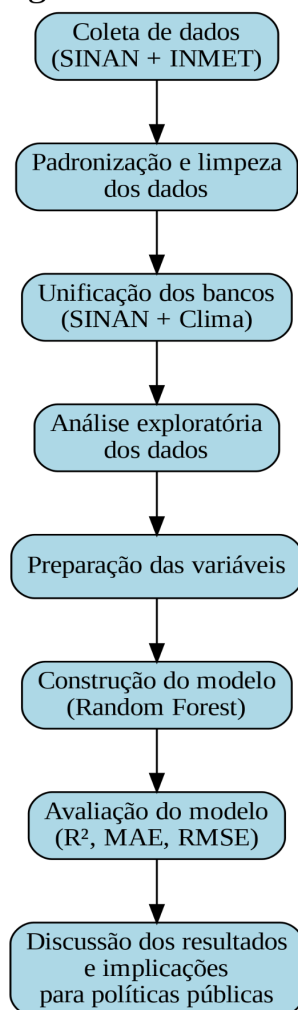


Figura 03 - Fluxograma das etapas da pesquisa:

Fonte: a pesquisa (2025)

3.2 Árvore de decisão

Como apresentado, tem-se a necessidade de realizar uma análise mais detalhada de um grande conjunto de dados com várias informações inseridas; logo processar essas informações torna-se complexo. Para fazer este modelo, faz-se o uso de um modelo de aprendizado de máquina, em formato de árvore de decisão.

De acordo com Breiman et al. (1984), essa ferramenta representa, por meio de um diagrama, as diferentes opções e resultados das decisões, assim como a probabilidade de cada

um ocorrer, com o objetivo principal de entender as variáveis e suas interações que levam ao fenômeno em estudo. A árvore de decisão é graficamente exibida por nós e ramificações, assemelhando-se a uma árvore invertida. Han e Kamber (2001). Kuhn e Johnson (2013) completam essa afirmação onde apresentam instruções encadeadas que separam os dados dentro das partições, aplicam previsões, reforça a noção de divisão sucessiva dos dados até atingir maior homogeneidade, no conceito de árvore de decisão. Além disso, conforme afirmam Kuhn & Johnson (2013), as árvores de decisão podem ser vistas como um conjunto de diretrizes do tipo "se-então", estruturadas de maneira hierárquica e interligada, cujos objetivos são direcionar a previsão dos resultados com base nos valores dos preditores. Dessa forma, cada divisão em um nó leva ao processo até a obtenção dos chamados nós terminais ou folhas, onde se encontra a resposta final do modelo.

Kuhn e Johnson (2013) também salientam que a elaboração da árvore ocorre por meio da divisão recursiva, ou seja, da separação contínua dos dados em subconjuntos cada vez mais homogêneos em relação à variável de interesse. Entende-se que essa característica assegura que o modelo se adapte de forma mais eficaz às particularidades de cada grupo identificado, aprimorando a pureza dos nós e, por consequência, a eficácia preditiva da árvore, conforme figura 4.

No contexto específico das árvores de classificação, a lógica mantém a mesma estrutura utilizada nas árvores de regressão: a meta é dividir os dados em agrupamentos mais coesos, onde cada nó exiba uma predominância clara de uma classe (Kuhn; Johnson, 2013). Desta maneira, a árvore de decisão se revela como uma abordagem que combina facilidade de interpretação com uma sólida capacidade de representar interações complexas entre variáveis.

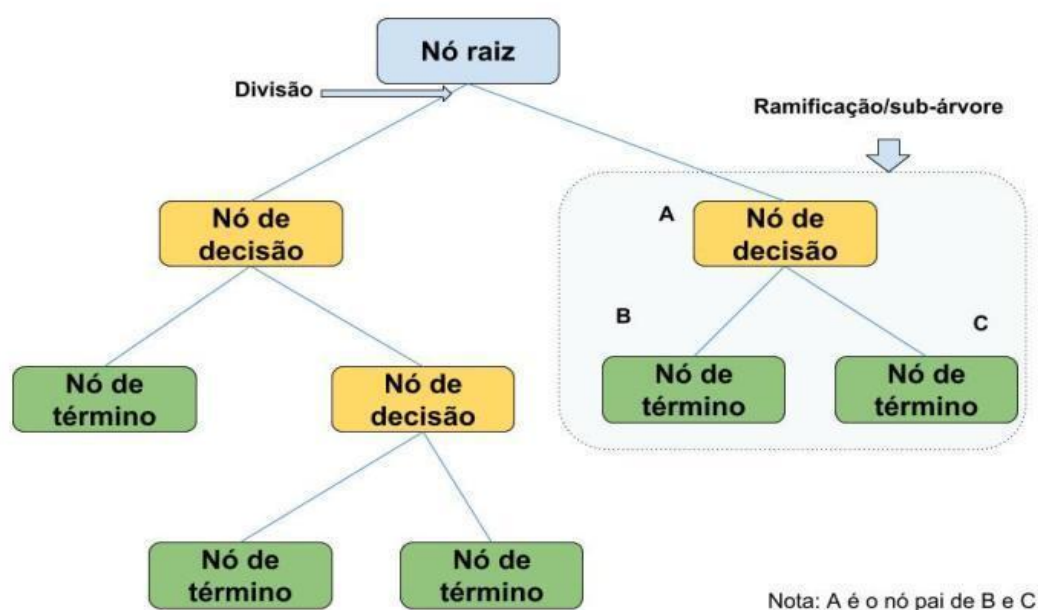


Figura 04: Ramificação de uma árvore de decisão
 Fonte: adaptação de Kuhn & Johnson (2013).

3.3 Random Forest

Para essa pesquisa, o algoritmo *Random Forest* foi utilizado como método para a modelagem preditiva da ocorrência de casos de dengue na cidade de Maringá. Essa opção é baseada em sua reconhecida solidez, adaptabilidade e habilidade para lidar com dados complexos, atributos fundamentais ao lidar com, por exemplo, variáveis climáticas e epidemiológicas, que possuem alta correlação e comportamento não linear, conforme mencionado por Kuhn e Johnson (2013).

O *Random Forest* é uma técnica de aprendizado de máquina que utiliza árvores de decisão e se baseia no conceito de *bagging* (*bootstrap aggregating*) (Kuhn; Johnson 2013). O modelo gera diversas árvores de forma autônoma, utilizando subconjuntos de dados obtidos por meio de reamostragem aleatória com reposição. Ademais, para a divisão em cada nó da árvore, considera-se apenas uma parte aleatória das variáveis preditoras. Kuhn e Johnson (2013) e Breiman (2001) concordam que o resultado final é obtido por meio da combinação dos desempenhos individuais de cada árvore, seja pela média (em casos de regressão) ou pela votação majoritária, em casos de classificação.

Breiman (2001) completa essa afirmação, quando menciona que essa técnica permite reduzir a correlação entre árvores individuais e, conseqüentemente, o erro de generalização do

modelo. O autor ainda destaca que o *Random Forest* apresenta robustez frente a ruídos nos dados, além de fornecer mecanismos internos, que permite calcular o erro de predição e a importância das variáveis sem necessidade de conjuntos de validação. Esses elementos reforçam a confiabilidade do método e embasam sua ampla aplicação em cenários complexos, como séries temporais epidemiológicas, que se utilizou neste trabalho.

Kuhn e Johnson (2013) também comentam que as principais vantagens do *Random Forest* é sua resistência ao sobreajuste (*overfitting*), um problema comum em árvores de decisão individuais. Ao mesclar várias árvores, o algoritmo cria um modelo mais robusto e menos suscetível a variações nos dados de treinamento. Outra característica relevante é a capacidade de calcular medidas de importância das variáveis, o que facilita a identificação de quais atributos (como temperatura, precipitação ou defasagens de casos anteriores) têm maior impacto no número de casos previstos.

Conforme mencionado, o *Random Forest* é o algoritmo mais apropriado para esta pesquisa, pois combina desempenho, robustez e interpretabilidade. Essas características são essenciais para fornecer previsões confiáveis que possam apoiar a tomada de decisões na criação e execução de políticas públicas de combate à dengue em nossa cidade.

3.4 Validação de Modelos

A validação é um passo fundamental na modelagem, pois possibilita a avaliação da habilidade de generalização do algoritmo e diminui a probabilidade de sobreajuste (*overfitting*). Kuhn e Johnson (2013) afirmam que a validação cruzada é uma das técnicas mais empregadas nesse contexto, uma vez que permite estimar de maneira robusta o desempenho preditivo de modelos em dados não observados.

Contudo, ao lidar com séries temporais, o uso de técnicas convencionais, como o *k-fold cross-validation*, podem gerar viés, pois essas técnicas dividem os dados de forma aleatória, ignorando a sequência temporal. Nesse tipo de dado, é essencial manter a ordem temporal para impedir que informações futuras afetem de forma inadequada o treinamento do modelo (Hyndman; Athanasopoulos, 2018).

Nesse contexto apresentado, o método *Leave-One-Year-Out* (LOYO) foi utilizado neste estudo. Segundo Tashman (2000) essa abordagem envolve a segmentação da base de dados por períodos anuais, usando todos os anos disponíveis para o treinamento e destinando um ano específico para o teste, de maneira rotativa. Dessa forma, cada ano é analisado como

um conjunto de validação separado, o que garante maior confiabilidade na avaliação do desempenho

Segundo Bergmeir e Benítez (2012), a principal vantagem do LOYO é sua capacidade de acompanhar o aspecto temporal dos dados epidemiológicos, o que possibilita a avaliação do modelo em variados contextos sazonais sem alterar a sequência natural dos registros. Ao tratar anos completos como uma unidade de teste, o método permite uma simulação mais fiel do uso do modelo em cenários futuros, refletindo a aplicação prática na vigilância epidemiológica.

Com esse embasamento, entende-se que a utilização do LOYO neste trabalho é justificada tanto pela literatura especializada quanto pela natureza dos dados analisados, considerando que a dengue exhibe variações sazonais e interanuais que poderiam ser ocultadas por outros métodos de validação.

4- PREPARAÇÃO DOS DADOS E ELABORAÇÃO DO MODELO

4.1 Análise Exploratória dos dados

Antes de iniciar qualquer modelagem, os dados obtidos do SINAN online e do INMET, foram preparados para o treinamento do modelo *Random Forest*. A primeira fase deste procedimento envolveu a normalização dos dados, sendo esta essencial para impedir que as diferentes escalas das variáveis afetasse de maneira desproporcional os resultados do modelo. A normalização (ou padronização) foi executada utilizando a fórmula básica:

$$x' = (x - \mu)/\sigma$$

onde x' é o valor normalizado, x é o valor original, μ é a média e σ é o desvio padrão da variável. Essa conversão garante que todas as variáveis estejam na mesma escala, facilitando assim o aprendizado da rede neural. Importante salientar que, como mencionado por Kuhn e Johnson (2013), a normatização e padronização dos dados é essencial para um modelo mais assertivo. Morettin e Singer (2019) complementam essa afirmação quando afirmam a importância de padronizar uma variável x com média μ e desvio padrão σ por meio da transformação $x' = (x - \mu)/\sigma$. Além disso, os dados foram segmentados em dois grupos: um destinado ao treinamento e outro ao teste, 70/30, com a finalidade de avaliar a capacidade de generalização do modelo.

Conforme apresentado anteriormente nesta pesquisa, para a investigação preliminar dos dados epidemiológicos relacionados à dengue, foram utilizados dados do SINAN, no formato Data Base File (DBF), abrangendo os anos de 2008, 2009, 2010, 2011, 2012, 2015, 2016, 2017, 2018 e 2019, que são considerados anos sem epidemias na cidade de Maringá-PR.

Após a padronização dos dados de cada ano, os conjuntos de informações foram agrupados em um único *DataFrame*, utilizando o Google Colab, abrangendo todas as notificações de dengue de Maringá durante o intervalo de tempo selecionado, organizadas por semana epidemiológica no formato AAAASS (ano e semana). As colunas categóricas restantes foram convertidas para um formato numérico, garantindo que os dados estivessem prontos para análise estatística e modelagem preditiva.

Para a análise exploratória, foi gerada a matriz de correlação entre as variáveis numéricas, que foi apresentada através de um gráfico de calor. Esta fase possibilitou a

identificação de correlações entre os sintomas, as características demográficas e a classificação final dos casos, ajudando na escolha de variáveis pertinentes para o modelo preditivo.

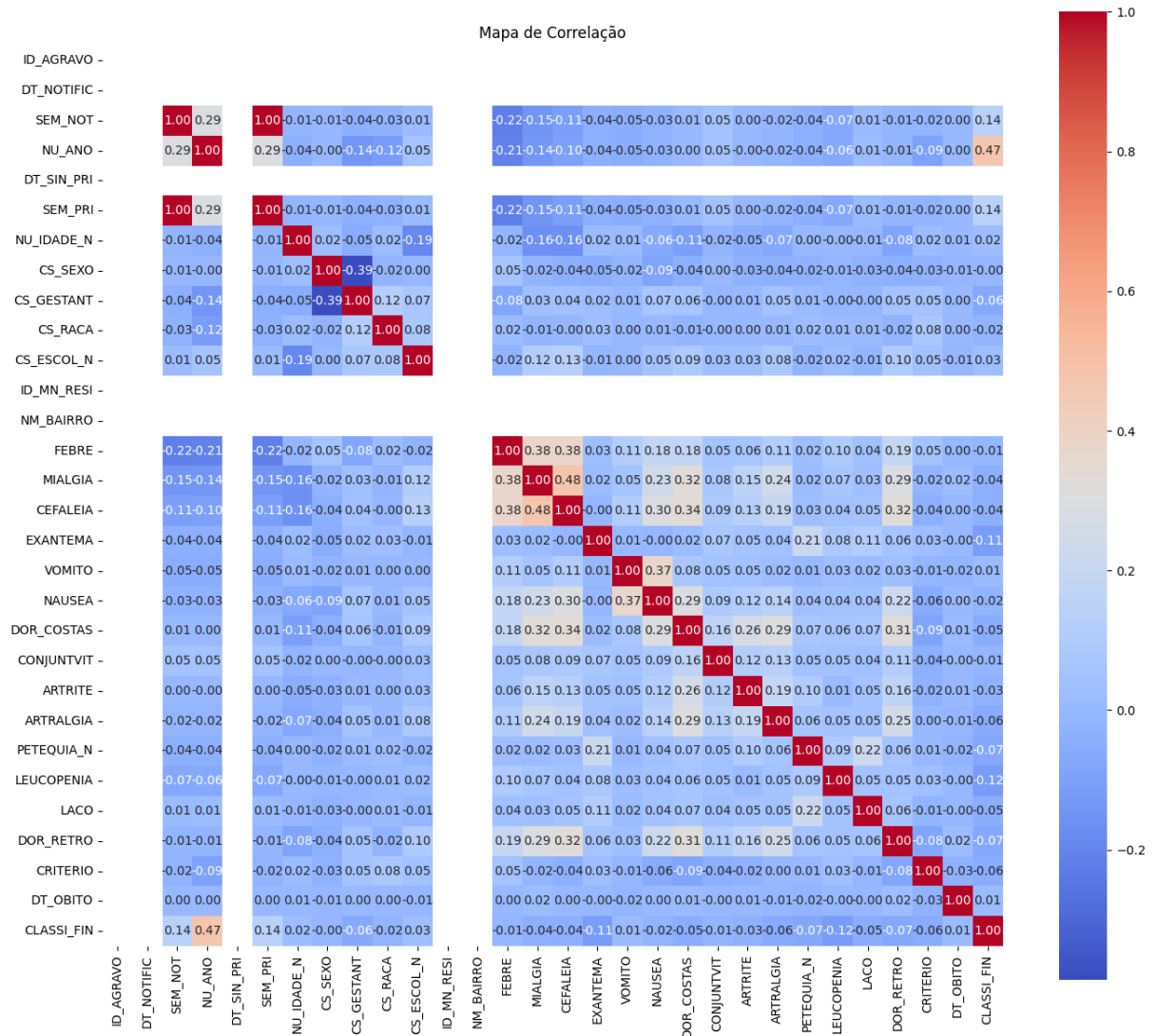


Figura 05: Correlação entre variáveis epidemiológicas

Fonte: a pesquisa (2025).

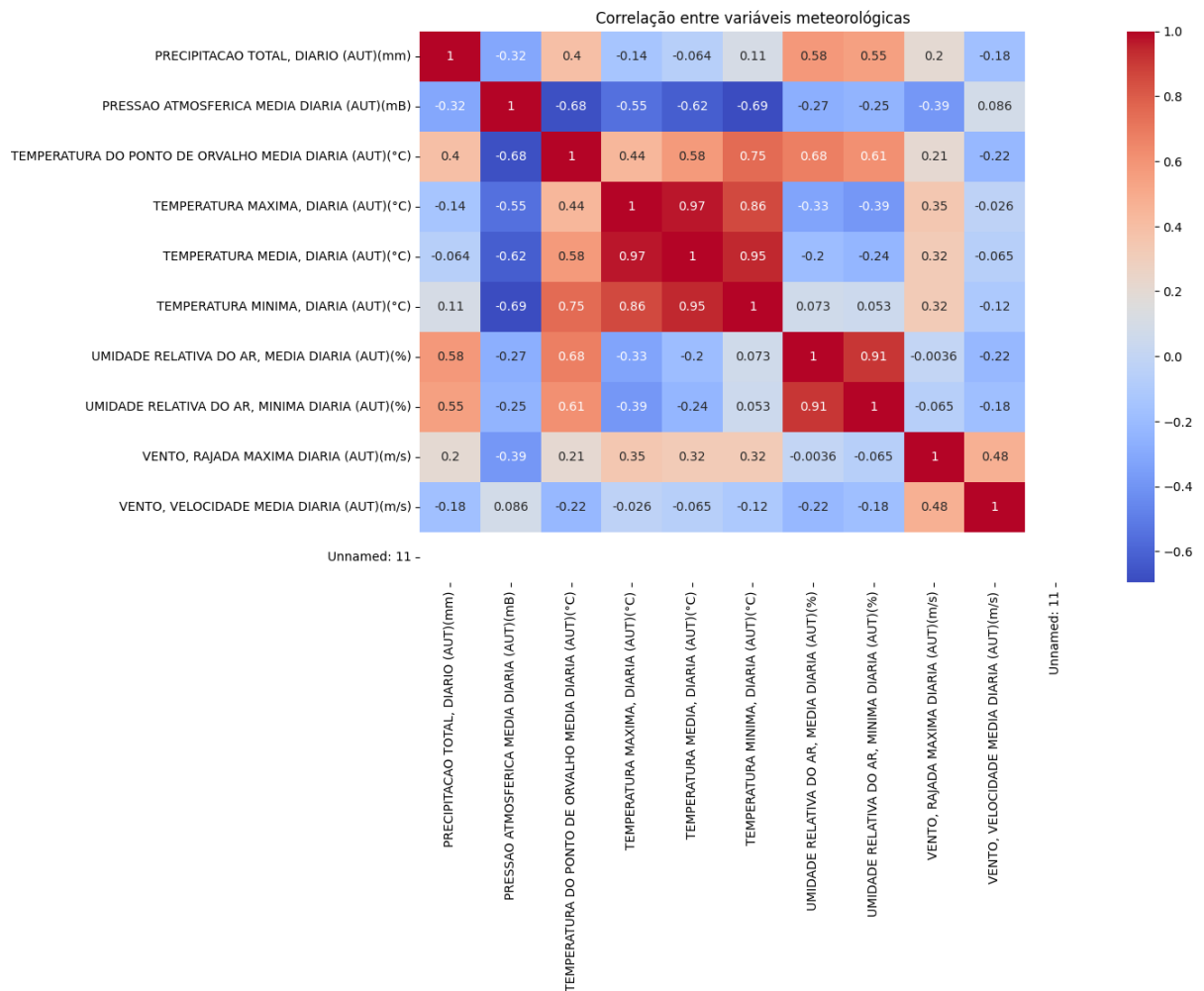


Figura 06 - Correlação entre variáveis meteorológicas

Fonte: a pesquisa (2025)

Para complementar a modelagem, foram utilizados dados diários de clima do período entre 2008 e 2019, obtidos do INMET e disponibilizados para uso. As datas das medições foram convertidas para o formato de data e hora, o que permitiu a criação de uma coluna identificadora da semana epidemiológica (SEMANA_EPI) que é compatível com os dados do SINAN. Essa coluna foi formada por uma combinação do ano e da semana epidemiológica, gerando um código único para cada semana (AAAASS), facilitando a integração com os dados de epidemiologia no futuro.

Para que os dados fossem adequados à análise semanal, os registros diários foram agrupados por semana epidemiológica, calculando-se as médias de variáveis contínuas como temperatura máxima, média e mínima, umidade relativa média, média da velocidade do vento e pressão atmosférica, além do total de precipitação. Essa estratégia permitiu a transformação

da desagregação dos dados de diária para semanal, preservando as informações climáticas relevantes que poderiam correlacionar com os casos de dengue.

Os dados foram ainda filtrados para incluir apenas os anos que são relevantes para a análise (2008–2012 e 2015–2019), assegurando a consistência temporal com o conjunto de dados epidemiológicos. Para a análise exploratória, foram observadas as distribuições das variáveis climáticas através de histogramas, além de serem calculadas as correlações entre elas, resultando em um mapa de calor. Essa fase possibilitou a identificação de relações entre várias variáveis climáticas, como temperatura, umidade, precipitação e vento, o que ajudou na seleção das variáveis mais importantes para a modelagem preditiva.

O objetivo inicial da pesquisa era criar um modelo que previsse casos de dengue em nível de bairro, permitindo uma análise espacial detalhada da ocorrência de casos. Entretanto, a análise preliminar mostrou que uma boa parte dos registros de endereços continham erros, faltas ou informações equivocadas. Essa questão comprometeria a qualidade das inferências espaciais, resultando em previsões que poderiam ser errôneas e distorcidas caso fossem aplicadas diretamente no modelo. A possibilidade de incluir dados geográficos imprecisos tornava impraticável a elaboração de um modelo confiável por bairros, já que a falta de precisão na localização afetaria diretamente a exatidão das previsões e a interpretação dos dados.

Diante dessa limitação, decidiu-se por prever os casos de dengue em uma escala municipal, onde foi adotado o município de Maringá como a unidade de análise. Para isso, os casos de dengue foram agregados semanalmente e combinados com dados climáticos coletados também por semanas. Essa estratégia possibilita o uso de séries temporais semanais que eram consistentes e confiáveis, mantendo a validade estatística do modelo de previsão.

4.2 Configuração do Modelo

O modelo preditivo foi desenvolvido com o uso do Random Forest, levando em conta um período de previsão de doze semanas. Para entender a dinâmica temporal da dengue, foram geradas variáveis a partir de *lags* e médias móveis das séries de casos e das variáveis climáticas, possibilitando que o modelo integrasse dados anteriores sem a necessidade de informações futuras. O objetivo do modelo (*target*) foi estabelecido como a quantidade de casos prevista para doze semanas adiante, em conformidade com o período de intervenção epidemiológica.

Também se pode comentar sobre o uso de um otimizador de hiperparâmetros, chamado *GridSearchCV*. É uma técnica para melhorar modelos que examina todas as

combinações viáveis de hiperparâmetros determinados pelo usuário, utilizando validação cruzada para medir o desempenho de cada combinação. Esse método fornece a configuração que gera o melhor resultado médio, assegurando que o modelo seja afinado de maneira eficaz e segura. Entre outras palavras, o *GridSearchCV* encontra automaticamente a melhor combinação de parâmetros do modelo para melhorar sua performance.

4.3 Execução do Modelo

O primeiro modelo foi ajustado para prever casos de dengue com um horizonte de doze semanas. Para isso, foram empregadas variáveis defasadas (*lags*) dos casos reportados, além de variáveis climáticas (precipitação, temperatura, umidade, vento e pressão atmosférica), já padronizadas por médias móveis semanais, acompanhando os casos de dengue por semana, do município de Maringá. Os dados foram divididos de maneira temporal, com o último ano reservado para o teste final e o restante usado para o treinamento.

Os resultados mostraram que, apesar de o modelo ter capturado parcialmente a dinâmica temporal dos casos, sua capacidade preditiva foi limitada em horizontes longos. O coeficiente de determinação (R^2) foi baixo, e os erros médios (MAE e RMSE) mostraram discrepância considerável entre os valores previstos e observados. A análise da relevância das variáveis mostrou que os *lags* epidemiológicos (especialmente CASOS_lag8 e CASOS_lag12) tiveram maior importância, seguidos por fatores climáticos relacionados à temperatura e precipitação. Esses resultados indicam que previsões em janelas temporais extensas, como doze semanas, tendem a ter maior incerteza, o que justifica a adaptação do modelo para horizontes mais curtos.

O modelo de previsão ajustado para um horizonte de doze semanas mostrou um desempenho restrito em relação à capacidade de explicação e à precisão. O coeficiente de determinação ($R^2 = 0,336$) sugere que o modelo foi capaz de explicar apenas aproximadamente 33,6% da variabilidade dos casos analisados, demonstrando uma performance preditiva modesta ao se considerar o intervalo temporal mais extenso. Os valores de erro corroboram essa limitação: o erro absoluto médio (MAE) foi de cerca de 38,9 casos, ao passo que a raiz do erro quadrático médio (RMSE) alcançou 64,07 casos. Isso indica diferenças consideráveis entre os valores previstos e os reais, principalmente em semanas com um volume maior de notificações:

```
Horizonte: 12 semanas à frente
```

```
R² : 0.336
```

```
MAE : 38.90 casos
```

```
RMSE : 64.07 casos
```

```
Top 15 importâncias:
```

```
umid_media_rollmean4 0.599014
```

```
umid_media_lag0 0.046930
```

```
temp_media_rollmean4 0.030877
```

```
umid_media_lag1 0.028588
```

```
umid_media_lag2 0.022208
```

```
temp_media_lag0 0.021580
```

```
umid_media_rollmean2 0.020806
```

```
temp_media_rollmean2 0.018668
```

```
CASOS_lag8 0.017233
```

```
pressao_atm_lag0 0.014974
```

```
CASOS_lag12 0.014176
```

```
temp_media_lag1 0.014095
```

```
temp_media_lag2 0.012704
```

```
umid_media_lag3 0.012115
```

```
pressao_atm_lag2 0.011442
```

```
dtype: float64
```

Figura 07: Resultado do modelo preditivo com 12 semanas de horizonte.

Fonte: a pesquisa (2025).

Já o modelo preditivo criado para um horizonte de 8 semanas mostrou resultados positivos, demonstrando uma boa habilidade para identificar os padrões temporais da dengue na cidade de Maringá. O coeficiente de determinação (R^2) mostrou que o modelo foi capaz de explicar uma parte considerável da variação nos casos notificados. Ao mesmo tempo, os valores do erro absoluto médio (MAE) e da raiz do erro quadrático médio (RMSE) indicaram um nível de precisão satisfatório, com diferenças relativamente pequenas entre os valores previstos e os valores observados.

Na avaliação da relevância das variáveis, um conjunto mais amplo de fatores significativos foi destacado. As defasagens dos casos notificados em semanas anteriores (CASOS_lag) tiveram um impacto significativo, indicando que o histórico recente da série epidemiológica é um preditor robusto do comportamento futuro. Dentre as variáveis climáticas, a umidade e a temperatura médias, em suas versões defasadas e médias móveis, mostraram-se relevantes, sugerindo que as condições meteorológicas têm um impacto direto no aumento ou redução da incidência de casos. O modelo exibiu um desempenho extremamente satisfatório. O coeficiente de determinação ($R^2 = 0,917$) indica que o modelo conseguiu explicar 91% da variabilidade dos casos notificados, demonstrando um ajuste excelente aos dados. Os valores de erro também corroboram a precisão obtida: um MAE de

cerca de 19 casos e um RMSE de aproximadamente 31 casos sugerem que as previsões se alinham de forma consistente com os registros reais, mesmo em um horizonte temporal relativamente extenso.

```

Horizonte: 8 semanas à frente
R²      : 0.917
MAE     : 18.98 casos
RMSE    : 31.54 casos

Top 15 variáveis mais importantes:
umid_media_roll8      0.092646
precipitacao_roll8    0.059821
umid_media_roll2      0.045126
CASOS_lag0            0.043282
CASOS_roll2           0.039959
CASOS_lag1            0.038289
CASOS                 0.037022
umid_media_roll4      0.034365
CASOS_lag2            0.033787
umid_media_lag8       0.031170
CASOS_roll4           0.028259
pressao_atm_roll4     0.025169
pressao_atm_roll8     0.024286
CASOS_roll8           0.020308
umid_media_lag4       0.019942
dtype: float64

```

Figura 08 - Resultado do modelo preditivo com 08 semanas de horizonte.

Fonte: a pesquisa (2025).

4.4 Validação do Modelo

Para avaliar a eficácia do modelo utilizou-se o *leave-one-year-out*, (LOYO), *cross-validation*. Izbicki e Dos Santos (2020) afirmam que este método é um dos mais eficientes para medir a competência de um modelo preditivo:

$$R^2(g) = \frac{1}{n} \sum_{i=1}^n (Y_i - g_{-i}(X_i))^2$$

Onde que g_{-i} é ajustado utilizando-se todas as observações exceto a i -ésima delas. Essa metodologia possibilitou a avaliação da habilidade de generalização em diversos períodos, levando em conta os anos não epidêmicos de 2008 a 2019 na cidade de Maringá.

De modo geral, o desempenho permaneceu estável ao longo dos anos, apresentando altos valores de R^2 e erros relativamente baixos. Isso sugere que o modelo foi capaz de identificar padrões consistentes, mesmo quando treinado sem dados do ano em questão. Essa

análise é especialmente importante porque simula casos em que o modelo é usado em períodos sem registros, o que é fundamental para cenários de vigilância e previsão.

Os gráficos de previsão em comparação com os dados observados para cada ano mostraram uma boa correspondência entre os valores estimados e os casos reportados. Embora alguns desvios tenham sido notados em semanas de baixa incidência (quando pequenas variações relativas têm maior impacto no erro percentual), os padrões gerais de subida e descida foram corretamente identificados, o que comprova a eficácia do modelo para sinalizar tendências.

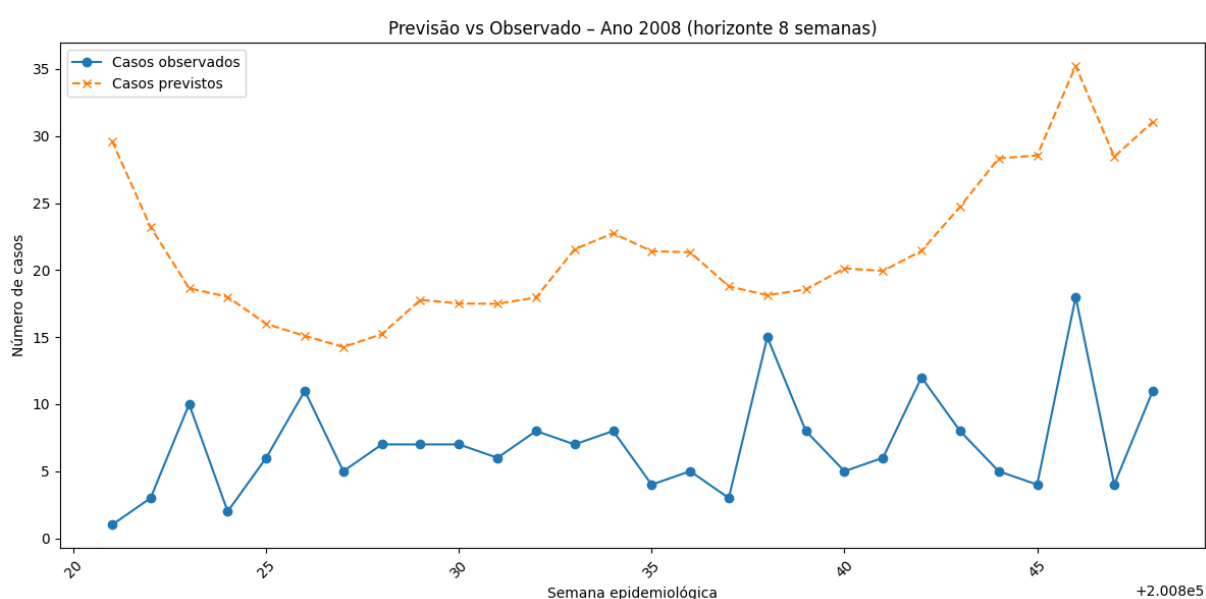


Figura 09 - Validação do modelo Previsão vs Observados, com dados do ano de 2008.
Fonte: a pesquisa (2025).

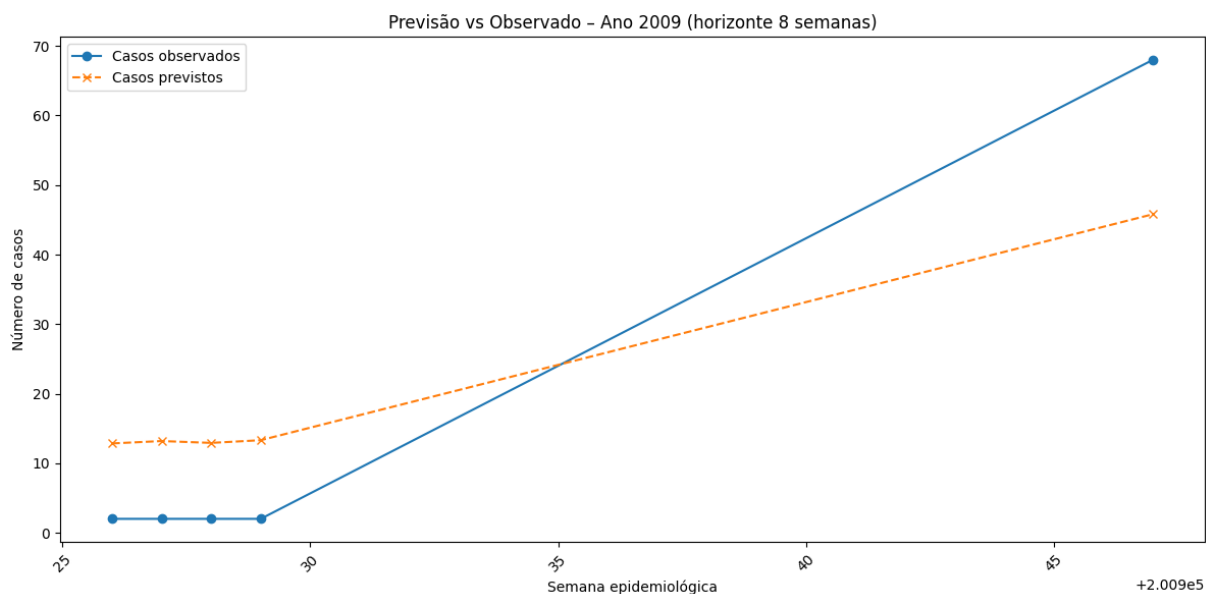
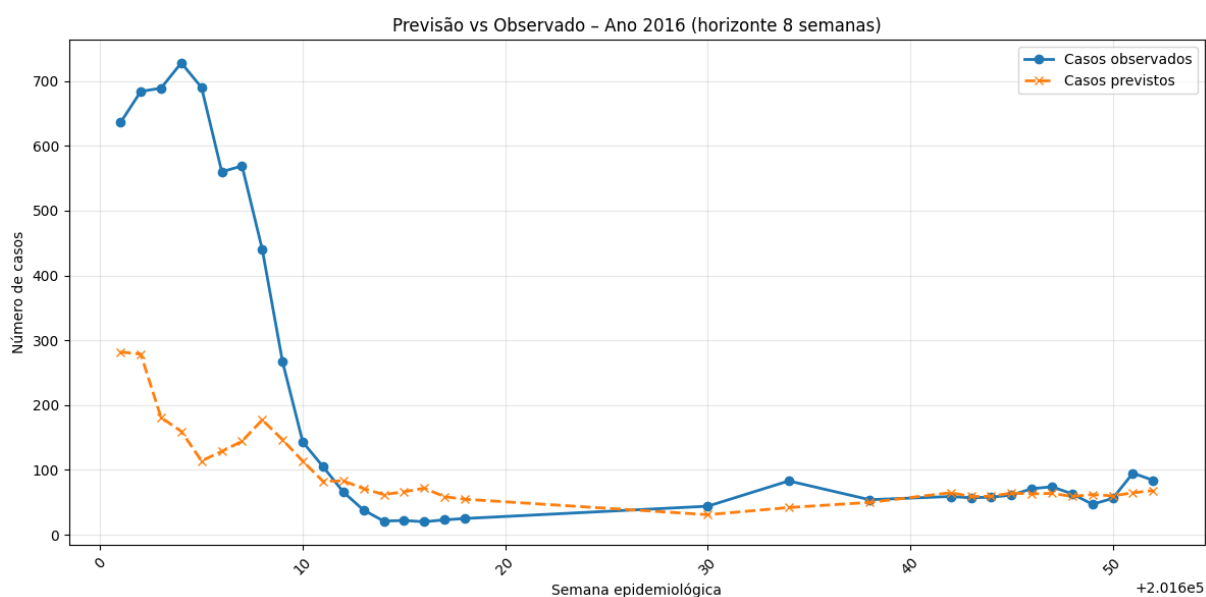


Figura 10 Validação do modelo Previsão vs Observados, com dados do ano de 2009.
Fonte: a pesquisa (2025).



5- RESULTADOS

O modelo preditivo desenvolvido demonstrou uma grande capacidade de previsão dos casos semanais de dengue na cidade de Maringá. O Coeficiente de Determinação (R^2) atingiu 0,917, sugerindo que cerca de 91,7% da variação no número de casos foi explicada pelo modelo. Esse resultado indica um elevado grau de precisão em modelos baseados em séries epidemiológicas, que são marcadas por grande variabilidade e impacto de fatores externos.

Quando abordados os erros de previsão, o modelo exibiu um Erro Absoluto Médio (MAE) de 18,98 casos. Isso indica que, em média, as estimativas divergiram dos valores observados em menos de 20 casos por semana epidemiológica. O Erro Quadrático Médio (RMSE) foi de 31,54 casos, indicando que o desempenho permaneceu satisfatório, mesmo durante os períodos de maior oscilação da série.

A comparação entre diferentes horizontes de previsão foi outro aspecto muito importante. O modelo com oito semanas de antecedência demonstrou um desempenho significativamente superior ao do modelo de doze semanas, tanto no que diz respeito à precisão quanto à estabilidade dos resultados. O horizonte mais curto demonstrou uma elevada capacidade preditiva ($R^2 = 0,917$), ao passo que o horizonte de doze semanas exibiu uma maior variação nos erros e uma menor confiabilidade ($R^2 = 0,336$). Esse resultado confirma que as previsões de médio prazo (até 8 semanas) são mais apropriadas para auxiliar a administração municipal de Maringá, pois oferecem tempo adequado para o planejamento das ações sem prejudicar a precisão das estimativas.

Tabela X - Comparação do desempenho do modelo Random Forest em diferentes horizontes de previsão

Horizonte de previsão	R^2	MAE (casos)	RMSE (casos)
8 semanas	0.917	18.98	31.54
12 semanas	0.336	38.98	64.07

Figura 12 - Comparativo de Horizontes de Previsão de oito e doze semanas.

Fonte: a pesquisa (2025)

Embora essa abordagem não possibilite uma análise minuciosa por bairros, ela garantiu que o modelo capturasse corretamente a dinâmica temporal da dengue no município de Maringá, permitindo previsões robustas em um nível municipal. As limitações dos dados

de endereço e a adaptação do modelo foram levadas em conta no planejamento da metodologia e na interpretação dos resultados, assegurando que as conclusões obtidas a partir da modelagem fossem consistentes.

Para avaliar a robustez do modelo preditivo, utilizou-se a técnica de validação cruzada *Leave-One-Year-Out* (LOYO). Os resultados da validação LOYO mostraram que o modelo teve um desempenho consistente ao longo dos anos, indicando que ele não só se ajustou bem aos dados de treinamento, mas também preservou sua capacidade preditiva em períodos que não foram usados para o ajuste. Essa evidência aumenta a confiança no modelo para uso prático no planejamento de políticas públicas.

Verifica-se que a criação deste modelo preditivo exibiu limitações significativas, gargalos, decorrentes da qualidade dos dados inseridos no sistema. Notadamente, o campo de endereço dos pacientes apresentou inconsistências. Durante o processamento do banco de dados do SINAN, constatou-se um volume considerável de registros incompletos, inconsistentes ou ausentes nesta variável. Tal fato impediu a execução de análises mais precisas em nível de bairro. Esta questão restringiu a capacidade de espacialização da previsão de casos no município de Maringá. Consequentemente, o modelo ficou restrito à escala municipal.

A ausência de dados de endereço corretos, em alguns casos verificados nos bancos, comprometeu a solidez do modelo preditivo inicialmente proposto por esta pesquisa. A padronização, qualificação e automação do processo de coleta e registro dos dados de residência dos pacientes constituem um investimento municipal interessante. Porém não foi iniciado até o presente momento. Tal ação deve ocorrer mediante a capacitação das equipes de saúde e a implementação de sistemas eletrônicos. Estes sistemas devem minimizar erros de digitação e omissões. Com registros consistentes, aprimoram-se os modelos preditivos futuros. Isso viabilizará análises espaciais futuras mais assertivas por bairro e oferece subsídios precisos às políticas públicas de enfrentamento da dengue.

6- PRODUTO E PROPOSTA DE INTERVENÇÃO

O produto principal deste estudo consiste em um modelo preditivo para a incidência de dengue no município de Maringá, empregando o algoritmo Random Forest. Este modelo foi desenvolvido com base em dados do SINAN, abrangendo períodos não epidêmicos entre 2008 e 2019. Variáveis meteorológicas, fornecidas pelo INMET, também foram incorporadas. O desempenho preditivo do modelo demonstrou-se elevado em um horizonte de oito semanas, com um coeficiente de determinação (R^2) de 0,917. Tal resultado indica uma capacidade robusta de antecipar padrões de aumento e diminuição nos casos de dengue em Maringá.

O modelo preditivo desenvolvido por essa pesquisa está disponível em formato digital e pode ser executado em qualquer sistema computacional que suporte linguagens de programação e bibliotecas de ciência de dados, como *Python*, desde que os dados de entrada estejam devidamente padronizados (anexos 3, 4 e 5). O gestor municipal pode ter acesso ao algoritmo e à sua interface de execução por meio de um notebook interativo (*Jupyter Notebook* ou *Google Colab*), de um *script* automatizado ou aplicação *web* que esteja integrada ao sistema de vigilância epidemiológica da cidade. O modelo não gera custos diretos com licenciamento, pois utiliza *softwares* de código aberto. Apenas recursos computacionais já existentes na administração pública são necessários.

A intervenção proposta emprega este modelo como ferramenta de suporte ao Programa Municipal de Controle da Dengue (PMCD). Tal aplicação visa auxiliar a gestão pública municipal no planejamento de ações preventivas, com maior antecipação e exatidão. O resultado desta pesquisa integra-se, portanto, diretamente no ciclo de políticas públicas. A pesquisa pode ser alocada nas fases de formulação, implementação e avaliação. Segundo mencionado por Capella (2018) e Saraiva e Ferrarezi (2006), o ciclo de políticas públicas compreende etapas interativas. Estas incluem definição de problemas, formação de agenda, formulação, tomada de decisão, implementação e avaliação. O modelo desenvolvido fortalece as fases de formulação e implementação de forma prática.

Na fase de formulação, o modelo preditivo fornece subsídios técnicos para a determinação de prioridades e estratégias. Isso fundamenta a seleção de medidas de prevenção e combate à dengue, baseada em critérios técnicos. Durante a implementação, o modelo opera como ferramenta de monitoramento e direcionamento das ações em campo. Ele indica períodos críticos com elevado risco de surtos, possibilitando o planejamento antecipado de mutirões de limpeza, campanhas educativas de orientação e intensificação do trabalho dos Agentes Comunitários de Endemias (ACE). Na avaliação, o modelo pode ser empregado

como recurso complementar para aferir a eficácia das ações adotadas. Isso ocorre ao comparar as previsões com os resultados reais, permitindo ajustes e aperfeiçoamentos contínuos.

A importância deste produto está na relevância estratégica de políticas públicas de saúde. A dengue demanda respostas rápidas e objetivas do sistema de saúde, alinhadas aos princípios da integralidade e universalidade do Sistema Único de Saúde (SUS). Nesse cenário, ferramentas de inteligência artificial e ciência de dados podem potencializar a capacidade estatal de atuação preventiva, resultando em redução de custos, otimização de recursos e, primordialmente, preservação de vidas.

Este modelo representa um instrumento inovador para Maringá. Ele se integra ao PMCD e fortalece a gestão pública da saúde. O modelo apoia a tomada de decisão dos gestores locais. Adicionalmente, demonstra replicabilidade em outros municípios brasileiros. Isso se deve ao uso de bases públicas e de fácil acesso, alinhando-se sempre aos princípios de transparência e eficiência. Tais princípios são previstos nas diretrizes nacionais de vigilância epidemiológica. A adoção deste produto visa transformar dados em conhecimento estratégico. Espera-se um impacto positivo na saúde populacional. Adicionalmente, busca-se o fortalecimento das políticas públicas de enfrentamento às arboviroses.

A proposta de intervenção apresentada neste estudo está em consonância com a visão da ciência de dados aplicada a políticas públicas, conforme discutido por De Toni e Dorneles (2022). Segundo os autores, o uso de modelos analíticos no setor público ajuda a converter grandes volumes de dados administrativos em conhecimento útil, o que fortalece a habilidade do governo em planejar, monitorar e avaliar políticas.

Nesse contexto, o modelo preditivo desenvolvido neste estudo vai além de uma aplicação técnica de aprendizado de máquina, funcionando como uma ferramenta de suporte para a administração pública municipal. Ao permitir a previsão de cenários epidemiológicos da dengue, o modelo auxilia na melhoria das decisões dentro do Programa Municipal de Controle da Dengue, fortalecendo práticas de gestão baseadas em evidências e em conformidade com os princípios SUS.

6.1 Relatório Técnico Conclusivo

Uma sugestão ao município de Maringá é padronizar o preenchimento automático da ficha de investigação de Dengue e Febre de Chikungunya, qual alimenta o banco de dados do SINAN, conforme estabelecida e atualizada pela Nota Informativa nº59/2016

CGPNCD/DEVIT/SVS/MS, constituindo uma recomendação prática para o município de Maringá. Adicionalmente, a aplicação de recursos de georreferenciamento para os endereços de residência pode diminuir erros de digitação e omissões. Esta medida associa o endereço do paciente a coordenadas geográficas corretas. A implementação destas ações qualifica as informações registradas no SINAN. Isso as torna mais consistentes para análises espaciais. Tal ação potencializa a correlação de dados e, conseqüentemente, aumentará o potencial dos futuros modelos preditivos. Concomitantemente, subsidiará políticas públicas mais direcionadas para a vigilância e o controle da dengue no nível intramunicipal. Isso permitirá a identificação de áreas críticas e a alocação estratégica de recursos.

REFERÊNCIAS

AGÊNCIA CÂMARA. Disponível em

<https://www.camara.leg.br/noticias/969316-ministerio-da-saude-estima-que-vacina-contradengue-esteja-disponivel-no-sus-em-um-ano-e-meio> Acesso em: 19 maio. 2025, 19h51.

BARRETO, Maurício L.; **TEIXEIRA**, Maria Glória. Dengue no Brasil: situação epidemiológica e contribuições para uma agenda de pesquisa. **Estudos Avançados**, São Paulo, v. 22, n. 64, p. 53-72, dez. 2008. Disponível em:

<https://doi.org/10.1590/S0103-40142008000300005>. Acesso em: 16 junho de 2025.

BERGMEIR, Christoph ; BENÍTEZ, José Manuel. On the use of cross-validation for time series predictor evaluation., v. 191, p. 192–213, 2012.

BRAGA, Ima Aparecida; VALLE, Denise. Aedes aegypti: histórico do controle no Brasil. Brasília, DF: Ministério da Saúde, Secretaria de Vigilância em Saúde, Programa Nacional de Controle da Dengue; Fundação Instituto Oswaldo Cruz, 2007.

BRASIL. Ministério da Saúde. Fundação Nacional de Saúde. Dengue: aspectos epidemiológicos, diagnóstico e tratamento. Brasília: Fundação Nacional de Saúde, 2002.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde e Ambiente. Dengue: diagnóstico e manejo clínico – adulto e criança. 6. ed. Brasília: Ministério da Saúde, 2024.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde e Ambiente. Departamento de Doenças Transmissíveis. **Diretrizes nacionais para prevenção e controle das arboviroses urbanas: vigilância entomológica e controle vetorial** [recurso eletrônico]. Brasília, DF: Ministério da Saúde, 2025.

BRASIL. <https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/estudos-e-notas-informativas/2025/nota-informativa-no-3-2025-cgarb-dedt-svsa-ms.pdf/view> acesso em 09 de junho de 2025 às 11h 23.

BREIMAN, L. *Random Forests*. **Machine Learning**, v. 45, 2001.

BONITA, R.; BEAGLEHOLE, R.; KJELLSTRÖM, T. Epidemiologia básica. 2. ed. Tradução e revisão científica de Juraci A. Cesar. São Paulo: Santos, 2010.

CAPELLA, Ana Cláudia Niedhardt. **Formulação de políticas públicas.** Brasília: Enap, 2018.

CHAGAS, Marcelle; CODEÇO, Cláudia; ALMEIDA, Iasmim; OLIVEIRA, Sara. Enfrentando a dengue nas favelas e periferias. 1. ed. Rio de Janeiro: InfoDengue/Fiocruz, 2023.

CHOLLET, François. Deep Learning with Python. Shelter Island: Manning Publications Co., 2021.

DE TONI, Jackson; DORNELES, Rachel . **Ciência de dados em políticas públicas: uma experiência de formação.** Brasília: Escola Nacional de Administração Pública (Enap), 2022. Disponível em: <http://repositorio.enap.gov.br/handle/1/7472>. Acesso em: 08 de dezembro de 2025, 17h40.

FARIA, Luerbio; OLIVEIRA, Fabiano de Souza; PINTO, Paulo Eustáquio Duarte; SZWARCFITER, Jayme Luiz. Ciência de dados: algoritmos e aplicações. 1. impressão. Rio de Janeiro: IMPA, 2021.

FERREIRA, Daniel Cassiano. **Limpeza e tratamento de dados.** [S. l.: s. n.], 2020.

GASPAR, Juliano de Souza et al. Introdução à análise de dados em saúde com Python. Belo Horizonte: Universidade Federal de Minas Gerais, Centro de Inovação em Inteligência Artificial para a Saúde, 2023.

GODOY, Arilda Schmidt. Introdução à pesquisa qualitativa e suas possibilidades: uma revisão histórica dos principais autores e obras que refletem esta metodologia de pesquisa em Ciências Sociais. Revista de Administração de Empresas, São Paulo, 1995.

HAN, JiaWei; KAMBER, Micheline. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Inc., San Francisco, 2001.

HYNDMAN, Rob J.; ATHANASOPOULOS, George. *Forecasting: principles and practice.* 2. ed. Melbourne: OTexts, 2018. Disponível em: <https://otexts.com/fpp2/>. Acesso em 09 set. 2025 às 08h e 47min

IBM. Data Science. Disponível em: <https://www.ibm.com/br-pt/topics/data-science>. Acesso em: 07 jan. 2025, 13h20.

IZBICKI, Rafael; DOS SANTOS, Tiago Mendonça. Aprendizado de máquina: uma abordagem estatística [livro eletrônico]. 1. ed. São Carlos, SP, 2020.

LARA, Jorge Tibilletti de. A emergência da dengue como desafio virológico: de doença-fantasma à endemia “de estimação”, 1986-1987. História, Ciências, Saúde – Manguinhos, Rio de Janeiro, v.29, n.2, 2022.

KUHN, Max & JOHNSON, Kjell *Applied Predictive Modeling*. Springer. - Saline, Michigan - USA, 2013.

MARINGÁ, dados obtidos da Gerência Financeira da Secretaria de Saúde de Maringá-
[Elotech](#) em 04 maio de 2025, 13h30.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. Fundamentos da metodologia científica. 5. ed. São Paulo: Atlas, 2005.

MORETTIN, Pedro A.; SINGER, Julio M. Introdução à ciência de dados: fundamentos e aplicações. Versão parcial preliminar. São Paulo, 2019.

MOUTINHO, Sônia Oliveira Matos; MARTINS, Paulo George Miranda; ALENCAR, Danila Fernandes; CONEGLIAN, Caio Saraiva. Ciência da Informação e Ciência de Dados: convergências interdisciplinares. Encontros Bibli, Florianópolis, v. 29, 2024, e99127. DOI: <https://doi.org/10.5007/1518-2924.2024.e99127> acesso em 10/04/2025 às 08:04.

MUELLER, John Paul; MASSARON, Luca. Aprendizado de máquina para leigos. Rio de Janeiro: Alta Books, 2019.

O GLOBO disponível em:
<https://oglobo.globo.com/economia/noticia/2024/03/06/economia-do-pais-deve-perder-mais-de-20-bi-com-epidemia-de-dengue-diz-estudo-de-industrias-de-minas.ghtml> acesso em
Acesso em: 20 de maio. 2025, 10h47min.

PEREIRA, Carlos; VEIGA, Nélío. A epidemiologia: de Hipócrates ao século XXI. Millenium, Viseu, 2014.

REIS, Marcelo Menezes. Estatística aplicada à administração. Florianópolis: Universidade Federal de Santa Catarina / Sistema UAB, 2008.

SARAVIA, Enrique; FERRAREZI, Elisabete. Políticas públicas: coletânea – volume 1. Brasília: ENAP, 2006.

SHAMAN, Jeffrey; KARSPECK, Alicia. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, v. 109, n. 50, p. 20425–20430, 2012.

TASHMAN, Leonard J. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, v. 16, n. 4, p. 437-450, 2000.

THOMAZ, Erika Bárbara Abreu Fonseca; PINHO, Judith Rafaelle Oliveira; ABREU, Thalita Queiroz; RODRIGUES, Vandilson Pinheiro. Conceitos e ferramentas da epidemiologia. São Luís: Edufma, 2015.

VIANA, Dione Viero; IGNOTTI, Eliane. A ocorrência da dengue e variações meteorológicas no Brasil: revisão sistemática. **Revista Brasileira de Epidemiologia**, São Paulo, v. 16, n. 2, p. 240-256, jun. 2013. Disponível em: <https://doi.org/10.1590/S1415-790X2013000200020>. Acesso em: 12 junho de 2025.

ZHAO, N.; CHARLAND, K.; CARABALI, M.; NSOESIE, E. O.; MAHEU-GIROUX, M.; REES, E.; YUAN, M.; GARCIA BALAGUERA, C.; JARAMILLO RAMIREZ, G.; ZINSZER, K. *Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia.* *PLOS Neglected Tropical Diseases*, v. 14, n. 9, e0008056, 2020. Disponível em: <https://doi.org/10.1371/journal.pntd.0008056>.

APÊNDICE

Anexo 1: Análise Exploratória dos Dados do SINAN (Epidêmico)

<https://colab.research.google.com/drive/1Y6P5CNFnWqHeYbeQa9AoPzKrOMFdK-UY?usp=sharing>

Anexo 2: Análise Exploratória dos Dados do INMET (Climatológico)

<https://colab.research.google.com/drive/1Jc9e-b8DRubKjywtHEv4QCzRolhl5P1u?usp=sharing>

Anexo 3: Modelo Preditivo

<https://colab.research.google.com/drive/1olAf4HcUqsTRKDyUoh3BJ8yGtu6VdiE3?usp=sharing>

Anexo 4: Validação do Modelo Preditivo com LOSO

<https://colab.research.google.com/drive/14oM8dfPWeEVTToSNCeiiGMc8kE1UEymO?usp=sharing>

Anexo 5: Guia de Implementação: Sistema de Predição de Dengue (passo à passo)

<https://docs.google.com/document/d/1AbLE353kuICYAWgCsYhZiBMe8t2g1IP8eohl56leADo/edit?usp=sharing>