UNIVERSIDADE PRESBITERIANA MACKENZIE

CIÊNCIA DE DADOS

ANTÔNIO CARLOS DOMANSKI DA SILVA – 10424144

BRUNA ALMEIDA DA SILVA - 10263278

LAURA ELOISE FERREIRA - 10424754

PREVISÃO DE CASOS DE DENGUE NO BRASIL POR MEIO DE SÉRIES TEMPORAIS

São Paulo - SP

2025

ANTÔNIO CARLOS DOMANSKI DA SILVA – 10424144

BRUNA ALMEIDA DA SILVA - 10263278

LAURA ELOISE FERREIRA - 10424754

PREVISÃO DE CASOS DE DENGUE NO BRASIL POR MEIO DE SÉRIES TEMPORAIS

Trabalho apresentado à disciplina de Projeto Aplicado IV como requisito parcial para a conclusão do Curso de Graduação Tecnológica em Ciência de Dados.

PROFESSOR: GUSTAVO SCALABRINI SAMPAIO

São Paulo – SP

2025

Sumário

Índice de Figuras

1. Introdução

A dengue é uma das doenças mais recorrentes no Brasil, representando um grave problema de saúde pública. Transmitida pelo mosquito *Aedes aegypti*, a doença apresenta surtos sazonais que afetam milhões de brasileiros todos os anos, sobrecarregando o sistema de saúde e impactando diretamente a qualidade de vida da população.

O crescimento urbano desordenado, as condições climáticas favoráveis e a dificuldade de controle dos criadouros contribuem para a disseminação da doença. Nesse contexto, a análise de séries temporais dos casos notificados torna-se uma ferramenta essencial para identificar padrões, tendências e prever surtos, possibilitando a adoção de medidas preventivas e de mitigação mais eficazes.

Diante desse cenário, este projeto propõe o desenvolvimento de um modelo analítico preditivo capaz de estimar o número de casos de dengue a partir de dados históricos, fornecendo informações úteis para órgãos públicos de saúde e gestores municipais.

1.1. Motivação

A escolha deste tema está diretamente relacionada à relevância social da dengue no Brasil. Anualmente, a doença gera altos custos para o sistema de saúde e coloca em risco a vida de milhares de pessoas. A capacidade de antecipar surtos pode auxiliar governos e instituições a direcionar recursos de forma mais eficiente, reduzindo impactos sociais e econômicos.

Além disso, do ponto de vista acadêmico e técnico, a previsão de séries temporais é uma aplicação prática e bastante consolidada na Ciência de Dados. O projeto, portanto, alia relevância social com a possibilidade de aplicar metodologias de aprendizado de máquina a um problema concreto.

1.2. Justificativa

A justificativa para a realização deste projeto se apoia tanto na relevância social quanto na relevância científica e técnica do tema. A dengue é uma das arboviroses mais comuns e preocupantes no Brasil, sendo reconhecida pela Organização Mundial da Saúde como uma prioridade global de saúde pública. A previsão de casos pode subsidiar ações preventivas, como campanhas de conscientização e reforço do combate ao mosquito transmissor, reduzindo os impactos da doença na população.

Do ponto de vista científico e técnico, a aplicação de métodos de previsão de séries temporais em dados epidemiológicos possibilita explorar modelos estatísticos e de aprendizado de máquina que contribuem para o avanço do uso da ciência de dados em problemas reais de saúde pública. Essa abordagem oferece soluções inovadoras e de impacto direto, ao mesmo tempo em que fortalece a integração entre análise de dados e tomada de decisão em contextos sociais relevantes.

1.3. Objetivo

Desenvolver um produto analítico capaz de prever o número de casos de dengue no Brasil, utilizando técnicas de séries temporais, de forma a apoiar gestores públicos e instituições de saúde no planejamento de ações preventivas e no enfrentamento de surtos da doença.

1.3.1. Objetivos Específicos

- Coletar e organizar dados históricos de casos de dengue a partir da base pública **InfoDengue**.
- Analisar e identificar padrões de tendência e sazonalidade nas séries temporais.
- Aplicar modelos de previsão (como ARIMA, Prophet ou redes neurais) e avaliar seu desempenho.
- Desenvolver visualizações que facilitem a interpretação dos resultados por gestores públicos.

- Produzir relatórios analíticos que possam subsidiar a tomada de decisão em saúde pública.

1.4. Descrição da Base de Dados

A base de dados utilizada neste projeto é proveniente do InfoDengue, uma iniciativa da Fundação Oswaldo Cruz (Fiocruz) em parceria com o Ministério da Saúde, que tem como objetivo monitorar de forma sistemática os casos de dengue no Brasil. Os dados são disponibilizados em arquivos no formato CSV, o que permite fácil manipulação e integração com ferramentas de análise de dados e softwares estatísticos.

Embora a plataforma disponibilize informações detalhadas por município e estado, não há um conjunto consolidado diretamente em nível nacional. Dessa forma, optou-se por selecionar as capitais de estados representativos, com elevada incidência histórica de casos da doença, de modo a refletir de maneira consistente a realidade brasileira. Foram escolhidas as capitais de São Paulo, Curitiba, Manaus, Goiânia e Salvador, que concentram grande parte dos registros de dengue no país e foram selecionadas de forma a representar as cinco regiões brasileiras, apresentando diversidade populacional e climática relevante para o estudo.

Cada registro corresponde a uma combinação de capital e semana epidemiológica, apresentando a data de notificação e o número de casos confirmados ou suspeitos de dengue. Além disso, a base inclui indicadores epidemiológicos, como a taxa de incidência, e variáveis ambientais, como temperatura média, umidade relativa e volume de precipitação, que permitem relacionar condições climáticas à propagação da doença.

A estrutura da base é organizada de forma tabular, com cada linha representando uma observação semanal e cada coluna representando uma variável específica, o que facilita a análise comparativa ao longo do tempo. Essa organização possibilita a identificação de padrões temporais, sendo adequada para análises exploratórias, aplicação de modelos de previsão e elaboração de visualizações que auxiliem a tomada de decisão em saúde pública.

Os dados do InfoDengue são coletados de notificações oficiais do Sistema de Informação de Agravos de Notificação (SINAN) e passam por processos de validação e padronização antes da disponibilização, garantindo consistência e confiabilidade. A fonte é pública e acessível para fins acadêmicos e de pesquisa, o

que torna a base ideal para estudos voltados ao entendimento da evolução da dengue no país, mesmo quando analisada a partir de estados selecionados, além de ser útil para o desenvolvimento de ferramentas analíticas que apoiem ações preventivas e estratégias de saúde pública.

2. Referencial Teórico

A análise de séries temporais é uma das técnicas mais utilizadas em estatística e ciência de dados, com aplicações em diversas áreas como economia, meteorologia e epidemiologia. Uma série temporal é definida como um conjunto de observações ordenadas no tempo, cuja análise permite identificar padrões de tendência, sazonalidade e variações aleatórias. A partir desses elementos, torna-se possível a construção de modelos matemáticos e computacionais para previsão de valores futuros (BOX; JENKINS, 2016).

No contexto da saúde pública, a utilização de séries temporais para previsão de doenças como a dengue é estratégica, pois possibilita antecipar surtos e subsidiar a tomada de decisão por parte dos gestores. O uso de modelos de previsão em epidemiologia permite compreender a dinâmica de transmissão das doenças e orientar políticas públicas mais eficazes (MORENO et al., 2018).

A literatura pertinente apresenta uma variedade de abordagens para a predição de arboviroses. Diversos trabalhos correlacionados já exploraram essa temática. Pesquisas realizadas pela Fundação Oswaldo Cruz (Fiocruz), por meio da plataforma InfoDengue, demonstram que a incorporação de variáveis climáticas como temperatura, umidade e precipitação aumenta a acurácia das previsões ao relacionar condições ambientais com a proliferação do mosquito *Aedes aegypti* (FUNDAÇÃO OSWALDO CRUZ, 2025). Estudos prévios frequentemente aplicam modelos da família ARIMA (Autoregressive Integrated Moving Average) ou suas variações sazonais (SARIMA), que se mostram robustos para capturar tendências e sazonalidades lineares. Outros pesquisadores têm obtido sucesso com algoritmos de *machine learning*, como *Support Vector Machines* (SVM) e *Random*

Forests, tratando a previsão como um problema de regressão supervisionada. Recentemente, modelos baseados em redes neurais, como as *Long Short-Term Memory* (LSTM), têm sido explorados para capturar dinâmicas não lineares complexas na propagação de epidemias, embora exijam maior volume de dados e custo computacional.

As alternativas de solução encontradas na literatura e aplicadas na prática podem ser agrupadas em diferentes abordagens. A primeira envolve os métodos estatísticos clássicos, como ARIMA (Autoregressive Integrated Moving Average), que modelam a série baseando-se em suas próprias observações passadas. Uma segunda abordagem, mais moderna, utiliza modelos aditivos, como o Prophet. Este modelo decompõe a série temporal em componentes de tendência, sazonalidade e feriados, sendo projetado para lidar de forma robusta com dados do mundo real, que frequentemente possuem sazonalidades múltiplas e dados faltantes, oferecendo previsões de alta qualidade com esforço reduzido (TAYLOR; LETHAM, 2018). A terceira categoria compreende a aplicação de algoritmos de aprendizado de máquina supervisionado. Nessa metodologia, a série temporal é reestruturada, utilizando-se janelas de tempo e valores passados (lags) como variáveis preditoras (features). Algoritmos como o XGBoost, um método de Gradient Boosting, são particularmente eficazes nesta tarefa, pois conseguem capturar relações complexas e não lineares entre as features (que podem incluir os lags, dados climáticos e componentes de calendário) e a variável alvo (o número de casos futuros) .

Dessa forma, observa-se que não existe um modelo universalmente superior: a escolha da metodologia depende do equilíbrio entre desempenho, viabilidade prática e interpretabilidade dos resultados. Para problemas de saúde pública, em que a clareza na comunicação é fundamental, muitas vezes é necessário optar por modelos que conciliem precisão, como a oferecida pelo XGBoost, com a facilidade de compreensão por parte dos gestores, uma vantagem de modelos como o Prophet.

3. Pipeline da Solução

A metodologia da pipeline computacional iniciou-se com a ingestão e consolidação dos dados, onde múltiplos arquivos CSV, correspondentes a diferentes localidades, foram unificados. Durante este processo, os dados foram padronizados, incluindo a normalização de nomes de colunas (como date e cases) e a conversão para os

tipos de dados apropriados (datetime e numérico).

Posteriormente, os registros foram agregados para formar uma única série temporal de frequência semanal em âmbito "mini-nacional". Esta série foi então submetida à engenharia de features, etapa na qual foram criadas variáveis explicativas baseadas no tempo (como mês e semana do ano) e componentes autorregressivos (lags e médias móveis).

Com o conjunto de dados estruturado, realizou-se a divisão cronológica em subconjuntos de treino e teste, reservando o período mais recente para a validação. A etapa de modelagem consistiu em uma avaliação comparativa entre os algoritmos Prophet e XGBoost, treinados no conjunto de treino e avaliados no de teste por meio das métricas MAE e RMSE para a seleção do modelo de melhor performance.

Identificado o modelo vencedor (XGBoost), este foi submetido a um processo de otimização de hiperparâmetros, utilizando a técnica de GridSearchCV com validação cruzada específica para séries temporais (TimeSeriesSplit). Por fim, o modelo otimizado foi retreinado com a totalidade dos dados históricos disponíveis e, então, utilizado para gerar a previsão final para as 12 semanas subsequentes.

3.1. Diagrama da Solução

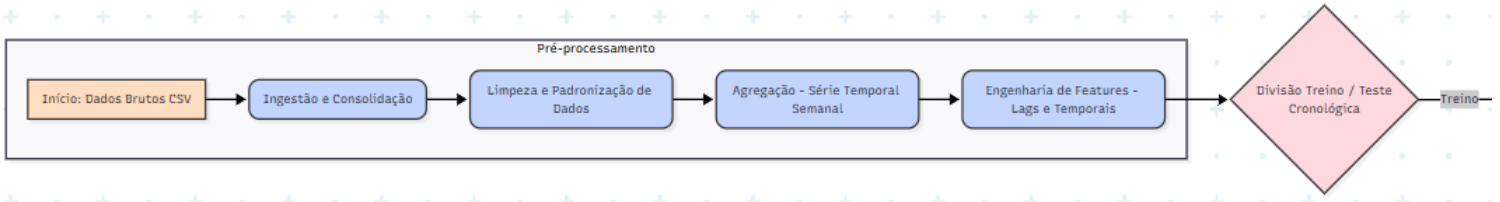

Figura 1. Diagrama do Pipeline de Solução - Parte 1

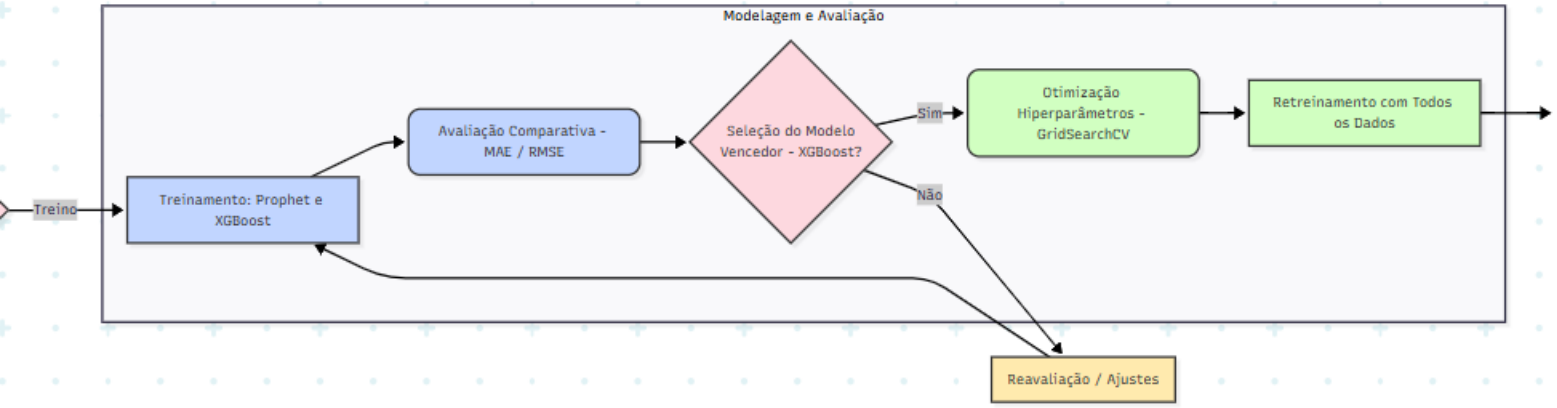

Figura 2. Diagrama do Pipeline de Solução - Parte 2

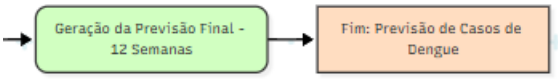

Figura 3. Digrama do Pipeline de Solução - Parte 3

4. EDA e Pré-processamento dos dados

A Etapa 1, "Upload e Carregamento dos CSVs", consistiu na ingestão inicial de dados e preparação do ambiente de execução no Google Colab, consolidando múltiplas fontes heterogêneas em um único dataset para análise subsequente. Foram utilizadas as bibliotecas os, shutil, glob, google.colab.files e pandas, e criado um diretório específico para armazenar os arquivos brutos.

Os arquivos CSV foram carregados via widget interativo do Colab, movidos para o diretório preparado e verificados. A função load_infodengue_files_safe foi implementada para leitura robusta, tratamento de diferentes codificações e separadores, e criação de colunas de metadados (capital e source_file) para rastreabilidade. Arquivos vazios eram descartados, garantindo a integridade dos dados.

Por fim, os DataFrames individuais foram consolidados em um único DataFrame mestre (df_raw), cujas dimensões (4100, 32) e visualização das primeiras linhas confirmaram a consistência e estrutura dos dados obtidos.

	data_iniSE	SE	casos_est	casos_est_min	casos_est_max	casos	p_rt1	p_inc100k	Localidade_id	nivel	...	tempmed	tempmax	casprov	casprov_est	casprov_est_mi
0	2025-09-14	202538	57.0	35	119.0	24	0.973635	3.045215	0	1	...	15.428950	20.897950	16.0	NaN	Na
1	2025-09-07	202537	43.0	36	60.0	32	0.935592	2.297267	0	1	...	14.318243	19.448586	24.0	NaN	Na
2	2025-08-31	202536	44.0	40	51.0	39	0.972978	2.350692	0	1	...	16.144029	20.368614	22.0	NaN	Na
3	2025-08-24	202535	24.0	22	29.0	22	0.147703	1.282196	0	1	...	13.401243	17.596443	11.0	NaN	Na
4	2025-08-17	202534	23.0	21	26.0	21	0.059572	1.228771	0	1	...	16.821800	22.361086	7.0	NaN	Na

5 rows × 32 columns

Figura 4 Etapa 1: Upload e Carregamento dos CSVs - Extraído do Notebook

A Etapa 2, "Ajuste de Colunas", consistiu na limpeza e normalização do schema do DataFrame consolidado (df_raw) obtido na etapa anterior, preparando os dados para análise exploratória e modelagem. Inicialmente, foi criada uma cópia de trabalho do dataset (df = df_raw.copy()) para preservar os dados originais e permitir re-execuções seguras do notebook.

Em seguida, realizou-se a renomeação semântica das colunas por meio de um dicionário de mapeamento, padronizando nomes em inglês e garantindo maior consistência. Entre as transformações principais, data_iniSE foi renomeada para date, casos para cases, tempmed para temp_mean e umidmed para humidity.

A conversão e tipagem de dados foi aplicada para assegurar a integridade do dataset. A coluna date foi convertida para o tipo datetime (datetime64[ns]) utilizando pd.to_datetime com tratamento de erros, enquanto cases passou por coerção numérica (pd.to_numeric), preenchimento de valores ausentes (fillna(0)) e conversão para inteiro (astype(int)).

Por fim, a validação estrutural do DataFrame foi realizada com df.info(), confirmando que as colunas foram renomeadas corretamente e que date e cases apresentavam o tipo de dados apropriado, totalizando 4100 registros não nulos.

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   data_iniSE        4100 non-null   object
 1   SE                4100 non-null   int64
 2   casos_est         4100 non-null   float64
 3   casos_est_min     4100 non-null   int64
 4   casos_est_max     4060 non-null   float64
 5   cases             4100 non-null   int64
 6   p_rt1             4100 non-null   float64
 7   p_inc100k         4100 non-null   float64
 8   Localidade_id     4100 non-null   int64
 9   nivel             4100 non-null   int64
 10  id                4100 non-null   int64
 11  versao_modelo     4100 non-null   object
 12  tweet             3515 non-null   float64
 13  Rt                4100 non-null   float64
 14  pop               4100 non-null   float64
 15  tempmin           4100 non-null   float64
 16  umidmax           4021 non-null   float64
 17  receptivo         4100 non-null   int64
 18  transmissao       4080 non-null   float64
 19  nivel_inc         4100 non-null   int64
 20  umidmed           4001 non-null   float64
 21  umidmin           4014 non-null   float64
 22  tempmed           4001 non-null   float64
 23  tempmax           4001 non-null   float64
 24  casprov           4095 non-null   float64
 25  casprov_est       0 non-null      float64
 26  casprov_est_min   0 non-null      float64
 27  casprov_est_max   0 non-null      float64
 28  casconf           0 non-null      float64
 29  notif_accum_year  4100 non-null   int64
 30  capital           4100 non-null   object
dtypes: float64(19), int64(9), object(3)
memory usage: 993.1+ KB
```

Figura 5. Etapa 2: Ajustes de Colunas - Extraído do Notebook.

O recorte amostral foi definido para cinco capitais brasileiras de interesse: curitiba, goiania, manaus, salvador e saopaulo. O DataFrame principal foi filtrado para reter apenas essas localidades, criando-se uma cópia do subset para preservar a integridade das operações subsequentes. Em seguida, foi realizada a agregação por capital e semana, somando os casos de dengue e calculando a média das variáveis climáticas (temp_mean e humidity). Esta consolidação permitiu observar o comportamento semanal dos casos de dengue em cada capital e servir como base para análises comparativas. (Etapa 3 – Delimitação Amostral e Agregação por Capital).

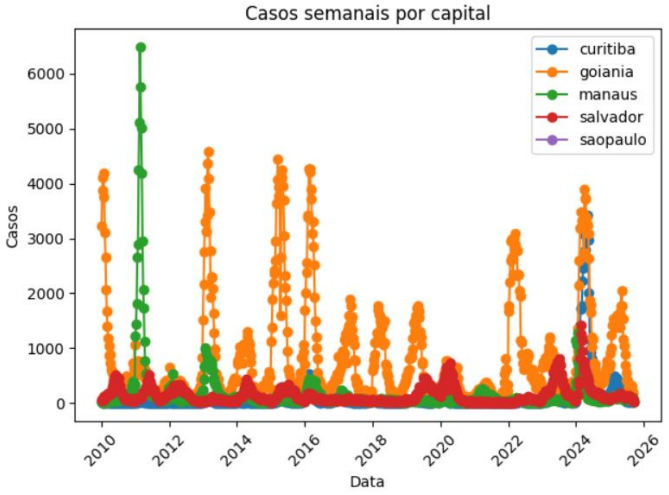

Figura 6. Etapa 3: Delimitação Amostral e Agregação por Capital - Extraído do Notebook

Com base nos dados agregados por capital, foi construída uma série temporal nacional agregada, somando os casos de todas as capitais e calculando a média das variáveis climáticas semanais. O DataFrame resultante foi ordenado cronologicamente e a coluna date foi definida como índice, garantindo compatibilidade com bibliotecas de análise de séries temporais. A visualização da série agregada permitiu identificar tendências gerais e padrões sazonais no nível nacional, servindo como insumo para a análise exploratória de dados. (Etapa 4 – Série Temporal Nacional Agregada).

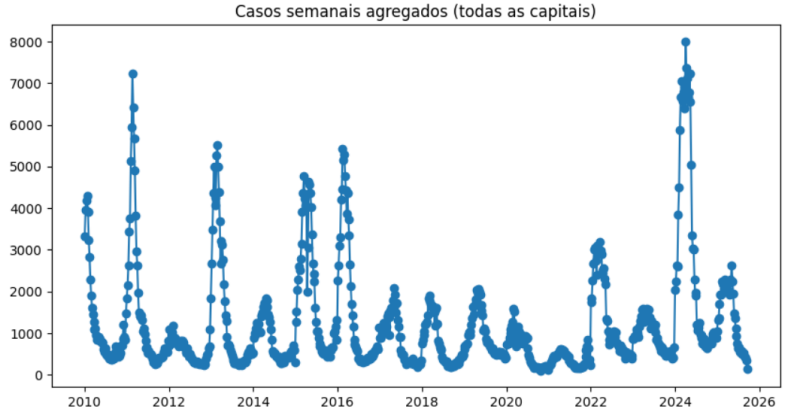

Figura 7. Etapa 4:Série Temporal Nacional Agregada - Extraída do Notebook

As Figuras 6 e 7 representam a etapa de consolidação da série temporal, elemento fundamental para compreender a dinâmica epidemiológica da dengue antes da aplicação de qualquer técnica de modelagem. A Figura 6 apresenta a evolução semanal dos casos para cada uma das cinco capitais selecionadas — Curitiba, Goiânia, Manaus, Salvador e São Paulo — revelando diferenças estruturais significativas entre as localidades. Observa-se que todas as capitais exibem um padrão de forte sazonalidade anual, caracterizado por elevação dos casos sobretudo nos primeiros meses do ano, período historicamente associado ao maior volume de chuvas e às condições climáticas favoráveis ao vetor Aedes aegypti.

Apesar do padrão sazonal comum, a magnitude e o comportamento dos surtos variam de acordo com as particularidades de cada região. Manaus e Salvador, por exemplo, apresentam oscilações mais abruptas e picos mais intensos, possivelmente relacionados ao clima tropical, com elevada temperatura e umidade ao longo de grande parte do ano. Curitiba, por outro lado, apresenta picos menos acentuados, coerentes com um clima mais ameno, ainda que mantenha periodicidade semelhante às demais capitais. Goiânia e São Paulo demonstram comportamento intermediário, mas com surtos expressivos em períodos específicos, o que reforça a influência de fatores locais, como densidade populacional, infraestrutura urbana, políticas de vigilância epidemiológica e ciclos de infestação do mosquito.

A Figura 7 sintetiza essas informações ao exibir a série temporal nacional agregada, construída pela soma dos casos semanais das capitais e pela média dos indicadores climáticos. O comportamento da série consolidada evidencia, de forma ainda mais clara, o ciclo anual da dengue no país, com picos recorrentes ao longo dos anos analisados. Entre os episódios mais marcantes, destacam-se os períodos de 2010/2011, 2013, 2015/2016 e o grande surto de 2024, cujas magnitudes permanecem elevadas mesmo após a agregação. Esses episódios refletem momentos críticos já documentados em relatórios epidemiológicos nacionais e confirmam que a série agregada captura fielmente tendências e variações estruturais do fenômeno.

A análise conjunta das séries por capital e da série agregada revela também a presença de forte autocorrelação temporal, característica comum em processos epidemiológicos. A repetição sistemática do padrão “ascensão rápida – pico – declínio gradual” reforça a importância da sazonalidade como componente dominante na dinâmica da doença, ao mesmo tempo em que revela a existência de variações estruturais de longo prazo, sugerindo mudanças no comportamento epidemiológico ao longo dos anos. Além disso, o pico de 2024 apresenta magnitude atípica quando comparado aos demais surtos, indicando um possível desvio estrutural que poderia desafiar modelos mais simples, especialmente aqueles baseados em componentes aditivos ou relações estritamente lineares.

Essas observações tornam evidente a relevância da série agregada como insumo para a modelagem, pois ela concentra padrões robustos e reduz o ruído associado às particularidades individuais de cada capital. A interpretação dos gráficos também estabelece fundamentos importantes para a Análise Exploratória subsequente (Etapa 5), na qual os componentes estruturais — tendência, sazonalidade e resíduo — serão extraídos e examinados de forma detalhada, contribuindo para a formulação e escolha dos modelos de previsão utilizados neste estudo.

A análise exploratória de dados (EDA) foi conduzida sobre a série temporal agregada nacional de casos de dengue, com o objetivo de compreender padrões, identificar sazonalidade, tendências e possíveis relações com variáveis exógenas. Inicialmente, foram extraídas estatísticas descritivas da variável cases utilizando a função .describe(), permitindo analisar medidas centrais, dispersão e amplitude dos dados, incluindo média, mediana, desvio padrão, quartis e valores extremos. Esse levantamento forneceu uma compreensão preliminar da distribuição dos casos semanais e evidenciou a presença de variabilidade sazonal significativa ao longo do período analisado.

Para aprofundar a análise temporal, aplicou-se a decomposição clássica de séries temporais por meio do método seasonal_decompose da biblioteca statsmodels. Optou-se por um

modelo aditivo, considerando que os componentes de tendência e sazonalidade se somam de maneira aproximadamente linear ao valor observado, e definiu-se um período de 52 semanas, correspondente ao ciclo anual de transmissão da dengue. A decomposição isolou três componentes distintos: a tendência de longo prazo, que revelou variações estruturais no comportamento da doença ao longo dos anos; a sazonalidade, evidenciando picos recorrentes em determinados meses; e os resíduos, que capturam variações não explicadas pelos padrões estruturais, permitindo identificar eventos atípicos ou ruído na série. Essa análise visual e quantitativa foi essencial para fundamentar decisões sobre a modelagem subsequente, sobretudo na escolha de métodos capazes de capturar tanto a tendência quanto os ciclos sazonais.

Além da decomposição, foi realizada uma investigação multivariada das relações entre a variável alvo cases e as variáveis climáticas exógenas temp_mean e humidity. Para isso, utilizou-se a função pairplot da biblioteca seaborn, que combina histogramas na diagonal para análise das distribuições individuais e gráficos de dispersão fora da diagonal para identificar possíveis correlações lineares ou não lineares. Essa análise visual permitiu perceber que mudanças na temperatura média e na umidade apresentam associações relevantes com a ocorrência de casos de dengue, reforçando a necessidade de incluir essas variáveis como preditores em modelos supervisionados.

Por fim, a EDA contemplou também a inspeção de outliers e valores ausentes, verificando consistência nos dados temporais e a adequação da série para operações de modelagem preditiva. Esta etapa foi crucial não apenas para validar a qualidade do dataset, mas também para fornecer insights sobre os padrões epidemiológicos e ambientais que influenciam a propagação da dengue, orientando a seleção de features e estratégias de pré-processamento para as etapas seguintes de modelagem. (Etapa 5 – Análise Exploratória de Dados).

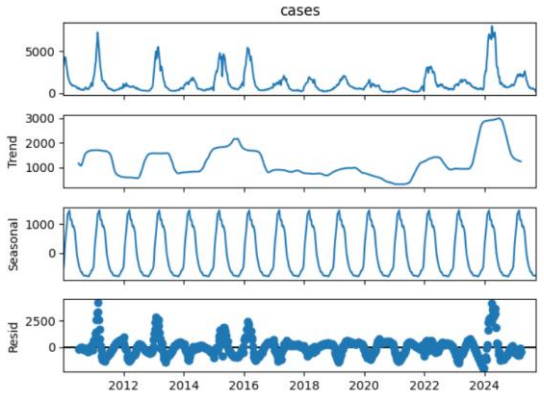

Figura 8. Etapa 5: Análise Exploratória de Dados - Extraída do Notebook

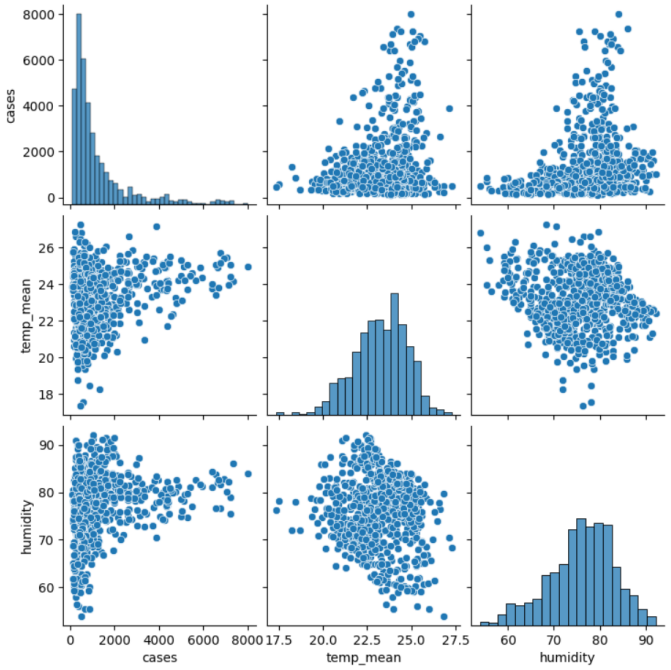

Figura 9. Etapa 5: Análise Exploratória de Dados - Extraída do Notebook

A partir das visualizações apresentadas nas Figuras 8 e 9, é possível aprofundar a análise exploratória iniciada anteriormente, interpretando de maneira mais detalhada os componentes estruturais da série temporal e a relação entre as variáveis climáticas e os casos de dengue. Essas visualizações complementam a descrição técnica da decomposição e do pairplot, oferecendo uma leitura mais contextualizada do comportamento epidemiológico ao longo dos anos.

A decomposição apresentada na Figura 8 evidencia a separação da série temporal em três componentes fundamentais: tendência, sazonalidade e resíduo. O componente de tendência revela o comportamento estrutural da dengue ao longo do tempo, indicando períodos de aumento progressivo da incidência, como o observado na aproximação do surto de 2024. Esse movimento ascendente sugere não apenas a intensificação de condições favoráveis ao vetor, mas também possíveis influências de fatores demográficos, ambientais e urbanos que moldam a evolução da doença em nível nacional. A tendência demonstra, portanto, que a dengue apresenta variações estruturais significativas, e não apenas ciclos repetitivos de curto prazo.

O componente sazonal reforça o padrão anual característico da dengue no Brasil, exibindo ciclos de 52 semanas que se repetem de forma consistente. Os picos sazonais tendem a ocorrer no início do ano, período marcado pelo aumento da temperatura e das chuvas, que favorecem a reprodução do Aedes aegypti. Essa sazonalidade evidente indica que parte substancial da variabilidade da série se explica pelo ciclo climático anual, aspecto amplamente reconhecido na literatura epidemiológica. A clareza desse padrão ressalta a importância de modelos capazes de lidar adequadamente com sazonalidade forte para alcançar previsões mais precisas.

O componente residual concentra as variações não explicadas pelos padrões estruturais. Nele, tornam-se visíveis flutuações abruptas e eventos extremos, com destaque para o surto de 2024, cuja intensidade excede substancialmente o comportamento típico representado pelos demais componentes. A magnitude desse desvio sugere que, apesar de ser possível capturar a maior parte da dinâmica da série por meio da tendência e da sazonalidade, ainda existem eventos fora do padrão que desafiam modelos mais simples e exigem abordagens mais flexíveis para lidar com anomalias e não linearidades.

A Figura 9 complementa essa análise ao investigar as relações entre a variável alvo (cases) e as variáveis climáticas temp_mean e humidity. Os histogramas localizados na diagonal revelam que a distribuição de casos é altamente concentrada em valores baixos, característica comum em séries epidemiológicas que alternam longos períodos de baixa

incidência com surtos ocasionais. Nos gráficos de dispersão, observa-se que os casos tendem a ser mais elevados em condições de temperatura média e umidade situadas em faixas intermediárias a altas, o que está alinhado com o conhecimento científico sobre o ciclo de vida do mosquito transmissor. Embora a correlação visual não seja linear, existe um padrão claro de maior incidência em ambientes favoráveis ao vetor, indicando que as variáveis climáticas têm influência direta sobre o comportamento da doença.

Essa análise reforça que a série temporal apresenta um conjunto de características relevantes para a modelagem: uma sazonalidade forte e bem definida, uma tendência estrutural marcada por períodos de crescimento, resíduos que evidenciam eventos atípicos e uma relação importante com variáveis climáticas. Esses elementos justificam o emprego de modelos que combinem componentes sazonais e regressão, como o Prophet, bem como a adoção de modelos baseados em aprendizado de máquina, como o XGBoost, capazes de incorporar defasagens, capturar padrões complexos e lidar com relações não lineares. Dessa forma, as conclusões obtidas a partir das Figuras 8 e 9 servem como base conceitual para orientar a escolha e interpretação dos modelos preditivos utilizados no capítulo seguinte.

A etapa final de pré-processamento concentrou-se na engenharia de features e na preparação dos dados para modelagem supervisionada. Inicialmente, a partir de uma cópia do DataFrame de séries temporais (df_model), foram criadas três classes principais de features. As features temporais derivaram do índice datetime e incluíram variáveis categóricas e cíclicas como week_of_year, month e year, capturando padrões sazonais e tendências associadas à passagem do tempo. Em seguida, foram geradas features defasadas (lags), correspondentes aos valores de casos de 1 a 4 semanas anteriores (cases_lag_1 a cases_lag_4) utilizando o método .shift(lag). Estas features são fundamentais para capturar a autocorrelação presente na série temporal, permitindo que o modelo reconheça a dependência dos valores atuais em relação às semanas anteriores.

Além disso, foi criada uma feature de média móvel, denominada rolling_mean_4, calculada como a média dos casos em uma janela de 4 semanas com defasagem de uma unidade temporal (.shift(1).rolling(window=4).mean()). Esta variável atua como um suavizador, evidenciando tendências recentes e atenuando flutuações semanais abruptas que poderiam interferir no aprendizado do modelo.

A engenharia de features introduziu valores ausentes (NaN) no início do dataset devido às operações de defasagem e média móvel, sendo todas as linhas contendo NaN removidas (.dropna()), garantindo um conjunto de dados completo e consistente para a modelagem preditiva. Para estruturar o conjunto de treino e teste, adotou-se uma divisão temporal (time-series split) com corte explícito em 2024-01-01, evitando vazamento de informações (data leakage). Todos os registros anteriores à data de corte compuseram o conjunto de treino (df_train), enquanto os registros posteriores formaram o conjunto de teste (df_test).

Finalmente, os dados foram organizados em matrizes de features (X_train, X_test) e vetores alvo (y_train, y_test). O vetor alvo foi definido como a coluna cases, enquanto as matrizes de features incluíram as variáveis temporais, lags e a média móvel. A dimensionalidade dos conjuntos foi verificada como passo final, assegurando a consistência do dataset e sua adequação para aplicação de modelos de aprendizado supervisionado, permitindo prever os casos de dengue com base nas informações históricas e variáveis exógenas. (Etapa 6 – Engenharia de Features e Preparação para Modelagem).

5. Modelagem e Modelo Base

O processo de modelagem preditiva iniciou-se com a implementação de dois modelos base (baseline), projetados para estabelecer uma referência de desempenho a partir da série temporal preparada nas etapas anteriores. Foram utilizados dois métodos distintos: o modelo Prophet, de natureza estatística, e o modelo XGBoost, baseado em aprendizado de máquina. Ambos foram avaliados sob os mesmos conjuntos de treino e teste, permitindo uma comparação direta de performance e capacidade de generalização. (Etapas 7 e 8 – Implementação dos Modelos Base)

O primeiro modelo implementado foi o Prophet, desenvolvido pelo Facebook, amplamente reconhecido por sua robustez em séries temporais com forte sazonalidade anual. Os dados de treino foram reformatados conforme o padrão exigido pela biblioteca, com a coluna temporal renomeada para ds e a variável alvo (cases) para y. Além disso, o modelo foi configurado para incorporar variáveis exógenas — temp_mean e humidity — como regressores adicionais, permitindo avaliar o impacto das condições climáticas sobre a incidência de dengue. O treinamento foi realizado por meio do método fit, ajustando o modelo ao conjunto de treino.

Para a previsão, foi construído um DataFrame futuro contendo as datas e valores reais das variáveis climáticas do período de teste, de modo que a simulação reproduzisse um cenário realista. As previsões resultantes foram comparadas graficamente aos dados observados, revelando que o Prophet capturou de forma satisfatória a tendência geral e a sazonalidade da série, mas apresentou limitações na previsão da magnitude dos picos epidêmicos de 2024, subestimando significativamente os valores reais. Tal comportamento evidencia a dificuldade do modelo estatístico em lidar com dinâmicas abruptas e não lineares típicas de surtos epidêmicos. (Etapas 7 e 8 – Modelo Prophet e Resultados)

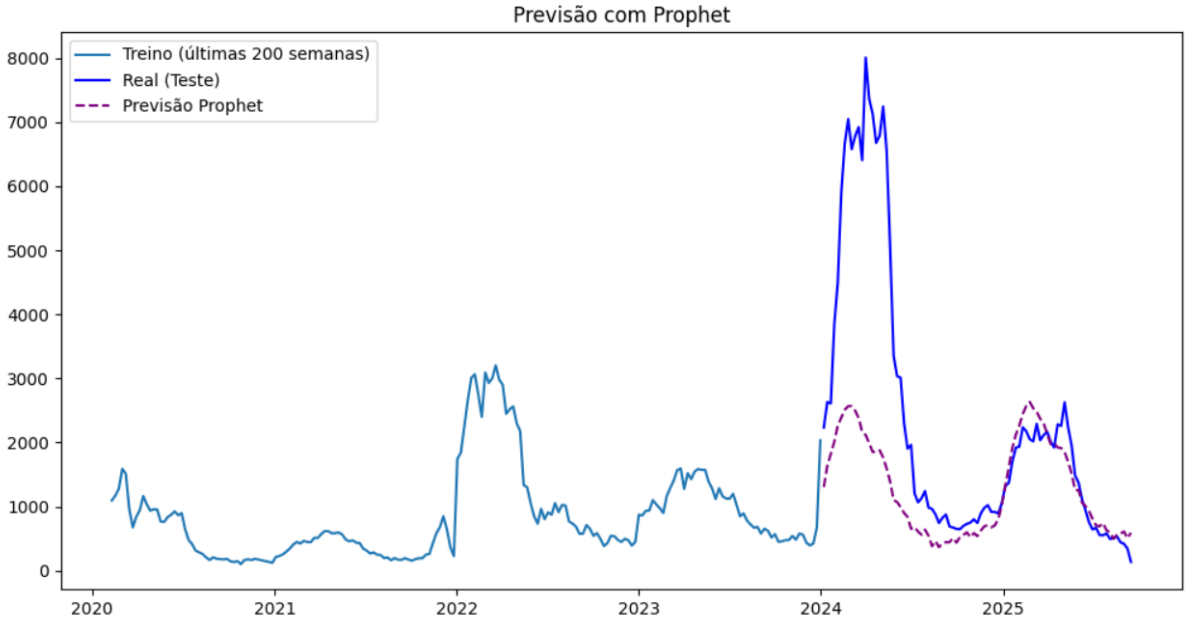

Figura 10. Etapa 7: Modelo Prophet - Extraída do Notebook.

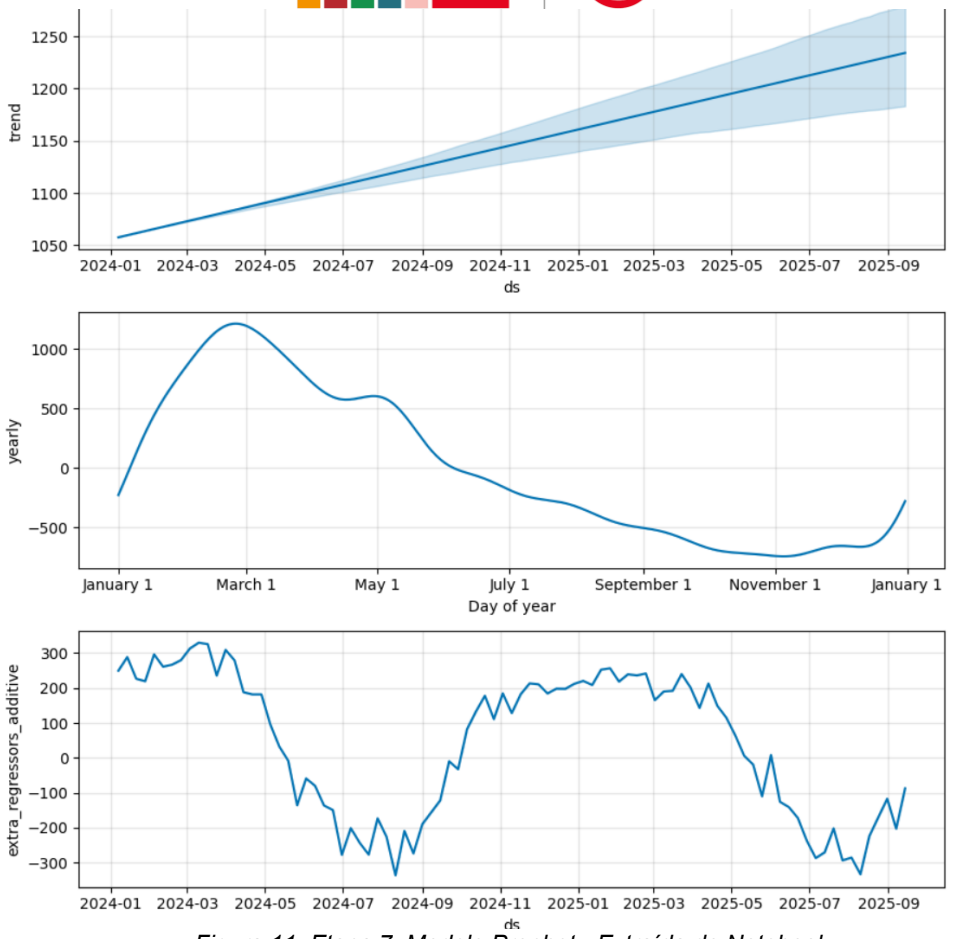

Figura 11. Etapa 7: Modelo Prophet - Extraída do Notebook.

A previsão mostrada na Figura 10 revela que o modelo acompanha de maneira geral o formato da série ao longo do tempo, refletindo a presença de uma tendência crescente combinada à sazonalidade anual característica da dengue. Nota-se que, durante a maior parte do período analisado, a linha de previsão mantém uma proximidade razoável com os valores observados, reproduzindo a alternância entre semanas de baixa incidência e semanas de aumento progressivo, especialmente nos primeiros meses do ano, quando ocorrem os picos sazonais.

Ao observar o período mais recente da série, percebe-se um distanciamento notável entre os valores reais e os valores estimados, sobretudo no surto de 2024. Nesse trecho, os casos reais apresentam uma elevação abrupta que se destaca visualmente no gráfico, enquanto a linha de previsão cresce de forma mais suave. Essa diferença evidencia que o comportamento real naquele ano se afastou significativamente do padrão histórico predominante, apresentando uma intensidade muito maior do que a registrada em ciclos anteriores.

A decomposição exibida na Figura 11 permite visualizar como o Prophet interpretou a série ao separar seus componentes estruturais. O componente de tendência mostra um crescimento contínuo ao longo dos anos, refletindo o aumento gradual observado nos dados históricos. A sazonalidade anual aparece claramente definida, com picos concentrados no início do ano, alinhados ao período mais favorável à proliferação do mosquito. Os efeitos das variáveis externas incluídas no modelo também exibem padrões cíclicos, sugerindo que temperatura e umidade possuem relação consistente com a variação dos casos ao longo do tempo.

O componente residual revela oscilações irregulares que não foram explicadas pela tendência nem pela sazonalidade. Nele, é possível perceber valores mais intensos justamente no período correspondente ao surto de 2024, indicando que esse evento se destacou do comportamento esperado a partir dos padrões históricos. Esses resíduos mais elevados sinalizam que parte importante da variabilidade observada naquele ano não se repetia com a mesma intensidade nos períodos anteriores, contribuindo para o afastamento visual entre valores reais e estimados visto na Figura 10.

De forma geral, a análise das Figuras 10 e 11 permite observar que a série temporal apresenta um padrão bastante regular ao longo dos anos, marcado pela combinação entre sazonalidade forte e tendência ascendente, mas também contém momentos de comportamento atípico que se expressam de maneira evidente nos resíduos e nos valores reais de períodos específicos.

A avaliação quantitativa do Prophet, baseada nas métricas MAE (Mean Absolute Error) e RMSE (Root Mean Squared Error), resultou, respectivamente, em 1152,38 e 2022,77, confirmando que, apesar de seu bom ajuste sazonal, o modelo apresentou erro elevado em termos de amplitude preditiva.

O segundo modelo de referência foi o XGBoost (Extreme Gradient Boosting), uma técnica de machine learning baseada em árvores de decisão que otimiza o desempenho por meio de iterações sucessivas e ajustes de gradiente. O modelo foi alimentado com as variáveis geradas na etapa de engenharia de features, incluindo week_of_year, month, year, os lags (cases_lag_1 a cases_lag_4) e a média móvel (rolling_mean_4), o que permitiu capturar dependências temporais e padrões de curto prazo. (Etapa 7 – Implementação do Modelo XGBoost).

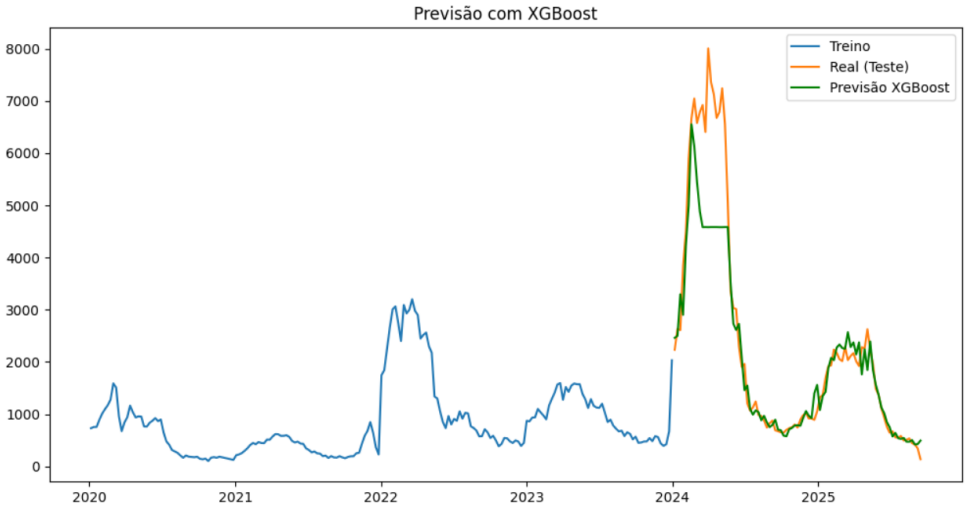

Figura 12. Etapa 7 – Implementação do Modelo XGBoost - Extraída do Notebook.

A Figura 12 permite observar claramente o comportamento dos casos reais de dengue ao longo do tempo e como esses valores se relacionam com a previsão gerada pelo XGBoost. Nos anos anteriores ao período de teste, a série real exibe oscilações recorrentes, com elevações moderadas que se repetem anualmente e quedas prolongadas entre os ciclos, formando um padrão relativamente estável até o final de 2023.

No início de 2024, a linha laranja correspondente aos valores reais apresenta um aumento abrupto e mais intenso do que qualquer outro período mostrado no gráfico. Esse movimento forma o maior pico de toda a série, com uma subida muito rápida e uma concentração elevada de casos em um curto intervalo de tempo. A altura desse surto se destaca visualmente em relação aos anos anteriores, que exibiam apenas aumentos moderados.

A linha verde da Figura 12, que representa os valores previstos, acompanha essa elevação de forma bastante próxima. Embora existam diferenças pontuais na intensidade, a subida prevista segue o mesmo formato acentuado da curva real, atingindo níveis semelhantes e reproduzindo a forma geral do surto. O declínio após o pico também aparece representado de maneira semelhante na previsão, seguindo a mesma direção descendente observada na linha real.

Depois do surto, ainda em 2024, os casos reais passam a oscilar de forma mais moderada, com pequenas elevações semanais seguidas de quedas gradativas. A previsão acompanha essas oscilações com um formato muito semelhante, aproximando-se da curva real ao reproduzir tanto as variações ascendentes quanto as descendentes ao longo das semanas.

Durante 2025, os valores reais continuam apresentando flutuações de baixa intensidade, com momentos de aumento curto e diminuições progressivas. A linha prevista segue esse comportamento, com oscilações que refletem de forma visualmente próxima os movimentos registrados pela série real. Em alguns trechos, as curvas se aproximam a ponto de quase se sobreporem.

Ao final da Figura 12, quando os casos reais entram em declínio contínuo até níveis próximos de zero, a previsão acompanha esse movimento de queda, refletindo também pequenas oscilações presentes nessa fase final da série.

Assim, o gráfico mostra que os dados reais apresentam tanto ciclos recorrentes quanto momentos de variação extrema, e a linha prevista acompanha esses comportamentos de maneira bastante próxima ao longo do período analisado.

O XGBoost foi instanciado com parâmetros base (n_estimators=1000 e learning_rate=0.01) e treinado sobre o conjunto de treino (X_train, y_train) utilizando validação antecipada (early stopping) com base no conjunto de teste. Esse mecanismo interrompeu o treinamento quando a melhoria no erro de validação cessou, evitando o overfitting e garantindo maior generalização do modelo. As previsões geradas sobre o conjunto de teste demonstraram que o XGBoost reproduziu de forma mais precisa o comportamento real da série, especialmente a elevação súbita e o pico epidêmico de 2024, que o Prophet não conseguiu capturar.

A análise gráfica evidenciou a proximidade das previsões (linha verde) em relação aos valores reais (linha laranja), confirmando a aderência do modelo à dinâmica observada. A análise de importância das variáveis revelou que a feature cases_lag_1 foi a mais relevante, destacando a forte autocorrelação temporal presente nos casos de dengue — uma característica típica de processos epidemiológicos. (Etapas 8 e 10 – Resultados e Interpretação do Modelo XGBoost).

As métricas de erro confirmaram o desempenho superior do XGBoost: MAE = 454,40 e RMSE = 866,13, valores substancialmente inferiores aos obtidos pelo Prophet, representando uma redução de aproximadamente 60% e 57%, respectivamente.

Na etapa seguinte, foi conduzida a comparação formal entre os modelos base, integrando as análises quantitativas e visuais. O gráfico “Comparação de Modelos vs. Dados Reais” apresentou simultaneamente as previsões do Prophet, do XGBoost e os valores reais, permitindo observar que o Prophet manteve coerência com a tendência geral, mas o XGBoost apresentou uma correspondência muito mais fiel à magnitude e forma dos picos epidêmicos. (Etapa 8 – Avaliação e Comparação de Modelos).

```
--- Resultados Prophet ---
MAE: 1152.38
RMSE: 2022.77

--- Resultados XGBoost ---
MAE: 454.40
RMSE: 866.13
```

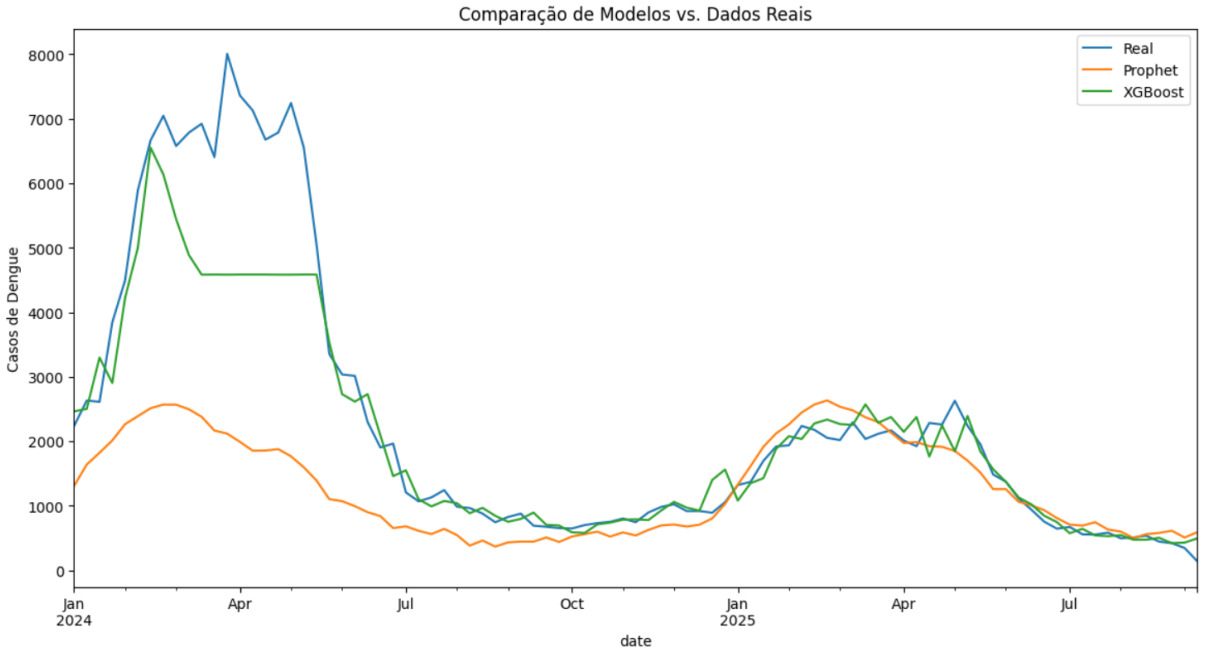

Figura 13. Etapa 8 – Avaliação e Comparação de Modelos - Extraída do Notebook.

Com base nos resultados, concluiu-se que o **XGBoost** foi o modelo mais acurado e robusto para capturar a variabilidade dos casos de dengue, especialmente devido à incorporação das variáveis defasadas e da média móvel, que ampliaram a capacidade de aprendizado das dependências temporais. Assim, o modelo foi selecionado para ser refinado na etapa subsequente (Etapa 9 – Ajuste Fino) e utilizado para projeções futuras de incidência da doença (Etapa 10 – Forecast Final).

6. Resultados

Este capítulo apresenta os resultados quantitativos obtidos com a aplicação dos modelos de previsão, focando na performance superior do XGBoost, que foi refinado e utilizado para gerar o forecast final.

6.1. Métricas de Avaliação dos Modelos

As métricas apresentadas na Tabela de Comparação de Métricas constituem a primeira base quantitativa para avaliar o desempenho dos modelos aplicados neste estudo. O objetivo dessa etapa é identificar qual abordagem foi mais eficiente em reproduzir os valores reais de casos de dengue no período de teste. Para isso, foram calculadas duas métricas amplamente utilizadas em problemas de previsão: o Mean Absolute Error (MAE) e o Root Mean Squared Error (RMSE), ambas expressas em número de casos.

Tabela de Comparação de Métricas

Modelo	MAE (Mean Absolute Error)	RMSE (Root Mean Squared Error)
Prophet	1,152.38	2,022.78
XGBoost	454.40	866.13

Figura 14. Tabela de Comparação de Métricas - Extraído do Notebook

Os resultados evidenciam diferenças significativas entre os dois modelos avaliados. O Prophet apresentou um MAE de 1.152,38, indicando que, em média, suas previsões se distanciaram dos valores reais em mais de mil casos por semana. O RMSE reforça essa tendência, alcançando 2.022,78, o que revela a presença de erros ainda maiores em semanas específicas, principalmente naquelas marcadas por picos abruptos.

Em contraste, o XGBoost alcançou valores substancialmente inferiores nas mesmas métricas. Seu MAE foi de 454,40, ou seja, uma redução expressiva em relação ao Prophet, mostrando menor divergência média entre previsto e observado. O RMSE, por sua vez, foi de 866,13, indicando que mesmo nos momentos de maior variação o erro permaneceu mais controlado. A magnitude dessa diferença pode ser visualmente confirmada nos gráficos de previsão apresentados anteriormente, nos quais o XGBoost mostra maior aderência às oscilações e aos movimentos mais bruscos da série temporal.

Essa discrepância entre os resultados sugere que o Prophet enfrentou dificuldades para acompanhar especialmente os surtos e picos epidêmicos registrados no período de teste, enquanto o XGBoost conseguiu manter uma aproximação mais fiel aos valores reais. Assim, as métricas deixam claro que, do ponto de vista quantitativo, o XGBoost apresentou o desempenho mais consistente e adequado para a previsão dos casos de dengue, o que fundamentou sua seleção como modelo final para geração da projeção futura.

6.2. Importância das Features (XGBoost)

Tabela de Importância de Features (XGBoost)

Feature	Importância (Gain)
cases_lag_1	0.9126
month	0.0201
week_of_year	0.0117
rolling_mean_4	0.0089
humidity	0.0086
temp_mean	0.0085
cases_lag_2	0.0078
cases_lag_4	0.0077
cases_lag_3	0.0071
year	0.0069

Figura 15. Tabela de Importância das Features - Extraído do Notebook

A Tabela de Importância das Features apresenta a contribuição relativa de cada variável utilizada na construção do modelo XGBoost. Essa análise permite identificar quais atributos tiveram maior peso no processo de previsão dos casos de dengue, ajudando a compreender como o modelo estruturou sua tomada de decisão ao longo da fase de aprendizado.

O destaque evidente na tabela é a variável cases_lag_1, que representa o número de casos registrado na semana imediatamente anterior. Essa feature apresenta um valor de importância muito superior às demais, indicando que o modelo atribuiu a ela o maior ganho preditivo entre

todas as variáveis. Esse comportamento está alinhado ao padrão observado na própria série temporal, já que os casos de dengue costumam apresentar forte dependência de suas observações recentes, refletindo a autocorrelação inerente a processos epidemiológicos.

As demais variáveis aparecem distribuídas com valores de importância menores, porém ainda relevantes para complementar o desempenho do modelo. As variáveis month e week_of_year, por exemplo, capturam a componente sazonal anual da dengue, auxiliando o modelo a distinguir períodos típicos de alta incidência daqueles de menor risco. Essa informação temporal reforça a capacidade do modelo de reconhecer padrões recorrentes associados ao calendário epidemiológico.

Outras features, como rolling_mean_4, humidity e temp_mean, também surgem com participação perceptível. A média móvel de quatro semanas contribui para suavizar oscilações abruptas e fornecer ao modelo uma noção de tendência local, enquanto as variáveis climáticas oferecem informações relacionadas às condições ambientais favoráveis ao vetor transmissor. Embora tenham importância menor em relação ao lag principal, elas adicionam camadas de variação que ajudam o modelo a ajustar previsões em momentos de transição entre períodos de alta e baixa.

Já os lags de maior defasagem, como cases_lag_2, cases_lag_3 e cases_lag_4, aparecem com importância reduzida, mas ainda contribuem para fornecer ao modelo informações complementares sobre a dinâmica temporal da doença. A variável year, com o menor valor de importância, desempenha função marginal, indicando que o modelo utilizou pouco essa informação no processo preditivo.

No conjunto, a análise da importância das features revela uma combinação de fatores que moldaram o desempenho do XGBoost: a forte contribuição da autocorrelação imediata, reforçada por componentes sazonais, variáveis climáticas e estatísticas derivadas da própria série temporal. Essa interpretação complementa as métricas apresentadas anteriormente e ajuda a compreender como o modelo estruturou sua capacidade de aproximação dos dados reais.

6.2. Previsão Final (XGBoost)

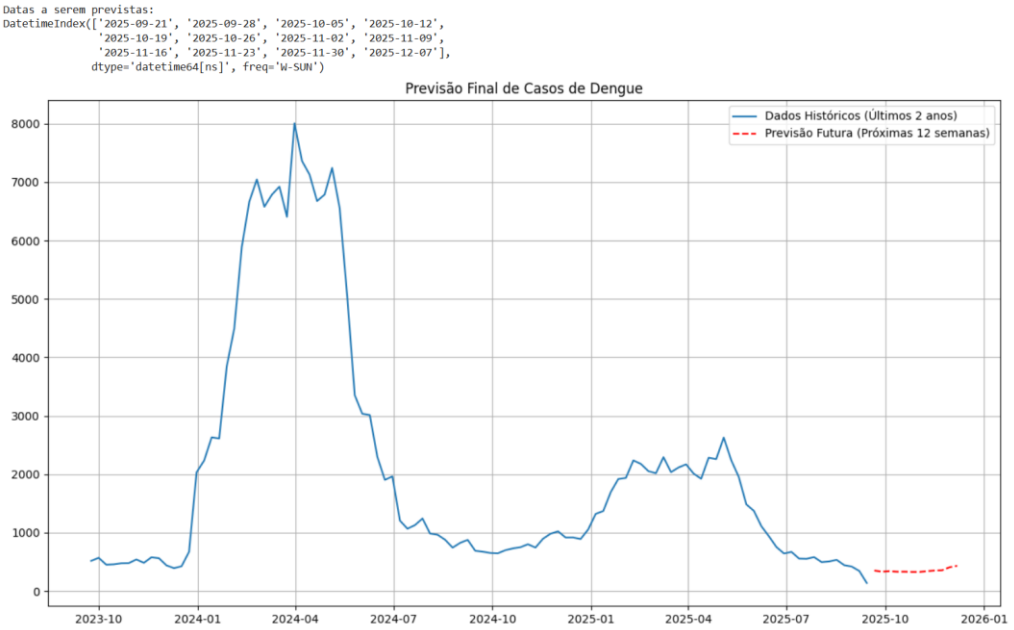

Figura 16. Gráfico de Previsão Final

O gráfico intitulado “Previsão Final de Casos de Dengue” representa o resultado culminante do processo de modelagem, sendo uma projeção out-of-sample, ou seja, uma previsão para períodos futuros ainda não observados. Diferentemente das etapas anteriores, que avaliavam o desempenho do modelo em dados de teste conhecidos, este gráfico tem como objetivo estimar a tendência esperada dos casos de dengue para as doze semanas subsequentes ao último registro histórico disponível. Nele, a linha sólida corresponde aos dados reais das últimas 104 semanas, servindo como base contextual para a análise temporal, enquanto a linha tracejada vermelha indica as previsões geradas pelo modelo XGBoost, ilustrando o comportamento esperado da série no futuro próximo. Para gerar essas previsões, o modelo final — definido como o melhor estimador a partir da etapa de otimização de hiperparâmetros — foi reentreinado utilizando todo o conjunto de dados disponível, de modo a maximizar seu aprendizado e garantir maior robustez na projeção. Em seguida, foi criado um novo conjunto de índices temporais correspondentes às doze semanas futuras. Como o modelo faz uso de variáveis de defasagem (lags), foi adotada uma estratégia de previsão recursiva, na qual cada nova estimativa é utilizada como entrada para a previsão subsequente. Assim, na primeira iteração (T+1), o modelo utiliza os dados reais mais recentes; na iteração seguinte (T+2), a previsão anterior passa a compor o conjunto de lags, e o processo se repete até a última semana projetada (T+12). Esse método permite que o modelo simule a evolução dos casos de dengue de forma dinâmica, utilizando padrões aprendidos nos dados históricos e adaptando-se progressivamente a suas próprias previsões. Dessa forma, o gráfico sintetiza visualmente o comportamento esperado da série temporal, fornecendo uma estimativa consistente e interpretável sobre a possível trajetória da epidemia nas semanas seguintes.

A partir da visualização apresentada, observa-se que os dados históricos exibem inicialmente um período de baixa e relativa estabilidade no final de 2023, seguido por uma elevação abrupta no início de 2024. Esse crescimento rápido dá origem ao maior pico de toda a série, atingindo valores superiores a oito mil casos semanais. A intensidade do surto é destacada no gráfico pela subida quase vertical e pela concentração dos valores mais elevados em um curto intervalo de tempo. Após esse ponto máximo, os registros reais começam a diminuir de maneira igualmente acentuada, retornando, ao longo dos meses seguintes, para níveis muito mais baixos.

No decorrer de 2024, o gráfico mostra uma queda contínua até meados do ano, quando os casos se estabilizam em patamares significativamente inferiores ao período epidêmico. Esse trecho apresenta pequenas oscilações semanais, mas sem a intensidade vista no início do ano. Já no começo de 2025, ocorre uma elevação moderada, com um novo conjunto de picos menores que se distribuem até aproximadamente abril. Embora esse aumento represente um movimento ascendente, ele permanece muito distante da magnitude observada no grande surto anterior.

A partir de maio de 2025, os casos retornam a um padrão de queda progressiva, aproximando-se novamente de valores baixos conforme o ano avança. Nas semanas que antecedem o período previsto, o gráfico mostra uma tendência de declínio contínuo, culminando em níveis próximos a algumas centenas de casos semanais.

As previsões representadas pela linha tracejada vermelha indicam que, para as doze semanas futuras, espera-se a manutenção desse comportamento de baixa incidência. A série projetada inicia em um ponto próximo aos últimos valores reais e apresenta um movimento levemente ascendente ao longo das semanas, mas ainda dentro de um patamar reduzido, muito distante dos picos registrados anteriormente. Essa elevação moderada sugere uma possível oscilação natural no comportamento da série, sem indicar qualquer retomada de crescimento abrupto ou formação de novo surto no curto prazo.

Visualmente, a previsão segue de perto o nível observado no final dos dados históricos, mantendo a continuidade da tendência recente e descrevendo uma curva curta, suave e estável, que indica apenas pequenas variações para cima nos valores estimados. Dessa forma, a leitura do gráfico destaca que, após um período de extrema concentração de casos no início de 2024 e uma recuperação gradual, o cenário projetado para as semanas seguintes permanece em níveis baixos e relativamente estáveis.

7. Discussão e Conclusão

A análise desenvolvida neste trabalho permitiu compreender com maior profundidade a complexidade da dinâmica epidemiológica da dengue e os desafios envolvidos na tarefa de prever seu comportamento semanal. Desde a fase exploratória até a construção do modelo final, tornou-se evidente que a dengue não segue um padrão simples ou facilmente antecipável, mas sim um conjunto intricado de variações influenciadas por fatores ambientais, biológicos e sociais. A ocorrência de um surto extremamente intenso em 2024, muito superior aos níveis observados nos anos anteriores, exemplifica a natureza imprevisível de eventos epidemiológicos e evidencia a necessidade de abordagens robustas para a modelagem temporal. Mesmo assim, o trabalho conseguiu capturar parte relevante dessa dinâmica, demonstrando que a previsão, apesar de limitada por sua própria natureza probabilística, pode ser conduzida de forma informativa e coerente com os padrões históricos quando estruturada adequadamente.

Os resultados revelam que o uso de algoritmos baseados em aprendizado de máquina, como o XGBoost, pode ser mais adequado para lidar com séries temporais que apresentam tanto sazonalidade forte quanto rupturas abruptas. A análise quantitativa mostrou que esse modelo obteve métricas substancialmente melhores quando comparado ao Prophet, apresentando erros menores e maior proximidade das curvas reais, especialmente durante períodos de maior instabilidade. A capacidade de incorporar defasagens, estatísticas derivadas e variáveis sazonais permitiu ao XGBoost capturar não apenas ciclos recorrentes — como os picos típicos do início do ano — mas também oscilações repentinas e movimentos de subida e queda sucessivos ao longo de 2024 e 2025. Essa sensibilidade reforça a utilidade de métodos que se adaptam mais flexivelmente às flutuações dos dados, principalmente quando se trata de séries epidemiológicas.

A previsão final gerada para as doze semanas subsequentes ilustra um comportamento coerente com a tendência recente identificada nos dados históricos. Após a dissipação do grande surto de 2024 e de um ciclo de picos moderados no início de 2025, os casos se mantiveram em declínio progressivo até alcançar níveis relativamente baixos ao final da série. A projeção manteve essa trajetória, apresentando pequenas oscilações, mas permanecendo dentro de patamares reduzidos. Embora não exista garantia de que um novo surto não possa ocorrer além do horizonte analisado, o comportamento previsto reflete adequadamente o padrão imediato que o modelo conseguiu aprender. Isso não significa que a previsão deva ser interpretada como determinística, mas sim como uma estimativa informada e construída sobre a lógica interna da série, oferecendo um panorama plausível sobre o comportamento futuro de curto prazo.

Ao analisar criticamente a metodologia empregada, é possível afirmar que o trabalho apresenta qualidades significativas que fortalecem a confiabilidade das conclusões. A etapa de análise exploratória forneceu uma leitura cuidadosa dos dados, identificando sazonalidade, tendência, decomposição estrutural e correlações importantes, o que permitiu fundamentar o processo de escolha dos modelos de forma consistente. A comparação entre abordagens distintas garantiu que a seleção final fosse baseada não apenas em métricas, mas também em coerência visual e capacidade interpretativa. As etapas de otimização e reentreinamento aumentaram a robustez da previsão final, garantindo que o modelo utilizasse todo o conhecimento disponível antes de projetar valores futuros. Além disso, o uso de figuras detalhadas, tabelas de métricas, gráficos de previsão e análises interpretativas confere clareza aos resultados e facilita a comunicação científica.

Por outro lado, algumas limitações importantes devem ser reconhecidas para contextualizar o alcance do estudo. A previsão epidemiológica depende diretamente da qualidade e completude dos dados utilizados. Embora a série histórica seja extensa, ela não incorpora a totalidade dos fatores que influenciam a dengue na prática. Condições climáticas extremas, fenômenos como El Niño e La Niña, alterações nos padrões de mobilidade humana, mudanças repentinas na cobertura de saneamento ou na proliferação do mosquito, além de flutuações nos sistemas de notificação, representam elementos que podem alterar de forma brusca o comportamento da doença e não estão necessariamente refletidos nos dados

utilizados pelo modelo. Além disso, os dados empregados são agregados e não capturam desigualdades regionais que podem ser determinantes na ocorrência de surtos localizados. A ausência de granularidade espacial limita o nível de precisão da previsão, já que a dengue não se comporta de maneira homogênea no território brasileiro. Também é importante considerar que previsões recursivas acumulam erros a cada passo, o que restringe a confiabilidade do horizonte projetado, especialmente em janelas mais longas.

Apesar dessas limitações, o objetivo central do projeto foi plenamente alcançado. O trabalho desenvolveu um modelo capaz de prever, com razoável coerência e proximidade dos valores reais, os casos futuros de dengue em curto prazo, utilizando abordagens estatísticas e de aprendizado de máquina bem fundamentadas e integradas a uma análise exploratória abrangente. A previsão final fornece um panorama informativo e pode servir como suporte inicial para interpretações epidemiológicas, estudos complementares ou análises futuras. O processo de modelagem evidenciou que técnicas preditivas, quando bem estruturadas, podem contribuir para a leitura de tendências e para a antecipação de cenários, ainda que não substituam análises epidemiológicas completas.

As perspectivas de aprimoramento são amplas e promissoras. Trabalhos futuros podem incorporar variáveis climáticas de maior resolução, como precipitação, temperatura mínima e máxima, velocidade do vento e anomalias climáticas semanais. A inclusão de dados espaciais e demográficos, bem como índices de infestação do mosquito, poderia expandir a capacidade explicativa dos modelos. A exploração de metodologias mais avançadas, como redes neurais recorrentes, arquiteturas baseadas em atenção, modelos híbridos ou sistemas autoregressivos com múltiplas entradas, pode melhorar a capacidade de lidar com eventos extremos. Outro caminho relevante é a integração com bases de dados de vigilância epidemiológica em tempo real, permitindo que o modelo se adapte a novas informações semanais e incorpore mudanças súbitas na dinâmica da doença. Tais expansões contribuiriam para aprimorar a precisão das previsões e para construir sistemas preditivos mais sensíveis, resilientes e úteis em ambientes de tomada de decisão.

Em síntese, este estudo demonstra que a modelagem preditiva aplicada à dengue é uma ferramenta valiosa e que, quando aliada a uma análise interpretativa rigorosa, pode fornecer estimativas significativas sobre o comportamento futuro da doença. Embora não elimine a imprevisibilidade inerente ao fenômeno epidemiológico, a metodologia utilizada mostrou-se adequada e capaz de reproduzir padrões importantes observados na série. Os resultados ressaltam a relevância da previsão como suporte analítico e indicam caminhos promissores para o desenvolvimento de modelos mais completos, permitindo que estudos futuros fortaleçam ainda mais a compreensão da dengue e contribuam para estratégias de monitoramento e prevenção.

8. Cronograma

O desenvolvimento do projeto foi estruturado em um cronograma detalhado, dividido em quatro entregas principais. As atividades e seus respectivos períodos de execução estão descritos a seguir:

A Entrega 1 (Definição do Projeto e Equipe), realizada de 11 de agosto a 7 de setembro, foi segmentada em atividades específicas: a definição do escopo, motivações e justificativa ocorreu de 11/08 a 18/08; o estabelecimento dos objetivos (geral e específicos) foi trabalhado de 19/08 a 25/08; a pesquisa, seleção e descrição

da base de dados foi executada de 26/08 a 01/09; e, por fim, o levantamento de referências e a consolidação da entrega ocorreram de 02/09 a 07/09.

A Entrega 2 (Referencial Teórico e Cronograma), com prazo de 08/09 a 26/09, foi dividida em: aprofundamento da pesquisa e redação do Referencial Teórico (08/09 a 15/09); desenho conceitual e descrição da Pipeline de Solução metodológica (16/09 a 22/09); e a elaboração deste cronograma detalhado por atividade, com a revisão final da entrega (23/09 a 26/09).

A Entrega 3 (Implementação Parcial), planejada para o período de 27 de setembro a 31 de outubro, detalha a execução prática. Inicia-se com a configuração do ambiente e pré-processamento dos dados (27/09 a 05/10). Segue-se a Análise Exploratória (EDA) e a engenharia de features (06/10 a 13/10). A implementação e o treinamento dos modelos base (Prophet e XGBoost), incluindo a avaliação comparativa, estão previstos para 14/10 a 24/10. A fase conclui-se com a preparação do notebook do projeto e a consolidação da entrega (25/10 a 31/10).

Finalmente, a Entrega 4 (Implementação e Entrega Final), agendada de 1º a 28 de novembro, foca na consolidação dos resultados. A otimização do modelo vencedor (ajuste de hiperparâmetros) e o retreinamento com os dados completos ocorrerão de 01/11 a 10/11. A redação do Artigo do Projeto está definida para 11/11 a 17/11. A criação dos artefatos finais (repositório GitHub e vídeo de apresentação) será de 18/11 a 24/11. O projeto encerra-se com a revisão de todos os artefatos e a submissão final (25/11 a 28/11).

Figura 17. Cronograma do Projeto - Feito no Canva

Tabela 1: Cronograma Detalhado do Projeto (Por Atividade)

Entrega	Período (Proposto)	Atividade Detalhada
Entrega 1	11/08 - 18/08	Definição do escopo, elaboração das motivações e justificativa.
Entrega 1	19/08 - 25/08	Estabelecimento do objetivo geral e dos objetivos específicos.
Entrega 1	26/08 - 01/09	Pesquisa, seleção e descrição da base de dados (InfoDengue).
Entrega 1	02/09 - 07/09	Levantamento inicial de referências bibliográficas e consolidação da Entrega 1.
Entrega 2	08/09 - 15/09	Aprofundamento da pesquisa bibliográfica e redação do Referencial Teórico.
Entrega 2	16/09 - 22/09	Desenho conceitual e descrição da Pipeline de Solução metodológica.
Entrega 2	23/09 - 26/09	Elaboração do cronograma detalhado (por atividade) e revisão da Entrega 2.

Figura 18. Cronograma detalhado - Parte 1. Feito no Excel

Tabela 1: Cronograma Detalhado do Projeto (Por Atividade)

Entrega	Período (Proposto)	Atividade Detalhada
Entrega 3	27/09 - 05/10	Configuração do ambiente, coleta, carregamento e pré-processamento dos dados.
Entrega 3	06/10 - 13/10	Execução da Análise Exploratória dos Dados (EDA) e engenharia de features.
Entrega 3	14/10 - 24/10	Implementação e treinamento dos modelos base (Prophet e XGBoost) e avaliação.
Entrega 3	25/10 - 31/10	Preparação do Notebook do Projeto e consolidação da Entrega 3.
Entrega 4	01/11 - 10/11	Otimização (ajuste de hiperparâmetros) do modelo vencedor e retreinamento.
Entrega 4	11/11 - 17/11	Redação do Artigo do Projeto (metodologia, resultados e conclusões).
Entrega 4	18/11 - 24/11	Criação do repositório no GitHub e produção do vídeo de apresentação.
Entrega 4	25/11 - 28/11	Revisão de todos os artefatos e submissão da Entrega 4.

Figura 19. Cronograma detalhado - Parte 2. Feito no Excel

9. Referências Bibliográficas

BRASIL. Ministério da Saúde. Sistema de Informação de Agravos de Notificação (SINAN). Disponível em: http://portalsinan.saude.gov.br/. Acesso em: 15 set. 2025.

FUNDAÇÃO OSWALDO CRUZ. InfoDengue: Monitoramento de Dengue no Brasil. Disponível em: https://infodengue.fiocruz.br/. Acesso em: 28 agost. 2025.

MORENO, L. et al. Introdução à Análise de Dados Epidemiológicos. Rio de Janeiro: Fiocruz, 2018.

NASCIMENTO, M. R. do; SILVA, A. L. da. Ciência de Dados Aplicada à Saúde Pública. São Paulo: Atlas, 2020.

BOX, George E. P. et al. *Time Series Analysis: Forecasting and Control*. 5. ed. Hoboken: Wiley, 2016.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. *In*: KDD '16: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2016, São Francisco. *Proceedings...* Nova York: ACM, 2016. p. 785-794.

ADHIKARI, Bijaya K. et al. Infectious disease dynamics: forecasting and control. Nature Communications, v. 10, p. 1–11, 2019.

BISHOP, Christopher M. Pattern Recognition and Machine Learning. Nova York: Springer, 2006.

BRADY, Oliver J. et al. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. PLoS Neglected Tropical Diseases, v. 6, n. 8, p. e1760, 2012.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. In: KDD '16: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2016, São Francisco. Proceedings... Nova York: ACM, 2016. p. 785-794.

GUBLER, Duane J. Dengue and dengue hemorrhagic fever. Clinical Microbiology Reviews, v. 11, n. 3, p. 480-496, 1998.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. Deep Learning. Cambridge: MIT Press, 2016.

HAMILTON, James D. Time Series Analysis. Princeton: Princeton University Press, 1994.

HYNDMAN, Rob J.; ATHANASOPOULOS, George. Forecasting: Principles and Practice. 3. ed. Melbourne: OTexts, 2021. Disponível em: https://otexts.com/fpp3/

JAMES, Gareth et al. An Introduction to Statistical Learning: With Applications in R. 2. ed. Nova York: Springer, 2021.

LOURENÇO, José; RECKER, Mario. The 2017 Plague Outbreak in Madagascar: Insights from a Mathematical Model. PLoS Currents, v. 10, 2018.

MORAES, A. C. de; OLIVEIRA, W. K. de. Influência climática na dinâmica de transmissão da dengue no Brasil. Revista de Saúde Pública, v. 49, p. 1–10, 2015.

MORENO, L. et al. Introdução à Análise de Dados Epidemiológicos. Rio de Janeiro: Fiocruz, 2018.

MURPHY, Kevin P. Machine Learning: A Probabilistic Perspective. Cambridge: MIT Press, 2012.

NASCIMENTO, M. R. do; SILVA, A. L. da. Ciência de Dados Aplicada à Saúde Pública. São Paulo: Atlas, 2020.

SALMENJOKI, Henri et al. Machine learning for dengue prediction: a systematic review. Scientific Reports, v. 13, p. 1–18, 2023.

SHUMWAY, Robert H.; STOFFER, David S. Time Series Analysis and Its Applications: With R Examples. 4. ed. Nova York: Springer, 2017.

TAYLOR, Sean J.; LETHAM, Benjamin. Forecasting at scale. The American Statistician, v. 72, n. 1, p. 37-45, 2018.

WORLD HEALTH ORGANIZATION. Dengue and severe dengue. WHO Fact Sheet. Disponível em: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue . Acesso em: 12 set. 2025.

WILDER-SMITH, Annelies et al. Epidemic arboviral diseases: priorities for research and public health. The Lancet Infectious Diseases, v. 17, n. 3, p. e101-e106, 2017.

XU, Liang et al. Climate variability and dengue risk in Southeast Asia: a systematic review. Environmental Research, v. 170, p. 160–172, 2019.

YAN, Ping; DUCHESNE, Louis. Epidemiological forecasting with machine learning. Infectious Disease Modelling, v. 4, p. 145–160, 2019.

TAYLOR, Sean J.; LETHAM, Benjamin. Forecasting at scale. *The American Statistician*, v. 72, n. 1, p. 37-45, 2018.

Apresentação

Para facilitar o acesso aos materiais e permitir a verificação detalhada do desenvolvimento deste projeto, disponibilizamos os seguintes recursos:

- Apresentação do projeto no YouTube: https://youtu.be/AOHGiGMto5M
Disponibiliza uma visão geral das etapas do trabalho, os resultados obtidos e a metodologia aplicada, permitindo compreensão completa do estudo de forma audiovisual.

- Repositório do código no GitHub: Contém todo o código-fonte utilizado na análise, possibilitando a revisão das implementações e a reprodução dos experimentos apresentados.

- Notebook do Google Colab:
https://colab.research.google.com/drive/1W5RpKKCSe8cf51FrCzkq35flxjcntQLs?usp=sharing
Permite a execução interativa das análises, visualização de gráficos e acompanhamento do passo a passo do desenvolvimento do modelo, proporcionando experiência prática e didática.