UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE COMPUTAÇÃO - IC

Avaliação de modelos para predição de séries temporais

Candidato: Fabio Fogliarini Brolesi

Orientador: Prof. Dr. Hélio Pedrini

Abril 2024

Sumário

Resumo

A compreensão e a predição de séries temporais desempenham um papel crucial em uma variedade de campos, incluindo meteorologia, finanças e medicina. Nesse contexto, este estudo se propõe a investigar e comparar modelos para predição de séries temporais, explorando as abordagens ARIMA, XGBoost e modelos baseados em *Transformers*. Os dados utilizados neste estudo abrangem meteorologia, finanças e dados médicos. Essa diversidade de fontes de dados reflete a importância de adaptar os modelos de predição a diferentes contextos e necessidades. Inicialmente, foram coletados conjuntos de dados preliminares de cada domínio, destacando-se a necessidade de futuras análises mais detalhadas para garantir a qualidade e a representatividade dos dados. Em seguida, implementou-se modelos de predição baseados em ARIMA, XGBoost para cada conjunto de dados. Os trabalhos com mecanismos baseados em *Transformers* serão feitos em etapa posterior. Durante a fase experimental, conduziu-se uma análise comparativa de desempenho entre os modelos, utilizando métricas relevantes para cada domínio, tais como erro médio absoluto e raiz quadrada do erro quadrático médio. Os resultados preliminares sugerem que o desempenho dos modelos varia significativamente de acordo com o domínio dos dados. No entanto, é importante ressaltar que este estudo está em estágio preliminar e análises mais detalhadas são necessárias para validar e refinar os resultados. Futuras etapas incluirão a otimização dos hiperparâmetros dos modelos, a incorporação de mais atributos discriminativos e a realização de testes em conjuntos de dados adicionais para uma avaliação mais abrangente do desempenho dos modelos de predição de séries temporais em diferentes domínios.

Lista de Abreviações e Acrônimos

API	*Application Programming Interface*
AR	*Autoregressive*
ARIMA	*Autoregressive Integrated Moving Average*
B3	Brasil, Bolsa, Balcão
CEPAGRI	Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura
CNN	*Convolutional Neural Network*
COVID-19	Doença do Coronavírus 2019
DWT	*Discrete Wavelet Transform*
ECG	Eletrocardiograma
EQM	Exame de Qualificação de Mestrado
FB	*Forecast Bias*
FEAGRI	Faculdade de Engenharia Agrícola da UNICAMP
FIIs	Fundos de Investimento Imobiliário
GPU	*Graphics Processing Unit*
HFRS	*Hemorrhagic Fever with Renal Syndrome*
Ibovespa	Índice Bovespa
IFIX	Índice de Fundos de Investimentos Imobiliários
MA	*Moving Averages*
MAE	*Mean Absolute Error*
MASE	*Mean Absolute Scaled Error*
MSE	*Mean Squared Error*
RNN	*Recurrent Neural Network*
sMAPE	*Symmetric Mean Absolute Percent Error*
TFT	*Temporal Fusion Transformer*
UNICAMP	Universidade Estadual de Campinas
XGBoost	*Extreme Gradient Boosting*

Capítulo 1

Introdução

No cenário atual da pesquisa em séries temporais, a análise e a predição de dados têm desempenhado um papel crucial em diversas áreas, como medicina [4], mercado financeiro [13] e meteorologia [5], impulsionando avanços significativos em métodos e técnicas. Nesse contexto, a avaliação de modelos torna-se essencial para identificar a eficácia e a adaptabilidade de diferentes abordagens diante de conjuntos de dados complexos e multifacetados [16]. A predição precisa de tendências de séries temporais e a compreensão da dinâmica subjacente aos dados de tempo são essenciais para a tomada de decisões estratégicas e eficazes em diferentes contextos.

O presente trabalho propõe uma investigação comparativa entre modelos tradicionais (por exemplo, o método ARIMA), técnicas de aprendizado de máquina (por exemplo, o XGBoost) e uma abordagem mais recente baseada em *Transformers* [17].

A motivação para escolher os presentes modelos vem da necessidade de identificar como diferentes abordagens respondem a desafios específicos apresentados por séries temporais oriundas de áreas distintas do conhecimento. Com o objetivo de executar uma análise abrangente e aplicada, serão explorados conjuntos de dados representativos de três domínios distintos do conhecimento: meteorologia, finanças e saúde.

Para o domínio de meteorologia, será utilizada uma série temporal obtida do Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura (CEPAGRI[1]), englobando dados como temperatura, umidade e precipitação. No âmbito financeiro, o presente trabalho abordará séries temporais derivadas de índices de papéis do mercado brasileiro, especificamente Ibovespa e IFIX, para mercado de ações e fundos imobiliários, respectivamente. Já para o campo médico, serão consideradas séries temporais provenientes de exames como o eletrocardiograma (ECG).

Este estudo propõe não apenas a aplicação de modelos consolidados em cada domínio, mas também uma avaliação comparativa de seu desempenho. Propõe-se então abranger diferentes tipos de modelagem: desde métodos estatísticos tradicionais até abordagens baseadas em aprendizado de máquina profundo. O entendimento de como cada modelo responde aos desafios específicos de cada conjunto de dados proporcionará percepções valiosas sobre a generalização de técnicas de séries temporais em diferentes contextos, contribuindo para o avanço da pesquisa e aplicação dessas metodologias.

[1] `https://www.cpa.unicamp.br/`

1.1 Caracterização do Problema

As séries temporais são representações sequenciais de dados ao longo do tempo. Elas estão presentes em diversas áreas do conhecimento e desempenham um papel fundamental em diversas disciplinas. A complexidade presente nesses conjuntos de dados como sazonalidades, tendências e possíveis anomalias desafiam a capacidade de modelos analíticos tradicionais em fornecer predições precisas.

Ao aprofundar na análise de séries temporais, encontra-se a necessidade de compreender e modelar as variáveis ao longo do tempo. Como exemplo, em meteorologia, fenômenos climáticos sazonais apresentam outros tipos de desafios a serem enfrentados, enquanto em finanças, a volatilidade dos mercados e eventos econômicos podem desencadear padrões complexos; já para a medicina, a utilização dos modelos apresenta desafios tais como a necessidade de robustez, interpretabilidade e capacidade de lidar com padrões sutis.

A falta de uma abordagem universal para lidar com a complexidade e heterogeneidade de dados temporais impulsiona a busca por métodos avançados. A detecção e adaptação a mudanças abruptas, a consideração de múltiplas escalas temporais e a interpretação de padrões não lineares são aspectos cruciais que permeiam a caracterização do problema, destacando a necessidade de métodos analíticos mais sofisticados.

Assim, a compreensão aprofundada desses desafios permite não apenas a melhoria das predições, mas também abre portas para a descoberta de padrões intrincados, auxiliando na tomada de decisões mais informadas em diferentes domínios. Essa caracterização procura estabelecer a base para a investigação aprofundada das estratégias analíticas que enfrentam tais desafios nas séries temporais.

1.2 Objetivos e Contribuições

O cerne deste estudo reside na busca pela melhoria substancial na análise de séries temporais, destacando-se quatro objetivos principais:

1. **Comparação de Modelos**
    - Investigar e comparar o desempenho de modelos clássicos com abordagens mais modernas baseadas em aprendizado de máquina.
    - Avaliar a eficácia relativa desses modelos em diferentes cenários de séries temporais, considerando características como sazonalidade, tendências e possíveis irregularidades nos dados.
2. **Exploração de Modelos *Transformers***
    - Investigar a aplicabilidade e eficácia de modelos baseados em *Transformers* na análise de séries temporais.
    - Avaliar como esses modelos lidam com desafios específicos, como a captura de padrões de longo prazo e a adaptação a mudanças repentinas nas séries temporais.
3. **Contribuição para o Conhecimento**
    - Fornecer compreensão aprofundada sobre as nuances da análise de séries temporais, identificando padrões e comportamentos distintos em diferentes contextos.

- Contribuir para o avanço do conhecimento ao oferecer uma perspectiva comparativa entre modelos clássicos, técnicas de aprendizado de máquina e abordagens baseadas em *Transformers*.
- Proporcionar uma compreensão mais refinada dos fatores que impactam a eficácia dos modelos analisados, permitindo aprimoramentos contínuos nas estratégias de predição.

4. **Aplicabilidade Prática**

   - Buscar implicações práticas das descobertas, visando fornecer orientações úteis para profissionais e pesquisadores que lidam com análise de séries temporais em diversos campos.

Esses objetivos não apenas buscam aprimorar a eficiência preditiva, mas também almejam uma compreensão mais profunda das dinâmicas temporais subjacentes. Ao atingir esses objetivos, este estudo pretende contribuir significativamente para o avanço do estado da arte na análise de séries temporais, abrindo caminho para futuras inovações e aprimoramentos nas práticas analíticas.

1.3 Questões de Pesquisa

As questões de pesquisa que se pretende responder são as seguintes:

1. **Comparação de Desempenho entre Modelos Clássicos e de Aprendizado de Máquina**

   - Como o desempenho dos modelos clássicos compara-se aos modelos de aprendizado de máquina, como técnicas de *boosting*, em termos de predições de séries temporais?

2. **Avaliação da Eficácia de Modelos baseados em *Transformers***

   - Qual é a eficácia dos modelos baseados em *Transformers* em relação aos métodos tradicionais e de aprendizado de máquina em diferentes contextos de séries temporais?

3. **Impacto das Características das Séries Temporais no Desempenho dos Modelos**

   - Como diferentes características das séries temporais, como sazonalidade e tendências, afetam a precisão dos modelos analisados?

4. **Identificação de Padrões Específicos**

   - Existem padrões que podem ser identificados ao comparar o desempenho desses modelos em cenários específicos de séries temporais?

As questões de pesquisa formuladas orientam a análise comparativa e visam aprofundar a compreensão dos fatores que influenciam o desempenho dos modelos. Ao abordar essas questões, espera-se fornecer uma base sólida para contribuições significativas no campo da análise de séries temporais.

1.4 Organização do Texto

O presente documento está organizado em seis capítulos. O Capítulo 1 apresenta e caracteriza o problema sendo tratado, as motivações, objetivos, questões de pesquisa principais e as contribuições

esperadas. O Capítulo 2 aborda a literatura relacionada à pesquisa e apresenta os conceitos importantes sobre o tópico investigado. O Capítulo 3 descreve o processo proposto, desde a obtenção dos dados e sua preparação para ser utilizado no problema, apresentação do fluxo de trabalho e testes conduzidos além da validação dos modelos apresentados. O Capítulo 4 descreve os testes já conduzidos até o momento e os resultados obtidos até aqui. O Capítulo 5 apresenta o plano para os trabalhos a serem conduzidos a partir de agora, e um cronograma de execução das atividades. O Capítulo 6 contém observações finais a respeito do projeto de pesquisa.

Capítulo 2

Revisão Bibliográfica

Este capítulo descreve alguns conceitos e trabalhos relacionados ao tema investigado no projeto de pesquisa.

2.1 Conceitos e Técnicas

Nesta seção, serão abordados alguns dos conceitos utilizados ao longo do trabalho e que são essenciais para o entendimento do problema em questão. Serão detalhados os modelos naïve, ARIMA, XGBoost e baseado em *Transformers*, além de uma discussão sobre detecção de anomalias.

2.1.1 Método Naïve

O método naïve de predição para séries temporais é uma abordagem simples e intuitiva para prever valores. Na sua forma mais básica, o método assume que o próximo valor na série temporal será igual ao último valor observado. Ele supõe que não haverá mudança na tendência ou padrão da série temporal e que o valor futuro será simplesmente uma repetição do último valor conhecido.

Embora seja fácil de entender e calcular, o método naïve tem algumas limitações: ele geralmente não leva em conta padrões sazonais, tendências de longo prazo ou variações aleatórias na série temporal. Portanto, sua precisão pode ser bastante limitada, especialmente em séries temporais com padrões complexos.

Apesar de suas limitações, o método ingênuo pode ser útil como uma linha de base simples para comparação com técnicas de predição mais avançadas. Ele pode fornecer uma referência rápida para avaliar o desempenho de modelos mais sofisticados e identificar se eles estão oferecendo melhorias significativas em relação a uma abordagem tão simples.

2.1.2 ARIMA

Um tipo de sinal muito popular no mundo real é um sinal autorregressivo (AR). Um sinal AR refere-se ao valor de uma série temporal, ou seja, para o intervalo de tempo corrente, ele depende dos valores da série temporal em tempos anteriores. Esta correlação serial é uma propriedade chave do sinal AR e é parametrizada por:

- ordem de correlação serial, ou seja, o número de etapas de tempo anteriores do sinal.

- coeficientes para combinar os intervalos de tempo anteriores.

Os modelos de média móvel integrada autoregressiva (ARIMA) são um dos métodos clássicos de predição mais populares. A família de métodos ARIMA depende de autocorrelação (a correlação de y_t com y_{t-1}, y_{t-2} e assim por diante).

Os mais simples da família são os modelos AR(p), que utilizam regressão linear com p intervalos de tempo anteriores ou seja, p *lags*. Segundo Joseph [7], eles são definidos conforme a Equação 2.1:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{2.1}$$

em que ϕ é um fator de amortecimento, c é o intercepto e ε_t é o ruído ou erro no intervalo de tempo t.

Os próximos na família são os modelos MA(q), nos quais, em vez de valores observados no passado, são usados os q passados erros na predição (supondo neles essencialmente o ruído branco) para chegar a uma predição conforme a Equação 2.2:

$$y_t = c + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_p \varepsilon_{t-p} + \varepsilon_t \tag{2.2}$$

em que ε é o ruído branco e c é o intercepto.

Isso normalmente não é empregado individualmente, mas em conjunto com modelos $AR(p)$, o que torna o próximo em nossa lista de modelos ARMA(p, q). Os modelos ARMA são definidos como $y_t = \text{AR}(t) + \text{MA}(t)$.

Em todos os modelos ARMA, existe uma suposição subjacente – a série temporal é estacionária. Há muitas maneiras de tornar a série estacionária, mas tomando a diferença de valores sucessivos é uma dessas técnicas. Isso é conhecido como diferenciação. Às vezes, é necessário realizar a diferenciação uma vez, enquanto outras vezes, é necessário fazer diferenças sucessivas antes da série temporal torna-se estacionária. O número de vezes que realizamos a operação diferencial é chamado de ordem de diferenciação. A letra "I" no método ARIMA, e a peça final do quebra-cabeça, significa *Integrado*. Ele define a ordem de diferenciação que se realizar antes que a série se torne estacionária e é denotada por d.

Portanto, o modelo ARIMA(p, d, q) representa a execução da d-ésima ordem de diferenciação e então considerados os últimos p termos de maneira autoregressiva e, em seguida, incluídos os últimos q termos de média móvel para que se tenha a predição.

O ARIMA padrão é apropriado para modelar séries temporais com tendências e padrões de sazonalidade, entretanto, quando os dados exibem sazonalidade em intervalos regulares, como diariamente ou mensalmente, o SARIMA, definido por Box and Jenkins [2], é considerado mais apropriado.

A principal diferença entre o ARIMA e o SARIMA é a adição de termos sazonais. Enquanto o ARIMA possui termos para modelar tendências e componentes autoregressivos e de média móvel, o SARIMA inclui termos sazonais adicionais para lidar com as flutuações sazonais nos dados.

O modelo SARIMA é definido por quatro conjuntos de parâmetros:

- parâmetros não sazonais (p, d, q): são os mesmos que no ARIMA e são usados para modelar a tendência e a variabilidade não sazonal nos dados.
- parâmetros sazonais (P, D, Q): são semelhantes aos não sazonais, mas são aplicados às diferenças sazonais dos dados.
- parâmetro de frequência sazonal (s): indica a periodicidade dos dados sazonais. Por exemplo, para dados mensais, *s = 12*, para dados trimestrais, *s = 4*.

O modelo SARIMA é então expresso como SARIMA$(p, d, q)(P, D, Q)[s]$, em que *p, d, q* são as ordens não sazonais, *P, D, Q* são as ordens sazonais e s é a periodicidade sazonal. A formulação que representa o modelo é a Equação 2.3:

$$\begin{aligned} Y_t =& \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} + \\ & \Phi_1 Y_{t-s} + \Phi_2 Y_{t-2s} + \ldots + \Phi_P Y_{t-Ps} - \Theta_1 \varepsilon_{t-s} - \Theta_2 \varepsilon_{t-2s} - \ldots - \Theta_Q \varepsilon_{t-Qs} + \varepsilon_t \end{aligned} \tag{2.3}$$

em que:

- Y_t é o valor da série temporal no tempo t.
- $\phi_1, \phi_2, \ldots, \phi_p$ são os coeficientes dos termos autoregressivos não sazonais.
- $\theta_1, \theta_2, \ldots, \theta_q$ são os coeficientes dos termos de média móvel não sazonais.
- $\Phi_1, \Phi_2, \ldots, \Phi_P$ são os coeficientes dos termos autoregressivos sazonais.
- $\Theta_1, \Theta_2, \ldots, \Theta_Q$ são os coeficientes dos termos de média móvel sazonais.
- ε_t é o termo de erro no tempo t.
- s é o período sazonal.

Esta equação representa a estrutura geral do modelo SARIMA. Os termos autoregressivos (ϕ) e de média móvel (θ) capturam a dependência serial e os efeitos de choques anteriores nos dados. Os termos sazonais (Φ e Θ) capturam padrões sazonais específicos.

2.1.3 XGBoost

XGBoost, cunhado por Chen and Guestrin [3], significa *"eXtreme Gradient Boosting"* e é um algoritmo de aprendizado de máquina que se destaca em tarefas de classificação e regressão. Ele pertence à categoria de modelos *ensemble*, que combina vários modelos mais fracos para formar um modelo mais robusto e preciso.

A principal ideia por trás do XGBoost é a técnica de impulsionar árvores de decisão, um método que cria uma série de modelos de árvores de decisão fracos e os combina para formar um modelo mais forte. O processo de impulsionar envolve treinar novos modelos para corrigir os erros dos modelos anteriores, focando nas instâncias mal classificadas.

Algumas características distintivas do XGBoost incluem:

1. **Regularização**: ele incorpora técnicas de regularização para evitar *overfitting*, como a penalização de termos de complexidade do modelo.
2. **Função de perda personalizada**: permite que se defina e otimize a função de perda com base nas características específicas do problema.
3. **Processamento paralelo**: pode ser paralelizado eficientemente, tornando-o rápido e escalável para conjuntos de dados grandes.
4. **Dados ausentes**: pode lidar automaticamente com valores ausentes nos dados de treinamento.

5. **Importância das características**: o XGBoost fornece *insights* sobre a importância das diferentes características (*features*), permitindo que se identifiquem os fatores mais influentes nos dados de séries temporais. Isso ajuda a focar nas variáveis mais relevantes para melhorar a precisão das predições.

2.1.4 Métodos Baseados em *Transformers*

Transformers, inicialmente desenvolvidos para tarefas de processamento de linguagem natural, foram posteriormente adaptados para uma variedade de outras tarefas, incluindo a predição de séries temporais. Os modelos de séries temporais baseados em *Transformers* aproveitam os mecanismos de atenção para capturar relações temporais complexas nos dados.

Algumas das características relevantes sobre *Transformers* para séries temporais incluem:

- **Mecanismos de Atenção:** os *Transformers* usam mecanismos de atenção para capturar dependências entre diferentes partes dos dados de entrada. Isso permite que o modelo pondere adequadamente a importância de diferentes pontos no histórico da série temporal ao fazer predições futuras.
- **Capacidade de Modelagem de Longo Prazo:** os *Transformers* têm uma capacidade inerente de modelar dependências de longo alcance nos dados. Isso é crucial para a predição de séries temporais, onde padrões complexos podem se estender por longos períodos.
- **Camadas de Codificação e Decodificação:** assim como nos modelos *Transformer* usados para tradução de linguagem natural, os modelos de séries temporais baseados em *Transformers* geralmente consistem em camadas de codificação e decodificação. As camadas de codificação processam o histórico da série temporal, enquanto as camadas de decodificação geram as predições futuras.
- **Posições de Codificação:** os *Transformers* requerem informações sobre a posição relativa dos pontos de dados na série temporal para capturar a ordem temporal. Isso é feito adicionando vetores de posição aos dados de entrada, o que permite ao modelo inferir a posição de cada ponto de dados na sequência temporal.
- **Dimensionalidade de Entrada Ajustável:** os modelos de séries temporais baseados em *Transformers* podem acomodar entradas de dimensionalidade variável, permitindo que sejam aplicados a uma ampla gama de problemas de predição de séries temporais.
- **Treinamento e Inicialização**: assim como em outras aplicações de *Transformers*, esses modelos normalmente são treinados usando técnicas de aprendizado supervisionado, onde os dados de entrada são sequências temporais históricas e os rótulos são as sequências temporais futuras que se deseja prever.
- **Arquiteturas Específicas:** existem várias arquiteturas específicas de *Transformer* projetadas para predição de séries temporais, como o TFT de Lim et al. [8], por exemplo. Cada uma das arquiteturas pode ter suas próprias características e vantagens específicas.

Além do TFT, identificou-se o *iTransformer* [11], que é uma arquitetura proposta para predição de séries temporais, sendo ele uma variação do *Transformer*, focando na aprendizagem de representação

e correlação adaptativa entre séries temporais. Ele emprega uma abordagem de *tokenização* das séries temporais para representação das variáveis, aplicando autoatenção para interações mútuas e redes de alimentação para representações das séries.

A predição é realizada por camadas lineares. As características específicas incluem a normalização em camadas para séries individuais, redes de alimentação aplicadas às representações das séries, e autoatenção que considera cada série de variável como um processo independente.

2.1.5 Detecção de Anomalias

A detecção de anomalias em séries temporais desempenha um papel crucial em diversas aplicações: em muitos cenários, a identificação precoce de padrões incomuns ou comportamentos anômalos pode ajudar a prevenir falhas, reduzir custos e melhorar a eficiência dos processos.

A detecção destas anomalias não é trivial, principalmente dada a ambiguidade em torno da definição do que é uma anomalia para um conjunto de dados. Portanto, ter conhecimento do domínio é vital para fazer o julgamento adequado ao detectar valores discrepantes.

As anomalias são geralmente consideradas ocorrências raras – e que podem ser chamadas de contaminação. Pode-se então assumir que uma pequena fração dos dados são discrepantes em um grande conjunto de dados. No entanto, esta complexidade requer métodos concebidos para encontrar padrões nos dados. As técnicas de detecção de valores discrepantes não supervisionados são adequadas para encontrar padrões em ocorrências raras.

O uso de algoritmos como o *Isolation Forest* Liu et al. [9] para detecção de anomalias em séries temporais é particularmente relevante devido a sua eficácia em identificar padrões incomuns em conjuntos de dados complexos e de alta dimensionalidade. O *Isolation Forest* é um método de detecção de anomalias baseado em árvores de decisão que se destaca por sua capacidade de isolar anomalias de forma eficiente, mesmo em conjuntos de dados grandes. Ao identificar observações que são isoladas em relação à maioria dos dados, o *Isolation Forest* pode detectar anomalias de forma rápida e eficaz, fornecendo *insights* valiosos para a tomada de decisões em tempo real. Essa abordagem é particularmente útil em séries temporais, onde padrões anômalos podem ser sutis e variáveis ao longo do tempo, permitindo uma detecção robusta e adaptável de anomalias em diferentes contextos e cenários.

Algumas vantagens que podem ser destacadas do *Isolation Forest* são:

1. **Robustez**: *Isolation Forest* é um método que combina os resultados de múltiplas árvores de isolamento. O escore final de anomalia é calculado pela média dos escores das árvores individuais. Essa abordagem de conjunto ajuda a reduzir os efeitos da aleatoriedade e das variações, tornando o algoritmo mais robusto e confiável. Tipicamente, quanto mais árvores houver, mais estável o algoritmo se torna.

2. **Escalabilidade**: *Isolation Forest* pode ser usado em conjuntos de dados com uma quantidade massiva de dados, pois utiliza uma estratégia de divisão aleatória para criar árvores de isolamento, o que o torna menos sensível ao aumento do tamanho dos dados em comparação com outros algoritmos tradicionais de detecção de anomalias.

3. **Capacidade de lidar com dados de alta dimensão**: *Isolation Forest* pode lidar eficientemente com conjuntos de dados de alta dimensão, pois utiliza seleção aleatória de características em cada divisão, reduzindo o impacto de características irrelevantes ou redundantes e focando nas mais informativas.

4. **Insensibilidade à distribuição dos dados**: *Isolation Forest* não assume nenhuma distribuição específica dos dados. Ele funciona bem com dados que podem exibir padrões complexos e não lineares. É menos afetado por *outliers*, ruído ou presença de características irrelevantes.

5. **Complexidade linear**: além das características de divisão aleatória e seleção aleatória de características, *Isolation Forest* não requer o cálculo de medidas de distância ou densidade entre os pontos de dados como outros métodos de detecção de anomalias, o que pode melhorar significativamente a velocidade e reduzir o custo computacional.

Algumas desvantagens do *Isolation Forest* são:

1. **Sensibilidade ao número de anomalias**: o desempenho do *Isolation Forest* pode ser influenciado pela proporção de anomalias no conjunto de dados. Quando a porcentagem de anomalias é alta, o algoritmo pode ter dificuldade em isolá-las efetivamente e distingui-las dos pontos de dados normais. A sintonia do parâmetro de contaminação (proporção de anomalias) torna-se crucial para obter resultados precisos.

2. **Ineficácia na detecção de anomalias contextuais**: *Isolation Forest* foca na identificação de pontos de dados que são diferentes em termos de valores de atributos. No entanto, pode não ter bom desempenho na detecção de anomalias contextuais que são definidas por suas relações com outros pontos de dados.

3. **Falta de consciência temporal**: *Isolation Forest* trata cada ponto de dados independentemente e não considera explicitamente a ordenação temporal e as dependências em séries temporais. Ele pode não capturar efetivamente anomalias que dependem da natureza sequencial dos dados ou exibem padrões temporais complexos.

2.2 Trabalhos Correlatos

O trabalho de Poggi et al. [14] discutiu uma abordagem para gerar séries temporais de velocidade do vento usando um modelo autoregressivo. O estudo aplicou o modelo a três locais na Córsega e gerou séries temporais sintéticas a cada 3 horas para esses locais. As séries sintéticas foram comparadas com as séries experimentais para verificar a capacidade do modelo em preservar as propriedades estatísticas das séries de velocidade do vento. O objetivo foi criar um gerador de dados para construir um ano de referência para simulação de sistemas de energia eólica na Córsega. A metodologia desenvolvida por Poggi et al. [14] envolveu a modelagem do vento em escala mensal, removendo as variações sazonais e diárias e transformando a série de velocidade do vento em uma variável gaussiana padronizada. Os resultados mostraram que o modelo AR(2) foi capaz de simular de forma precisa as séries de velocidade do vento em intervalos de 3 horas.

Por sua vez, Ho and Xie [6] investigaram o uso de modelos ARIMA na predição de confiabilidade de sistemas reparáveis. A técnica de séries temporais utilizada faz poucas suposições e é flexível, fundamentada estatisticamente e teoricamente sólida, não exigindo a priori de modelos na análise dos dados de falha. O estudo apresenta um exemplo ilustrativo de falhas de um sistema mecânico. Concluise que o modelo ARIMA é uma alternativa viável que apresenta resultados satisfatórios em termos de desempenho preditivo. Neste caso, o uso de modelos ARIMA para análise e predição de falhas em sistemas reparáveis permitiu explorar a autocorrelação nos dados de falha, obtendo estimativas mais precisas.

Já o trabalho de Zhang et al. [20] abordou a análise de predição de séries temporais de volume de vendas usando o algoritmo XGBoost. O problema abordado é a demanda inadequada de pessoal no setor de varejo, causada pela complexidade e dinamicidade do ambiente de mercado. A metodologia utilizada nesse estudo envolve a aplicação do algoritmo XGBoost, que é uma implementação do *Gradient Boosting*.

A aplicação da técnica XGBoost para a análise de séries temporais e a comparação do desempenho do modelo proposto com outros modelos de referência são relevantes para o estudo de predições de séries temporais no contexto, pois fornecem uma abordagem mais precisa e eficaz para o dimensionamento adequado das equipes.

Um estudo que leva em conta os métodos tanto de ARIMA quanto de XGBoost é o de Wang and Guo [18]. O trabalho propõe um modelo híbrido para predição de preços de ações. A metodologia envolve a decomposição dos dados em partes aproximadas e de erro usando a transformada *wavelet* discreta (DWT), a aplicação do modelo ARIMA nas partes aproximadas e do modelo XGBoost nas partes de erro. Os resultados das predições são combinados por meio da reconstrução wavelet. Os experimentos mostraram que o modelo híbrido obteve erros menores do que outros modelos individuais, como ARIMA e XGBoost. O modelo proposto também demonstrou habilidade de aproximação e generalização, além de se ajustar bem aos preços de abertura de índices de ações. A combinação de diferentes modelos contribuiu para a redução de erros de predição e a seleção adequada de parâmetros do modelo híbrido foi uma questão de pesquisa importante. Além disso, o uso do XGBoost para prever a parte de erro de dados de estoque explorou as características não-lineares presentes nessa parte, o que melhorou a precisão da predição.

Lv et al. [12] apresentaram uma análise de séries temporais da febre hemorrágica com síndrome renal (HFRS) na China, utilizando modelos de predição baseados em ARIMA e XGBoost. O estudo propõe uma estratégia de predição de múltiplos passos baseada no algoritmo XGBoost, comparando sua precisão de ajuste e predição com o modelo ARIMA. Os dados de incidência de HFRS foram coletados de 2004 a 2018 e divididos em conjuntos de treinamento e teste. O modelo XGBoost mostrou uma precisão de predição significativamente melhor do que o modelo ARIMA, especialmente na predição de dados complexos e não lineares como a HFRS. Além disso, o estudo destaca a importância dos modelos de predição de múltiplos passos em doenças infecciosas. As descobertas contribuem para o trabalho que será apresentado, fornecendo uma metodologia robusta e comparativa para análise de predição de séries temporais de doenças infecciosas, com ênfase na aplicação dos modelos ARIMA e XGBoost.

Por sua vez, Rahman et al. [15] apresentaram um estudo que compara a precisão preditiva dos modelos ARIMA e XGBoost na análise de séries temporais relacionadas à incidência de casos confirmados e mortes por COVID-19 em Bangladesh. O estudo tem como objetivo modelar a tendência geral dos casos confirmados e mortes, gerar predições de curto prazo e comparar a acurácia preditiva dos dois modelos. Os modelos ARIMA e XGBoost foram estabelecidos usando os dados de treinamento, e os conjuntos de teste foram utilizados para avaliar a capacidade de predição de cada modelo. Foram utilizadas medidas de erro médio absoluto, erro médio percentual, erro quadrático médio e erro percentual absoluto médio para avaliar a acurácia dos modelos. O modelo ARIMA apresentou um desempenho melhor do que o modelo XGBoost na predição de casos confirmados e mortes por COVID-19 em Bangladesh. Esses resultados sugerem que o modelo ARIMA pode desempenhar um papel crítico na estimativa da propagação de uma pandemia em Bangladesh e em países semelhantes.

Lim et al. [8] apresentaram o *Temporal Fusion Transformer* (TFT), um modelo de aprendizado profundo baseado em atenção para predição de séries temporais multi-horizonte com alta eficácia e

interpretabilidade. Ele utiliza componentes especializados, como mecanismos de seleção de variáveis, e outros codificadores para capturar relacionamentos temporais em diferentes escalas de tempo. Além disso, o modelo é capaz de identificar variáveis relevantes, padrões temporais persistentes e eventos significativos, tornando-se uma ferramenta poderosa para predição de séries temporais em uma variedade de cenários. Ele traz melhorias significativas de desempenho em comparação com *benchmarks* existentes e conclusões valiosas para interpretabilidade de predições de séries temporais.

O trabalho de Liu et al. [10] abordou a questão da não estacionariedade em séries temporais e propõe um novo método chamado *Non-stationary Transformers* para melhorar a capacidade de predição de séries temporais. A abordagem consiste em duas partes interdependentes: *Series Stationarization*, que normaliza as estatísticas de cada série de entrada para melhorar a previsibilidade, e *De-stationary Attention*, que recupera as informações não estacionárias inerentes nas dependências temporais. Através desses componentes, os *Non-stationary Transformers* conseguem melhorar a previsibilidade dos dados e manter a capacidade do modelo ao mesmo tempo. As contribuições do estudo incluem a refinamento da importância da capacidade preditiva de séries não estacionárias, a proposta de uma abordagem genérica para lidar com a estacionariedade das séries e evitar a *over-stationarization*, e a melhoria significativa no desempenho de quatro variantes do *Transformer* em seis *benchmarks* de séries temporais. Essas descobertas são relevantes para o projeto em questão, pois fornecem uma metodologia para lidar com a não estacionariedade em predições de séries temporais, potencialmente aprimorando a precisão e a robustez do modelo de predição a ser desenvolvido.

Woo et al. [19] desenvolveram uma nova arquitetura de *Transformer*, chamada *ETSformer*, para predição de séries temporais, que combina os princípios do alisamento exponencial com redes *Transformers*. A metodologia incorpora camadas de decomposição de nível, crescimento e sazonalidade para extrair padrões temporais relevantes. Por meio dos mecanismos de Atenção de Alisamento Exponencial e de Frequência, o modelo consegue melhorar a eficiência e a eficácia na modelagem de dependências temporais, além de oferecer interpretabilidade aos resultados. As contribuições do *ETSformer* incluem uma abordagem mais eficaz para lidar com dados de séries temporais, com resultados superiores a outras abordagens em diversos conjuntos de dados reais.

O trabalho de Liu et al. [11], que abordou o *iTransformer*, coloca a questão do recente aumento de modelos de predição, questionando a busca contínua por modificações arquiteturais baseadas em *Transformers*. Os autores propõem o *iTransformer*, que inverte a estrutura do *Transformer* sem modificar seus módulos nativos. O *iTransformer* considera séries independentes como *tokens* variados para capturar correlações multivariadas por meio de atenção e utiliza normalização de camada e redes *feed-forward* para aprender representações de séries. Experimentos mostram que o *iTransformer* alcança desempenho do estado da arte e exibe notável generalidade de estrutura, com análises promissoras.

Capítulo 3

Material e Métodos

Este capítulo apresenta uma descrição detalhada dos materiais e métodos utilizados na condução da pesquisa, destacando as fontes de dados e os procedimentos adotados para analisar as variáveis de interesse. A Figura 3.1 ilustra a metodologia utilizada na pesquisa.

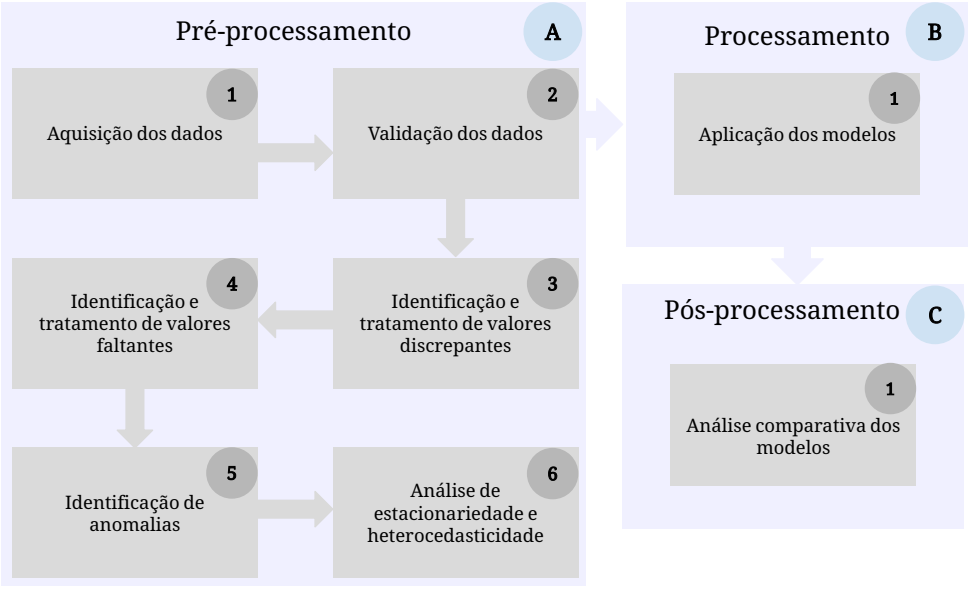

Figura 3.1: Diagrama das etapas da metodologia do projeto. O processo de treinamento e de teste deverá ser repetido para cada modelo avaliado.

3.1 Metodologia

A metodologia proposta no presente trabalho é apresentada na Figura 3.1. Ela está estruturada em três grandes partes a saber:

1. **Pré-processamento:** compreende a aquisição dos dados, e um conjunto de etapas para torná-los adequados para análise. Isso inclui validação, identificação de valores discrepantes, identificação de dados ausentes, identificação de anomalias e análise de estacionariedade e heterocedasticidade da série temporal.

2. **Processamento:** os dados pré-processados são processados para extrair informações úteis. Isso envolverá a aplicação de algoritmos de aprendizado de máquina, técnicas estatísticas e abordagens para identificação de padrões.

3. **Pós-processamento:** após o processamento dos dados, os resultados serão interpretados e a agregação de resultados será feita em forma de relatórios e visualizações para finalizar a análise e aplicação dos resultados.

3.2 Coleta e Preparação dos Dados

Aqui será feita a descrição dos conjuntos de dados a serem trabalhados, forma de obtenção e pré-processamentos.

3.2.1 Obtenção dos Dados

Para obtenção de dados, diferentes origens e formas de recuperá-los foram utilizadas. A seguir, para cada conjunto de dados, considerando dados meteorológicos e dados financeiros, é descrita a forma com que cada um deles foi coletado, bem como uma breve caracterização dos conjuntos de dados.

Dados Meteorológicos

Para os dados meteorológicos, este trabalho utiliza dados do CEPAGRI. Para obtenção dos dados enviou-se uma comunicação por *e-mail* e após definição de limites de datas e formalização, foi enviado um arquivo contendo os dados, em intervalos te 10 em 10 minutos.

A coleta de dados baseou-se em informações provenientes de estações meteorológicas distribuídas na região do campus em Campinas. Uma sumarização dos dados recebidos do é mostrada na Tabela 3.1. O período de coleta e a taxa de amostragem dos dados foram:

- data inicial: 01/jan/2013.
- data final: 31/dez/2023.
- frequência: a cada 10 minutos.

As informações dadas a respeito da base de dados fornecida pelo CEPAGRI no momento de disponibilização da mesma foram as seguintes:

- Os dados não passaram por testes de consistência e detecção de erros.
- Sensores defeituosos `vento a 5m` de nov/21 a fev/22 e `umidade` de abr/23 a out/23, e parte dos dois primeiros anos de `vento a 2m`) tiveram seus dados removidos pelo pesquisador do CEPAGRI.
- Mudança da estação do pátio da FEAGRI para o Museu Exploratório de Ciências em 14/05/2013 - S 22°48'56"-W47°03'28" - Altitude 664m. O pesquisador notou que houve uma mudança somente na intensidade dos ventos, pois o local atual é mais elevado; isso ocorreu no começo de 2013.
- Dados ausentes devem estar rotulados com um "-6999".

Tabela 3.1: Dados recebidos do CEPAGRI.

	Média	Desvio Padrão	Min	25%	50%	75%	Max
Ano	2018.004836	3.207171	2013.0	2015.0	2018.0	2021.0	2023.0
Dia Juliano	180.465257	105.438971	1.0	88.0	181.0	270.0	366.0
HoraMinuto	1191.507207	692.785919	0.0	600.0	1200.0	1800.0	2400.0
Vento a 5m (m/s)	2.537840	1.464628	0.0	1.477	2.16	3.344	12.78
Vento a 2m (m/s)	1.846405	1.001328	0.0	1.108	1.6395	2.438	10.14
Direção do vento (º)	161.072593	88.644614	-25.84	105.4	158.9	219.0	462.2
Umidade relativa (%)	70.146111	18.697351	5.689	56.76	74.0	84.8	100.0
Temperatura do ar (ºC)	21.765935	21.485673	-6999.0	18.4	21.32	25.33	38.95
Índice UV	-0.195043	104.625650	-7999.0	-0.035	0.027	1.787	14.52
Pressão atmosférica (hPa)	939.449850	111.889829	-6999.0	937.0	941.0	945.0	1433.0
Chuva (mm/10min)	0.023543	0.327608	0.0	0.0	0.0	0.0	111.0
Tmáx instantânea	21.906237	11.720936	-6999.0	18.42	21.36	25.52	39.4

- As medições registradas são resultado de uma média, a cada 10 minutos, de amostragens em intervalos de 10 segundos (60 valores pra compor a média) - exceto direção do vento, que era um valor instantâneo e chuva, que é o total no intervalo.
- Em 2015, começaram a medir máximos e mínimos instantâneos de temperatura (ou seja o maior valor dos 60 coletados nos intervalos de 10 min).
- A umidade relativa perde a precisão desde 2017 e a recupera em outubro de 2023, pois houve a inserção de um fator de correção desnecessário. Os dados mas ficam um pouco abaixo (entre 7 a 10% no valor da umidade relativa real).
- A pressão atmosférica tem uma precisão menor do que o ideal para o pesquisador, sendo esta uma característico do sensor.
- A direção do vento vai de 0 a 360º. Qualquer valor diferente disso será desconsiderado.
- O índice UV (A e B) tem uma escala que vai de 0 a 16. Qualquer valor diferente disso será desconsiderado.
- A pressão atmosférica será considerada em uma faixa de 930 a 960 hPa. Qualquer valor diferente disso será desconsiderado.

Economia

Para séries temporais de dados econômicos, dois conjuntos de dados foram selecionados para este trabalho: índice da bolsa de valores de São Paulo – Ibovespa e o índice de fundos de investimentos imobiliários – IFIX.

O Ibovespa é o principal indicador do desempenho médio das cotações das ações listadas na B3 (Brasil, Bolsa, Balcão), a bolsa de valores do Brasil. Ele reflete o desempenho das empresas mais negociadas e representativas do mercado acionário brasileiro.

O cálculo do Ibovespa é feito a partir da variação dos preços dos papéis das empresas que compõem este índice. A variação percentual dos preços das ações em relação a esse valor-base é o que determina os pontos do Ibovespa.

O índice Ibovespa mostra o desempenho médio do mercado acionário brasileiro e serve como referência para analistas financeiros, investidores e gestores de fundos.

Para coleta dos dados, utilizou-se a biblioteca *yfinance*[1] do Python que recupera dados do site *Yahoo! Finance*[2], onde selecionamos os limites de datas.

Dados sumarizados da coleta do índice podem ser vistos na Tabela 3.2. O período de coleta e a taxa de amostragem dos dados foram:

- data inicial: 27/abr/1993.
- data final: 31/jan/2024.
- frequência: diária.

Tabela 3.2: Dados do Ibovespa coletados a partir da biblioteca `yfinance` da linguagem Python.

	Média	Desvio Padrão	Min	25%	50%	75%	Max
Open	47034.65	35845.03	23.70	12813.5	48059.0	66707.5	134194.0
High	47514.49	36145.62	24.20	12957.5	48824.0	67274.5	134392.0
Low	46553.23	35553.88	23.70	12620.5	47493.0	66122.5	133832.0
Close	47046.00	35856.74	23.70	12754.5	48074.0	66717.5	134194.0
Volume	6099007.37	19919283.63	0.0	0.0	1658000.0	4176250.0	232265300.0
Dividends	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Stock Splits	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Já o índice IFIX é um indicador que tem como objetivo medir o desempenho médio dos Fundos de Investimento Imobiliário (FIIs) listados na B3. Diferentemente do Ibovespa, que reflete o desempenho das ações, o IFIX tem foco nos fundos imobiliários. Os fundos imobiliários são veículos de investimento coletivo cujo objetivo principal é aplicar recursos em ativos relacionados ao mercado imobiliário.

O cálculo considera a variação dos preços das cotas dos fundos e, similar ao Ibovespa, é baseado em um valor inicial, conhecido como "pontos". A variação percentual dos preços em relação a esse valor-base determina a pontuação do índice.

Os dados do IFIX foram extraídos do site `infomoney`[3], cujas principais características, listadas na Tabela 3.3. O período de coleta e a taxa de amostragem dos dados foram:

- data inicial: 14/fev/2022.
- data final: 09/fev/2023.
- frequência: diária.

[1] `https://github.com/ranaroussi/yfinance`
[2] `https://finance.yahoo.com/`
[3] https://infomoney.com.br

Tabela 3.3: Dados do IFIX coletados do site `infomoney`.

Índice	Contagem	Média	Desvio Padrão	Min	25%	50%	75%	Max
ABERTURA	495.0	2.968	0.189	2.705	2.802	2.921	3.176	3.341
FECHAMENTO	497.0	2.969	0.190	2.705	2.802	2.921	3.177	3.352
MÍNIMO	497.0	2.963	0.189	2.698	2.799	2.915	3.174	3.340
MÁXIMO	497.0	2.975	0.189	2.715	2.809	2.928	3.183	3.354

Medicina

Reconhecer a importância de contextos médicos e validar dados de séries temporais nos modelos propostos é importante na medida em que a confiabilidade das predições podem ter um impacto direto na saúde e no bem-estar dos pacientes.

A proposta é realizar a aplicação dos modelos desenvolvidos neste estudo em séries temporais médicas, como os eletrocardiogramas (ECG). Esta será conduzida futuramente, uma vez que neste momento ainda não foram identificadas bases de dados consistentes para a análise considerando os modelos e a frequência de coleta de dados.

3.2.2 Pré-Processamento

Para o pré-processamento, será feita uma análise inicial dos dados para exploração, análise e catalogação dos dados, conforme item A.1 da Figura 3.1.

Em seguida, serão identificados e tratados valores limites possíveis dos dados (por exemplo, uma marcação inconsistente de temperatura negativa em Campinas, ou um valor de vento negativo). Para este caso, os dados serão tratados como nulos. O indicativo desta etapa encontra-se no item A.2 da Figura 3.1.

Na etapa seguinte, será realizada a análise dos dados faltantes, bem como o seu tratamento. Algumas possíveis formas de tratamento podem ser a remoção destes dados, ou o preenchimento com o último valor válido da série, a média dos valores da série ou o preenchimento com o primeiro valor válido, após os dados inconsistentes. Esta etapa está ilustrada no item A.3 da Figura 3.1.

Em seguida, será realizada a análise de estacionariedade e heterocedasticidade das séries temporais de modo a entender como tratá-las antes de executar o processamento dos dados propriamente dito. Este é o item A.4 da Figura 3.1.

A detecção de anomalias nas séries temporais está representada no item A.5 da Figura 3.1. Tal identificação tem por objetivo encontrar observações que desviam significativamente do comportamento típico dos dados coletados. Para o trabalho corrente, utilizou-se o método de detecção *Isolation Forest* descrito na Seção 2.1.5 e com um valor de contaminação de 0.05. Potencialmente, pode-se tratar as anomalias identificadas removendo-as e interpolando os valores anteriores e posteriores. Para o presente trabalho, apenas identificou-se e caracterizou-se os conjuntos de anomalias.

Já a etapa de identificação de estacionariedade e heterosedasticidade é representada pela Etapa A.6. A avaliação da estacionariedade é importante uma vez que com a garantia da estacionariedade da série, é possível desenvolver modelos mais robustos. Também no contexto de uma série temporal, afirmar que ela heteroscedástica significa dizer que a variabilidade ou dispersão da série temporal varia com o tempo. A suposição de não heteroscedasticidade é importante em termos de modelo de regressão para trazer robustez a ele.

3.3 Processamento

Para o processamento (etapa B.1 da Figura 3.1), serão aplicados os modelos ARIMA com parâmetros $p, d,$ e q distintos de modo identificar qual a melhor combinação para o modelo. Também será utilizado o modelo XGBoost a partir do mesmo conjunto de treinos e realizando uma busca em grade (*grid search*) para identificar os melhores parâmetros para a predição da série temporal.

Por fim, será utilizada uma técnica de predição de série temporal baseada em *Transformers*. O presente trabalho não trata deste tema neste momento, mas a proposta de utilizar esta técnica vem do fato de os *Transformers* terem, entre outros, o mecanismo de atenção, que será de grande valor para a análise futura.

3.4 Pós-Processamento e Métricas de Avaliação

Na etapa C.1 da Figura 3.1, há o pós-processamento com a análise comparativa dos modelos. Para o trabalho corrente, serão utilizadas as métricas a seguir:

- **Erro Médio Absoluto**: esta métrica (do inglês, *Mean Absolute Error* - MAE) considera o erro entre a predição no tempo $t(f_t)$ e o valor observado no tempo $t(y_t)$. Segundo Joseph [7], a métrica é calculada conforme a Equação 3.1:

$$\text{MAE} = \frac{1}{N \times L} \times \sum_{N}^{i} \sum_{L}^{j} |f_{i,j} - y_{i,j}| \tag{3.1}$$

  em que N é o número de séries temporais, L é a duração da série temporal e f são os valores previstos e y são os valores observados.

- **Erro Quadrático Médio**: esta métrica (do inglês, *Mean Squared Error* - MSE) é a média do erro quadrático entre o tempo (f_t) predito e o tempo (y_t) observado. Segundo Joseph [7], a métrica é calculada conforme a Equação 3.2:

$$\text{MSE} = \frac{1}{N \times L} \times \sum_{N}^{i} \sum_{L}^{j} (f_{i,j} - y_{i,j})^2 \tag{3.2}$$

  em que N é o número de séries temporais, L é a duração da série temporal e f são os valores preditos e y são os valores observados.

- **Erro Médio Absoluto Escalado**: esta métrica (do inglês, *Mean Absolute Scaled Error* - MASE) é uma medida de predição de desempenho usada para comparar o erro médio absoluto do modelo definido com o erro médio do método naïve. Valores de MASE inferiores a 1 indicam que o modelo em questão está superando o modelo de referência, enquanto valores superiores a 1 indicam que o modelo está tendo um desempenho pior do que o modelo de referência. O cálculo,

apresentado em [7] é dado pela Equação 3.3:

$$\text{MASE} = \frac{\frac{1}{L} \times \sum_{i}^{L} |f_i - y_i|}{\frac{1}{L-1} \times \sum_{j=2}^{L} |y_j - y_{j-1}|} \tag{3.3}$$

em que L é a duração da série temporal e f são os valores preditos e y são os valores observados.

- **Forecast Bias**: FB ou viés de predição permite um entendimento do viés do modelo, ou seja, ajuda a entender se a predição está acima ou abaixo dos valores reais. Ela é calculado como a soma dos valores observados, como uma percentagem, sobre a soma dos valores reais. Segundo Joseph [7], a métrica é calculada pela Equação 3.4:

$$\text{FB} = \frac{\sum_{i}^{N}\sum_{j}^{L} f_{i,j} - \sum_{i}^{N}\sum_{j}^{L} y_{i,j}}{\sum_{i}^{N}\sum_{j}^{L} y_{i,j}} \tag{3.4}$$

  em que N é o número de séries temporais, L é a duração da série temporal e f são os valores preditos e y são os valores observados.

  Se o resultado for positivo, o modelo está subestimando os valores reais. Se for negativo, o modelo está superestimando os valores reais. Um viés próximo de zero indica que o modelo faz predições sem tendências sistemáticas.

- **sMAPE**: *Symmetric Mean Absolute Percent Error*, proposto por Armstrong [1], é uma medida de baseada em erros percentuais, e definida conforme a Equação 3.5:

$$\text{SMAPE} = \frac{100}{n} \sum_{i=1}^{n} \frac{|f_i - y_i|}{\frac{|f_i| + |y_i|}{2}} \tag{3.5}$$

  em que f são os valores preditos, y são os valores observados e n é o número de observações da série temporal.

  Ele é especialmente útil em contextos onde as observações possuem variações significativas ou quando é necessário dar igual peso tanto a valores baixos quanto a valores altos. Além disso, o sMAPE tem a característica de ser simétrico, o que significa que não favorece sub ou superestimações. Isso é especialmente útil em situações onde é desejável penalizar tanto os erros positivos quanto os negativos de forma equitativa.

3.5 Modelos

Para os modelos já descritos neste trabalho, executaremos as etapas descritas na Seção 3.1 com os diversos conjuntos de dados descritos na Seção 3.2.1.

A utilização das técnicas de modelos clássicos, baseados em aprendizado de máquina tradicional e baseados em *Transformers* terá como finalidade comparar o desempenho do tipo de modelo aplicado a cada um dos diferentes conjuntos de dados. Com isso, pretende-se compreender quais são as melhores formas de abordar cada tipo de problema.

3.6 Recursos Computacionais

Nesta seção, os materiais e os recursos computacionais que serão utilizados ao longo do trabalho são elencados, incluindo *hardware* e *software* de terceiros.

A implementação do projeto será feita majoritariamente na linguagem de Programação Python, na sua versão 3.11. Entretanto, outras linguagens (por exemplo, R) e versões podem ser utilizadas, de acordo com a necessidade.

A linguagem Python é bastante difundida em termos de análise de problemas de aprendizado de máquina e reconhecimento de padrões, existindo uma vasta quantidade de pacotes e referências para o problema de séries temporais. Entre as principais bibliotecas que serão utilizadas, destacam-se: Pandas[4], NumPy[5], scikit-learn[6], yfinance[7], Darts[8], XGBoost[9], Matplotlib[10] e PyTorch[11].

Para o ambiente de execução dos experimentos, atualmente há o uso da plataforma *Google Colaboratory*[12], que disponibiliza unidades gráficas de processamento (*Graphics Processing Units* - GPUs) NVIDIA A100, NVIDIA V100 e NVIDIA T4.

[4] https://pandas.pydata.org/
[5] https://numpy.org/
[6] https://scikit-learn.org/
[7] https://github.com/ranaroussi/yfinance
[8] https://unit8co.github.io/darts/
[9] https://xgboost.readthedocs.io/
[10] https://matplotlib.org/
[11] https://pytorch.org/
[12] https://colab.research.google.com/

Capítulo 4

Resultados Preliminares

Até o momento da elaboração deste documento, alguns resultados foram obtidos, considerando-se os conjuntos de dados do CEPAGRI, Ibovespa e IFIX. Para todos os casos, utilizou-se um recorte de 200 dias para treinamento e 7 dias para teste. A data de corte para início dos dados de teste foi o dia 02/out/2023, escolhida arbitrariamente.

4.1 Dados Meteorológicos

Para os dados meteorológicos, utilizou-se informações vindas do CEPAGRI. Os resultados preliminares encontram-se a seguir, com a identificação das anomalias e também os valores das métricas de avaliação.

4.1.1 CEPAGRI

Os dados do CEPAGRI foram processados e divididos em conjuntos de treinamento e teste. Para o treinamento utilizou-se 210 registros e para o teste, 30. No caso do ARIMA, utilizou-se um autoARIMA com sazonalidade. Para o caso do XGBoost, utilizou-se uma busca em grade (*grid search*) para identificar os melhores parâmetros e aplicá-los no modelo. Com isso, os resultados são apresentados na Tabela 4.1.

Tabela 4.1: Resultados preliminares a partir dos dados do CEPAGRI.

	MSE	MAE	MASE	Forecast Bias	sMAPE
ARIMA	6.9944	2.1269	1.348	8.3619	9.3815
XGBoost	0.9815	0.838	0.5207	3.0751	3.5848

Os dados de anomalias detectadas considerando *Isolation Forest* com contaminação de 5% estão ilustrados na Figura 4.1. Os dados de treinamento e teste, bem como os resultados do modelo considerando ARIMA são apresentados na Figura 4.2. Para o XGBoost, treinamento, teste e resultado do modelo são apresentados na Figura 4.3.

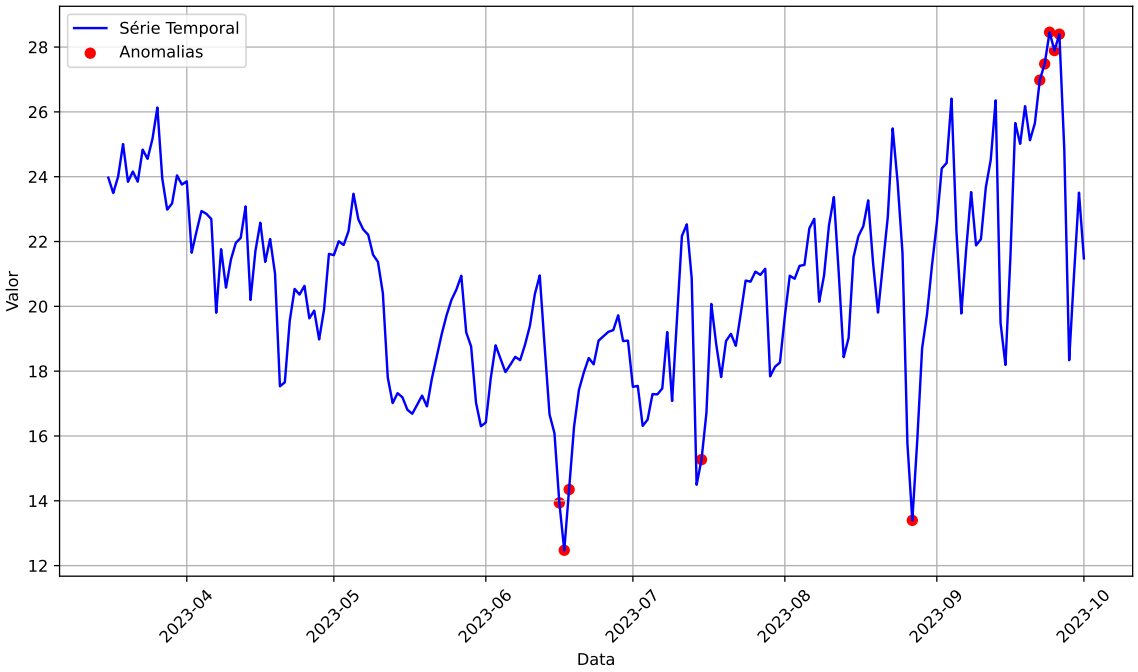

Figura 4.1: Gráfico de valores do CEPAGRI de teste e anomalias detectadas utilizando *Isolation Forest*.

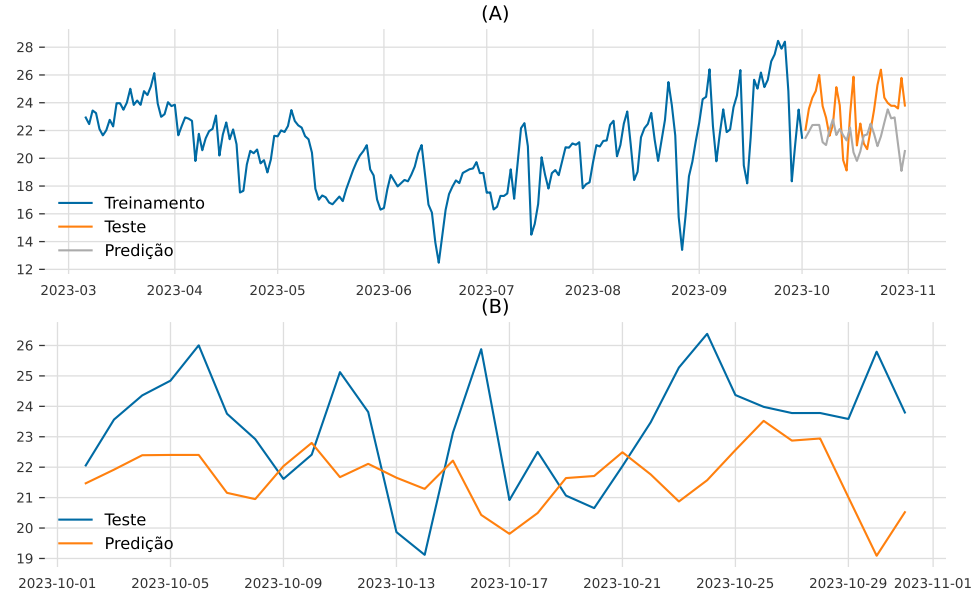

Figura 4.2: Gráfico de valores do CEPAGRI e predição do modelo utilizando ARIMA. Em (A), uma visão geral do recorte do conjunto de dados. Em (B), uma visão dos dados de teste e a predição.

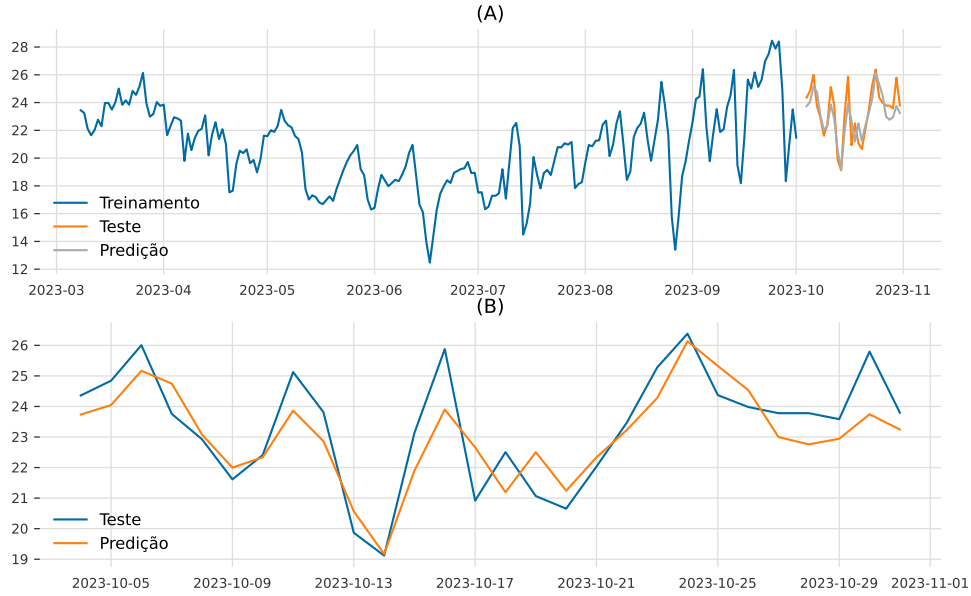

Figura 4.3: Gráfico de valores do CEPAGRI e predição do modelo utilizando XGBoost. Em (A), uma visão geral do recorte do conjunto de dados. Em (B), uma visão dos dados de treinamento e a predição.

4.2 Dados Financeiros

Para os dados financeiros, considerando Ibovespa e IFIX, os resultados preliminares encontram-se a seguir, com a identificação das anomalias e também as métricas de avaliação.

4.2.1 Ibovespa

Para o caso do Ibovespa, utilizou-se o método ARIMA e o XGBoost. Claramente a série é não-estacionária e isso traz desafios ainda maiores para a estruturação de estacionariedade, bem como a definição correta de parâmetros dos modelos.

Os dados de anomalias detectadas considerando *Isolation Forest* com contaminação de 5% são ilustrados na Figura 4.4. Os dados de treinamento e teste, bem como os resultados do modelo considerando ARIMA são apresentados na Figura 4.5. Para o XGBoost, treinamento, teste e resultado do modelo são apresentados na Figura 4.6.

Tabela 4.2: Resultados preliminares a partir dos dados do Ibovespa.

	MSE	MAE	MASE	Forecast Bias	sMPAE
ARIMA	1.0161×10^7	2.4187×10^3	2.3640	0.675	2.082
XGBoost	4.8792×10^5	5.5212×10^2	0.5216	-0.0011	0.4726

O uso de ambos os modelos carecem de uma análise mais profunda, de modo a identificar possíveis lacunas que precisam ser abordadas para que os modelos tenham maior robustez. Um outro ponto a ser considerado é o de que como a linha de tempo é relativamente grande e com cenários econômicos

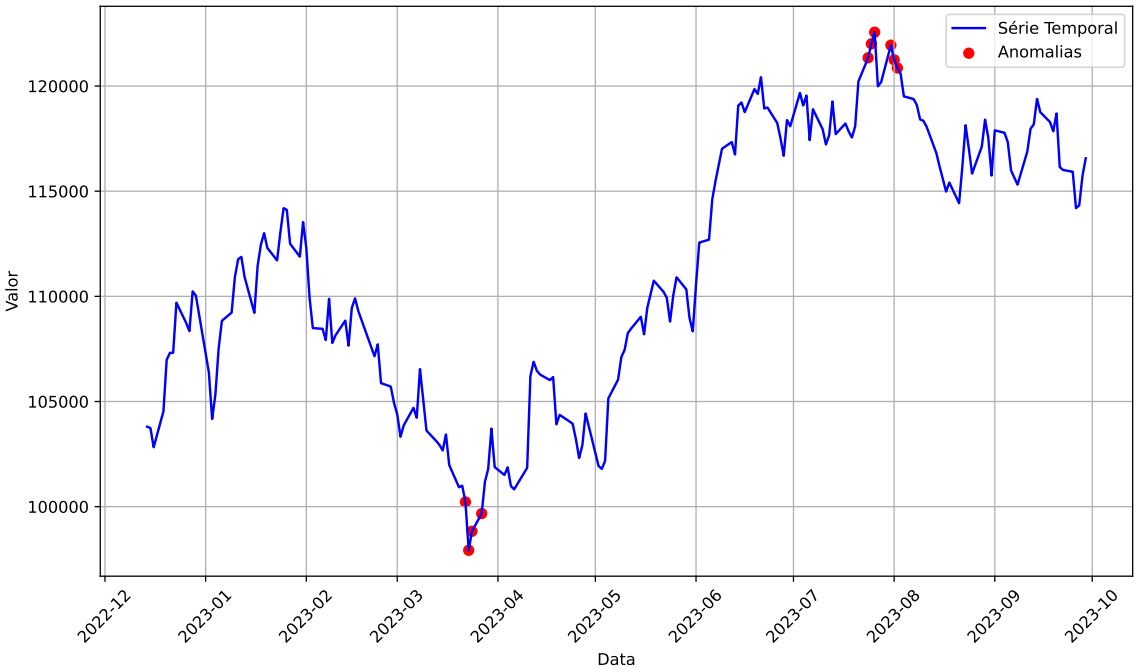

Figura 4.4: Gráfico de valores do Ibovespa de teste e anomalias detectadas utilizando *Isolation Forest*.

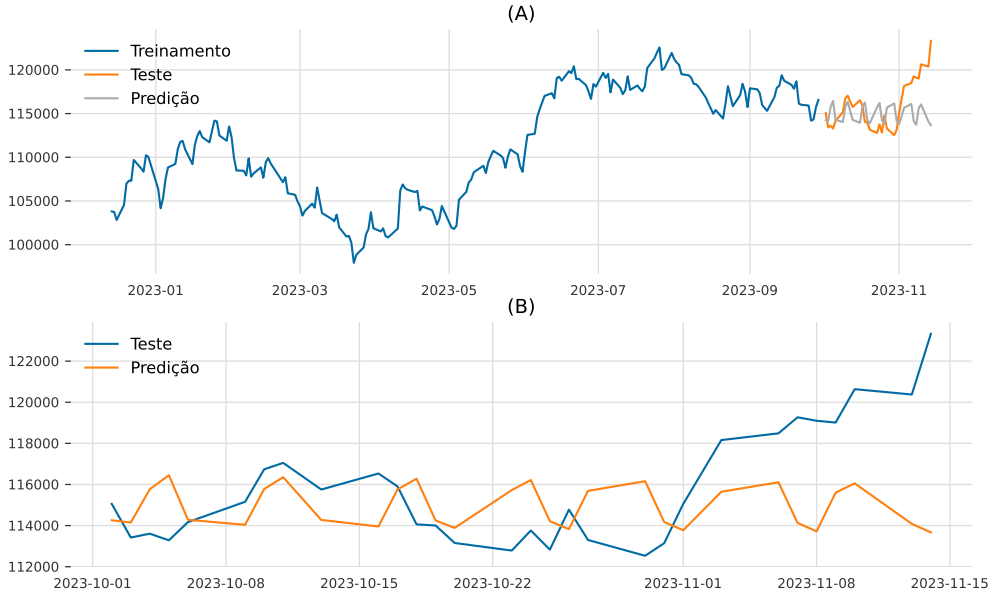

Figura 4.5: Gráfico de valores do Ibovespa e predição do modelo utilizando ARIMA. Em (A), uma visão geral do recorte do conjunto de dados. Em (B), uma visão dos dados de treinamento e a predição.

diversos ao longo deste tempo.

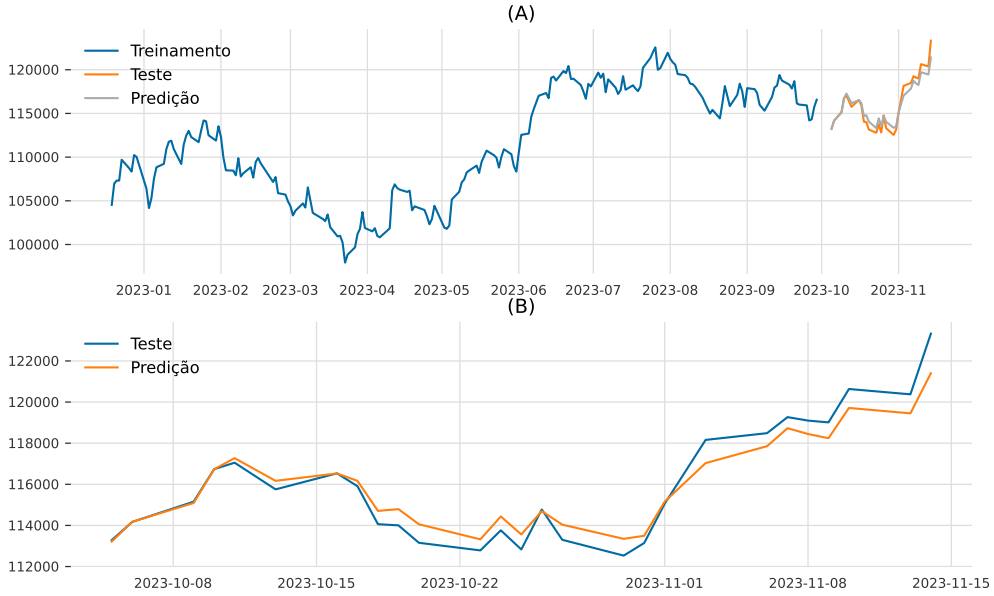

Figura 4.6: Gráfico de valores do Ibovespa e predição do modelo utilizando XGBoost. Em (A), uma visão geral do recorte do conjunto de dados. Em (B), uma visão dos dados de teste e a predição.

4.2.2 IFIX

Igualmente para o caso do IFIX, como no Ibovespa, a série não é estacionária. A Tabela 4.3 mostra o comparativo dos dois modelos aplicados ao conjunto de dados. Os dados de anomalias detectadas considerando *Isolation Forest* com contaminação de 5% estão ilustrados na Figura 4.7. Os dados de treinamento e teste, bem como os resultados do modelo considerando ARIMA, são apresentados na Figura 4.8. Para o XGBoost, o treinamento, o teste e o resultado do modelo são apresentados na Figura 4.9.

Tabela 4.3: Resultados preliminares a partir dos dados do IFIX.

	MSE	MAE	MASE	Forecast Bias	sMAPE
ARIMA	0.0051	0.0689	12.4177	-2.0604	2.1502
XGBoost	0.0001	0.0079	1.3961	0.5124	0.2492

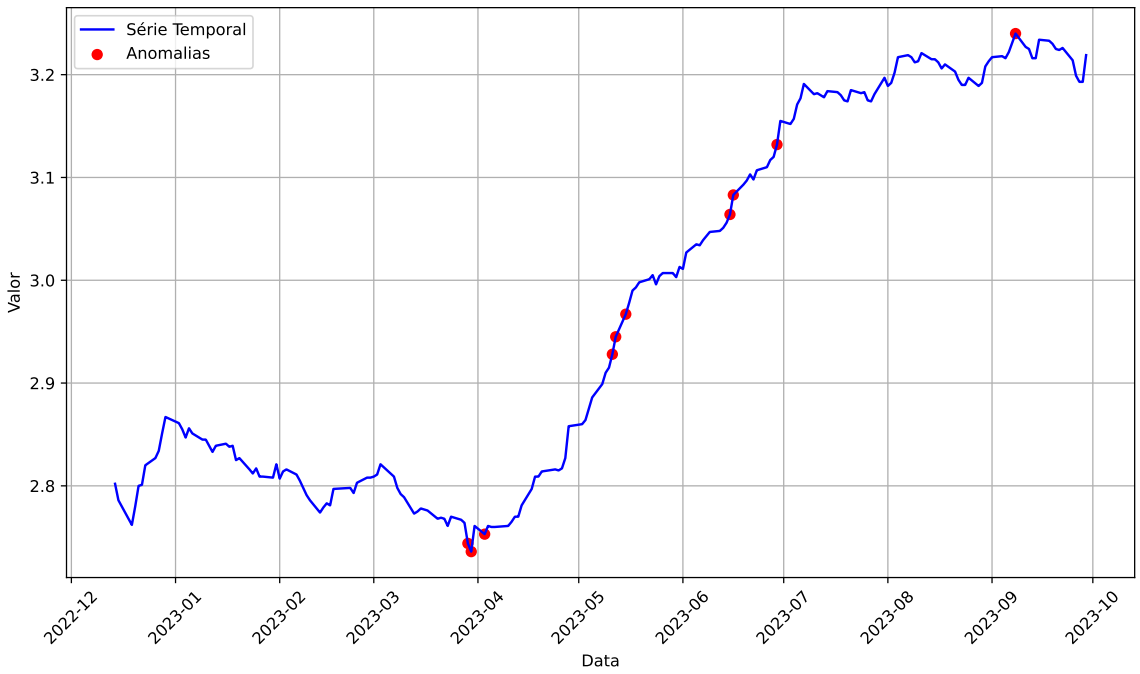

Figura 4.7: Gráfico de valores do IFIX de teste e anomalias detectadas utilizando Isolation Forest

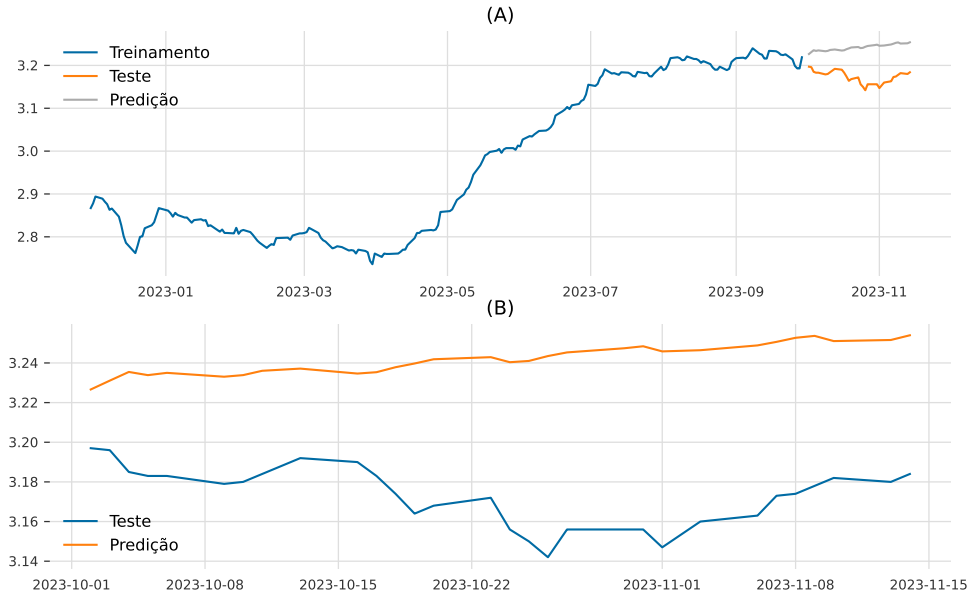

Figura 4.8: Gráfico de valores do IFIX e predição do modelo ARIMA. Em (A), uma visão geral do recorte do conjunto de dados. Em (B), uma visão dos dados de teste e a predição.

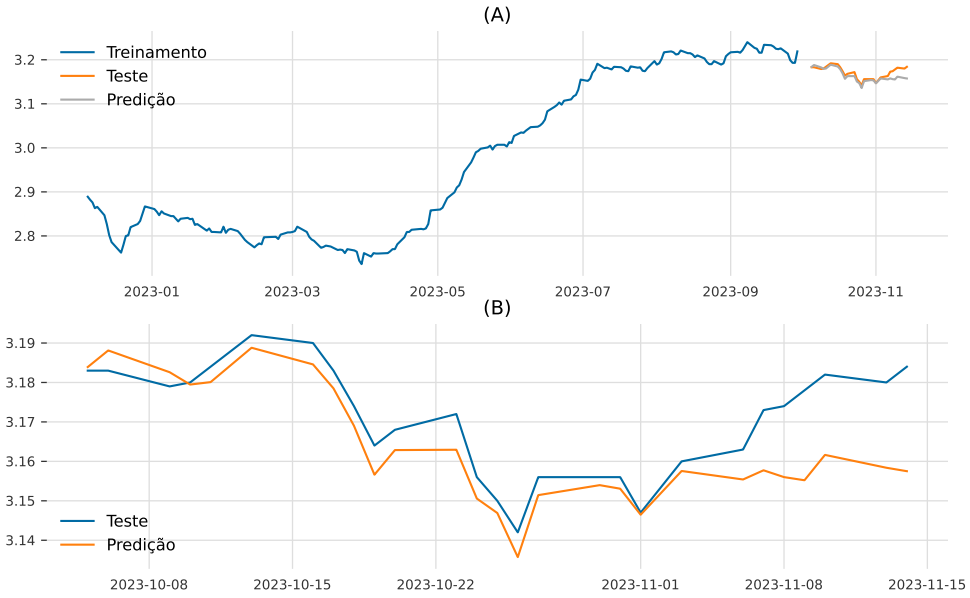

Figura 4.9: Gráfico de valores do IFIX e predição do modelo XGBoost. Em (A), uma visão geral do recorte do conjunto de dados. Em (B), uma visão dos dados de treinamento e a predição.

Capítulo 5

Plano de Trabalho e Cronograma de Execução

O plano de pesquisa que abordará atividades teóricas e também experimentais, é listado aqui:

1. Revisão bibliográfica.
2. Análise exploratória dos dados.
3. Desenvolvimento inicial de modelos (prova de conceito).
4. Exame de Qualificação do Mestrado (EQM).
5. Evolução das bases de dados.
6. Desenvolvimento e testes com modelo baseado em Transformers.
7. Evolução nos modelos levantados.
8. Execução de testes e análise dos resultados.
9. Documentação e publicação dos resultados.
10. Escrita da dissertação de mestrado.
11. Apresentação da dissertação de mestrado.

O plano de trabalho estimado para um período de 24 meses dividido em trimestres é mostrado na Figura 5.1.

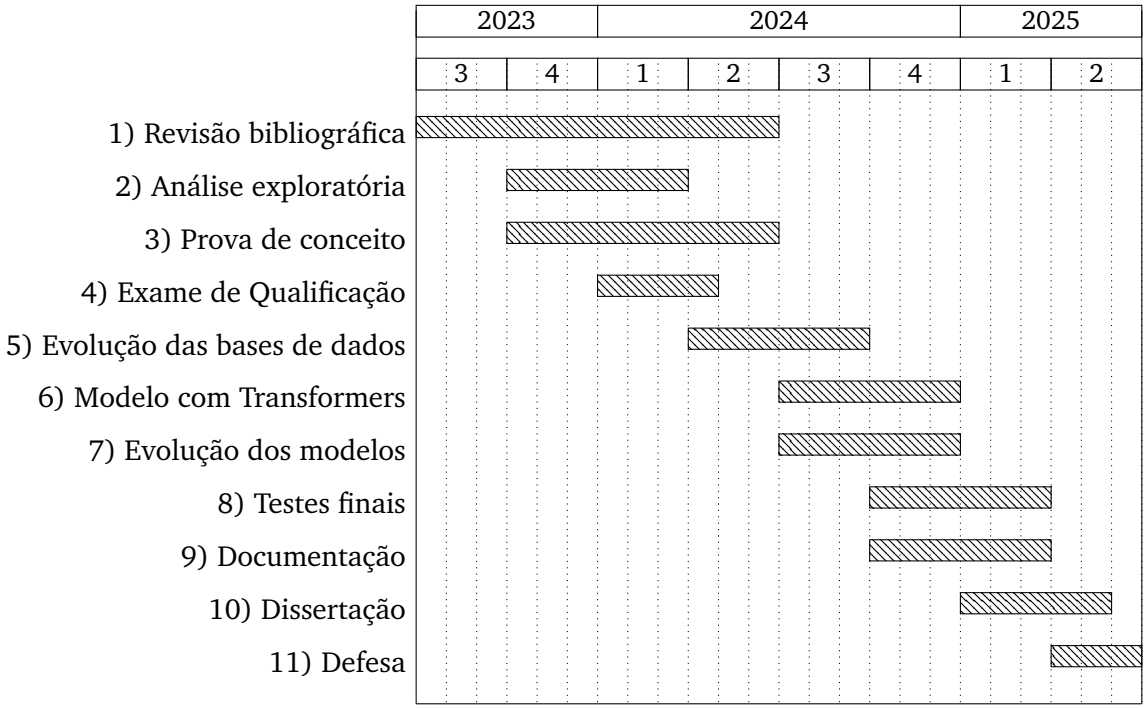

Figura 5.1: Cronograma de atividades da proposta.

Capítulo 6

Considerações Finais

Os resultados previstos e as contribuições esperadas ao final deste estudo incluem:

1. Disponibilização de conjuntos de dados preparados para o problema de predição de séries temporais, abrangendo diversos domínios além de ciências da natureza, finanças e dados médicos. Isso permitirá não apenas a reprodução dos experimentos realizados, mas também a aplicação dos modelos em novos conjuntos de dados que possam surgir.

2. Disponibilização de implementações de modelos de predição de séries temporais baseados em ARIMA, XGBoost e *Transformers*, juntamente com os parâmetros ajustados para cada domínio específico. Isso facilitará a utilização desses modelos por outros pesquisadores e profissionais interessados em predição de séries temporais em diferentes contextos.

3. Avaliação e comparação detalhadas dos modelos de predição de séries temporais, permitindo a identificação dos mais eficazes para cada domínio de aplicação. Isso ajudará a orientar futuras pesquisas e aplicações práticas no campo da predição de séries temporais.

Além disso, é importante destacar alguns desafios e limitações encontrados durante este estudo. Por exemplo, a seleção e a preparação adequadas dos conjuntos de dados, bem como a otimização dos hiperparâmetros dos modelos, são processos complexos e exigem cuidadosa consideração. Além disso, a disponibilidade de recursos computacionais para o treinamento dos modelos pode ser uma limitação, especialmente ao lidar com grandes volumes de dados.

Apesar desses desafios, acredita-se que este estudo contribui significativamente para o avanço do conhecimento no campo da predição de séries temporais e pode servir como base para pesquisas futuras nesta área. A disponibilização de conjuntos de dados e modelos implementados pode facilitar a replicação e extensão deste trabalho por outros pesquisadores interessados em explorar e melhorar as técnicas de predição de séries temporais em diversos domínios.

Referências Bibliográficas

[1] J. Armstrong. *Long-range Forecasting: From Crystal Ball to Computer*. A Wiley interscience publication. Wiley, 1978. 19

[2] G. Box and G. Jenkins. Time series analysis, forecasting and control. *Journal of the American Statistical Association*, 134, 01 1971. 6

[3] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. 7

[4] A. Chhabra, S. K. Singh, A. Sharma, S. Kumar, B. B. Gupta, V. Arya, and K. T. Chui. Sustainable and intelligent time-series models for epidemic disease forecasting and analysis. *Sustainable Technology and Entrepreneurship*, 3(2):100064, May 2024. 1

[5] J. Dong, W. Zeng, L. Wu, J. Huang, T. Gaiser, and A. K. Srivastava. Enhancing short-term forecasting of daily precipitation using numerical weather prediction bias correcting with xgboost in different regions of china. *Engineering Applications of Artificial Intelligence*, 117:105579, Jan. 2023. 1

[6] S. Ho and M. Xie. The use of arima models for reliability forecasting and analysis. *Computers and Industrial Engineering*, 35(1–2):213–216, Oct. 1998. 10

[7] M. Joseph. *Modern Time Series Forecasting with Python*. Packt Publishing, Birmingham, England, June 2022. 6, 18, 19

[8] B. Lim, S. O. Arık, N. Loeff, and T. Pfister. Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting. *International Journal of Forecasting*, 37(4):1748–1764, Oct. 2021. 8, 11

[9] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008. 9

[10] Y. Liu, H. Wu, J. Wang, and M. Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 9881–9893. Curran Associates, Inc., 2022. 12

[11] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. *arXiv 2310.06625*, 2023. 8, 12

[12] C.-X. Lv, S.-Y. An, B.-J. Qiao, and W. Wu. Time series analysis of hemorrhagic fever with renal syndrome in mainland china by using an xgboost forecasting model. *BMC Infectious Diseases*, 21(1), Aug. 2021. 11

[13] T. Muhammad, A. B. Aftab, M. Ibrahim, M. M. Ahsan, M. M. Muhu, S. I. Khan, and M. S. Alam. Transformer-based deep learning model for stock price prediction: A case study on bangladesh stock market. *International Journal of Computational Intelligence and Applications*, 22(03), Apr. 2023. 1

[14] P. Poggi, M. Muselli, G. Notton, C. Cristofari, and A. Louche. Forecasting and simulating wind speed in corsica by using an autoregressive model. *Energy Conversion and Management*, 44(20):3177–3196, Dec. 2003. 10

[15] M. S. Rahman, A. H. Chowdhury, and M. Amrin. Accuracy comparison of arima and xgboost forecasting models in predicting the incidence of covid-19 in bangladesh. *PLOS Global Public Health*, 2(5):e0000495, May 2022. 11

[16] I. Sardar, M. A. Akbar, V. Leiva, A. Alsanad, and P. Mishra. Machine Learning and Automatic ARIMA/Prophet Models-based Forecasting of COVID-19: Methodology, Evaluation, and Case Study in SAARC Countries. *Stochastic Environmental Research and Risk Assessment*, 37(1):345–359, 2023. 1

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 1

[18] Y. Wang and Y. Guo. Forecasting method of stock market volatility in time series data based on mixed model of arima and xgboost. *China Communications*, 17(3):205–221, Mar. 2020. 11

[19] G. Woo, C. Liu, D. Sahoo, A. Kumar, and S. Hoi. Etsformer: Exponential smoothing transformers for time-series forecasting, 2022. URL `https://arxiv.org/abs/2202.01381`. 12

[20] L. Zhang, W. Bian, W. Qu, L. Tuo, and Y. Wang. Time series forecast of sales volume based on XGBoost. *Journal of Physics: Conference Series*, 1873(1):012067, Apr. 2021. 11