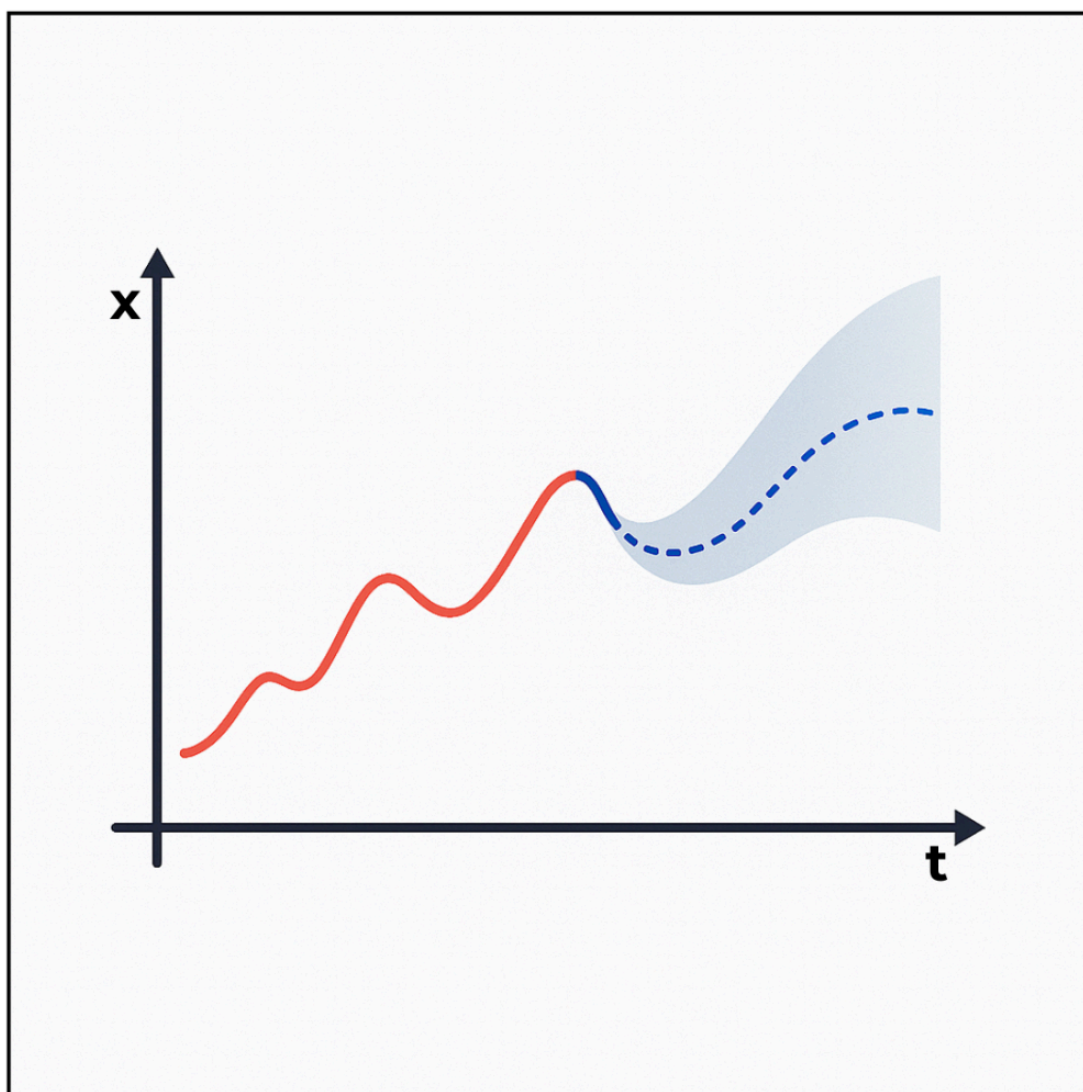


# Forecasting em Séries Temporais

Da Teoria à Aplicação em Dados Epidemiológicos

Davi de Jesus Teixeira



UNIVERSIDADE FEDERAL DE GOIÁS (UFG)  
INSTITUTO DE INFORMÁTICA (INF)

DAVI DE JESUS TEIXEIRA

## **Forecasting em Séries Temporais**

Da Teoria à Aplicação em Dados Epidemiológicos

Goiânia  
2025



UNIVERSIDADE FEDERAL DE GOIÁS  
INSTITUTO DE INFORMÁTICA

## TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

### 1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): DAVI DE JESUS TEIXEIRA

Título do trabalho: Forecasting em Séries Temporais

Da Teoria à Aplicação em Dados Epidemiológicos

### 2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [ X ] SIM [ ] NÃO<sup>1</sup>

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

#### Casos de embargo:

- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro.

**Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.**



Documento assinado eletronicamente por **Davi De Jesus Teixeira, Discente**, em 05/02/2026, às 23:53, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Fernando Marques Federson, Professor do Magistério Superior**, em 13/03/2026, às 11:27, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site [https://sei.ufg.br/sei/controlador\\_externo.php?acao=documento\\_conferir&id\\_orgao\\_acesso\\_externo=0](https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0), informando o código verificador **5956396** e o código CRC **0B11F790**.

---

**Referência:** Processo nº 23070.005489/2026-25

SEI nº 5956396

DAVI DE JESUS TEIXEIRA

**Forecasting em Séries Temporais**  
Da Teoria à Aplicação em Dados Epidemiológicos

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.  
Orientador: Prof. Dr. Fernando Marques Federson

Goiânia  
2025

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UFG.

TEIXEIRA, DAVI DE JESUS  
Forecasting em Séries Temporais [manuscrito]: Da Teoria à Aplicação  
em Dados Epidemiológicos / DAVI DE JESUS TEIXEIRA. - 2025.  
87 f.: 2025

Orientador: Prof. Dr. Fernando Marques Federson  
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de  
Goiás, Instituto de Informática (INF), Inteligência Artificial, Goiânia, 2025.

1. Inteligência Artificial. 2. Séries Temporais. 3. Previsão.

I. Federson, Fernando Marques , orient. II. Título.

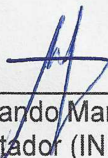
CDU 004

DAVI DE JESUS TEIXEIRA

**Forecasting em Séries Temporais**  
Da Teoria à Aplicação em Dados Epidemiológicos

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.

Data da Aprovação: 09 de dezembro de 2025.



---

Prof. Dr. Fernando Marques Federson  
Orientador (INF-UFG)



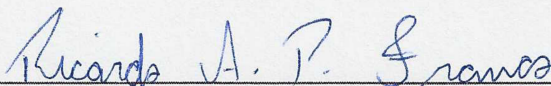
---

Prof. Dr. Aldo André Díaz Salazar  
Coordenador de TCC do BIA (INF-UFG)



---

Prof. Dr. Anderson da Silva Soares  
Coordenador do BIA (INF-UFG)



---

Prof. Dr. Ricardo Augusto Pereira Franco  
(INF-UFG)

DAVI DE JESUS TEIXEIRA

## **Forecasting em Séries Temporais**

Da Teoria à Aplicação em Dados Epidemiológicos

### **RESUMO**

Este Relatório de Conclusão de Curso tem como objetivo reunir os resultados da minha jornada para me tornar um especialista em **Forecasting e Análise de Séries Temporais**. Uma ilustração e sua narrativa descrevem os períodos de trabalho. Os Apêndices contêm os Termos de Aceite de Entrega e os resultados obtidos durante cada período de trabalho.

Palavras-chave: Inteligência artificial; Séries temporais; Previsão.

### **ABSTRACT**

This Course Completion Report aims to bring together the results of my journey to become an expert in **Forecasting and Time Series Analysis**. An illustration and its narrative describe the work periods. The Appendices contain the Delivery Acceptance Terms and the results obtained during each work period.

Keywords: Artificial intelligence; Time series; Forecast.

Goiânia

2025

# Minha Jornada



Davi de Jesus Teixeira

Especialista em: Forecasting e Análise de Séries Temporais

---

## MINHA JORNADA

**Nome:** Davi de Jesus Teixeira

**Especialidade:** Forecasting e Análise de Séries Temporais

### Objetivo deste documento

Durante o processo da disciplina Residência em IA<sup>1</sup>, foram gerados diversos resultados na construção da minha especialização. A cada semana, um conjunto de resultados foi formalizado por um Termo de Aceite de Entrega e avaliado por uma banca, considerando o planejado e o realizado para o período. Este documento tem como objetivo descrever esses resultados obtidos, fazendo referência aos Termos de Aceite de Entrega e seus documentos associados.

### Minha Jornada

Iniciei a Residência buscando clareza sobre qual caminho seguir dentro do amplo universo das Séries Temporais. A decisão não surgiu de imediato; demandou leitura, revisão de conceitos e uma análise cuidadosa das direções que a área tem tomado nos últimos anos. Esse movimento inicial foi fundamental para compreender onde minha pesquisa poderia se posicionar e quais temas ofereciam profundidade suficiente para sustentar uma jornada de forma consistente.

Por isso, durante a **Semana 1**, concentrei-me em posicionar o projeto no cenário das conferências internacionais que serviam como referência para a Residência neste ano. A pesquisa nos principais congressos da área de Inteligência Artificial e Dados (como o **ICAI'25**, **ICDATA'25** e **ACC'25**) permitiu entender com mais precisão como meu tema se conecta às discussões atuais em IA e Ciência de Dados. Esse mapeamento inicial ajudou a definir não apenas o que estudar, mas também como direcionar meus esforços ao longo das **Semanas** seguintes. Na **Semana 2**, dei início à construção da minha base conceitual, que

---

<sup>1</sup> Dez Semanas, entre setembro de 2025 e dezembro de 2025.

serviria como guia para todo o processo. Um dos destaques desse período foi a leitura dos capítulos iniciais de dois livros: *Forecasting: Principles and Practice*, de Hyndman & Athanasopoulos, e *Mastering Modern Time Series Forecasting*, de Valeriy Manokhin. A partir dessas leituras, elaborei uma linha do tempo com marcos chave da área e o impacto de cada um deles na forma como hoje entendemos a previsão de Séries Temporais. A leitura de artigos de revisão, como Casolaro et al. (2023) e Kong et al. (2025), também se mostrou bastante útil, pois permitiu identificar as principais tendências do campo, as tecnologias mais utilizadas atualmente e diversos casos de aplicação relevantes. Assim, ao final das **Semanas 1 e 2**, minha jornada já não era apenas a escolha de um projeto. Eu havia construído um escopo claro, contextualizado nas conferências de interesse, além de uma visão histórica estruturada da área e um conjunto organizado de conceitos e tendências que servirão de base para os próximos passos, quando a ênfase passará gradualmente da fundamentação para a experimentação prática em Séries Temporais. Toda a revisão Histórica e das principais tendências de forecasting em Séries Temporais pode ser encontrada no **Apêndice 1**.

Na **Semana 3**, avancei para a formalização do tema e do meu objeto de estudo — as Séries Temporais. Formalmente, uma série temporal consiste em uma sequência de observações  $\{y_1, y_2, \dots, y_t\}$  indexadas cronologicamente, onde cada observação  $y_t$  corresponde a um valor medido no instante  $t$ . A característica distintiva é a dependência temporal: observações sucessivas não são independentes, mas frequentemente correlacionadas, o que torna a ordem dos dados fundamental para a análise e modelagem. Diferentemente de dados tabulares tradicionais (onde linhas podem ser embaralhadas sem perda de informação), as séries temporais carregam estrutura temporal que deve ser preservada. Busquei organizar de forma mais precisa o que caracteriza uma série temporal e como essa estrutura se diferencia de outras estruturas de série, como séries matemáticas, por exemplo. O meu principal achado durante essa reflexão foi a compreensão das propriedades fundamentais (decorrentes da variável temporal) que caracterizam uma série temporal, como tendência, sazonalidade, ruído, ciclos, autocorrelação, entre outras. Dessa forma, deixei de apenas introduzir e pesquisar sobre o assunto e passei a estruturar, de fato, o campo conceitual que sustentará os próximos passos, garantindo uma base sólida antes

de avançar para a prática em forecasting. A formalização do meu tema e do objeto de estudo pode ser encontrada no **Apêndice 2**.

Nas **Semanas 4 e 5**, aprofundi meus estudos com foco em desenvolver uma visão mais crítica sobre os métodos de forecasting. Retomando minha coleção de referências coletadas até então, encontrei um trabalho que questiona a divisão tradicional entre “métodos estatísticos” e “métodos de machine learning”. Os autores propõem uma classificação mais rica, baseada em dimensões objetivas (como linearidade, complexidade computacional e previsões locais ou globais), além de dimensões subjetivas (como interpretabilidade, garantias estatísticas e a distinção entre abordagens data-driven e model-driven). Essa classificação foi fundamental para mim, pois por meio dela, ampliei minha compreensão sobre como escolher corretamente abordagens diferentes para cada contexto de previsão. Observações detalhadas sobre esse artigo podem ser encontradas no **Apêndice 3**.

Paralelamente, avancei também no estudo da previsibilidade, propriedade que expressa o limite teórico de quão bem valores futuros podem ser estimados. Essa análise é fundamental para determinar se uma série contém informações suficientes para ser modelada. Então, buscando entender melhor como quantificar essa propriedade, estudei métricas como Expoente de Hurst, ACF (Autocorrelation Function), PACF (Partial Autocorrelation Function), Variance Ratio e medidas baseadas em entropia.

Na **Semana 6**, finalizei o estudo de previsibilidade lendo *Enhancing Predictability Assessment: An Overview and Analysis of Predictability Measures for Time Series and Network Links*. O artigo consolidou dois conceitos centrais: previsibilidade intrínseca (limite teórico imposto pela estrutura da série) e previsibilidade realizada (o que os modelos efetivamente alcançam). Essa leitura mostrou que o desempenho observável (ou previsibilidade realizada) está condicionado à informação disponível na série, reforçando a importância de se medir a previsibilidade da série, mas ainda mais importante, mostrando a importância de avaliar corretamente os modelos. O **Apêndice 4** reúne parte do trabalho das

semanas 5 e 6 onde deixei minhas principais notas sobre o tópico de previsibilidade em Séries Temporais.

Para complementar esse estudo, ainda na **Semana 6** aprofundi-me nas métricas de avaliação em forecasting, organizando-as em determinísticas (MAE, RMSE, MAPE, sMAPE, MASE etc.) e probabilísticas (CRPS, Pinball Loss, Winkler Score, MSIS). O objetivo foi entender o que cada métrica mede, suas limitações e quando usar, garantindo que a escolha esteja alinhada ao objetivo analítico e às características da série — algo crucial porque problemas de Séries Temporais diferem de classificação/regressão tradicionais. Os materiais contendo o estudo e as minhas considerações sobre as principais métricas de avaliação de modelagens de Séries Temporais estão no **Apêndice 5**.

Na **Semana 7**, estudei pipelines de pré-processamento a partir do *Survey Time-series Data Preprocessing: A Survey and an Empirical Analysis*. Reforcei que, em Séries Temporais, a qualidade do modelo é limitada pela qualidade dos dados e pelo respeito à ordem temporal — o que exige cuidados que não aparecem em tarefas de classificação/regressão “estáticas”. Para Séries Temporais, o pré-processamento precisa lidar tanto com problemas comuns a outros pipelines de IA (amostragem irregular, dados faltantes, outliers) quanto com desafios específicos (não-estacionariedade, mudanças de regime, sazonalidades múltiplas), exigindo métodos adaptados à natureza temporal dos dados. Além das etapas de limpeza, dediquei atenção especial às técnicas de normalização e transformação. Diferentemente de dados tabulares tradicionais, a normalização em Séries Temporais deve preservar a estrutura temporal e considerar se será aplicada globalmente (em toda a série) ou por janelas deslizantes — escolha que impacta diretamente a capacidade do modelo de capturar padrões de longo prazo. Estudei também transformações para estabilização de variância e linearização de tendências, como transformações logarítmicas, Box-Cox e diferenciação. A diferenciação, em particular, revelou-se fundamental para tornar séries não-estacionárias em estacionárias, requisito essencial para diversos modelos estatísticos clássicos como ARIMA. O principal aprendizado é que a escolha entre essas transformações não é arbitrária: depende das características da série e dos pressupostos do modelo que será utilizado. Explicações sobre os processos de

normalização e transformações em Séries Temporais podem ser encontrados no **Apêndice 6**.

Para fechar a linha de raciocínio sobre modelagem, durante a **Semana 8**, consolidei um *pipeline* completo de pré-modelagem que vai além da EDA e abrange pré-processamentos, validações estatísticas e critérios de escolha de modelos (estatísticos, ML e DL). O *pipeline* ficou estruturado em etapas sequenciais: (1) entendimento inicial da série, (2) análise de qualidade dos dados, (3) Análise Exploratória Visual dos dados, (4) transformações e pré-processamento, (5) avaliação/tratamento da estacionariedade, (6) análise de autocorrelação (ACF/PACF e testes), (7) relações com variáveis exógenas, (8) detecção de anomalias e rupturas, (9) verificação de ruído branco, (10) quantificação da previsibilidade intrínseca, (11) modelagem e, finalmente, (12) quantificação da previsibilidade realizada (avaliação da modelagem). Além de mapear caminhos alternativos conforme as propriedades da série e a técnica de modelagem escolhida, registrei boas práticas para reforçar a consistência das análises. uma visão mais aprofundada sobre o *pipeline* está descrita no **Apêndice 7**. No final dessa **Semana**, concentrei-me na leitura do artigo *Assessing dengue forecasting methods: a comparative study of statistical models and machine learning techniques in Rio de Janeiro, Brazil*. O trabalho apresenta uma aplicação prática de forecasting de dengue, exatamente o tipo de problema que escolhi para validar o pipeline proposto. A partir dessa leitura, pude compreender padrões metodológicos da área, como escolha de modelos, janelas de avaliação e uso de variáveis exógenas. Decidi então aplicar meu pipeline ao estado de Goiás, utilizando dados extraídos da API oficial do **InfoDengue** (<https://info.dengue.mat.br>) — a mesma fonte do artigo de referência — para possibilitar comparações consistentes entre as implementações.

Para concluir a minha Jornada, durante as **Semanas 9 e 10**, concentrei meus esforços em aplicar todo o conhecimento adquirido em um caso prático de forecasting. Então, com os dados que havia extraído do **InfoDengue** em mãos, foquei em compreender como foi feita a modelagem do problema e quais foram os critérios de comparação entre métodos. Os principais pontos metodológicos observados incluíram: avaliação de horizontes de previsão de 1, 2, 3, 4, 8 e 12 semanas de dados, comparação entre modelos estatísticos

e modelos de machine learning e deep learning, além de experimentos com e sem variáveis climáticas (temperatura e umidade). Aplicando essa abordagem ao contexto de Goiás, comecei com a análise exploratória da série. Após verificar a integridade das datas e interpolar uma semana ausente via spline, a série revelou padrão sazonal claro, com picos concentrados no início de cada ano, especialmente em 2010, 2013, 2016, 2022 e 2024. Embora outliers tenham sido identificados por critérios estatísticos (Z-score e IQR), optei por mantê-los, pois representam epidemias reais e não erros de medição. A decomposição mostrou tendência cíclica e sazonalidade proporcional ao nível de casos, com surtos mais intensos em anos de maior incidência. Os mapas de calor e médias semanais confirmaram que os surtos ocorrem predominantemente entre fevereiro e maio (período chuvoso), enquanto julho a novembro marcam o intervalo seco de baixa incidência. Para estabilizar a variabilidade, testei transformações logarítmicas, raiz quadrada e Box-Cox, sendo esta última a que apresentou melhor simetria. Após normalização via MinMaxScaler, os testes ADF e KPSS confirmaram estacionariedade da série transformada. O expoente de Hurst ( $H=0.701$ ) indicou persistência temporal — períodos de alta tendem a se manter — um bom indicativo da previsibilidade da série. Iniciei a modelagem univariada com XGBoost, criando 52 features de lags semanais, features de médias e desvios móveis desses lags, além de features sazonais (mês, semana, ano) para capturar padrões anuais e semianuais. O modelo foi treinado em dados de 2017–2021, validado em 2022–2023 e testado em 2024. Na validação, alcançou  $MAE=0.04$ ,  $RMSE=0.07$  e  $R^2=0.84$ . No teste, manteve desempenho sólido ( $MAE=0.053$ ,  $RMSE=0.063$ ,  $SMAPE=6.8\%$ ,  $R^2=0.85$ ), com boa captura das nuances do pico epidêmico de 2024. A análise residual confirmou normalidade e ausência de forte autocorrelação. Para avaliar a robustez preditiva, assim como no artigo, apliquei previsão direta multi-horizonte, treinando modelos independentes para 1, 4, 8 e 12 semanas à frente. Como esperado em séries epidêmicas, houve degradação conforme o horizonte aumentou: o SMAPE passou de 11.5% (1 semana) para 23.8% (12 semanas) e o  $R^2$  caiu até valores negativos em horizontes longos. Em seguida, implementei o método Walk-Forward, que re-treina o modelo a cada semana incorporando novas observações. Essa estratégia alcançou  $MAE=0.032$ ,  $RMSE=0.040$  e  $R^2=0.94$ , com resíduos centrados e sem padrão, comprovando estabilidade temporal. Por fim, estendi o experimento para Walk-Forward

---

Multi-Horizonte ( $h = 1, 4, 8, 12$ ), obtendo resultados que mostraram excelente equilíbrio: 1 semana (SMAPE=6.5%,  $R^2=0.88$ ), 4 semanas (SMAPE=10.8%,  $R^2=0.66$ ), 8 semanas (SMAPE=11.8%,  $R^2=0.59$ ) e 12 semanas (SMAPE=11.5%,  $R^2=0.57$ ). O erro aumentou de forma suave e controlada, indicando previsões coerentes em diferentes janelas — um desempenho relativamente bom para uma abordagem puramente univariada. O XGBoost demonstrou eficiência para previsões de curto prazo e desempenho aceitável em janelas maiores. Mais detalhes sobre o pipeline implementado e sobre os resultados estão descritos no **Apêndice 8**.

Finalizadas as **Semanas**, gostaria de refletir sobre toda essa trajetória. A Residência foi uma experiência extremamente enriquecedora, que me desafiou a sair da zona de conforto e mergulhar em um tema sobre o qual não tinha experiência prévia, e, com certeza, fechar esse ciclo desenvolvendo um projeto que integra todo o conhecimento adquirido durante esse processo foi especialmente gratificante. Mais do que o conteúdo técnico, aprendi a ser mais resiliente, perseverante e determinado a galgar continuamente novos conhecimentos — aprendizados que levarei tanto para minha carreira profissional quanto para minha vida pessoal.

Para finalizar, reconheço o privilégio de ter colegas e professores tão especiais que sempre me orientaram com muita dedicação e generosidade. Atribuo a vocês grande parte do meu sucesso nesta jornada acadêmica e levarei sempre comigo cada momento ao lado de vocês. Estendo meus agradecimentos à minha família — pai, mãe, meu irmão Pedro Ernesto e à minha irmã Paula — e à minha namorada Esther, que, em todos os momentos, desde a escolha do meu curso até a minha formação, sempre estiveram do meu lado, me apoiando durante as crises, me incentivando nas minhas batalhas e comemorando comigo até as minhas pequenas vitórias.

## APÊNDICE 1

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“Gate”) de aprovação:** 11 de set. de 2025

**Participantes da Entrega** [matriculados em Residência em IA]:

DAVI DE JESUS TEIXEIRA

**Entrega:** [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Durante esse Stage, iniciei meus estudos para me tornar especialista no tema: **Técnicas Avançadas de Forecasting e Quantificação de Incerteza (Conformal Prediction) em Séries Temporais**. Para isso, comecei revisitando a **história do forecasting** e alguns conceitos fundamentais.

Li os capítulos iniciais de:

- **Hyndman & Athanasopoulos – Forecasting: Principles and Practice (FPP3)**
- **Valeriy Manokhin – Mastering Modern Time Series Forecasting**

Com base nessas leituras, organizei um resumo histórico em `HISTORIA TSF.md`, criando uma linha do tempo com marcos relevantes e o impacto desses marcos para a área. Também mapeei e pedi para o GPT ranquear conforme importância geral, conceitos-chave que apareceram conforme a minha leitura progredia para me aprofundar/revisar nas próximas semanas, antes de ir para temas mais complexos em `CONCEITOS PRINCIPAIS.md`. Só para citar, alguns desses conceito são:

- Cadeias de Markov, Modelos Autoregressivos, Decomposição de Wold, ARIMA, Autocorrelação (ACF/PACF), Não Estacionariedade, Médias móveis, Etc...

Além disso, li dois artigos de revisão recentes que tratam das principais tendências em TSF:

- **Casolaro et al. (2023)** – *Deep Learning for Time Series Forecasting: Advances and Open Problems* - <https://www.mdpi.com/2078-2489/14/11/598>
- **Kong et al. (2025)** – *Deep learning for time series forecasting: a survey* - <https://arxiv.org/abs/2503.10198>

A partir deles, elaborei minhas considerações e anotações sobre os avanços da área e o que há de mais

“quente” atualmente: [☰ REVISÃO PRINCIPAIS TENDÊNCIAS EM TSF.md](#) .

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima Semana, pretendo continuar com os estudos dos conceitos que mencionei acima a partir das [☰ REFERENCIAS](#) que separei (e continuo atualizando), além de começar a explorar melhor a área de **Conformal Prediction** lendo o artigo: **A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification**, de **Anastasios N. Angelopoulos** e **Stephen Bates** a fim de entender como a estrutura de Conformal Prediction pode ajudar a fornecer intervalos de previsão confiáveis e estatisticamente válidos para um modelo de DL.

**Observação:** [caso precise fazer alguma observação, de qualquer “natureza”]

Como eu faço as minhas anotações em um programa chamado Obsidian.md (por isso do formato .md dos arquivos de resumos), a renderização desses arquivos aqui para o Google Docs pode não ser ótima (principalmente das fórmulas matemáticas que eu escrevo em LaTeX). Grato pela compreensão.

**ACEITE DA ENTREGA:**

**CEDRIC LUIZ DE CARVALHO:** [Go! ▾](#)

## SOBRE O APÊNDICE 1:

Antes de apresentar os apêndices, é importante esclarecer que, embora tenha identificado interesse em explorar Conformal Prediction nesse período inicial, não consegui aprofundar nesse tema ao longo das semanas, pois outros conceitos fundamentais mostraram-se prioritários para a construção completa do pipeline de análise. A decisão de postergar esse tema para estudos após o TCC permitiu estabelecer uma base conceitual mais sólida sobre Séries Temporais, Forecasting e suas propriedades.

Partindo de nenhuma experiência prévia, durante as Semanas 1 e 2, concentrei-me em pesquisar o contexto geral da área de Forecasting em Séries Temporais. Considerei fundamental compreender a história e as principais tendências do campo antes de avançar para tópicos mais específicos.

A seguir, apresento os materiais e estudos realizados nessas semanas:

- Sobre a História do Forecasting:
  - No documento abaixo, apresento o meu estudo inicial sobre a Área de Forecasting em Séries Temporais. Inicialmente, trago uma definição sobre o conceito de Forecasting e um pouco do histórico da área apresentado de forma cronológica.
    - <https://drive.google.com/file/d/1oM91CL6YVvQI9sNvXASIVhgDzpphm-P/view?usp=sharing> (Acessado em 03/12/2025)
- Revisão sobre as principais tendências:
  - Nesse documento, apresento pontos relevantes sobre as tendências em forecasting que extraí dos artigos de Casolaro et al. (2023) e Kong et al. (2025).
  - Alguns dos pontos anotados abordam os modelos usados, formas de representação dos dados e extração de características, aplicações, etc..
  - Essa revisão se mostrou relevante pois serviu de motivação para as minhas pesquisas durante as próximas semanas.
    - <https://drive.google.com/file/d/11Wx8Zyuul-x3XCqfKqrnsYSQL1CkFKZN/view?usp=sharing> (Acessado em 03/12/2025)
  - Artigos estudados:
    - Casolaro et al. (2023). Deep Learning for Time Series Forecasting: Advances and Open Problems. (<https://www.mdpi.com/2548358>) (Acessado em 03/12/2025)
    - Kong et al. (2025). Deep learning for time series forecasting: a survey. (<https://doi.org/10.1007/s13042-025-02560-w>) (Acessado em 03/12/2025)

## APÊNDICE 2

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“Gate”) de aprovação:** 2 de set. de 2025

**Participantes da Entrega** [matriculados em Residência em IA]:

DAVI DE JESUS TEIXEIRA

**Entrega:** [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Durante a última Semana, prossegui com as seguintes atividades dentro do tema: **Forecasting em Séries Temporais com Quantificação de Incerteza.**

### Formalização do Tema

Criei um documento a fim de refletir sobre o meu tema a fim de tentar organizar melhor as ideias na minha cabeça. [Formalização do tema](#)

Resumindo o conteúdo do documento:

#### Série matemática

- Sequência ordenada, definida por regras formais e estáticas.
- Simples e previsível na interpretação.

#### Série temporal

- Sequência de dados empíricos coletados ao longo do tempo, normalmente em intervalos regulares.
- Mais complexa por fatores como:
  - **Autocorrelação:** valores atuais dependem de instantes passados.
  - **Ruídos externos:** fatores aleatórios e fora de controle.
  - **Não-estacionariedade:** tendências e sazonalidades mudam ao longo do tempo.
- Diferença central: o tempo introduz **dependências e incertezas.**

#### Onde entra a Quantificação de Incerteza

- Todo modelo de previsão carrega incerteza.
- Medir incerteza vai além de prever um valor único → envolve intervalos de confiança.

- Importância:
  - Finanças, políticas públicas, medicina → decisões **não podem depender de um único valor**.
  - É preciso considerar **riscos e variabilidade**.

### Conformal Prediction (CP)

- Método de quantificação de incerteza.
- Características principais:
  - Cria intervalos de confiança **ajustados ao comportamento dos dados**.
  - Garante uma **taxa de acerto definida** (ex.: 90% ou 95%).
  - Entrega não só o valor previsto, mas também a **faixa provável**.
- Vantagens: previsões mais **seguras, transparentes e confiáveis**.

### Integração CP + Forecasting em Séries Temporais

- Une **capacidade preditiva + quantificação de incerteza**.
- Oferece:
  - Intervalos claros de confiança.
  - Apoio robusto à tomada de decisão.

### Pesquisa: Introdução ao CP

Segui então para o **Research Rabbit** a fim de encontrar um ponto de partida amigável para aprender os conceitos iniciais de CP.

Para isso, criei uma coleção a partir de artigos sugeridos após o input dos trabalhos dos inventores desse método, **Vladimir Vovk** e **Alexander Gammerman**, apenas para formar um grafo de relacionamentos entre referências bem citadas e que mostram a evolução da técnica.

Olhando para o grafo, a partir da identificação de fortes relacionamentos, decidi que um bom ponto de partida pode ser o artigo:

#### **A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification** *Autores: Anastasios N. Angelopoulos e Stephen Bates*

Esse artigo se mostrou adequado porque cita os principais trabalhos da área e, ao fazer um screening inicial, percebi que de fato apresenta um bom overview do tema, e li o início para entender para que e por que usar essa técnica.

### O que estudei no artigo:

Segue link com algumas ponderações que tive ao ler o início do artigo e fazer outras pesquisas complementares

[A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification](#)

## Ferramentas para aplicar o estudo

Existem diversos frameworks consolidados para aplicar os métodos que quero estudar. Alguns que levantei e pretendo começar a explorar nas próximas semanas:

- **Darts**
- **XGBoost e CatBoost**
- **PyTorch Forecasting**
- **Nixtla, Hugging Face** → implementações de modelos SOTA como NHITS, NBEATS, TimeMixers, TimeGPT etc.
- **MAPIE** → Conformal Prediction

## Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima Semana, pretendo dar continuidade ao estudo dos conceitos fundamentais de **forecasting em séries temporais**, unindo a revisão da teoria com a prática em código usando os frameworks que levantei.

Esse estudo será baseado principalmente nos livros listados em [REFERÊNCIAS](#), por apresentarem abordagens mais didáticas do que artigos, o que facilitará a consolidação dos tópicos que preciso visitar e aprender.

Além disso, retornarei ao Research Rabbit e ao meu DOC de referências para organizar melhor a coleção de artigos focados em modelos e métodos para forecasting em séries temporais, de modo a explorar melhor soluções SOTA da área.

## Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

---

## ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: [Go!](#)

## SOBRE O APÊNDICE 2:

Durante a Semana 3, após refletir mais profundamente sobre o meu tema, decidi reorganizar as ideias sobre Forecasting, consolidando o conhecimento adquirido durante as Semanas 1 e 2. Esse processo me permitiu compreender melhor as propriedades que definem essencialmente o que é uma Série Temporal.

Busquei então explicar o que caracteriza uma Série Temporal a partir da seguinte comparação entre séries matemáticas e séries temporais:

**Séries Matemáticas:** Na matemática, uma série é, de forma simples, uma sequência ordenada definida por regras formais e estáticas, o que garante simplicidade e previsibilidade na interpretação.

**Séries Temporais:** São sequências de dados empíricos coletados ao longo do tempo, normalmente em intervalos regulares. Sua análise é mais complexa devido a alguns fatores:

- **Autocorrelação:** valores em um instante dependem de instantes anteriores, nem sempre de forma óbvia ou uniforme.
- **Ruídos externos:** variações aleatórias e fatores fora do controle dificultam a separação entre padrões genuínos e flutuações casuais.
- **Não-estacionariedade:** tendências e sazonalidades podem mudar ao longo do tempo, exigindo tratamentos específicos.

Assim, ao contrário das séries matemáticas, nas séries temporais o tempo introduz dependências e incertezas que tornam a análise e a previsão consideravelmente mais desafiadoras.

Diante dessa compreensão, aprofundei-me em propriedades fundamentais como Estacionariedade e Autocorrelação. Esse estudo gerou as seguintes anotações e reflexões:

- **Anotações sobre Autocorrelação:**
  - <https://drive.google.com/file/d/1oO-MJDOxMd7ZV3TP2fMXo2NEUmrCRlj/view?usp=sharing> (Acessado em 03/12/2025)
- **Anotações sobre Estacionariedade:**
  - <https://drive.google.com/file/d/1hWObcwliQvdJkr1AYQaRw6CPjD2C-z81/view?usp=sharing> (Acessado em 03/12/2025)

Essa semana foi fundamental para consolidar conceitos essenciais ao desenvolvimento de análises rigorosas em Séries Temporais.

## APÊNDICE 3

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“Gate”) de aprovação:** 24 de set. de 2025

**Participantes da Entrega** [matriculados em Residência em IA]:

DAVI DE JESUS TEIXEIRA

**Entrega:** [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Durante esta Semana, avancei nas atividades relacionadas ao tema **“Forecasting em Séries Temporais com Quantificação de Incerteza”**.

Conforme planejado anteriormente, segui a mesma estratégia adotada nos meus estudos sobre Quantificação de Incerteza: revisitei minha coleção de referências sobre Forecasting no Research Rabbit, desta vez com filtros mais rigorosos e uma leitura mais crítica do que ia adicionando à coleção.

Nesse processo, encontrei um artigo que discute a tendência crescente na literatura em classificar métodos de previsão apenas em duas categorias: **“estatísticos”** e **“machine learning”**. Os autores criticam bastante essa abordagem reducionista, destacando que ela é enganosa e incapaz de representar a real diversidade de técnicas disponíveis.

Como alternativa, o artigo propõe que a análise dos métodos seja realizada a partir de múltiplas dimensões, tanto **objetivas** (de natureza matemática e computacional) quanto **subjetivas** (metodológicas e culturais). Essa abordagem permite uma avaliação mais abrangente e fiel sobre como os métodos de forecasting realmente se comportam e em quais contextos oferecem melhor desempenho.

As principais subdivisões propostas são:

**Dimensões Objetivas:** dizem respeito às propriedades matemáticas e computacionais dos modelos.

- Métodos Globais vs. Locais
- Previsões Probabilísticas vs. de Ponto
- Complexidade Computacional
- Linearidade e Convexidade

**Dimensões Subjetivas:** refletem aspectos metodológicos e culturais que orientam a concepção e a aplicação dos métodos.

- Data-driven vs. Model-driven
- Modelos de Conjunto (Ensemble) vs. Modelos Únicos
- Discriminativos vs. Generativos
- Garantias Estatísticas
- Explicativos/Interpretáveis vs. Preditivos

Documentei tudo que o artigo discute em:

[RESUMO ARTIGO CRITERIA CLASSIFYING FORECASTING MODELS.md](#)

Segundo os autores, essas dimensões oferecem uma estrutura de análise mais precisa e informativa, permitindo comparar métodos de forma criteriosa e evitar classificações superficiais. E com essa classificação, podemos aumentar o leque de possibilidades para selecionar abordagens de forecasting mais adequadas a cada contexto, equilibrando fatores como acurácia, custo, interpretabilidade e aplicabilidade.

Com base nisso, irei criar uma tabela com essas subdivisões para ir incrementando com os algoritmos/abordagens que serão estudados e testados nas próximas semanas a fim de estabelecer um comparativo sólido entre as técnicas que explorarei durante a residência, e criar uma “toolbox” de técnicas que conheço e que entendo o escopo de aplicação.

Além disso, continuei com a leitura do livro de forecasting e do artigo *The M4 Competition: 100,000 time series and 61 forecasting methods*, mas não consegui progredir muito.

### Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima semana irei prosseguir com o início do preenchimento da minha tabela de métodos, aplicando os conceitos que eu estudar nos datasets abaixo.

- **AirPassengers**
- **Wineind (Australian Wine Sales)**
- **Electricity Monthly (EIA)**

Pretendo continuar com o estudo em forecasting pelo livro, mas irei retornar ao artigo *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification* (que trouxe semana passada) para retomar os estudos também nessa frente.

### Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

---

## ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: [Go!](#)

## SOBRE O APÊNDICE 3:

Neste apêndice, trato de apresentar em parte o conteúdo da Semana 4 do processo. Durante essa semana, foi realizada a leitura do artigo de Januschowski, T. et al. *Criteria for classifying forecasting methods*. *International Journal of Forecasting*, 2020. Disponível em: <https://doi.org/10.1016/j.ijforecast.2019.05.008>. (Acessado em 03/12/2025)

O principal ponto do artigo é questionar a divisão tradicional e simplista entre "métodos estatísticos" e "métodos de machine learning" no contexto de previsão de séries temporais. Os autores argumentam que essa categorização binária é enganosa e superficial, frequentemente reduzindo "métodos estatísticos" a variantes de suavização exponencial e ARIMA, enquanto "métodos de ML" se tornam sinônimos de redes neurais e random forests. Em vez dessa dicotomia, o artigo propõe um conjunto de dimensões alternativas — tanto objetivas quanto subjetivas — para classificar os métodos de forecasting de forma mais rigorosa e útil.

Mais detalhes sobre essa classificação proposta pelo artigo, podem ser encontrados nas minhas anotações em:

- <https://drive.google.com/file/d/1o4odgGenUhoADs3j2iKY8Ya4vG7YigDN/view?usp=sharing> (Acessado em 03/12/2025)

## APÊNDICE 4

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“Gate”) de aprovação:** 1 de out. de 2025

**Participantes da Entrega** [matriculados em Residência em IA]:




DAVI DE JESUS TEIXEIRA

**Entrega:** [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]


Continuando com os estudos sobre o meu tema "**Forecasting em Séries Temporais com Quantificação de Incerteza**", durante a Semana, prossegui com o seguinte plano:

Leitura do Capítulo 2 (88 páginas) do livro que estou baseando meus estudos "**Mastering Modern Time Series Forecasting**". Centrei meus estudos na parte de metrificar a previsibilidade em séries temporais (métricas como Expoente de Hurst, interpretação da ACF e PACF, Variance Ratio, Taxa de Entropia, ApEn, SampEn, etc.). Quando necessário (aparição de um conceito muito importante que bloqueava meu avanço), afunilei em um estudo mais específico do conceito.

O estudo resultou em vários resumos e “resuminhos”, alguns dos mais relevantes para a semana são:

-  SOBRE PROBLEMAS DE PREVISÃO.md
-  EXPOENTE DE HURST.md
-  PREVISIBILIDADE EM SÉRIES TEMPORAIS.md

Vale ressaltar que utilizo a plataforma de escrita [Obsidian.md](#). Essa plataforma permite que eu crie links entre as minhas anotações e esses links são organizados em um grafo que relaciona as minhas anotações, como no exemplo da foto abaixo:

 GRAFO OBSIDIAN.png

Essa organização das minhas notas me deu uma idéia interessante.

Por fim, decidi ser mais cauteloso em relação ao tópico de quantificação de Incerteza e adiei os estudos dessa frente no meu cronograma, estudando mais a fundo os conceitos principais de séries acho que ainda tenho um caminho básico mais importante a trilhar agora.

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

1. Ler o artigo: **BEZBOCHINA. Enhancing Predictability Assessment: An Overview and Analysis of Predictability Measures for Time Series and Network Links.**
2. Ler o próximo capítulo do livro que aborda de forma bastante completa as **métricas principais de forecasting.**
3. Entrar a fundo no **ARIMA** e **Exponential Smoothing** e suas variações. (com entrar a fundo quero dizer finalmente aplicar os algoritmos, utilizando as métricas corretas, avaliando a previsibilidade de um série, e seguindo outras boas práticas adquiridas com meus estudos)
4. Classificar os modelos e suas variações de acordo com a classificação de modelos de forecasting que apresentei Semana passada na tabela:  
**CLASSIFICAÇÃO DE TÉCNICAS PARA FORECASTING**
5. (???) Tentar organizar de alguma forma o meu grafo de materiais em um mapa mental / trilha do que estou estudando (???)

**Observação:** [caso precise fazer alguma observação, de qualquer “natureza”]

---

**ACEITE DA ENTREGA:**

**CEDRIC LUIZ DE CARVALHO:** Go! ▾

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“Gate”) de aprovação:** 8 de out. de 2025

**Participantes da Entrega** [matriculados em Residência em IA]:

DAVI DE JESUS TEIXEIRA

**Entrega:** [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Durante esta Semana, avancei nos meus estudos sobre **Forecasting em Séries Temporais** focando em compreender mais sobre **métricas de avaliação em forecasting**.

Comecei lendo e resumindo o artigo **Enhancing Predictability Assessment: An Overview and Analysis of Predictability Measures for Time Series and Network Link**, publicado na revista *Entropy* (2023).

O estudo propõe uma distinção conceitual entre:

- **Previsibilidade Realizada:** o que os modelos efetivamente alcançam (métricas como MAE, RMSE, MAPE,  $R^2$ );
- **Previsibilidade Intrínseca:** o limite teórico de previsibilidade dos dados, medido por entropia e complexidade

A maior contribuição do artigo é a **validação da hipótese de que a previsibilidade intrínseca limita a previsibilidade realizada, independentemente do modelo usado**. Reforçando o papel das **medidas de complexidade como ferramentas diagnósticas pré-modelagem** e permitindo avaliar se vale a pena investir esforço computacional em uma série ou se ela é inerentemente imprevisível.

Depois dessa etapa, entrei no estudo das métricas de avaliação em forecasting, dividindo-as em duas grandes frentes: **métricas determinísticas** e **métricas probabilísticas**.

Na primeira frente, aprofundi-me nas métricas determinísticas, que avaliam diretamente o erro entre o valor previsto e o observado. Estudei as principais métricas utilizadas, como **MAE, MSE, RMSE, ME, CFE, TS, MAPE, WAPE, sMAPE, MAAPE, MASE, RMSSE e FVA**. O foco foi compreender suas diferenças, aplicações e limitações práticas.

Dediquei atenção especial ao **MAPE**, uma das métricas mais populares em contextos de negócios (Sales & Operations Planning), analisando suas inconsistências matemáticas — como distorções em séries com valores próximos de zero e falta de simetria entre erros positivos e negativos. A leitura do artigo [LINK](#) ajudou a compreender por que essa métrica deve ser substituída por alternativas mais robustas, como

### sMAPE, MAAPE e WAPE.

Na segunda frente, estudei as métricas probabilísticas, **voltadas para modelos que fazem previsões em forma de distribuições ou intervalos de confiança.**

Enquanto as métricas determinísticas respondem “**o quanto a previsão está errada**”, as probabilísticas respondem “**o quão bem o modelo expressa sua incerteza**”.

Entre as principais métricas analisadas estão **CRPS (Continuous Ranked Probability Score), Pinball Loss, Winkler Score, MSIS, ICP, PINAW, Calibration Score e U de Theil.** Cada uma mede diferentes aspectos da qualidade das previsões, como a precisão dos intervalos, a cobertura e a calibração das incertezas.

Por fim, organizei tabelas de referência (dentro de cada markdown de estudo) sintetizando as propriedades, vantagens e contextos ideais de uso de cada métrica estudada, o que permitiu entender como escolher a métrica mais adequada ao objetivo analítico.

LINKS:

[METRICAS DETERMINISTICAS EM FORECASTING.md](#)

[METRICAS PROBABILISTICAS EM FORECASTING.md](#)

[RESUMO ARTIGO Enhancing Predictability Assessment An Overview and Analysis of Predictability Measures for Time Series and Network Links.md](#)

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima Semana, o objetivo é:

1. Ler e sintetizar o próximo capítulo do livro que é mais focado em modelagem.
2. Estudar melhor Pré-processamento de Séries Temporais
3. Corrigir notebook de aplicação
4. Marcar um Coach Career

**Observação:** [caso precise fazer alguma observação, de qualquer “natureza”]

---

## ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: [Go!](#)

## SOBRE O APÊNDICE 4:

Os termos de aceite do apêndice 4 cobrem em parte os trabalhos das Semanas 5 e 6. O foco principal foi o estudo da propriedade Previsibilidade — inerente de todas as Séries Temporais — que representa o limite teórico de quão bem os valores futuros podem ser previstos, dada a informação disponível. Compreender a previsibilidade é fundamental porque permite responder questões críticas antes mesmo de investir esforços em modelagem: vale a pena modelar esta série? Qual é a melhor precisão que podemos esperar alcançar? Devemos dedicar tempo à engenharia de recursos ou apenas usar uma linha de base ingênua? O conceito de previsibilidade foi dividido em duas perspectivas complementares: a **previsibilidade teórica**, que é uma propriedade intrínseca dos dados e representa o nível máximo de previsibilidade possível independentemente do modelo utilizado, e a **previsibilidade prática**, que mede o desempenho alcançado na prática com um modelo específico. A diferença entre essas duas formas é crucial: se a previsibilidade teórica é baixa, não adianta trocar de modelo, pois a série é próxima de aleatória; se a previsibilidade teórica é alta mas a prática é baixa, o problema está no modelo escolhido ou em sua parametrização.

Com base nessa motivação, realizei o estudo e anotações acerca do tópico de previsibilidade em Séries Temporais. Os links e uma breve descrição do que será encontrado no acesso estão documentados abaixo:

- <https://drive.google.com/file/d/1ag4AqNVXbHZYVhQbIGM981SC2fStXFZI/view?usp=sharing> (Acessado em 03/12/2025)
  - Neste material, começo a introduzir o conceito de previsibilidade a partir da apresentação das premissas fundamentais da previsão de séries temporais e os critérios que determinam quando um problema é apropriado para forecasting, incluindo a necessidade de estrutura temporal, autocorrelação e estacionariedade, além de situações em que a previsão não é adequada.
- [https://drive.google.com/file/d/1U2OhuFeUh8LlrkntJpKZpyWU7Jp04T\\_X/view?usp=sharing](https://drive.google.com/file/d/1U2OhuFeUh8LlrkntJpKZpyWU7Jp04T_X/view?usp=sharing) (Acessado em 03/12/2025)
  - Neste material, é apresentado o conceito de previsibilidade em séries temporais e sua distinção em duas formas: **previsibilidade teórica** (propriedade intrínseca dos dados que define o limite máximo de previsão possível) e **previsibilidade prática** (desempenho real alcançado com um modelo específico). São detalhadas diversas métricas para quantificar a previsibilidade teórica, incluindo a estrutura ADI-CV para classificação de demanda, previsibilidade baseada em variância (NSE), testes de estrutura temporal, funções de autocorrelação (ACF e PACF), Variance Ratio, e um

conjunto abrangente de medidas baseadas em teoria da informação: Entropia de Shannon, taxa de entropia, complexidade de Lempel-Ziv, Entropia Aproximada (ApEn), Entropia Amostral (SampEn), Entropia de Permutação (PE) e Entropia Multiescala (MSE). O material enfatiza que a previsão exige presença de estrutura temporal, memória e autocorrelação nos dados, e que avaliar a previsibilidade antecipadamente é fundamental para evitar esforços desperdiçados em séries essencialmente aleatórias.

- <https://drive.google.com/file/d/1Mp4ALds6reCdQYlVxWqnHwZ5KnODZn0U/view?usp=sharing> (Acessado em 03/12/2025)
  - Para finalizar, neste material, é apresentado o **Expoente de Hurst (H)**, uma métrica fundamental para avaliar a dependência de longo prazo e a previsibilidade teórica de séries temporais.

Para finalizar os meus estudos sobre previsibilidade, na Semana 6 li o artigo de Bezbochina et al. (2023), "Enhancing Predictability Assessment: An Overview and Analysis of Predictability Measures for Time Series and Network Links", publicado na revista *Entropy*. Este trabalho apresenta uma revisão e análise sistemática sobre medidas de previsibilidade aplicadas a séries temporais e redes complexas, abordando uma lacuna importante na literatura: embora existam inúmeros métodos de previsão, raramente são discutidos os limites teóricos da previsibilidade dos dados em si, independentemente do modelo utilizado. Os autores formalizam a distinção entre **previsibilidade intrínseca** (propriedade teórica máxima dos dados, medida por entropia e complexidade, ou seja, a **previsibilidade teórica**) e **previsibilidade realizada** (desempenho efetivo alcançado por um modelo específico, medido por métricas de erro como MAE, RMSE e MAPE, ou seja, a **previsibilidade prática**). Os principais resultados confirmam empiricamente que há **uma correlação robusta e estatisticamente significativa** entre previsibilidade intrínseca e realizada, demonstrando que medidas de entropia podem servir como indicadores do limite teórico de precisão preditiva alcançável antes mesmo da modelagem, permitindo diagnóstico pré-modelagem, seleção eficiente de variáveis e modelos, e redução de custos computacionais ao evitar esforços em séries com baixo potencial preditivo.

Mais detalhes sobre o artigo podem ser encontrados nas minhas anotações em:

- <https://drive.google.com/file/d/1ldluUW9SHrILcoHCLxcKwt8BU8wSz2XZ/view?usp=sharing> (Acessado em 03/12/2025)

## APÊNDICE 5

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“Gate”) de aprovação:** 8 de out. de 2025

**Participantes da Entrega** [matriculados em Residência em IA]:

DAVI DE JESUS TEIXEIRA

**Entrega:** [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Durante esta Semana, avancei nos meus estudos sobre **Forecasting em Séries Temporais** focando em compreender mais sobre **métricas de avaliação em forecasting**.

Comecei lendo e resumindo o artigo Enhancing Predictability Assessment: An Overview and Analysis of Predictability Measures for Time Series and Network Link, publicado na revista Entropy (2023).

O estudo propõe uma distinção conceitual entre:

- **Previsibilidade Realizada:** o que os modelos efetivamente alcançam (métricas como MAE, RMSE, MAPE,  $R^2$ );
- **Previsibilidade Intrínseca:** o limite teórico de previsibilidade dos dados, medido por entropia e complexidade

A maior contribuição do artigo é a **validação da hipótese de que a previsibilidade intrínseca limita a previsibilidade realizada, independentemente do modelo usado**. Reforçando o papel das **medidas de complexidade como ferramentas diagnósticas pré-modelagem** e permitindo avaliar se vale a pena investir esforço computacional em uma série ou se ela é inerentemente imprevisível.

Depois dessa etapa, entrei no estudo das métricas de avaliação em forecasting, dividindo-as em duas grandes frentes: **métricas determinísticas** e **métricas probabilísticas**.

Na primeira frente, aprofundi-me nas métricas determinísticas, que avaliam diretamente o erro entre o valor previsto e o observado. Estudei as principais métricas utilizadas, como **MAE, MSE, RMSE, ME, CFE, TS, MAPE, WAPE, sMAPE, MAAPE, MASE, RMSSE e FVA**. O foco foi compreender suas diferenças, aplicações e limitações práticas.

Dediquei atenção especial ao **MAPE**, uma das métricas mais populares em contextos de negócios (Sales & Operations Planning), analisando suas inconsistências matemáticas — como distorções em séries com valores próximos de zero e falta de simetria entre erros positivos e negativos. A leitura do artigo [LINK](#) ajudou a compreender por que essa métrica deve ser substituída por alternativas mais robustas, como

### sMAPE, MAAPE e WAPE.

Na segunda frente, estudei as métricas probabilísticas, **voltadas para modelos que fazem previsões em forma de distribuições ou intervalos de confiança.**

Enquanto as métricas determinísticas respondem “o quanto a previsão está errada”, as probabilísticas respondem “o quão bem o modelo expressa sua incerteza”.

Entre as principais métricas analisadas estão **CRPS (Continuous Ranked Probability Score), Pinball Loss, Winkler Score, MSIS, ICP, PINAW, Calibration Score e U de Theil.** Cada uma mede diferentes aspectos da qualidade das previsões, como a precisão dos intervalos, a cobertura e a calibração das incertezas.

Por fim, organizei tabelas de referência (dentro de cada markdown de estudo) sintetizando as propriedades, vantagens e contextos ideais de uso de cada métrica estudada, o que permitiu entender como escolher a métrica mais adequada ao objetivo analítico.

LINKS:

[METRICAS DETERMINISTICAS EM FORECASTING.md](#)

[METRICAS PROBABILISTICAS EM FORECASTING.md](#)

[RESUMO ARTIGO Enhancing Predictability Assessment An Overview and Analysis of Predictability Measures for Time Series and Network Links.md](#)

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima Semana, o objetivo é:

1. Ler e sintetizar o próximo capítulo do livro que é mais focado em modelagem.
2. Estudar melhor Pré-processamento de Séries Temporais
3. Corrigir notebook de aplicação
4. Marcar um Coach Career

**Observação:** [caso precise fazer alguma observação, de qualquer “natureza”]

---

## ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: [Go!](#)

## **SOBRE O APÊNDICE 5:**

O Apêndice 5 cobre ainda parte do trabalho realizado na Semana 6, mas dessa vez, focado no estudo das métricas de avaliação de modelagens em Séries Temporais. O estudo foi dividido em duas grandes frentes: métricas determinísticas e métricas probabilísticas.

Na primeira frente, aprofundi-me nas métricas determinísticas, que avaliam diretamente o erro entre o valor previsto e o observado, estudando as principais métricas utilizadas, como MAE, MSE, RMSE, ME, CFE, TS, MAPE, WAPE, sMAPE, MAAPE, MASE, RMSSE e FVA, com foco em compreender suas diferenças, aplicações e limitações práticas.

- Minhas anotações acerca do tema podem ser encontradas em:  
<https://drive.google.com/file/d/12pFnGVRr0Kxt6L1WBCjElayFO4MhthFt/view?usp=sharing> (Acessado em 03/12/2025)

Na segunda frente, estudei as métricas probabilísticas, voltadas para modelos que fazem previsões em forma de distribuições ou intervalos de confiança. Enquanto as métricas determinísticas respondem "o quanto a previsão está errada", as probabilísticas respondem "o quão bem o modelo expressa sua incerteza". Entre as principais métricas analisadas estão CRPS (Continuous Ranked Probability Score), Pinball Loss, Winkler Score, MSIS, ICP, PINAW, Calibration Score e U de Theil, cada uma medindo diferentes aspectos da qualidade das previsões, como a precisão dos intervalos, a cobertura e a calibração das incertezas.

- Minhas anotações acerca do tema podem ser encontradas em:  
<https://drive.google.com/file/d/1jzZv8jovwvxpnJLQFpfYFEBvcgBC2EJy/view?usp=sharing> (Acessado em 03/12/2025)

## APÊNDICE 6

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“Gate”) de aprovação:** 15 de out. de 2025

**Participantes da Entrega** [matriculados em Residência em IA]:

DAVI DE JESUS TEIXEIRA

**Entrega:** [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Durante esta Semana, eu prossegui com as seguintes atividades acerca do tema **Forecasting em Séries Temporais**

Foquei em estudar sobre a parte de **pré-processamento de Séries Temporais**, motivação que surgiu após frustrações em colocar em prática os outros conceitos que havia aprendido até então.

Para isso, realizei a leitura do Survey - *Time-series data preprocessing: A survey and an empirical analysis*.

- Resumo: [RESUMO SURVEY](#)

O estudo debate como o **desempenho preditivo dos modelos depende fortemente da qualidade dos dados**, e por isso, propõe uma **“taxonomia” para as diferentes técnicas de pré-processamento**, além da **proposição de diversos pipelines estruturados** por meio de DAGs, com o intuito de **definir seqüências de transformações que poderiam ser aplicadas nos dados**.

Apesar da pegada bem mais voltada à engenharia de dados, o artigo me provocou **categorizando diversas técnicas de pré-processamento de dados de séries temporais, apresentando métodos** que até então eu não conhecia e **propondo pipelines de pré-processamento de forma lógica, correta e sequencial**.

Diante disso, produzi o material [PREPROCESSAMENTO EM SÉRIES TEMPORAIS](#), que aprofunda um pouco mais nas categorias de pré processamento apresentadas no survey (**fusão de dados, detecção de outliers, imputação de valores ausentes, normalização, seleção de características, discretização e compressão**) e nas técnicas de cada categoria, discutindo suas vantagens, limitações e quando aplicar cada uma delas.

Para essa semana, utilizei o **livro como leitura complementar** para esse estudo, ou seja, após ler o survey, dei uma olhada sobre o que o livro tinha a acrescentar para as minhas anotações.

Após esse estudo, consegui **corrigir os problemas que tive no meu notebook**, feito durante a semana passada, que tinha como objetivo explorar a previsibilidade e as métricas de uma série. Agora, pré-processando de maneira mais inteligente e correta os meus dados, consegui resultados condizentes e realistas no meu notebook.

Por fim, comecei a decidir qual tema quero “trabalhar” e levar como “projeto” durante o processo, que é fazer o forecasting de Dengue no Centro-Oeste. Para isso, dei uma olhada e separei diversos artigos e teses sobre o tema para me aprofundar nas próximas semanas. A ideia é: à medida que eu for estudando os conceitos que não domino (por exemplo, nesta semana o pré-processamento de Séries Temporais), ir identificando nesses artigos a forma como foi feita a aplicação, a fim de achar um padrão em comum e avaliar pontos de melhoria.

Além de reunir artigos, reuni também fontes de dados candidatas (o governo tem várias fontes bem boas para esse problema) e desejo testar e explorar a extração desses dados futuramente.

### Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

- **Estudar melhor** como fazer uma “boa” **EDA Temporal**, abordando seus principais tópicos: decomposição de séries, técnicas de visualização de padrões, análise de autocorrelação (ACF/PACF), testes de estacionariedade (ADF, KPSS, PP) e análise espectral via transformada de Fourier.
- **Aprofundar** nas **transformações de sinal e estabilização de séries**, aplicando e comparando log, Box–Cox, Yeo–Johnson, diferenciação, médias móveis e filtros (Hodrick–Prescott e Savitzky–Golay), avaliando o impacto de cada uma na variância e estacionariedade.
- **Escolher e ler** alguns dos artigos separados sobre **forecasting de Dengue**, identificando como e onde esses estudos aplicaram os conceitos que já estudei até então.
- Finalizar a Semana iniciando a **organização dos dados da Dengue localmente**, estruturando as fontes que reuni e preparando um dataset-base para as próximas etapas.

### Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

---

## ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go! ▾

## SOBRE O APÊNDICE 6:

O Apêndice 6 documenta o trabalho realizado na Semana 7. O foco da semana foi em pré-processamento de Séries Temporais.

Essa motivação surgiu após frustrações ao tentar colocar em prática os conceitos aprendidos nas semanas anteriores, quando percebi que o desempenho preditivo dos modelos depende fortemente da qualidade e do tratamento adequado dos dados. Para abordar essa lacuna, realizei a leitura do survey de Tawakuli et al. (2024), 'Time-series data preprocessing: A survey and an empirical analysis (<https://doi.org/10.1016/j.jer.2024.02.018> - Acessado em 03/12/2025), que propõe uma taxonomia abrangente para as diferentes técnicas de pré-processamento e apresenta diversos pipelines estruturados por meio de DAGs (Directed Acyclic Graphs), definindo sequências lógicas e corretas de transformações que podem ser aplicadas nos dados.

Apesar da abordagem mais voltada à engenharia de dados, o artigo foi provocador ao categorizar técnicas que até então eu desconhecia e ao propor pipelines de forma estruturada e sequencial, ajudando na compreensão geral das técnicas estudadas. Um resumo do artigo foi feito e pode ser acessado pelo link: <https://drive.google.com/file/d/12DNUuPlcTjY5nZCMgFKuCOzQQ7uorvH1/view?usp=sharing> (Acessado em 03/12/2025)

Durante esta semana, utilizei o livro *Mastering Modern Time Series Forecasting*, de Valeriy Manokhin como leitura complementar, verificando o que ele poderia acrescentar às minhas anotações após a leitura do survey. Com base nesse estudo, produzi o material <https://drive.google.com/file/d/1yxsO8HnxfYJh2B9kSHV1zsF-mn4LnNKp/view?usp=sharing> (Acessado em 03/12/2025), que aprofunda nas categorias de pré-processamento apresentadas no survey — fusão de dados, detecção de outliers, imputação de valores ausentes, normalização, seleção de características, discretização e compressão — e nas técnicas específicas de cada categoria, discutindo suas vantagens, limitações e contextos ideais de aplicação.

Mais anotações sobre cada subtema/parte do Pré Processamento em Séries Temporais podem ser encontradas nos seguintes links:

- NORMALIZAÇÃO DE DADOS:
  - [https://drive.google.com/file/d/1tUAYNO4sLM\\_jGkuP\\_Chb7OutXvM334h8/view?usp=sharing](https://drive.google.com/file/d/1tUAYNO4sLM_jGkuP_Chb7OutXvM334h8/view?usp=sharing) (Acessado em 03/12/2025)
- IMPUTAÇÃO DE VALORES FALTANTES:
  - [https://drive.google.com/file/d/1vTc8oj7F2Ms\\_eP4LREIQ9rbqwN7RVk3x/view?usp=sharing](https://drive.google.com/file/d/1vTc8oj7F2Ms_eP4LREIQ9rbqwN7RVk3x/view?usp=sharing) (Acessado em 03/12/2025)
- DETECÇÃO E REMOÇÃO DE OUTLIERS
  - [https://drive.google.com/file/d/1vmCzWR3-TOkjqlZOVD6r-xvywTWE\\_73c/view?usp=sharing](https://drive.google.com/file/d/1vmCzWR3-TOkjqlZOVD6r-xvywTWE_73c/view?usp=sharing) (Acessado em 03/12/2025)
- FEATURE SELECTION E EXTRACTION:
  - <https://drive.google.com/file/d/1Pga1tCkNtlKylae9VvH6wt6h2XAZB4VM/view?usp=sharing> (Acessado em 03/12/2025)
- DISCRETIZAÇÃO E REDUÇÃO DE DIMENSIONALIDADE:
  - <https://drive.google.com/file/d/1hvYSM9I-nBO1m4FjJHfaT8is6GHTEpPg/view?usp=sharing> (Acessado em 03/12/2025)
- COMPRESSÃO E FUSÃO DE DADOS:
  - [https://drive.google.com/file/d/1rXl2jv\\_5SY\\_nwH-K21DWFoz\\_-ZFgNmp3/view?usp=sharing](https://drive.google.com/file/d/1rXl2jv_5SY_nwH-K21DWFoz_-ZFgNmp3/view?usp=sharing) (Acessado em 03/12/2025)

## APÊNDICE 7

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“Gate”) de aprovação:** 23 de out. de 2025

**Participantes da Entrega** [matriculados em Residência em IA]:

DAVI DE JESUS TEIXEIRA

**Entrega:** [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Durante esta Semana de estudos em **Forecasting em Séries Temporais** finalizei meus estudos **Pré-Modelagem em Séries Temporais**, consolidando um pipeline completo que vai além da EDA tradicional e cobre pré-processamentos, validações estatísticas e critérios de escolha de modelos (estatísticos, ML e DL).

O principal material gerado nesta semana foi o markdown [PRÉ MODELAGEM.md](#), no qual escrevi, em formato de “passo a passo”, descreve um **pipeline robusto que antecede a implementação de modelos de previsão em séries temporais**. O documento organiza todo o processo pré-modelagem e orienta decisões conforme as propriedades da série e a técnica escolhida para a modelagem.

Os tópicos abordados foram:

Entendimento Inicial da Série → Análise de Qualidade dos Dados → EDA Visual → Transformações e Pré-processamento → Avaliação e Tratamento da Estacionariedade → Análise de Autocorrelação (ACF/PACF e testes) → Relação entre Séries (exógenas) → Anomalias e Rupturas → Verificação de Ruído Branco

Nesse material, além de explorar diversos caminhos que podem ser tomados de acordo com as propriedades de cada série / modelagem escolhida, ressaltai ainda boas práticas para fortalecer as minhas análises e garantir melhores resultados.

Além disso, efetuei o screening de alguns artigos e eventualmente, também a leitura de um desses artigos sobre o tema Forecasting de Dengue, temas que me agrada para teste/desenvolvimento utilizando o que estou estudando durante o meu processo. O artigo lido foi: ***Assessing dengue forecasting methods: a comparative study of statistical models and machine learning techniques in Rio de Janeiro, Brazil.***

A leitura contribuiu para entender melhor como diferentes abordagens (estatísticas e de ML) têm sido aplicadas nesse contexto, mas não sei se achei útil produzir resumo sobre...

Por fim, validei a **extração dos dados** que utilizarei nos experimentos, confirmando que a procedência, qualidade e formato estão adequados às minhas necessidades. Dados acessados em:  
<https://info.dengue.mat.br>

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Aplicar todos meus conhecimentos no dataset que encontrei.
- Aprofundar sobre os modelos, principalmente os que considero mais interessantes (modelos de IA).
- Rodar e avaliar pelo menos um modelo de acordo com esse primeiro pipeline que construí
- Tentar elencar pontos de melhoria para refatorações

**Observação:** [caso precise fazer alguma observação, de qualquer “natureza”]

**ACEITE DA ENTREGA:**

CEDRIC LUIZ DE CARVALHO: Go! ▾

## **SOBRE O APÊNDICE 7:**

A entrega da Semana 8, coberta neste apêndice, foi focada em esquematizar e propor enfim um pipeline completo para a pré-modelagem de problemas de Forecasting em Séries Temporais, desde a fase de entendimento da série até o processo de modelagem propriamente dita. O objetivo foi consolidar todo o conhecimento adquirido nas semanas anteriores em um fluxo estruturado que vai além da EDA tradicional e abrange pré-processamentos, validações estatísticas e critérios de escolha de modelos (estatísticos, ML e DL).

O principal material gerado nesta semana foi o documento [https://drive.google.com/file/d/1\\_I9lfo0GMzauYDqNLTCd\\_YsBzzQgzNrM/view?usp=sharing](https://drive.google.com/file/d/1_I9lfo0GMzauYDqNLTCd_YsBzzQgzNrM/view?usp=sharing) (Acessado em 03/12/2025), escrito em formato de "passo a passo", que descreve um pipeline robusto orientando decisões conforme as propriedades da série e a técnica escolhida para modelagem. Além de explorar diversos caminhos que podem ser tomados de acordo com as propriedades de cada série e a modelagem escolhida, o material ressalta boas práticas essenciais para fortalecer as análises e garantir resultados mais confiáveis e robustos em projetos de forecasting.

## APÊNDICE 8

## Termo de Aceite de Entrega

### Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

**Data da Reunião (“Gate”) de aprovação:** 12 de nov. de 2025

**Participantes da Entrega** [matriculados em Residência em IA]:

DAVI DE JESUS TEIXEIRA

**Entrega:** [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Durante a última Semana de Processo, tentei reunir todos os conhecimentos que adquiri para tentar modelar o problema de forecasting de casos de Dengue no estado de Goiás. Inspirado pelo artigo **“Assessing dengue forecasting methods: a comparative study of statistical models and machine learning techniques in Rio de Janeiro, Brazil”**, detalharei abaixo o pipeline que propus, dificuldades enfrentadas e um pouco da minha experiência com o forecasting na prática.

Comecei com a análise exploratória da série de **2010 a 2025**, que possuía 818 registros e 98% de cobertura. Após verificar a integridade das datas e interpolar uma semana ausente via *spline*, a série revelou um **padrão sazonal claro**, com picos concentrados no início de cada ano, especialmente em 2010, 2013, 2016, 2022 e 2024. A presença de **outliers** foi confirmada por critérios estatísticos (Z-score e IQR), mas mantive esses pontos, já que representam **epidemias reais** e não erros de medição.

A decomposição da série mostrou uma **tendência cíclica** e uma sazonalidade proporcional ao nível de casos — ou seja, surtos mais intensos em anos de maior incidência. Os **mapas de calor e médias semanais** reforçaram que os surtos ocorrem predominantemente entre fevereiro e maio, coincidindo com o período chuvoso, enquanto julho a novembro marcam o intervalo seco de baixa incidência.

Para estabilizar a variabilidade, testei transformações logarítmicas, raiz quadrada e **Box-Cox**. A Box-Cox apresentou a melhor simetria, tornando a série mais adequada para modelagem. Após normalizar via MinMaxScaler, verifiquei a **estacionariedade** com os testes ADF e KPSS. Ambos testes confirmaram que

a série transformada é estacionária. O **expoente de Hurst (H=0.701)** indicou **persistência temporal**, ou seja, períodos de alta tendem a se manter, um bom indicativo da previsibilidade da série.

Com a base transformada e normalizada, iniciei a **modelagem univariada** com XGBoost. Foram criadas **52 features de lags semanais (uma para cada semana do ano)**, features de médias e desvios móveis desses lags e **features sazonais (mês, semana, ano)** para tentar capturar padrões anuais e semi anuais.

**O modelo foi treinado em dados de 2017–2021, validado em 2022–2023 e testado em 2024.**

Na **validação (2022–2023)**, o modelo alcançou **MAE=0.04, RMSE=0.07 e R<sup>2</sup>=0.84**, mostrando boa generalização dentro da amostra. No **teste (2024)**, manteve desempenho sólido (**MAE=0.053, RMSE=0.063, SMAPE=6.8%, R<sup>2</sup>=0.85**). O gráfico de previsão mostrou um bom alinhamento com a série real, capturando nuances do pico e da queda do período epidêmico do início de 2024. A análise residual confirmou uma certa **normalidade e ausência de forte autocorrelação**, indicando que o modelo ajustou relativamente bem a estrutura temporal.

Para avaliar a robustez preditiva, apliquei a abordagem de **previsão direta multi-horizonte**, treinando modelos independentes para prever 1, 4, 8 e 12 semanas à frente. Como esperado, houve **degradação do desempenho** conforme o horizonte aumentou: o **SMAPE** passou de **11.5% (1 semana)** para **23.8% (12 semanas)** e o **R<sup>2</sup>** caiu até valores negativos, mostrando comportamento completamente aleatório em horizontes de predição longos. Esse comportamento é comum em séries epidêmicas, nas quais a incerteza aumenta rapidamente com o tempo.

Em seguida, implementei o método **Walk-Forward**, que simula o uso do modelo em cenário real: o XGBoost é **re-treinado a cada semana**, incorporando as novas observações e prevendo apenas o próximo ponto. Essa estratégia eliminou o viés temporal e melhorou a adaptação local, alcançando **MAE=0.032, RMSE=0.040 e R<sup>2</sup>=0.94**. Os resíduos se mantiveram centrados e sem padrão, comprovando a **estabilidade e a consistência temporal** do modelo ao longo do tempo.

Por fim, estendi o experimento para o **Walk-Forward Multi-Horizonte (h = 1, 4, 8, 12)**, combinando re-treinamento incremental e previsão de múltiplas semanas à frente. Os resultados mostraram excelente equilíbrio entre estabilidade e alcance preditivo:

- **1 semana** → SMAPE = 6.5%, R<sup>2</sup> = 0.88
- **4 semanas** → SMAPE = 10.1%, R<sup>2</sup> = 0.66
- **8 semanas** → SMAPE = 11.1%, R<sup>2</sup> = 0.59
- **12 semanas** → SMAPE = 11.5%, R<sup>2</sup> = 0.57

O erro aumentou de forma **suave e controlada**, indicando que o modelo manteve previsões coerentes em diferentes janelas, indicando um desempenho relativamente bom para uma abordagem puramente univariada.

O **XGBoost univariado** demonstrou ser **eficiente para previsões de curto prazo de 1 semana e “ok” para previsões em janelas maiores**. O próximo passo natural seria integrar **variáveis exógenas climáticas e epidemiológicas** (como temperatura, umidade e precipitação) e testar mais modelos. Tentei, mas infelizmente não consegui desenvolver uma modelagem multivariada satisfatória e nem outros modelos.

---

NOTEBOOK COM A MODELAGEM COMPLETA: [MONO.ipynb](#)

**Planejamento:** [descrever o que pretende fazer para realizar a próxima ENTREGA]

Organizar tudo que estudei para pensar como escrever o meu TCC.

**Observação:** [caso precise fazer alguma observação, de qualquer “natureza”]

---

**ACEITE DA ENTREGA:**

CEDRIC LUIZ DE CARVALHO:

---

## SOBRE O APÊNDICE 8:

Por fim, as entregas das Semanas 9 e 10, cobertas neste apêndice, focam na aplicação prática dos meus conhecimentos para a modelagem de um problema real: O Forecasting de casos de dengue no estado de Goiás.

A modelagem completa, desde a exploração dos dados até a avaliação dos resultados do algoritmo foi feita em formato de Jupyter Notebook, disponibilizado no meu Github: <https://github.com/dvzk1/TCC-IA/blob/main/CODES/MONO.ipynb> (Acessado em: 03/12/2025)

Abaixo entrarei em detalhes da implementação e discutirei resultados com maior profundidade:

## Parte 1: DADOS

### Descrição dos Dados e Contexto da Base Utilizada

Os dados analisados foram obtidos diretamente da plataforma oficial **InfoDengue** (<https://info.dengue.mat.br> - Acessado em: 03/12/2025 ), um sistema de vigilância epidemiológica que consolida notificações de dengue no Brasil e produz indicadores de alerta baseados em clima, transmissibilidade e detecção de surtos.

Essa plataforma integra informações de múltiplas fontes — SINAN (notificações de saúde), dados meteorológicos, populações estimadas e estimativas de transmissão — resultando em um dataset robusto para análise epidemiológica e modelagem preditiva.

A base utilizada contém **818 registros semanais**, cobrindo o período de **janeiro de 2010 a agosto de 2025**, o que representa **15,6 anos de observações em frequência semanal**. A cobertura é muito alta (98,3%), com apenas uma semana faltante, o que indica boa integridade temporal.

### Estrutura da Base e Variáveis Disponíveis

---

Cada linha da base representa uma **semana epidemiológica**, identificada por `data_iniSE` (domingo da semana), contendo informações relacionadas a:

#### (A) Casos e estimativas epidemiológicas

- **casos** – número de casos notificados na semana
- **casos\_est**, **casos\_est\_min**, **casos\_est\_max** – nowcasting, com intervalo de credibilidade
- **casprov**, **casconf** – casos prováveis e confirmados (aqui, porém, sem preenchimento)
- **notif\_accum\_year** – acumulado anual de notificações

Esses campos representam o **comportamento real da dengue** ao longo do tempo, afetado por sazonalidade, ciclos epidêmicos e fatores ambientais.

#### (B) Indicadores de transmissão

- **Rt** – número reprodutivo instantâneo
- **p\_rt1** – probabilidade de  $R_t > 1$
- **nivel**, **nivel\_inc** – níveis oficiais de alerta
- **receptivo** – grau de receptividade climática
- **transmissao** – evidência de transmissão sustentada

Esses indicadores são importantes para modelagens preditivas ou para análises de risco, pois capturam a dinâmica epidemiológica de curto prazo.

#### (C) Variáveis Climáticas

- **tempmin**, **tempmed**, **tempmax**
- **umidmin**, **umidmed**, **umidmax**

As variáveis climáticas são altamente relevantes porque temperatura e umidade influenciam:

- reprodução do mosquito *Aedes aegypti*
- tempo de desenvolvimento das larvas
- taxa de picada
- sobrevivência do vetor

Essas variáveis tornam o dataset valioso para **modelagens multivariadas**, incluindo modelos ARIMAX, regressões de defasagens distribuídas, Prophet com regressões, LSTMs multivariadas, entre outros.

### Enfoque da Modelagem Neste Estudo

Embora a base ofereça um conjunto amplo de variáveis exógenas, a abordagem adotada neste notebook inicialmente se concentrou na **modelagem univariada**, utilizando apenas:

- **data\_iniSE** como eixo temporal
- **casos** como variável de interesse

ou seja, uma **série temporal clássica**, focada no comportamento histórico dos casos ao longo dos últimos 15 anos.

Essa escolha permite:

- entender o padrão bruto da dengue
- identificar ciclos, picos epidêmicos e sazonalidade
- construir um baseline de modelagem temporal
- preparar terreno para futuras extensões multivariadas

Ao mesmo tempo, ressalta-se que **uma modelagem multivariada é perfeitamente viável e recomendada futuramente**, incorporando as variáveis climáticas e indicadores de transmissibilidade para:

- melhorar a acurácia preditiva
- capturar relações não lineares entre clima e casos
- antecipar períodos de risco elevado
- desenvolver modelos de alerta antecipado (early warning)

### Inspeção Inicial dos Dados

Após carregar o dataset e exibir as primeiras/últimas linhas, foi verificado que:

- não há valores ausentes na variável casos
- algumas variáveis apresentam missing values (ex.: tweet, umidmed, etc.), o que exige cuidado em modelagens multivariadas
- a coluna temporal foi convertida para datetime
- o índice foi ordenado cronologicamente
- foram criados atributos temporais como:

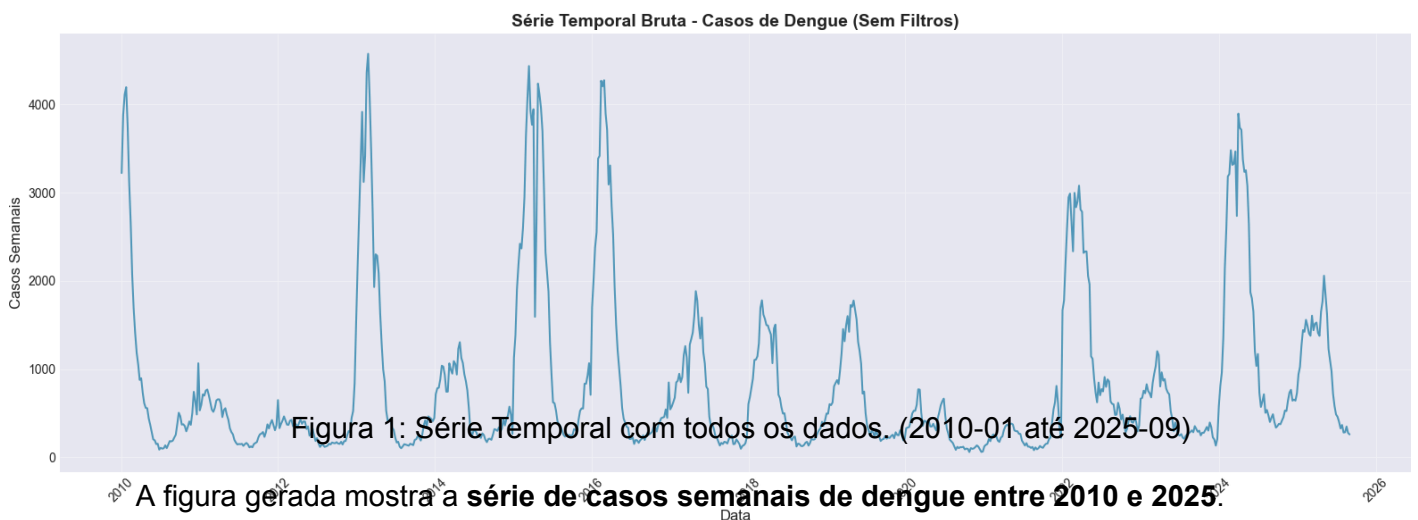
- year
- epiyear
- epiweek

A análise de integridade temporal confirmou:

- **Amostragem semanal correta**
- **Apenas 1 semana ausente em 15 anos**
  
- **Nenhum timestamp duplicado**

Essas verificações são fundamentais para garantir a consistência da série antes de qualquer modelagem.

### Visualização Bruta da Série Temporal:



Os principais padrões observados são:

- **Sazonalidade anual muito forte**
  - Picos se repetem principalmente no **primeiro semestre**, período mais quente e úmido no Brasil — ambiente ideal para o mosquito.
- **Flutuações de alta intensidade**
  - Existem anos com picos acima de **4000 casos/semana** (ex.: 2010, 2013, 2015, 2016) e outros com níveis mais moderados.

Isso reflete os ciclos epidêmicos característicos da dengue, influenciados por sorotipos circulantes, imunidade da população, condições ambientais, etc.

- **Quedas abruptas entre epidemias**
  - Após os grandes picos, há queda brusca, confirmando períodos de recrudescimento.
- **Últimos anos com forte crescimento**
  - A partir de 2023–2024 observa-se novo ciclo epidêmico relevante, consistente com a crise recente de dengue no país.

## PARTE 2: PRÉ MODELAGEM

### Pré-Modelagem: Análise de Qualidade e Comportamento da Série

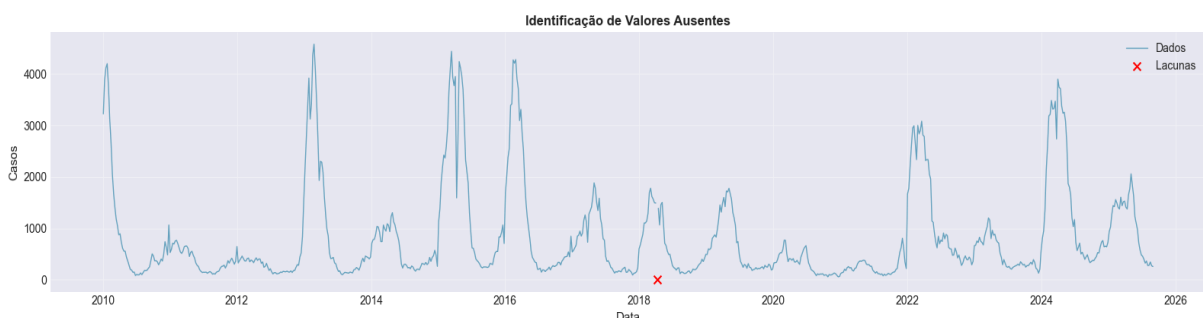
Após a inspeção inicial da série temporal, avancei para uma análise aprofundada da qualidade dos dados e do comportamento estrutural da série de casos de dengue. Essa etapa é fundamental antes de iniciar qualquer tentativa de modelagem, pois permite identificar padrões, rupturas, ruído, outliers, sazonalidades incomuns e possíveis mudanças de regime ao longo dos 15 anos de observação.

### Checagem de Integridade Temporal e Reindexação

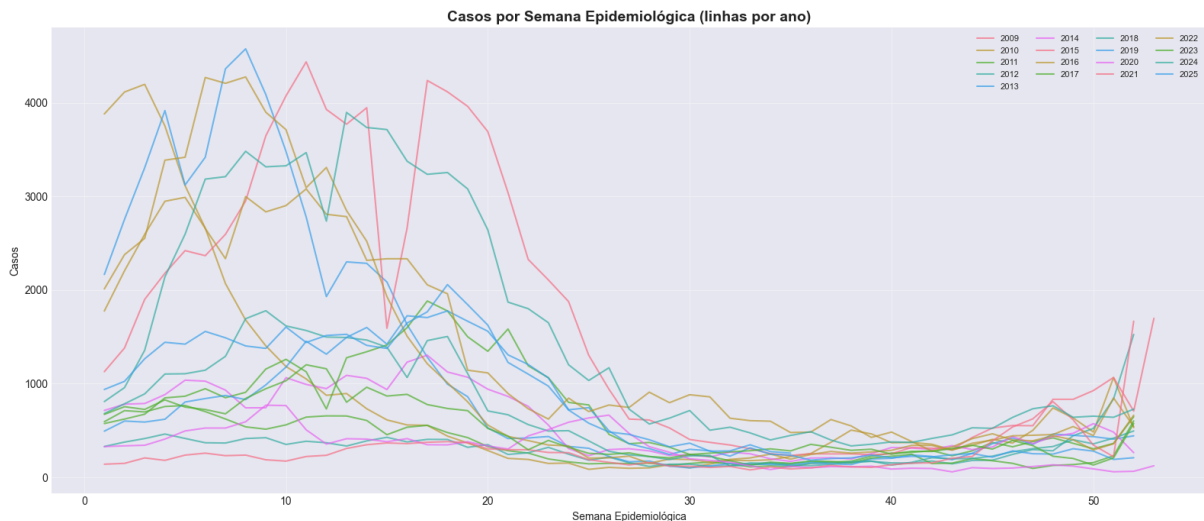
Para garantir que a série estivesse estruturada corretamente, reindexei o dataset em uma grade contínua de semanas epidemiológicas (W-SUN) entre 2010 e 2025. Após a reindexação:

- Semanas esperadas: 818
- Semanas com dados reais: 817
- Apenas 1 valor faltante
- Nenhuma duplicidade temporal
- Ordem cronológica preservada

A figura gerada destaca essa lacuna única (em abril de 2018) com um marcador vermelho sobre o eixo do tempo. Essa etapa assegura que métodos posteriores — decomposição, diferenciação, ARIMA, Prophet, ou redes neurais — operem sobre uma série uniforme e coerente.



## Visualização Ano-a-Ano por Semana Epidemiológica



Para investigar a sazonalidade intra-anual, projetei os casos por semana epidemiológica, com uma linha para cada ano. Essa visualização evidencia claramente que:

- Os picos epidêmicos se concentram nas primeiras 10 a 20 semanas do ano.
- Vários anos possuem surtos intensos (2010, 2013, 2015, 2016, 2024).
- A intensidade oscila de forma não trivial — característica típica de doenças transmitidas por vetores, influenciadas por sorotipos circulantes e imunidade da população.

Esse comportamento reforça a necessidade de modelagem que considere sazonalidade explícita e ciclos plurianuais.

## Identificação de Valores Ausentes

A análise confirmou:

- Apenas um valor ausente em toda a série (0,12%).
- A lacuna está isolada e não afeta intervalos maiores.

Isso indica excelente qualidade de notificação e permite estratégias simples de imputação caso necessário (p.ex., interpolação temporal ou uso de médias móveis).

## Detecção de Outliers

Em seguida, investiguei potenciais outliers estruturais, aplicando três métodos:

### (1) Z-Score ( $>|3|$ )

→ Detectados 26 pontos extremos.

Esses representam semanas com contagens extraordinariamente acima da média histórica.

### (2) IQR (limites whisker $\pm 1.5 \cdot \text{IQR}$ )

→ Detectados 89 outliers.

Como a série possui picos epidêmicos muito altos, métodos robustos como IQR classificam quase todos os surtos fortes como outliers — o que, na prática epidemiológica, são eventos reais e esperados.

### (3) Rolling IQR (janela móvel de 52 semanas)

→ Apenas 2 outliers.

Esse método é mais apropriado para séries sazonais, pois compara cada ponto com seu contexto anual.

A conclusão é que os picos muito elevados não são “erros”, mas parte natural da dinâmica da dengue, que possui surtos intensos seguidos de quedas abruptas.

## Estatísticas Móveis (Médias e Desvios)

Calculei uma média móvel de 52 semanas com intervalo de  $\pm 1$  desvio padrão. Esse gráfico permite observar:

- A tendência anual suavizada da série.
- O aumento na amplitude dos surtos em anos específicos (2010, 2013, 2015–2016, 2024).
- A expansão da banda de desvio padrão nos períodos epidêmicos, indicando maior volatilidade.

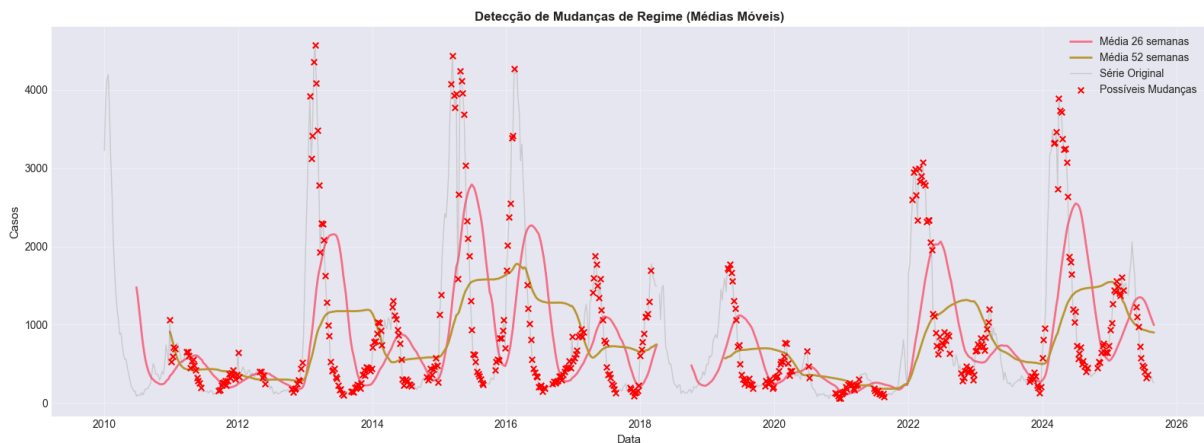
Essa visualização confirma que a dengue apresenta comportamento altamente sazonal, mas com amplitude variável, o que aponta para possíveis ciclos multianuais influenciados por imunidade populacional e mudanças no sorotipo dominante.

## Detecção de Mudanças de Regime

Para investigar rupturas estruturais, utilizei a diferença percentual entre duas médias móveis:

- Janela curta: 26 semanas ( $\approx 6$  meses)
- Janela longa: 52 semanas ( $\approx 1$  ano)

Classifiquei como mudança de regime qualquer período em que a média curta difere mais que 30% da média longa.



O resultado foi que 444 semanas (aprox. metade da série) foram marcadas como possíveis mudanças estruturais.

Esse valor alto não representa um erro, a série de dengue sofre mudanças abruptas recorrentes devido a:

- surtos intensos
- transição entre sorotipos dominantes
- condições climáticas excepcionais
- ciclos de imunidade
- impacto de medidas de controle.

No gráfico correspondente, os pontos vermelhos mostram semanas em que a média recente se descola fortemente da média histórica — principalmente durante grandes epidemias.

## Interpretação Integrada da Qualidade dos Dados

A análise conduzida permite consolidar alguns pontos importantes:

1. **A série é confiável e praticamente completa**
  - 1.1. Apenas um valor faltante e nenhum problema estrutural de indexação.
2. **Os “outliers” são na verdade componentes legítimos da dinâmica epidêmica**
  - 2.1. Modelos de previsão devem ser capazes de lidar com esse comportamento explosivo.
3. **Fortes mudanças de regime são parte natural da série**
  - 3.1. Isso indica que modelos lineares simples podem ser insuficientes.
4. **Há elevada sazonalidade anual e variações cíclicas plurianuais**
  - 4.1. Confirmando a necessidade de abordagens que consigam capturar tendência e sazonalidade.
5. **A variabilidade e amplitude dos surtos mudam ao longo dos anos**
  - 5.1. Significando que a série não é estacionária no sentido estrito.

## Tratamento de Valores Nulos na Variável Target

Antes de realizar qualquer análise avançada, tratei o único valor ausente encontrado na variável casos. Essa lacuna estava localizada em:

- 08/04/2018 (Semana Epidemiológica de 2018)

Ela foi corrigida por meio de uma interpolação temporal spline cúbica, aplicada somente na variável alvo.

Após a interpolação:

- 0 valores nulos restantes
- A semana crítica passou a apresentar valor coerente com a vizinhança temporal  
A continuidade da série foi preservada, sem impacto visual relevante

Esse cuidado é fundamental, pois algoritmos de decomposição e correlogramas não admitem NaNs.

## Tendência, Sazonalidade e Estruturas Temporais

Após a validação e limpeza da série temporal, avancei para uma análise aprofundada do comportamento temporal dos casos de dengue, explorando padrões de **sazonalidade, tendência, correlações temporais e decomposição estrutural**.

Esta etapa é crucial para orientar a futura modelagem, principalmente quando se trabalha com doenças que apresentam ciclos epidêmicos complexos, como a dengue.

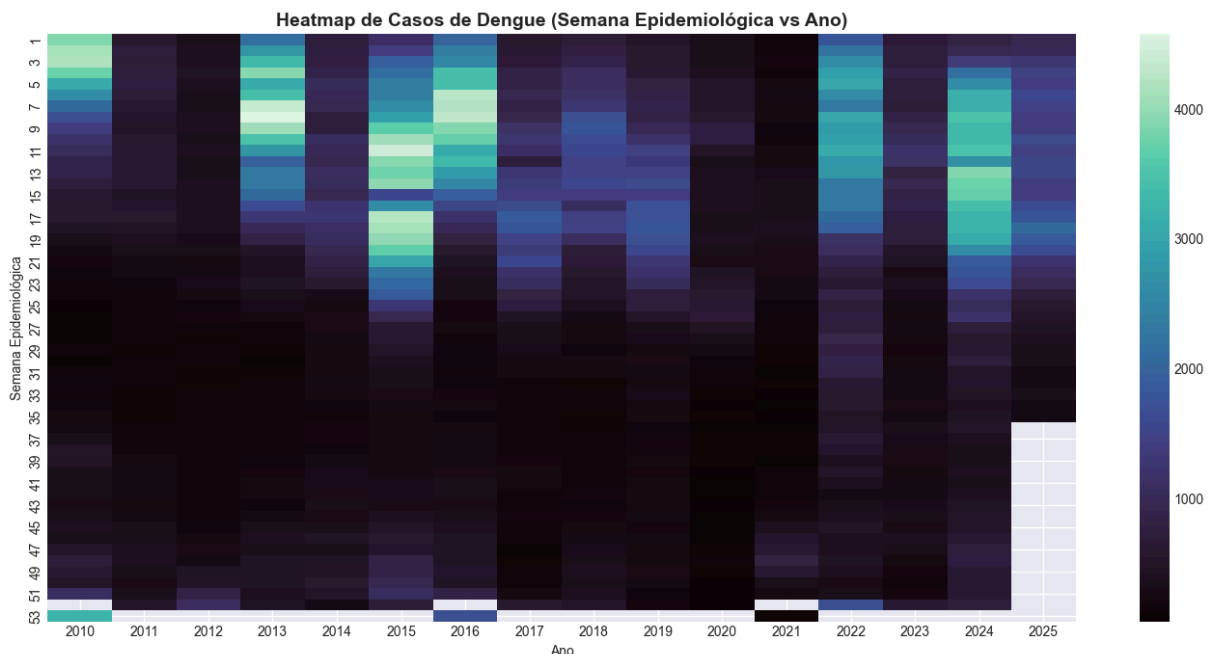
## Heatmap Mensal: Identificação Clara da Estação da Dengue

O primeiro heatmap agregando casos médios por **mês** e por **ano** revela de forma bastante nítida a **estação anual da dengue**.

Principais observações:

- Os meses de **janeiro a maio** concentram os maiores valores.
- Em anos epidêmicos (2010, 2013, 2015–2016 e 2024), observa-se um bloco intenso de calor logo no **início do ano**, especialmente em fevereiro e março.
- O período de **junho a novembro** apresenta consistentemente níveis muito baixos, caracterizando uma estação seca com baixa circulação viral.
- Os últimos anos confirmam esse padrão sazonal, com valores elevados voltando a se intensificar em 2024.

Esse comportamento reforça que a modelagem precisa incorporar sazonalidade explícita (p.ex., SARIMA, Prophet, TBATS ou Fourier terms).

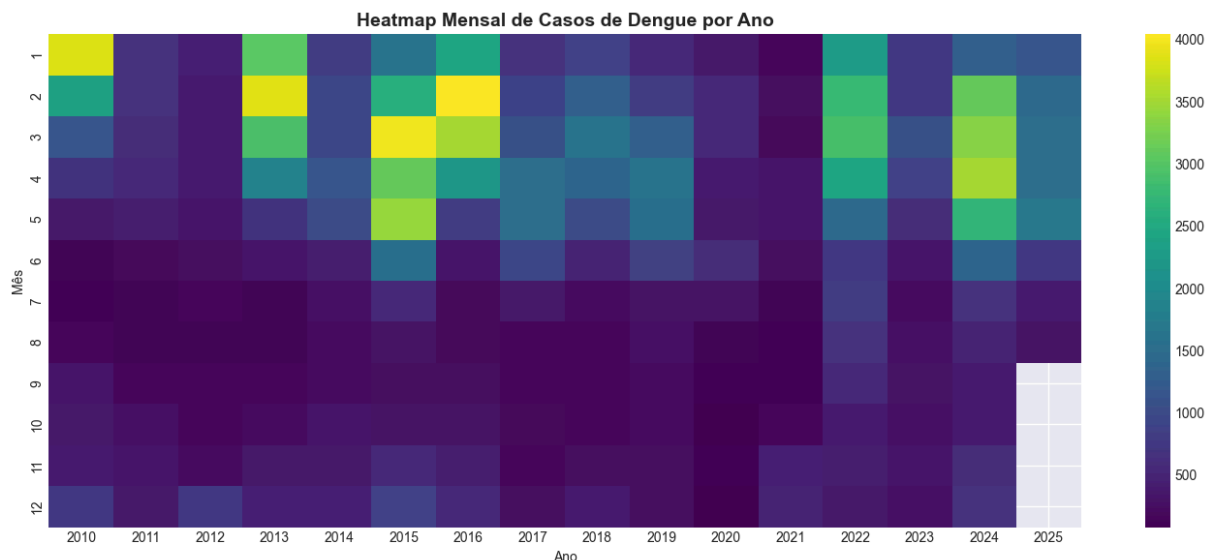


## Heatmap Semanal: Sazonalidade em Alta Resolução

O heatmap por **semana epidemiológica** evidencia com ainda mais clareza o padrão sazonal fino:

- As semanas **7 a 20** concentram a maior parte dos surtos intensos.
- Grandes epidemias mostram faixas verticais altas e bem definidas, especialmente nos anos de 2010, 2013, 2015, 2016 e 2024
- Entre as semanas **30 e 45**, os casos praticamente desaparecem — o que coincide com meses de inverno e clima desfavorável ao mosquito.

Essa visualização é essencial para orientar a escolha da **periodicidade** e das **diferenciações sazonais** na modelagem.

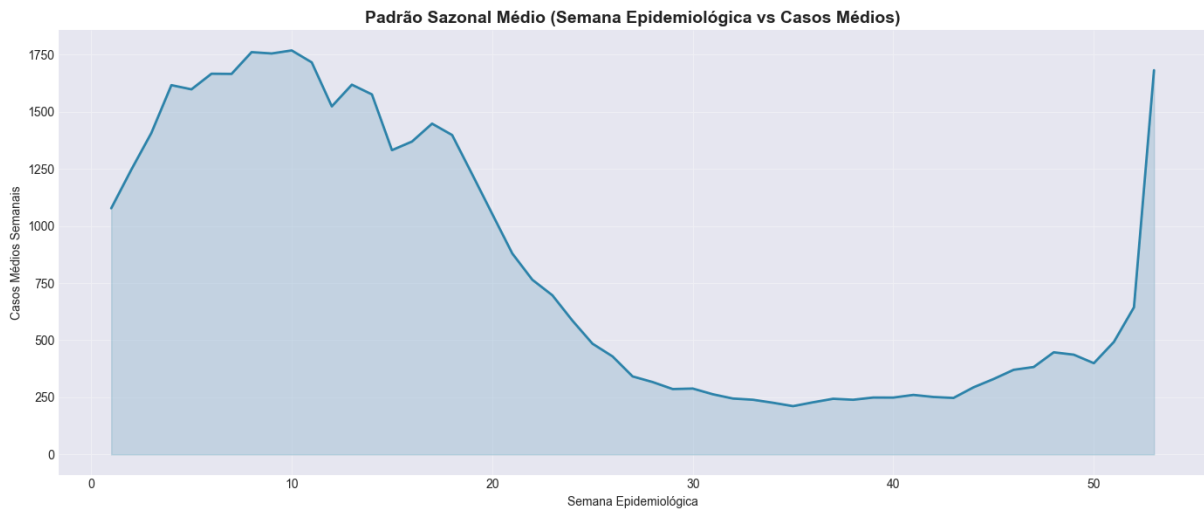


### Padrão Sazonal Médio por Semana Epidemiológica

Agrupando todos os anos e calculando a média semanal:

- A curva sazonal média sobe rapidamente após a semana **1**, atingindo seu pico entre as semanas **8 e 12**.
- Após a semana **20**, há queda progressiva e contínua até atingir o menor nível por volta das semanas **35–40**.
- Um pequeno aumento reaparece perto das semanas **50–52**, possivelmente ligado ao início da estação chuvosa.

É um padrão **extremamente consistente**, o que é favorável para modelos sazonais.



## Correlogramas (ACF e PACF)

A análise de autocorrelação (ACF) e autocorrelação parcial (PACF) revelou características importantes:

### ACF

- Desaceleração lenta e oscilatória, típica de séries com **forte componente sazonal**.
- Picos marcantes em defasagens múltiplas de **52 semanas**, confirmando periodicidade anual.
- Estrutura semelhante a uma função quase sinusoidal, sugerindo ciclos longos.

### PACF

- Queda abrupta nas primeiras defasagens, seguido de ruído irregular.
- Forte autocorrelação imediata (lag 1).
- Índícios de um componente AR não trivial — comum em séries epidemiológicas.

Esses resultados indicam que:

- A série **não é estacionária**,
- Tem forte sazonalidade anual,
- E apresenta memória temporal significativa, exigindo diferenciações adequadas.

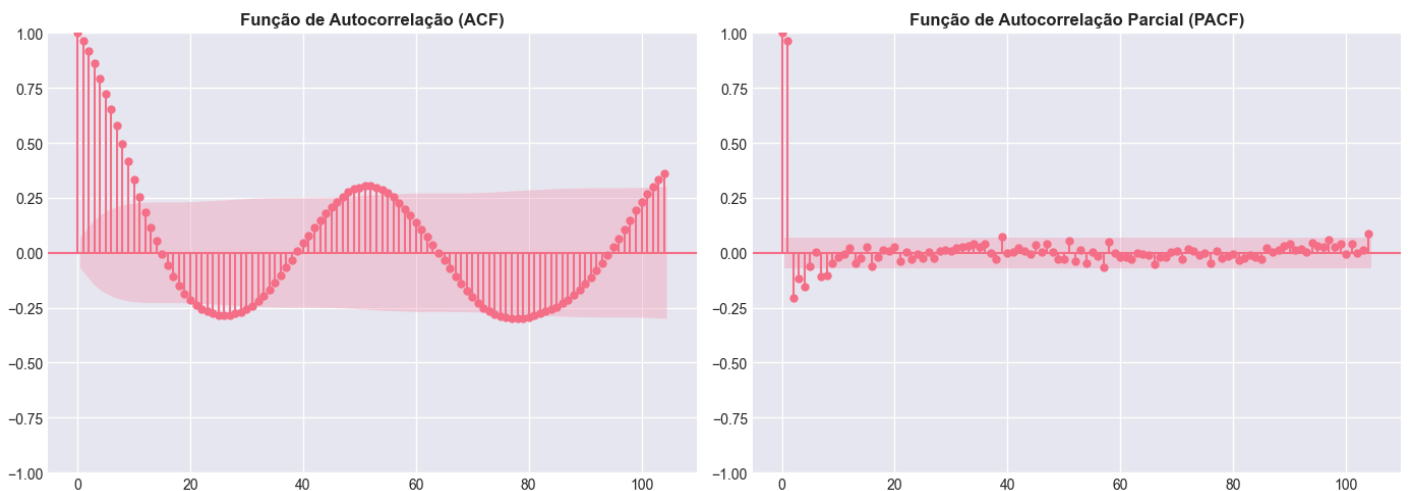
## Decomposição da Série Temporal

A decomposição foi realizada em dois modelos:

### Modelo Aditivo

A série é separada em:

- **Trend:**
  - Ciclo claro de epidemias plurianuais, com grandes picos em 2010, 2013 e 2015–2016, depois queda e novo aumento em 2024.
  - Períodos prolongados de baixa atividade entre epidemias.
- **Seasonal:**
  - Oscilação anual extremamente estável e suave.
  - Mostra amplitude relativamente constante independentemente do ano.
- **Residual:**

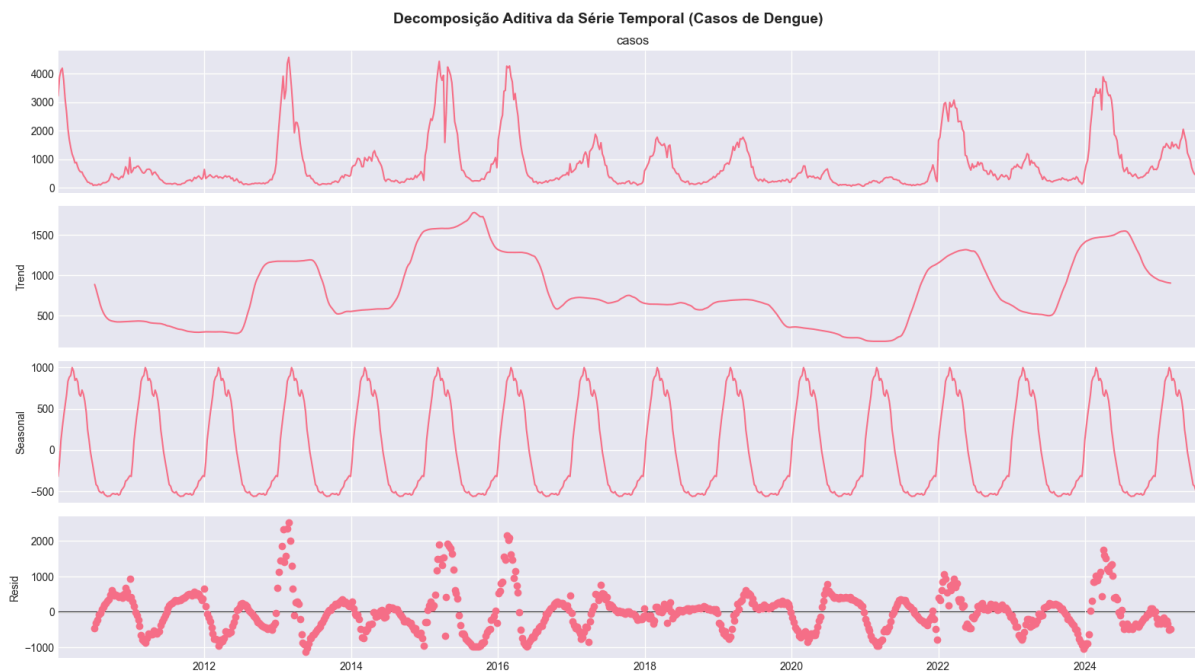


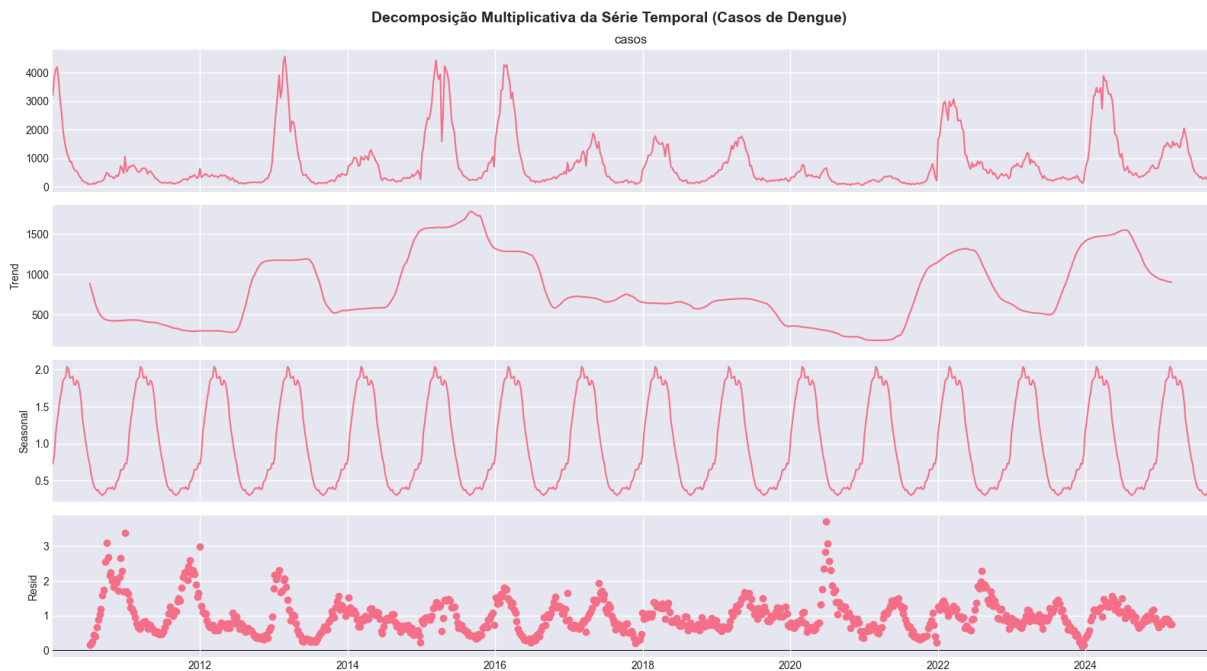
- Resíduos maiores nos períodos epidêmicos — esperado devido à volatilidade da transmissão.  
Períodos não epidêmicos possuem resíduos mais moderados.

### Modelo Multiplicativo

Mostrou-se particularmente adequado, pois a amplitude da sazonalidade aumenta proporcionalmente ao nível da tendência — fenômeno típico na dengue, onde:

- surtos maiores  $\Rightarrow$  maior amplitude sazonal,
- anos de baixa circulação  $\Rightarrow$  amplitude reduzida.





## Integração dos Resultados

Ao analisar em conjunto todos os gráficos produzidos — heatmaps, decomposições, correlogramas e padrão sazonal — algumas características marcantes emergem:

1. **A série é fortemente sazonal, com periodicidade anual bem definida.**
2. **Existe um ciclo plurianual associado às grandes epidemias.**
3. **A amplitude sazonal cresce e diminui ao longo dos anos, favorecendo uma modelagem multiplicativa.**
4. **Há dependência temporal significativa (ACF/PACF), indicando a presença de componentes AR e MA.**
5. **Os resíduos apresentam maior variabilidade nos períodos de surto, o que sugere que:**
  - modelos lineares podem ser insuficientes para capturar episódios epidêmicos extremos.
  - modelos não lineares (LSTM, modelos com regressão exógena, GAMs, etc.) podem ser explorados posteriormente.

## Conclusão da Análise Temporal

Essa etapa estabeleceu os elementos centrais para a futura modelagem:

- **Sazonalidade anual forte e consistente**
- **Tendência não linear com ciclos epidêmicos intensos**  
**Estruturas multiplicativas evidentes (amplitude proporcional ao nível)**
- **Memória temporal longa (evidenciada pela ACF)**
- **Resíduos com comportamento dinâmico**
- **Dependência de fatores climáticos e ambientais (para modelos multivariados)**

## Transformações e Normalizações da Série

Depois de validar a integridade dos dados e caracterizar o comportamento temporal da série, avancei para uma etapa essencial na preparação para modelos estatísticos e algoritmos de Machine Learning: a transformação da escala da variável alvo (casos) e a normalização dos valores.

Séries epidemiológicas como a de dengue apresentam distribuição altamente assimétrica, com caudas longas e picos epidêmicos muito fortes. Essa característica viola pressupostos de normalidade implícitos em diversos modelos e tende a prejudicar algoritmos sensíveis à escala. Por isso, a etapa de transformação é particularmente importante neste contexto.

## Distribuição Original da Série

A análise inicial da distribuição dos casos semanais revelou:

- **Assimetria positiva muito forte**, puxada pelos surtos epidêmicos.
- A maior parte dos valores concentra-se em níveis baixos, enquanto poucos pontos atingem valores extremamente elevados (acima de 3000 ou 4000 casos).
- O QQ-plot mostra uma curva que se descola completamente da linha normal, confirmando a ausência de normalidade.

Essa distribuição é típica de processos infecciosos e reforça a necessidade de transformar a escala.

## Aplicação de Transformações (log, sqrt, boxcox)

Foram testadas três transformações clássicas:

### Transformação Logarítmica (log1p)

- Reduz drasticamente a escala dos surtos.
- Diminui a variância.

- Ainda assim, não corrige completamente a assimetria.

### Transformação Raiz Quadrada (sqrt)

- Suaviza a variabilidade, mas tem desempenho inferior ao log em termos de simetrização.
- Apresentou assimetria ainda relativamente elevada.

### Transformação Box-Cox

- Permite ajustar a transformação através de um parâmetro  $\lambda$ .
- $\lambda$  encontrado: **-0.19605**, valor próximo do caso da transformação logarítmica, porém mais flexível.

Entre todas as opções, a **transformação Box-Cox** apresentou a **melhor simetrização**, com assimetria de **0.05**, praticamente perfeita.

**Conclusão:** A melhor transformação é a **Box-Cox**, pois aproxima a distribuição de uma forma quase simétrica e muito mais adequada a modelos lineares e redes neurais.

### Construção da Nova Variável Transformada

A coluna `casos_transf` foi adicionada ao `df_time` contendo a série transformada por Box-Cox.

Isso significa que todas as análises posteriores — testes de estacionaridade, modelagem, forecast — podem ser feitas tanto:

- em **escala original**, para interpretação epidemiológica,
- quanto na **escala transformada**, para fins de ajuste mais estável do modelo.

Essa separação permite manter o melhor dos dois mundos: rigor estatístico e clareza interpretativa.

### Normalização (Min-Max e Z-Score)

Após escolher a melhor transformação, apliquei dois métodos clássicos de normalização:

#### Min-Max (0 a 1)

- Reescala toda a série para o intervalo [0, 1].
- Útil para modelos sensíveis à amplitude, especialmente redes neurais.
- Preserva a forma da distribuição, mas comprime valores extremos.

### Z-Score

- Centraliza (média = 0) e padroniza (desvio = 1).
- Adequado para métodos lineares e métodos estatísticos.
- Mantém maior separação entre picos e semanas de baixa incidência.

As distribuições resultantes foram visualizadas, confirmando:

- Min-Max → distribuição comprimida e adequada para modelos não lineares.
- Z-Score → distribuição centralizada e mais informativa para análise estatística.

Então, com base nisso colunas `casos_minmax` e `casos_zscore` foram criadas e adicionadas ao dataframe.

### Conclusão da Etapa de Transformações

A quarta etapa do pipeline concluiu:

- A série original é altamente assimétrica e heteroscedástica
- A transformação **Box-Cox** foi identificada como a melhor opção
- Foram criadas três versões da série transformada/normalizada

### Testes de Estacionariedade e Expoente de Hurst

Após realizar a transformação Box-Cox e normalizar a escala dos dados, o próximo passo é avaliar se a série resultante apresenta **estacionariedade**, condição fundamental para a correta aplicação de vários modelos clássicos de séries temporais, especialmente SARIMA.

A estacionariedade indica que a série possui:

- média constante ao longo do tempo,
- variância constante,
- ausência de tendências persistentes,
- comportamento estável nas autocorrelações.

Como a série original de dengue contém fortes tendências epidêmicas e sazonalidade pronunciada, é natural que ela seja **não estacionária na sua forma bruta**. Por isso, a

avaliação ocorre **após a transformação Box-Cox**, que já estabiliza parte da variância e reduz assimetria.

### Teste ADF (Augmented Dickey-Fuller)

O teste ADF parte da hipótese nula de que a série possui **raiz unitária** (ou seja, não é estacionária).

Resultados obtidos:

- **ADF Statistic: -7.5237**  
**p-value = 0.0000**

Como o p-valor é menor que 0,05, rejeita-se a hipótese nula.

→ **Conclusão:**

**A série transformada é estacionária do ponto de vista do ADF.**

Isso indica que a transformação Box-Cox foi suficiente para retirar grande parte da tendência estocástica da série.

### Teste KPSS (Kwiatkowski–Phillips–Schmidt–Shin)

Esse teste funciona de forma complementar ao ADF.

Aqui, a **hipótese nula é que a série é estacionária**, e rejeitá-la significaria detectar tendência.

Resultados:

- **KPSS Statistic: 0.1375**
- **p-value = 0.1000**

Como o p-valor é **maior** que 0,05, **não rejeitamos** a hipótese nula.

→ **Conclusão:**

**O KPSS também indica estacionariedade.**

Isso é importante porque o KPSS costuma detectar estacionariedade com mais sensibilidade a tendências determinísticas.

A convergência entre os dois testes (ADF e KPSS) confirma que:

- A série Box-Cox transformada é estacionária na média
- Não há evidência significativa de tendência remanescente

- A variabilidade encontra-se mais estabilizada após a transformação

## Expoente de Hurst

Para complementar os testes formais, estimei o **Expoente de Hurst (H)**, que captura a *persistência temporal* da série.

Resultado:

- **H = 0.701**

Interpretação:

- **H < 0.5**: série antipersistente (alternância rápida, tendência reversível)
- **H = 0.5**: ruído branco
- **H > 0.5**: série persistente (tendências de longo prazo, memória longa)

→ **Conclusão:**

**H = 0.701 indica forte persistência**, ou seja:

- a série possui **memória de longo prazo**
- valores altos tendem a ser seguidos por valores altos
- valores baixos tendem a ser seguidos por baixos
- característica típica de séries epidemiológicas com ciclos epidêmicos.

Esse resultado complementa os testes clássicos:

- **Mesmo estacionária**, a série possui **dependência temporal significativa**, o que afeta a escolha dos modelos.  
Modelos simples ARIMA podem não capturar totalmente essa persistência.

## Conclusão da Etapa de Estacionariedade

- A transformação Box-Cox estabilizou adequadamente a série
- Testes ADF e KPSS são **ambos conclusivos** em favor da estacionariedade
- O expoente de Hurst revelou **persistência temporal** típica de doenças infecciosas
- A série está preparada para **modelagem paramétrica**
- Contudo, a memória longa indica que modelos mais ricos podem capturar melhor dependências de maior alcance

## PARTE 3: A MODELAGEM E A AVALIAÇÃO

### Modelagem Univariada com XGBoost

Após finalizar todas as etapas de limpeza, transformação e verificação de estacionariedade da série temporal, avancei para a modelagem univariada, utilizando o XGBoost como modelo preditivo. O objetivo é prever a evolução semanal de casos de dengue utilizando apenas a própria série como fonte de informação, incorporando atrasos temporais, estatísticas de janelas móveis e componentes sazonais.

### Definição do Período de Modelagem

Para garantir consistência temporal e evitar vazamento de informação, estabeleci três camadas de recorte no tempo:

- **Janela completa usada na modelagem: 2016–2024**  
Esse é o intervalo efetivo de dados utilizado após as transformações.
- **Treino + Validação (Treino expandido): 2016–2023**
- **Janela específica de validação temporal: 2022–2023**, com o objetivo de simular previsões genuínas no final da série.
- **Teste final (com modelo congelado): Ano de 2024**, totalmente fora do período de treino e sem re-treino — garantindo avaliação justa.

Esse desenho cria uma estrutura adequada para **forecast realista**, evitando validações ingênuas que misturam passado e futuro indevidamente.

### Pré-Processamento para o Modelo: Box-Cox + Normalização

A transformação aplicada anteriormente é refinada nesta etapa para evitar vazamento:

1. O  $\lambda$  da transformação **Box-Cox** é ajustado **somente no conjunto de treino (até 2023)**.
2. O mesmo  $\lambda$  é aplicado a toda a série.
3. Depois, apliquei **MinMaxScaler**, também ajustado apenas nos dados até 2023.

Essa abordagem assegura que o modelo não receba informações estatísticas futuras durante o ajuste das transformações — um cuidado essencial em séries temporais.

As variáveis transformadas alimentam o modelo XGBoost exclusivamente em sua versão normalizada (“casos\_minmax”).

---

## Construção das Features Para Aprendizado Supervisionado

Como o XGBoost não é um modelo intrinsecamente temporal, foi necessário reconstruir uma matriz de features que representasse o comportamento passado da série. Usei:

### Lags

- 52 atrasos semanais (lag\_1 até lag\_52) capturam até um ano completo de informação histórica.

### Estatísticas de janelas móveis (rolling features)

- rolling mean (4 e 12 semanas)
- rolling std (4 e 12 semanas)

Essas features adicionam suavização e volatilidade, enriquecendo o aprendizado.

### Features calendáricas

- mês
- semana epidemiológica
- ano

### Codificação sazonal com funções trigonométricas

- sin/cos anual (52 semanas)
- sin/cos semianual (26 semanas)

Essa estrutura permite que o XGBoost capte sazonalidade cíclica sem depender de modelos explícitos de Fourier.

Após o processamento, as linhas iniciais, que não possuem lags completos, foram removidas.

## Divisão Estrita em Treino, Validação Temporal e Teste

A divisão final ficou:

- **Treino puro:** 2017–2021

- **Validação temporal:** 2022–2023
- **Teste final:** 2024
- **Total de semanas no treino+val:** 366
- **Total no teste:** 52 semanas

Esse recorte garante um experimento realista, simulando previsão de anos completos sabendo apenas do histórico até 2023.

### **Validação Temporal com Backtesting Rolling**

A validação é feita de forma **sequencial**, semana a semana:

1. Para cada data da janela 2022–2023:
  - o modelo treina apenas no passado daquela semana
  - prevê um único ponto no futuro imediato
  - acumula erro real × previsto

Esse processo é o padrão-ouro para modelos de séries temporais, pois simula de forma fiel como uma previsão semanal ocorreria em produção.

### **Resultados de Validação (2022–2023)**

MAE = 0.04

RMSE = 0.07

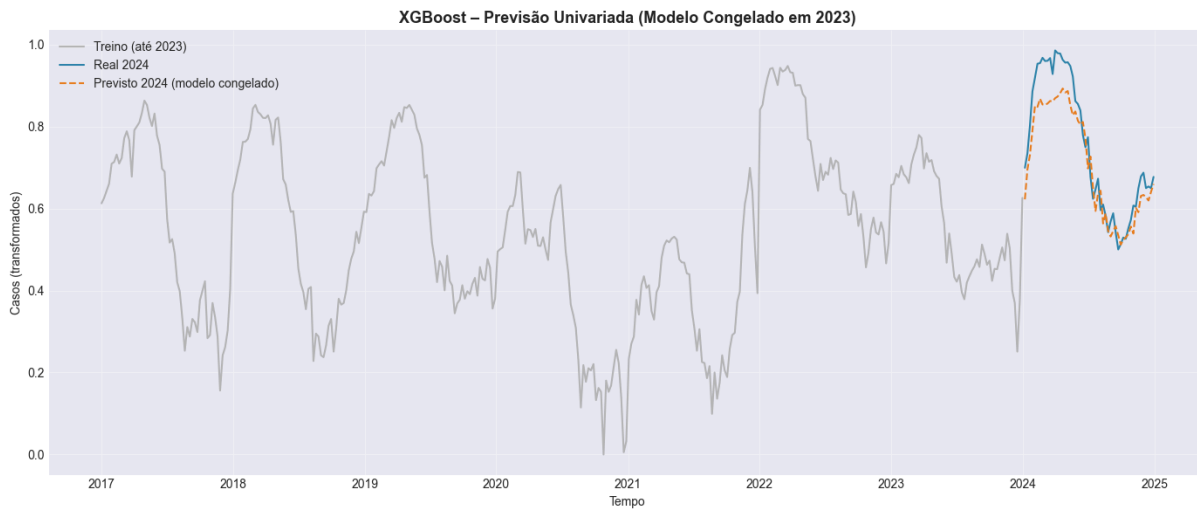
$R^2$  = 0.84

Esses erros estão na **escala normalizada**, mas representam boa aderência do modelo nas previsões semanais.

O conjunto de hiperparâmetros utilizados para treinar o modelo foi:

- **'n\_estimators': 400,**
- **'learning\_rate': 0.05,**
- **'max\_depth': 3,**
- **'min\_child\_weight': 3,**
- **'subsample': 0.8,**
- **'colsample\_bytree': 0.8,**
- **'reg\_lambda': 1.0,**
- **'reg\_alpha': 0.0**

Esse conjunto representa um modelo relativamente raso (depth=3), com boa regularização e taxa de aprendizado baixa.



## Treino Final (Modelo Congelado)

Após escolher o melhor conjunto de hiperparâmetros, ajustei o modelo final usando **todo o período até 2023**, congelando-o em seguida.

Não houve re-treino em 2024, o que preserva a integridade experimental.

## Avaliação no Teste Final – Ano de 2024

### Resultados

**MAE = 0.053**

**RMSE = 0.063**

**SMAPE = 6.80%**

**R<sup>2</sup> = 0.851**

Esses resultados indicam:

- excelente aderência global ( $R^2 \approx 0.85$ )
- erros proporcionalmente baixos (SMAPE < 10%)
- boa capacidade de seguir a tendência geral de 2024

A curva mostra que o modelo acompanha bem os movimentos principais, embora suavize alguns picos — comportamento típico de modelos em árvore.

## Diagnóstico de Resíduos

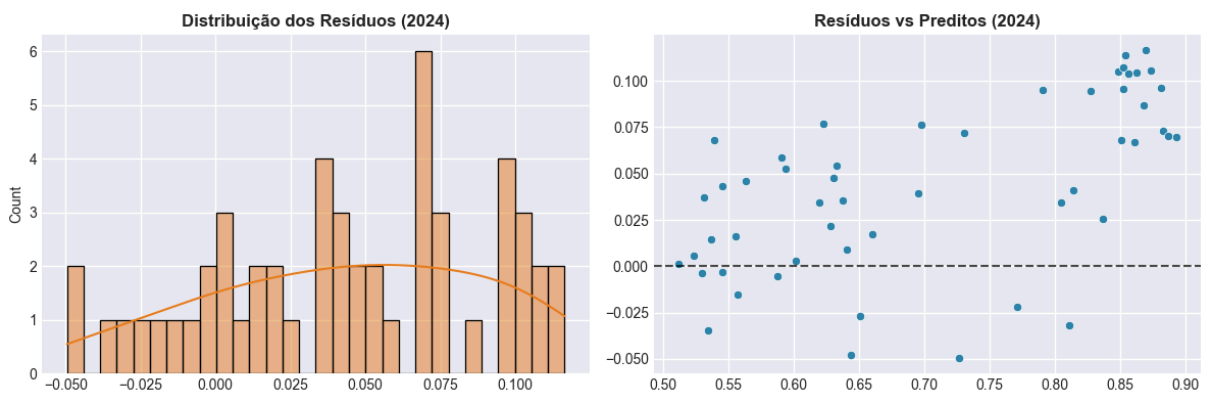
Os resíduos foram avaliados por três métodos:

### Distribuição dos resíduos

- aproximadamente simétrica
- sem caudas pesadas
- sem padrões evidentes de heterocedasticidade

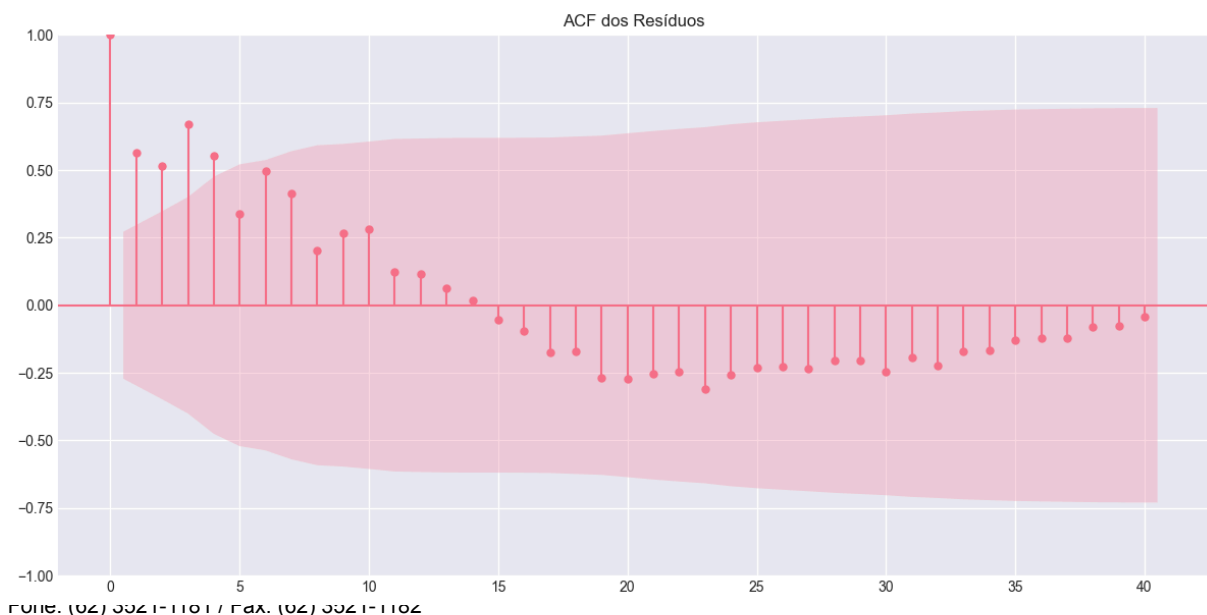
### Resíduos vs Preditos

- forma aleatória, sem estrutura — excelente sinal.



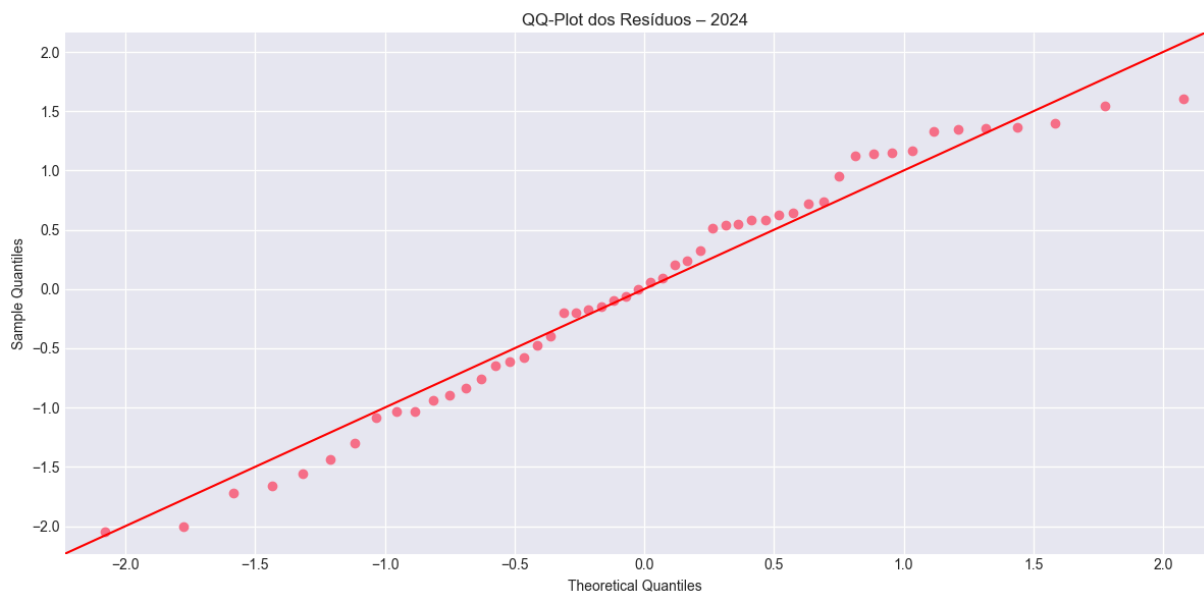
### ACF dos resíduos

- ausência de autocorrelação marcante
- indica que o modelo capturou bem a estrutura temporal dos dados



## QQ-Plot

- alinhamento razoável com a normalidade
- resíduos adequados para um modelo preditivo



Conclusão: **os resíduos são bem comportados**. Isso sugere que pouco dos padrões temporais significativos ficaram de fora.

## Desempenho por Horizonte de Previsão

O modelo também foi avaliado para prever vários passos no futuro (1, 4, 8 e 12 semanas):

**TABELA 1: XGBoost – desempenho por horizonte (targets em 2024):**

Horizonte	MAE	RMSE	R <sup>2</sup>	SMAPE
1	0.09	0.10	0.64	11.51
4	0.13	0.16	0.04	18.72
8	0.14	0.18	-0.15	18.72
12	0.17	0.21	-0.59	19.51

**Interpretação:**

- O modelo vai bem em **curto prazo** (1 semana).
- O desempenho cai rapidamente conforme aumenta o horizonte — comportamento típico de séries com forte sazonalidade e alta variabilidade epidêmica.
- Previsões acima de 8 semanas tornam-se pouco confiáveis com esse tipo de abordagem.

**Walk-Forward Online e Multi-Horizonte com XGBoost**

Após avaliar o modelo XGBoost congelado (treinado até 2023 e testado diretamente em 2024), implementei uma segunda estratégia de validação temporal mais próxima do uso real em produção: o **walk-forward online**, com re-treino incremental a cada semana.

Essa abordagem simula o funcionamento de um modelo operacional, que é atualizado continuamente conforme novos dados chegam. Além disso, ampliei o experimento para um cenário multi-horizonte (1, 4, 8 e 12 semanas à frente), avaliando a degradação preditiva conforme aumenta a distância temporal.

**Walk-Forward Online (1 passo à frente, re-treino semanal)**

Neste experimento, para cada semana de 2024:

1. O modelo é treinado usando **todos os dados disponíveis até aquela semana** (incluindo as semanas anteriores de 2024).
2. Ele prevê apenas o **próximo ponto futuro imediato**.
3. O processo se repete, expandindo o conjunto de treino a cada iteração.

Essa estratégia representa o cenário mais realista possível para aplicações operacionais, especialmente em epidemias como dengue, cujos dados são atualizados semanalmente.

## Desempenho agregado (2024)

**MAE = 0.032**

**RMSE = 0.040**

**R<sup>2</sup> = 0.941**

O ganho em relação ao modelo congelado é significativo:

- R<sup>2</sup> sobe de ~0.85 para ~0.94
- MAE cai praticamente pela metade
- Previsão semanal torna-se muito mais aderente à curva real

Isso mostra que o XGBoost se beneficia fortemente do re-treino incremental, atualizando-se em função dos padrões mais recentes da série.

## Visualização da Previsão Walk-Forward



O gráfico mostra:

- **série original até 2023** (cinza)
- **série real de 2024** (azul)
- **previsões walk-forward semanais** (tracejado laranja)

A leitura visual revela que **o modelo walk-forward supera claramente o modelo congelado**. Enquanto o modelo treinado apenas até 2023 apresentou dificuldade em acompanhar a fase epidêmica inicial de 2024 — subestimando o pico e reagindo com atraso às oscilações abruptas — o walk-forward acompanha de forma precisa o comportamento real semana a semana. No gráfico, observa-se que as previsões atualizadas semanalmente capturam tanto a rápida ascensão do surto quanto o declínio subsequente, mantendo-se aderentes ao padrão verdadeiro mesmo durante períodos de maior variabilidade. Além disso, não há atraso perceptível na captura das subidas e descidas, evidenciando que, quando atualizado dinamicamente, o XGBoost consegue explorar de maneira muito mais eficiente as defasagens e janelas móveis construídas no pré-processamento.

## Diagnóstico dos Resíduos do Walk-Forward

Dois gráficos complementam a análise:



### Distribuição dos resíduos

- aproximadamente simétrica
- centrada em zero
- sem caudas longas

- sem sinais de heterocedasticidade.

Isso indica que o modelo não está deixando para trás erros sistemáticos.

### Resíduos vs Preditos

O gráfico de dispersão mostra uma nuvem aleatória ao redor do eixo horizontal:

- nenhum padrão de forma “U” (erro crescente com magnitude)
- nenhum funil (que indicaria heteroscedasticidade)
- nenhuma autocorrelação visível

Esses pontos reforçam a **boa calibragem do modelo** no modo walk-forward.

### Walk-Forward Multi-Horizonte

Além do cenário de 1 passo à frente, avaliei o desempenho para horizontes maiores (4, 8 e 12 semanas). Nesse caso:

- cada horizonte possui seu **próprio modelo**,
- cada previsão é feita com re-treino online considerando apenas dados passados,
- **nunca** se usa como entrada um alvo futuro conhecido.

Os resultados são:

**TABELA 2: XGBoost – desempenho por horizonte (Walk-Foward):**

Horizonte	MAE	RMSE	R <sup>2</sup>	SMAPE
<b>1 semana</b>	0.045	0.055	0.885	<b>6.50%</b>
<b>4 semanas</b>	0.070	0.095	0.661	10.08%
<b>8 semanas</b>	0.080	0.104	0.590	11.08%
<b>12 semanas</b>	0.084	0.107	0.567	11.50%

## Interpretação:

- **1 semana à frente:**  
desempenho excelente, indicando previsões muito confiáveis no curtíssimo prazo.
- **4 semanas:**  
ainda aceitável, com leve degradação. O modelo captura bem a tendência mensal.
- **8 a 12 semanas:**  
a queda continua — comportamento esperado com epidemias.  
Mesmo assim, o modelo **ainda preserva bom sinal** ( $R^2 \sim 0.57$ ), algo raro em séries tão não lineares.

## Insight Final

### Melhor cenário para previsões operacionais semanais → Walk-Forward 1 passo

- $R^2 \approx 0.94$
- erros baixos
- resíduos bem comportados
- aderência visual quase perfeita

### Para previsões de médio prazo (1 mês) → horizonte 4 semanas

- SMAPE  $\approx 10\%$
- cenário útil para planejamento tático (insumos, recursos, equipes)

### Para previsões epidêmicas estruturais (2–3 meses) → horizonte 8–12 semanas

- o erro aumenta, mas ainda preserva informação de tendência
- pode ser usado como camada de apoio (por exemplo, modelos híbridos climáticos + ML)

## Conclusão

O walk-forward mostrou que o XGBoost:

- aprende muito bem com atualizações semanais
- acompanha picos epidêmicos com alta precisão
- mantém resíduos estáveis e aleatórios
- e apresenta degradação previsível e controlada no longo prazo.

---

## PARTE 4: CONCLUSÃO E POSSÍVEIS PRÓXIMOS PASSOS

### Conclusão Geral do Estudo

A análise conduzida ao longo deste notebook permitiu caracterizar de forma abrangente a dinâmica temporal dos casos semanais de dengue, incorporando procedimentos rigorosos de pré-processamento, avaliação da estrutura da série e modelagem preditiva. O estudo confirma a natureza altamente sazonal, não linear e marcada por ciclos epidêmicos intensos, alternados com períodos de baixa transmissão.

A etapa de pré-modelagem mostrou que:

- a série apresenta **picos epidêmicos extremos**, mas consistentes com a epidemiologia da dengue
- possui **sazonalidade anual clara**, com maior concentração de casos entre janeiro e maio; apresenta **mudanças frequentes de regime**, inerentes à alternância de sorotipos, variações climáticas e imunidade da população
- tem comportamento assimétrico e heterocedástico, exigindo **transformações adequadas de escala**
- a transformação **Box-Cox** produziu uma série muito mais estável e simétrica, permitindo modelagens estatísticas mais robustas
- após a transformação, a série tornou-se **estacionária**, embora mantenha **memória longa** (Hurst  $\approx 0.70$ ).

Com esse arcabouço consolidado, o XGBoost foi utilizado como modelo univariado, incorporando lags, estatísticas móveis e componentes sazonais. Duas estratégias foram avaliadas:

### Modelo Congelado (treino até 2023, previsão direta de 2024)

Esse modelo apresentou bom desempenho global ( $R^2 \approx 0.85$ ), capturando tendências gerais do ano de 2024. No entanto:

- suavizou excessivamente o surto epidêmico de início de ano;
- subestimou o pico;
- apresentou atraso nas inflexões ascendentes e descendentes.

Ou seja, embora as métricas fossem boas, o modelo congelado **não conseguiu reagir adequadamente às mudanças rápidas típicas de surtos epidêmicos**.

## Walk-Forward Online (re-treino semanal)

O cenário operacional de walk-forward apresentou desempenho significativamente superior:

- $R^2$  aumentou de  $\sim 0.85 \rightarrow 0.94$
- MAE reduziu quase pela metade
- as previsões acompanharam o comportamento real com **alta fidelidade**, inclusive o pico epidêmico
- resíduos tornaram-se ainda mais estáveis e aleatórios, indicando modelo bem calibrado

O gráfico mostrou que o walk-forward reproduz não apenas a forma geral da curva, mas também sua **dinâmica fina**, sem atrasos perceptíveis.

## Walk-forward multi-horizonte

O estudo multi-horizonte mostrou que:

- **curto prazo (1 semana): excelente desempenho**, SMAPE  $< 7\%$
- **médio prazo (4 semanas): desempenho bom**, SMAPE  $\sim 10\%$
- **longo prazo (8–12 semanas): degradação previsível**, porém ainda com preservação de tendência

Esse comportamento é típico de séries epidemiológicas e reforça que modelos semanais são ideais para monitoramento e resposta rápida, enquanto previsões de meses requerem abordagens híbridas ou variáveis exógenas.

## Síntese Final

Este estudo demonstrou alguns pontos bem interessantes como:

- a série de dengue possui padrões complexos, mas previsíveis em curto prazo;
- a transformação Box-Cox e o cuidado com estacionariedade foram essenciais;
- o XGBoost congelado oferece bom baseline;
- o **walk-forward supera amplamente o modelo congelado**, principalmente em períodos epidêmicos;
- previsões de curto prazo ( $< 4$  semanas) são altamente confiáveis;
- previsões de médio prazo ( $\sim 1$  mês) são úteis;
- previsões mais longas exigem modelos mais sofisticados e variáveis exógenas;

- há grande potencial de evolução para modelos multivariados e híbridos.

Com essas conclusões, o pipeline estabelece uma base sólida para previsão operacional da dengue e abre caminho para sistemas mais completos, robustos e integrados à vigilância epidemiológica moderna.