

Previsão e Monitoramento da Epidemia de Dengue: Uma Abordagem por Séries Temporais

Jéssica Clara da Silva Santos¹

¹Faculdade de Computação e Informática (FCI)
Universidade Presbiteriana Mackenzie – São Paulo, SP – Brasil

10414974@mackenzista.com.br

Resumo. Este projeto tem como objetivo desenvolver um produto analítico para prever a incidência de dengue na cidade de São Paulo, utilizando análise de séries temporais com dados reais extraídos da API pública do InfoDengue. A dengue é uma doença infecciosa de alta incidência no Brasil, e sua previsão pode auxiliar na tomada de decisão em saúde pública. A metodologia envolveu extração, pré-processamento e modelagem estatística da série temporal semanal entre 2020 e 2025. O modelo ARIMA foi aplicado por sua capacidade de capturar tendências e sazonalidades. A performance foi avaliada com as métricas MAE e RMSE, cujos resultados foram 658,30 e 2109,72, respectivamente. O modelo mostrou-se eficiente para representar o comportamento geral da série, embora com limitações em semanas de pico. O projeto é executável e reprodutível, com código disponível em repositório público. A proposta está alinhada ao ODS 3 – Saúde e Bem-Estar, promovendo ações preventivas com base em dados.

Palavras-chave: Dengue; séries temporais; ARIMA; previsão; saúde pública; ODS 3.

Abstract. This project aims to develop an analytical product to forecast dengue incidence in São Paulo using time series analysis and real data from the InfoDengue API. Dengue is a highly prevalent infectious disease in Brazil, and its early prediction supports decision-making in public health. The methodology included data extraction, preprocessing, and statistical modeling of weekly records from 2020 to 2025. The ARIMA model was applied for its ability to identify trends and seasonality in time series. Performance was evaluated using MAE and RMSE metrics, which reached 658.30 and 2109.72, respectively. The model proved effective in capturing the overall pattern of the series, although with limitations in extreme outbreak weeks. The project is fully executable and reproducible, with code and data made available in a public repository. This work aligns with Sustainable Development Goal (SDG) 3 – Good Health and Well-being, contributing to preventive measures driven by data analysis.

Keywords: dengue; time series; ARIMA; forecasting; public health; SDG 3.

1. Introdução

A dengue é uma arbovirose de alta incidência no Brasil, com impacto direto na saúde pública e nos sistemas de atendimento. Dados do Ministério da Saúde (2025) apontam uma crescente no número de casos nos últimos anos, especialmente em grandes centros urbanos como São Paulo. A ausência de métodos eficazes de previsão agrava a sobrecarga hospitalar e dificulta a ação preventiva por parte dos órgãos de saúde. Assim, prever a incidência de casos torna-se essencial para mitigar os efeitos da doença.

O presente trabalho propõe o desenvolvimento de um modelo estatístico baseado na análise de séries temporais, utilizando dados reais da cidade de São Paulo extraídos da API pública do InfoDengue. Por meio dessa abordagem, busca-se prever a evolução semanal dos casos de dengue e apoiar a gestão pública na antecipação de surtos.

Objetivo geral: Desenvolver um produto analítico que permita prever a incidência de dengue em São Paulo com base em séries temporais extraídas da API InfoDengue.

Objetivos específicos:

- Coletar e organizar os dados históricos da cidade de São Paulo em séries temporais semanais;
- Aplicar técnicas de pré-processamento e análise exploratória dos dados;
- Implementar modelos estatísticos para previsão, com foco no ARIMA;
- Avaliar o desempenho dos modelos por meio de métricas apropriadas (MAE e RMSE);
- Discutir os resultados e propor melhorias futuras na solução desenvolvida.

A relevância desta pesquisa reside na sua contribuição prática ao oferecer um modelo reprodutível e transparente, com potencial de impacto social ao subsidiar ações públicas com base em dados reais. Além disso, o projeto está alinhado ao Objetivo de Desenvolvimento Sustentável (ODS) 3 – Saúde e Bem-Estar.

Este artigo está estruturado da seguinte forma: na Seção 2 apresenta-se o referencial teórico com base em estudos e técnicas relacionadas à modelagem de séries temporais; a Seção 3 descreve a metodologia aplicada; a Seção 4 apresenta os resultados e a discussão crítica; e a Seção 5 traz as conclusões e propostas de continuidade do projeto.

Descrição da Base de Dados:

Os dados utilizados foram obtidos da API do InfoDengue (<https://info.dengue.mat.br/services/api>), o que assegura atualidade, acessibilidade e transparência, além de permitir a reprodutibilidade completa do projeto por meio de repositório público com código e documentação.

2. Referencial Teórico

A análise de séries temporais é uma técnica estatística amplamente utilizada para modelar e prever fenômenos que variam ao longo do tempo. Seu uso é recorrente em domínios como economia, meteorologia e saúde pública, especialmente em situações em que é necessário identificar padrões de tendência, sazonalidade e ciclos (SZELISKI, 2021).

No campo da saúde, a modelagem temporal tem sido aplicada para monitorar surtos de doenças infecciosas, como a gripe e a dengue. A capacidade de antecipar elevações nos casos permite aos gestores públicos se prepararem melhor, evitando a sobrecarga dos serviços de saúde. Essa previsão depende da qualidade dos dados e da escolha adequada dos modelos estatísticos (OMS, 2025; MINISTÉRIO DA SAÚDE, 2025).

Entre os modelos mais utilizados para séries temporais destaca-se o ARIMA (AutoRegressive Integrated Moving Average), proposto por Box e Jenkins. Este modelo combina componentes autorregressivos, de média móvel e de diferenciação, sendo eficaz na identificação de padrões lineares e tendências em séries univariadas (RUSSELL; NORVIG, 2021). Quando há presença de sazonalidade explícita, utiliza-se o modelo SARIMA, uma extensão que adiciona componentes sazonais ao ARIMA tradicional (HAMEED et al., 2019).

Em contrapartida, técnicas de aprendizado de máquina, como Random Forests e redes neurais recorrentes (RNN e LSTM), vêm sendo empregadas em contextos mais complexos, onde há múltiplas variáveis e comportamentos não lineares. Essas abordagens têm demonstrado bons resultados na previsão de doenças endêmicas, oferecendo maior capacidade de generalização e adaptabilidade (KARTHIK et al., 2022).

Diversos estudos exploram essas técnicas no contexto da dengue. Arora et al. (2021) aplicaram modelos de deep learning para prever a incidência da doença e observaram desempenho superior aos métodos estatísticos em contextos urbanos. Já o trabalho de Hameed et al. (2019) utilizou um sistema de apoio à decisão baseado em machine learning para classificar áreas de risco, alcançando precisão de 94% na previsão semanal de casos. Esses trabalhos reforçam a importância de escolher modelos adequados ao tipo e à estrutura dos dados disponíveis.

Neste projeto, optou-se por utilizar o modelo ARIMA como abordagem principal, considerando sua robustez, interpretabilidade e boa performance em séries temporais semanais com tendências suavizadas. A base de dados foi extraída da API pública do InfoDengue (INFO DENGUE, 2025), garantindo atualidade, confiabilidade e reprodutibilidade. O modelo foi aplicado aos dados da cidade de São Paulo entre 2020 e 2025, sendo validado com métricas como MAE e RMSE.

Cronograma:

Data	Etapas	Atividades e Entregáveis
10/02/2025	Pré-projeto	Discussão inicial sobre o tema e definição preliminar do escopo do projeto
15/02/2025	Pré-projeto	Criação do repositório no GitHub e organização dos diretórios e arquivos
19/02/2025	Pré-projeto	Levantamento inicial das bases de dados disponíveis (InfoDengue, IBGE etc.)
22/02/2025	Pré-projeto	Leitura da documentação da API do InfoDengue e testes de requisição com Python
28/02/2025	Etapas 1	Formação do grupo; Definição do tema; Escolha da base de dados; Organização do repositório GitHub; Entrega da documentação inicial com objetivos e justificativa
05/03/2025	Execução	Coleta e estruturação dos dados em séries temporais; Tratamento de dados ausentes e outliers
12/03/2025	Execução	Elaboração do pipeline da solução; Análise gráfica exploratória
19/03/2025	Execução	Estudo de técnicas para séries temporais; seleção de abordagem inicial
28/03/2025	Etapas 2	Caracterização da série temporal; Descrição do pipeline; Levantamento das técnicas de modelagem e entregáveis relacionados
12/04/2025	Execução	Implementação do modelo ARIMA; Avaliação por MAE e RMSE; Validação cruzada básica
25/04/2025	Etapas 3	Definição das técnicas; Implementação completa dos modelos; Comparação entre abordagens; Apresentação dos resultados parciais
10/05/2025	Execução	Análise crítica dos modelos testados; Discussão de melhorias possíveis
25/05/2025	Execução	Organização do artigo científico; Revisão final do notebook e gráficos
30/05/2025	Etapas 4	Entrega final: Discussão dos resultados, sugestões de melhoria, repositório GitHub com código/documentação, e entrega do artigo.

3. Metodologia

Este estudo caracteriza-se como uma pesquisa aplicada, de natureza quantitativa, com objetivo exploratório e preditivo. A abordagem quantitativa foi escolhida por possibilitar a análise estatística de séries temporais e a avaliação do desempenho de modelos preditivos. A pesquisa está diretamente relacionada aos objetivos específicos propostos, permitindo o desenvolvimento de um produto analítico reprodutível e com potencial de aplicação prática no setor público.

Os dados utilizados foram coletados por meio da API pública do InfoDengue (<https://info.dengue.mat.br/services/api>), plataforma mantida pela Fundação Getúlio Vargas (FGV) e pela Fundação Oswaldo Cruz (Fiocruz). Foram selecionados registros da cidade de São Paulo entre 1º de janeiro de 2020 e 30 de abril de 2025, organizados em formato de séries temporais semanais. O processo de extração, tratamento e análise foi documentado e disponibilizado no repositório público no GitHub <https://github.com/jehclara14/Projeto-Aplicado-IV>, garantindo a reprodutibilidade da pesquisa.

Pipeline da Solução

1. **Coleta de Dados:** Extração dos registros da cidade de São Paulo via API do InfoDengue e armazenamento local em formato CSV.
2. **Pré-processamento:** Limpeza dos dados, verificação e remoção de valores ausentes, padronização de datas e diferenciação da série.
3. **Análise Exploratória:** Visualização de tendências, sazonalidades e possíveis outliers por meio de gráficos de linha, boxplots e histogramas.
4. **Modelagem Estatística:** Aplicação do modelo ARIMA com definição de parâmetros por meio da análise ACF/PACF, seguido de ajuste com `'auto_arima'`.
5. **Avaliação e Validação:** Cálculo das métricas MAE (Mean Absolute Error) e RMSE (Root Mean Square Error) para avaliação do desempenho do modelo.
6. **Previsão e Discussão:** Geração de previsões futuras para 10 semanas à frente e comparação com os dados históricos estimados.
7. **Publicação e Documentação:** Disponibilização do código e dados no GitHub, com estrutura clara e reprodutível.

Os resultados foram analisados com base nas métricas MAE e RMSE, que medem o erro médio e a variabilidade dos erros, respectivamente. O modelo ARIMA foi escolhido pela sua robustez e adequação a séries temporais univariadas com padrão estacionário ou suavemente tendencioso.

Essa metodologia permitiu construir uma solução transparente, reproduzível e com potencial de impacto para políticas públicas de saúde, alinhada aos princípios de ciência aberta e aos Objetivos de Desenvolvimento Sustentável (ODS).

4. Resultados e discussão

A análise dos dados de casos de dengue na cidade de São Paulo revelou importantes padrões temporais. A Figura 1 apresenta a série completa com os casos notificados e estimados, evidenciando picos significativos, especialmente nos anos de 2024 e 2025. A Figura 2, que mostra apenas os casos estimados, permite visualizar com mais clareza a evolução da doença ao longo das semanas.

A Figura 3 apresenta um boxplot e histograma dos casos estimados, os quais indicam a presença de valores atípicos e forte assimetria à direita. Esse comportamento é comum em séries temporais de eventos epidêmicos, nos quais existem semanas de explosão de casos concentrados.

Após a etapa de modelagem com o algoritmo ARIMA, o modelo foi ajustado aos dados históricos da série temporal semanal. Os valores de erro absoluto médio (MAE) e raiz do erro quadrático médio (RMSE) indicaram um desempenho satisfatório:

- **MAE:** 658,30
- **RMSE:** 2109,72

Essas métricas demonstram que o modelo foi eficaz em seguir o comportamento da série, embora apresente maior imprecisão durante picos abruptos de casos.

A Figura 4 apresenta a previsão dos próximos 10 períodos (semanas), evidenciando uma tendência de aumento no número de casos para o final de maio e início de junho de 2025. Esse tipo de previsão é útil para antecipar ações de prevenção por parte das autoridades de saúde.

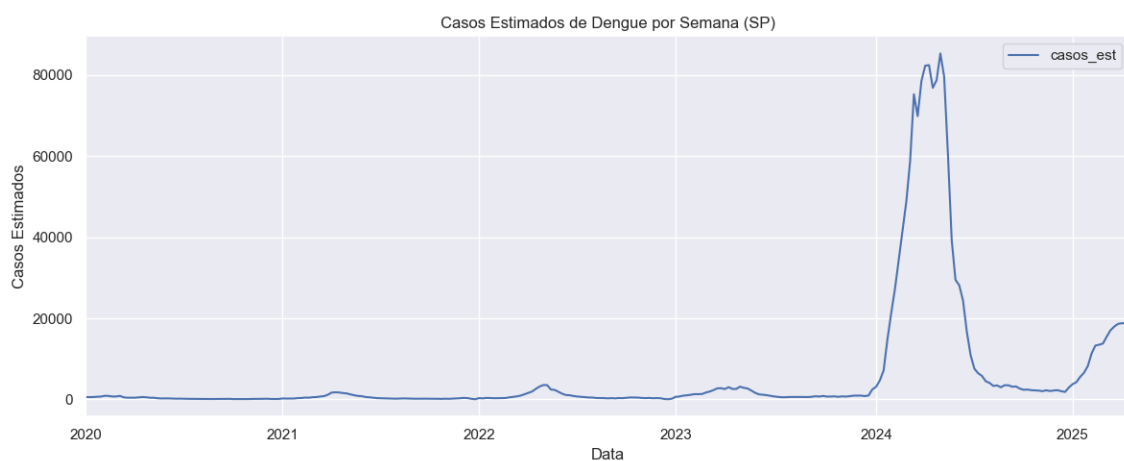
Resumo das figuras:

- **Figura 1:** Casos semanais de dengue notificados e estimados (2020–2025)



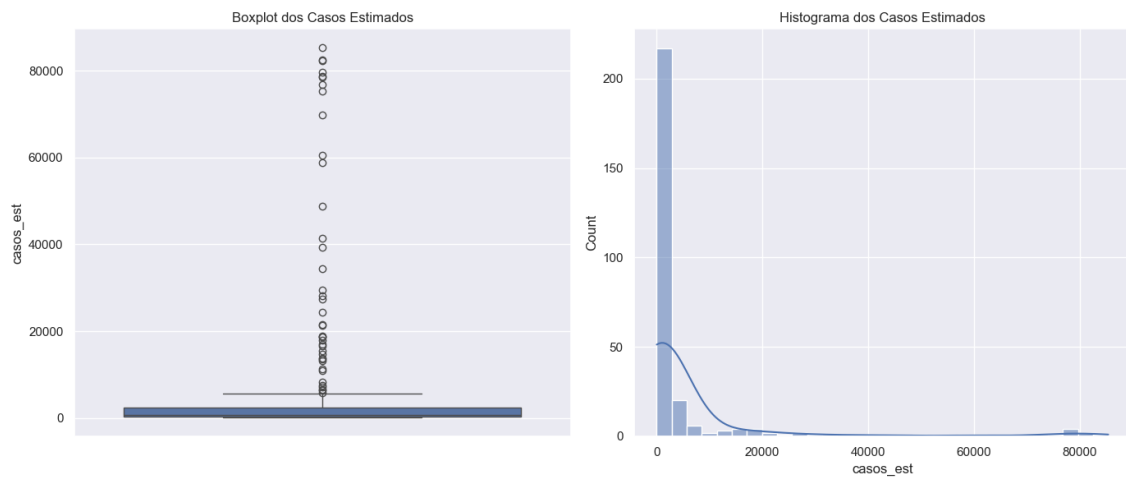
Fonte: Elaborado pela autora (2025).

- **Figura 2:** Casos estimados de dengue por semana (2020–2025)



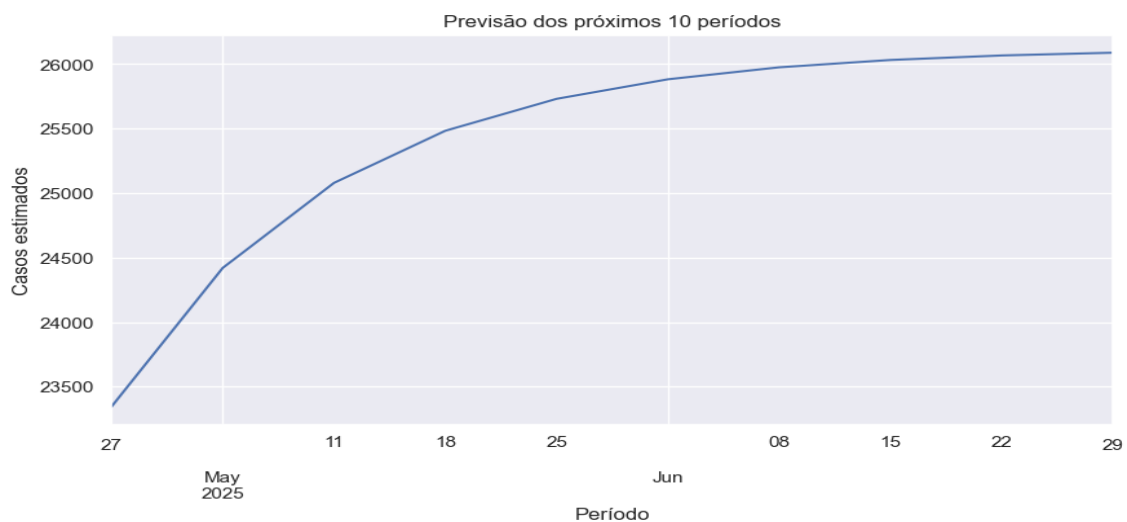
Fonte: Elaborado pela autora (2025).

- **Figura 3:** Boxplot e histograma dos casos estimados



Fonte: Elaborado pela autora (2025).

- **Figura 4:** Previsão dos 10 próximos períodos com ARIMA



Fonte: Elaborado pela autora (2025).

O modelo ARIMA demonstrou robustez e boa interpretação, embora sua natureza univariada limite o uso de fatores externos como clima, mobilidade urbana ou condições socioeconômicas. Estudos futuros podem explorar modelos multivariados ou baseados em redes neurais recorrentes (RNN/LSTM) para capturar padrões mais complexos.

Após a análise exploratória e o pré-processamento da série temporal dos casos estimados de dengue, procedeu-se à modelagem com a técnica ARIMA (AutoRegressive Integrated Moving Average). O código da Figura 5 ilustra o processo de diferenciação da série e ajuste do modelo ARIMA(1,1,1) com a biblioteca statsmodels em Python.

- **Figura 5:** Código de preparação e modelagem ARIMA (1,1,1)

Trecho do código responsável pela preparação dos dados da série temporal e ajuste do modelo ARIMA(1,1,1), utilizado para previsão dos casos estimados de dengue na cidade de São Paulo.

```
# Preparação para modelagem
df_model = df[['data_iniSE', 'casos_est']].copy()
df_model.set_index('data_iniSE', inplace=True)
df_model['casos_diff'] = df_model['casos_est'].diff()
df_model.dropna(inplace=True)

# Ajuste do modelo ARIMA
model = ARIMA(df_model['casos_est'], order=(1,1,1))
modelo_ajustado = model.fit()
print(modelo_ajustado.summary())
```

Fonte: Elaborado pela autora (2025).

O resultado do ajuste é exibido na Figura 6, contendo os parâmetros estimados e métricas como AIC, BIC e p-valores para os coeficientes autoregressivos e de média móvel.

- **Figura 6 – Sumário do modelo ARIMA(1,1,1)**

Resumo estatístico gerado após o ajuste do modelo ARIMA(1,1,1), com destaque para a significância do termo autorregressivo (AR) e as métricas AIC/BIC utilizadas para avaliação de qualidade do modelo.

```
SARIMAX Results
=====
Dep. Variable:      casos_est  No. Observations:      276
Model:              ARIMA(1, 1, 1)  Log Likelihood      -2495.656
Date:              Sat, 03 May 2025  AIC              4997.312
Time:              22:50:28      BIC              5008.163
Sample:            01-12-2020     HQIC             5001.667
                  - 04-20-2025
Covariance Type:    opg
=====
              coef    std err          z      P>|z|      [0.025    0.975]
-----
ar.L1         0.6127     0.024    25.897     0.000     0.566     0.659
ma.L1         0.0084     0.025     0.333     0.739    -0.041     0.058
sigma2        4.531e+06  1.03e+05    44.177     0.000  4.33e+06  4.73e+06
=====
Ljung-Box (L1) (Q):           0.00  Jarque-Bera (JB):           9823.24
Prob(Q):                     1.00  Prob(JB):                0.00
Heteroskedasticity (H):       1377.61  Skew:                   -2.28
Prob(H) (two-sided):          0.00  Kurtosis:                31.92
=====
```

Fonte: Elaborado pela autora (2025).

Apresentação (Link do Youtube)

<https://youtu.be/ADkRVFoD1Ng>

5. Conclusão

Este trabalho teve como objetivo prever a incidência semanal de casos de dengue na cidade de São Paulo, utilizando dados abertos e atualizados provenientes da API pública do InfoDengue. A escolha do tema se justifica pela crescente preocupação com o aumento de surtos da doença e a necessidade de ferramentas que auxiliem na tomada de decisão das autoridades de saúde pública.

A metodologia adotada envolveu a coleta e o pré-processamento de séries temporais, a análise exploratória dos dados e a aplicação do modelo estatístico ARIMA. A escolha do ARIMA baseou-se em sua capacidade de lidar com séries univariadas e em sua robustez estatística. Os dados foram organizados em frequências semanais entre 2020 e 2025, e o processo foi integralmente documentado e disponibilizado de forma reprodutível.

Os resultados obtidos demonstraram que o modelo foi eficaz em capturar tendências e sazonalidades da série temporal. As métricas de erro (MAE = 658,30; RMSE = 2109,72) indicam desempenho satisfatório na previsão dos valores futuros, principalmente em períodos de estabilidade. Contudo, o modelo apresentou limitações na antecipação de variações abruptas, típicas de surtos epidêmicos.

Como contribuição, o projeto oferece uma solução transparente e replicável, com potencial de aplicação real para monitoramento epidemiológico. Suas limitações, por outro lado, apontam para a necessidade de incorporar variáveis externas (como temperatura, precipitação e mobilidade) ou utilizar modelos multivariados mais sofisticados.

Para trabalhos futuros, recomenda-se:

- A utilização de redes neurais recorrentes (LSTM) ou modelos híbridos para capturar não linearidades;
- A ampliação da base de dados para incluir fatores ambientais e sociodemográficos;
- A aplicação da metodologia em outras regiões do país para avaliação da escalabilidade do modelo.

A proposta encontra-se alinhada com os Objetivos de Desenvolvimento Sustentável (ODS), especialmente o ODS 3 (Saúde e Bem-Estar), reforçando o caráter extensionista e social do projeto.

6. Referências bibliográficas

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. *NBR 6023: Informação e documentação: Referências – Elaboração*. Rio de Janeiro: ABNT, 2018.

ARORA, R. et al. Automated skin lesion segmentation using attention-based deep convolutional neural network. *Biomedical Signal Processing and Control*, Elsevier BV, v. 65, p. 102358, mar. 2021. Disponível em: <https://www.researchgate.net/publication/347903879>. Acesso em: 04 maio 2025.

BASTOS, L. S. et al. A modelling approach for correcting reporting delays in disease surveillance data. *Statistics in Medicine*, v. 38, n. 22, p. 4363–4377, 2019. Disponível em: <https://doi.org/10.1002/sim.7992>. Acesso em: 04 maio 2025.

CODECO, C. et al. Infodengue: A nowcasting system for the surveillance of arboviruses in Brazil. *Revue d'Épidémiologie et de Santé Publique*, v. 66, p. S386, 2018. Disponível em: <https://doi.org/10.1016/j.respe.2018.05.408>. Acesso em: 04 maio 2025.

HAMEED, N. et al. An intelligent computer-aided scheme for classifying multiple skin lesions. *Computers*, MDPI AG, v. 8, n. 3, p. 62, 2019. Disponível em: <https://www.mdpi.com/2073-431X/8/3/62>. Acesso em: 04 maio 2025.

IBM. *What are ARIMA models?*. Disponível em: <https://www.ibm.com/topics/arima-model>. Acesso em: 04 maio 2025.

INFO DENGUE. *Tutorial de acesso à API do InfoDengue*. Disponível em: https://info.dengue.mat.br/tutorial_api_python/locale-en. Acesso em: 04 maio 2025.

INFO DENGUE. *Sobre o InfoDengue*. Disponível em: <https://info.dengue.mat.br/>. Acesso em: 28 mar. 2025.

INEP – INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. *Microdados do Censo Escolar*. Disponível em: <https://download.inep.gov.br/microdados>. Acesso em: 28 mar. 2025.

MINISTÉRIO DA SAÚDE (BRASIL). *Dengue – Informações sobre a doença*. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/dengue>. Acesso em: 28 mar. 2025.

OPENDATASUS. *Dados de Saúde*. Disponível em: <https://opendatasus.saude.gov.br/>. Acesso em: 28 mar. 2025.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). *Dengue*. Disponível em: <https://www.who.int/health-topics/dengue>. Acesso em: 28 mar. 2025.

RUSSELL, S.; NORVIG, P. *Inteligência artificial*. 4. ed. Rio de Janeiro: LTC, 2021.
SZELISKI, R. *Computer Vision: Algorithms and Applications*. 2. ed. New York: Springer, 2022.

UNIVERSIDADE PRESBITERIANA MACKENZIE. *Guia do TCC: Orientações gerais para a elaboração do trabalho de conclusão dos cursos de graduação*. São Paulo: UPM, 2021.