UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

João Vitor de Araújo Costa

Predição de Preços de Ações do setor elétrico com aprendizado de máquina

Recife

Fevereiro de 2026

João Vitor de Araújo Costa

Predição de Preços de Ações do setor elétrico com aprendizado de máquina

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE
Departamento de Estatíistica e Informática
Curso de Bacharelado em Sistemas de Informação

Orientador: Gabriel Alves De Albuquerque Junior

Recife
fevereiro 2026

João Vitor de Araújo Costa

Predição de Preços de Ações do setor elétrico com aprendizado de máquina

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada em: 20 de Fevereiro de 2026.

BANCA EXAMINADORA

Nome do Orientador (Gabriel Alves de Albuquerque Junior)
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Lidiano Oliveira
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Predição de Preços de Ações do setor elétrico com aprendizado de máquina

João Vitor de Araújo Costa , Gabriel Alves de Albuquerque Júnior

Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco Rua Dom Manuel de Medeiros, s/n, - CEP: 52171-900 – Recife – PE – Brasil

[joaovitor.araujo@ufrpe.br, gabriel.alves@ufrpe.br]

Resumo. *A previsão de preços de ações é um desafio clássico no mercado financeiro, devido à alta volatilidade e à complexidade dos fatores que influenciam o comportamento dos ativos. Este trabalho tem como objetivo aplicar e comparar diferentes algoritmos de aprendizado de máquina para prever os preços das ações brasileiras TAEE11 (Taesa) e CMIG4 (Cemig), com foco exclusivo em problemas de regressão. Para isso, foram utilizados dados históricos de preços das ações combinados com variáveis macroeconômicas, como taxa de juros, IPCA, IEE e câmbio, além de indicadores técnicos como RSI e MACD. O conjunto de dados compreende o período de 2010 a 2019. Os resultados indicam que o modelo Random Forest apresentou melhor desempenho geral, observando-se as métricas, o modelo apresentou o menor RMSE, MAE, MAPE e maior R².*

Palavras-chave*: aprendizado de máquina, previsão de preços, ações,, séries temporais, XGBoost, LSTM, Random Forest, SVR.*

Abstract. *Stock price forecasting is a classic challenge in the financial market due to the high volatility and complexity of the factors influencing asset behavior. This work aims to apply and compare different machine learning algorithms to predict the prices of the Brazilian stocks TAEE11 (Taesa) and CMIG4 (Cemig), focusing exclusively on regression problems. To this end, historical stock price data combined with macroeconomic variables such as interest rates, IPCA (Brazilian inflation index), IEE (Brazilian economic index), and exchange rates, as well as technical indicators such as RSI and MACD, were used. The dataset covers the period from 2010 to 2019. The results indicate that the Random Forest model showed the best overall performance; observing the metrics, the model presented the lowest RMSE, MAE, MAPE, and the highest R².*

Keywords*: machine learning, stock price prediction, time series, XGBoost, LSTM, Random Forest, SVR.*

1. Introdução

No contexto do mercado financeiro, a previsão de preços de ações representa um problema complexo devido à alta volatilidade, à não linearidade das séries temporais e à influência de múltiplos fatores econômicos e setoriais. Métodos tradicionais de análise, embora amplamente utilizados, frequentemente apresentam limitações na captura de padrões dinâmicos e relações complexas presentes nos dados financeiros, o que motiva o uso de técnicas baseadas em aprendizado de máquina.

As soluções atuais na literatura incluem a aplicação de modelos estatísticos clássicos e algoritmos de machine learning, como árvores de decisão, máquinas de vetores de suporte e redes neurais, geralmente avaliados com conjuntos fixos de treinamento e teste. No entanto, essa abordagem pode gerar modelos excessivamente dependentes do histórico completo dos dados, resultando em suavização excessiva das previsões ou atraso na identificação de mudanças nos ciclos de preços, especialmente em séries temporais financeiras sujeitas a mudanças de regime.

Diante desse cenário, a proposta deste trabalho consiste em empregar uma abordagem de validação do tipo walk-forward, permitindo ajustar dinamicamente as janelas de treinamento e teste de acordo com os ciclos de preços das ações analisadas. Essa estratégia busca tornar os modelos mais sensíveis às mudanças temporais e mais alinhados às condições reais do mercado. O desempenho dos algoritmos é avaliado por meio de métricas específicas de regressão, possibilitando identificar o modelo mais adequado para cada ativo financeiro analisado, contribuindo para uma avaliação mais realista e robusta da capacidade preditiva das técnicas estudadas.

1.1. Objetivo geral

O objetivo central deste trabalho é realizar uma análise comparativa entre diferentes algoritmos de aprendizado de máquina aplicados à previsão da do preço de ações brasileiras, utilizando dados históricos de mercado, variáveis macroeconômicas e indicadores técnicos.

Abaixo destacam-se os objetivos específicos, que detalham as etapas necessárias e práticas para alcançar o objetivo da pesquisa:

- Coletar e organizar dados históricos das ações brasileiras TAEE11 e CMIG4, incluindo preços diários e volume;
- Incorporar variáveis macroeconômicas como IPCA, taxa de juros, câmbio e índice setorial (IEE);
- Calcular e aplicar indicadores técnicos, como RSI e MACD, como variáveis de entrada;
- Implementar os algoritmos LSTM, XGBoost, Random Forest e SVR para prever os preços;
- Avaliar os modelos com métricas de regressão (MAE, RMSE, MAPE e R^2)

- Comparar os resultados e discutir as vantagens e limitações de cada abordagem.

Este artigo está organizado da seguinte forma: na seção 2 apresenta-se a Fundamentação Teórica introduz os conceitos de aprendizado de máquina e descreve os algoritmos utilizados. Já na seção 3, os Trabalhos Relacionados, são apresentados estudos que aplicam algoritmos de aprendizado de máquina à previsão de preços de ações, destacando abordagens semelhantes à proposta deste trabalho de. Na seção 4, a Metodologia, são detalhados o conjunto de dados, os indicadores utilizados e os métodos aplicados para o treinamento e avaliação dos modelos. Na seção 5, os Resultados são apresentados e analisados os desempenhos obtidos. Por fim, em Considerações Finais na seção 6, são discutidas as conclusões do estudo e possíveis direções para trabalhos futuros.

A partir dessa proposta, a análise dos resultados foi conduzida de modo a responder às seguintes questões de pesquisa:

1. Qual algoritmo de aprendizado de máquina apresenta o melhor desempenho preditivo para cada ação analisada (CMIG4 e TAEE11), considerando métricas de erro e aderência às tendências de preços?
2. Como os diferentes modelos se comportam em relação à captura de variações de curto prazo, reconhecimento de mudanças de tendência e estabilidade das previsões ao longo do tempo?

Ao responder a essas questões, o trabalho contribui para a compreensão das vantagens e limitações de cada abordagem, auxiliando na escolha do modelo mais adequado para a previsão de preços de ações em diferentes contextos do mercado brasileiro.

2. Referencial Teórico

Nesta seção, são apresentados os principais conceitos teóricos que fundamentam o desenvolvimento deste trabalho. O objetivo é fornecer uma base sólida sobre os métodos, ferramentas e tecnologias utilizadas na análise e previsão de preços de ações por meio de algoritmos de aprendizado de máquina. Para isso, são abordados os fundamentos do Machine Learning, a linguagem de programação Python como principal ferramenta de desenvolvimento, e os algoritmos utilizados neste estudo: Árvores de Decisão, XGBoost, Random Forest, Support Vector Regression (SVR) e Long Short-Term Memory (LSTM). Cada tópico é explorado de forma a permitir a compreensão das técnicas aplicadas, suas vantagens, limitações e motivações para sua escolha neste projeto.

2.1. Machine Learning

O aprendizado de máquina (Machine Learning – ML) é um campo da inteligência artificial que permite que sistemas computacionais aprendam a partir de dados, ajustando seus comportamentos e tomando decisões sem a necessidade de programação explícita. Por meio da identificação de padrões em grandes volumes de dados, esses sistemas podem realizar previsões, classificações e análises complexas, sendo aplicáveis em diversos contextos, como saúde, transporte e mercado financeiro.

Dentro do aprendizado de máquina, existem diferentes abordagens. No aprendizado supervisionado, os dados de entrada são acompanhados de suas respectivas saídas ou

rótulos, permitindo que o modelo aprenda relações diretas entre variáveis. Esta abordagem é especialmente relevante para problemas de regressão, que buscam prever valores contínuos, como o preço futuro de uma ação. Diferentemente da classificação, que atribui categorias discretas, a regressão fornece estimativas numéricas precisas, possibilitando medir quantitativamente variações, tendências e padrões temporais nos dados. Por esse motivo, o aprendizado supervisionado com regressão é o foco deste trabalho.

O aprendizado não supervisionado refere-se a situações em que os dados não possuem rótulos, e o objetivo é identificar estruturas ocultas, padrões ou agrupamentos entre os registros. Técnicas como agrupamento (*clustering*) e redução de dimensionalidade permitem analisar grandes volumes de dados sem supervisão, revelando insights sobre relações entre variáveis que não seriam percebidas de outra forma. Embora valioso, esse tipo de abordagem não é diretamente adequado para previsão de preços de ações, pois não fornece valores contínuos estimados.

Já o aprendizado por reforço envolve agentes que aprendem interagindo com um ambiente e recebendo recompensas ou penalidades de acordo com suas ações. Esse tipo de aprendizado é eficiente em problemas de tomada de decisão sequencial, como jogos ou controle de sistemas, mas também pode ser aplicado em finanças para estratégias de trading. Diferentemente da regressão, o foco está em maximizar recompensas cumulativas, e não em prever valores contínuos diretamente.

Portanto, considerando a problemática deste trabalho, que busca prever o comportamento futuro de preços de ações, a abordagem supervisionada com regressão se mostra a mais adequada. Ela permite relacionar indicadores financeiros, como preços históricos e dados macroeconômicos, com os movimentos futuros do mercado, fornecendo estimativas precisas que podem ser utilizadas para análise quantitativa e tomada de decisão.

2.2. Séries Temporais

Séries temporais são sequências de dados observadas ao longo do tempo, geralmente em intervalos regulares, como dias, meses ou anos. Diferentemente de dados independentes, esse tipo de série apresenta dependência temporal, na qual valores passados influenciam diretamente o comportamento futuro. No contexto do mercado financeiro, os preços das ações constituem um exemplo clássico de série temporal, pois refletem tanto informações históricas quanto expectativas futuras dos agentes do mercado.

Uma série temporal pode ser decomposta em componentes fundamentais, como tendência, que representa o movimento de longo prazo dos dados; sazonalidade, que corresponde a padrões que se repetem em intervalos regulares; e componente aleatória, associada a ruídos e variações imprevisíveis do mercado. Em séries financeiras, a sazonalidade nem sempre é evidente, enquanto a volatilidade e a presença de ruído costumam ser elevadas, tornando o processo de previsão mais complexo.

A modelagem de séries temporais exige técnicas capazes de capturar essas dependências ao longo do tempo, bem como lidar com mudanças de padrão e ciclos de preços. Modelos tradicionais podem apresentar dificuldades nesse cenário, especialmente quando há não linearidade e variações abruptas. Por esse motivo,

algoritmos de aprendizado de máquina têm sido amplamente utilizados, pois conseguem aprender relações complexas a partir dos dados históricos.

Entre esses algoritmos, modelos recorrentes como a Long Short-Term Memory (LSTM) se destacam por sua capacidade de armazenar informações relevantes ao longo do tempo, permitindo capturar tanto dependências de curto quanto de longo prazo. Além disso, outros modelos baseados em regressão, como XGBoost, Random Forest e SVR, podem ser aplicados a séries temporais por meio da utilização de janelas deslizantes (*windowing*), transformando o problema temporal em um problema supervisionado de regressão.

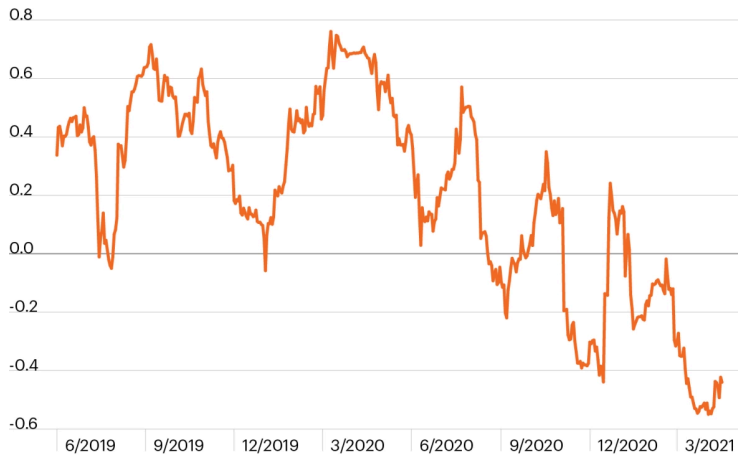

Figura 1: Exemplo de Série Temporal com dados aleatórios, o eixo X mostra o tempo e o Y o valor dos dados.

Fonte: EJFGV

2.3. Validação Walk-Forward

Em problemas de previsão de séries temporais, como a previsão de preços de ativos financeiros, os dados apresentam uma característica fundamental: dependência temporal. Isso significa que observações passadas influenciam diretamente observações futuras, tornando inadequado o uso de técnicas tradicionais de validação que assumem independência entre os dados, como a validação cruzada aleatória (*random split*).

Nesse contexto, surge a validação walk-forward, também conhecida como *rolling window validation* ou *forward chaining*, uma abordagem específica para avaliação de modelos em séries temporais. O principal objetivo dessa técnica é simular o cenário real de previsão, no qual o modelo é treinado apenas com dados passados e utilizado para prever dados futuros.

Na validação *walk-forward*, o conjunto de dados é dividido em janelas temporais sequenciais. Inicialmente, uma janela de dados históricos é utilizada para treinar o modelo. Em seguida, o modelo é testado em um intervalo de tempo imediatamente posterior. Após essa etapa, a janela de treinamento é deslocada para frente no tempo, incorporando novos dados, e o processo é repetido até o final da série.

Esse procedimento pode ser realizado de duas formas principais:

- Janela deslizante (*sliding window*): o tamanho da janela de treinamento permanece fixo, descartando os dados mais antigos a cada avanço.

- Janela expansiva (*expanding window*): a janela de treinamento cresce ao longo do tempo, acumulando todo o histórico disponível.

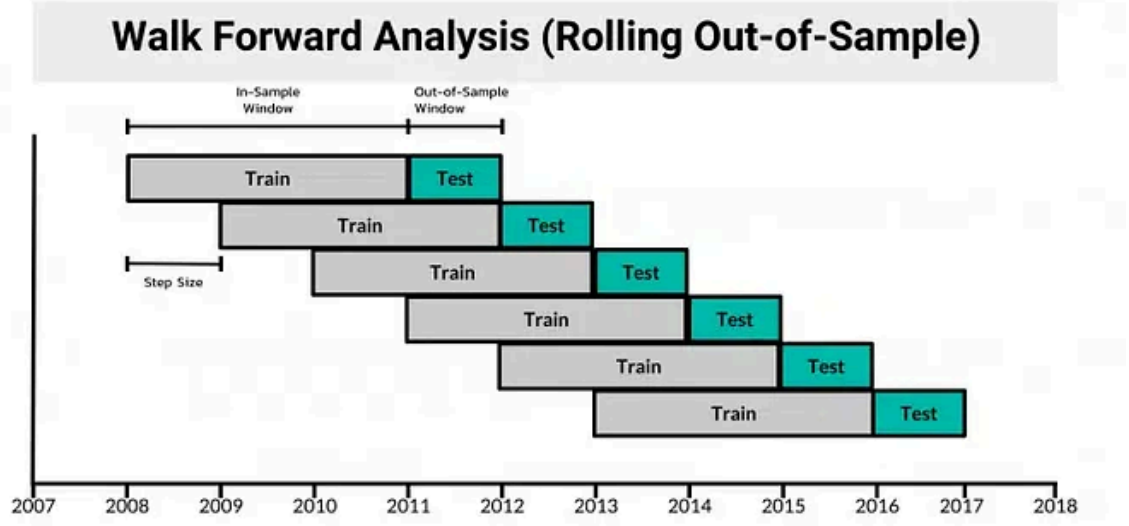

Figura 2: Demonstração da validação *Walk Forward* utilizada nos modelos

Fonte: Ahmed (2024)

Inicialmente, o modelo é treinado com um conjunto de dados históricos que vai do instante t_1 até t_n e, a partir desse treinamento, realiza a previsão para o próximo período, t_{n+1}. Em seguida, essa janela de treinamento é deslocada para frente, passando a considerar os dados de t_2 até t_{n+1}, e o modelo é novamente treinado para prever o valor em t_{n+2}. Esse procedimento se repete sucessivamente, permitindo que o modelo seja constantemente atualizado com as informações mais recentes disponíveis, simulando de forma mais realista o processo de previsão em séries temporais.

A principal vantagem da validação *walk-forward* é que ela preserva a ordem temporal dos dados, evitando o vazamento de informação futura (*data leakage*), que ocorre quando dados do futuro são utilizados, direta ou indiretamente, no treinamento do modelo. Além disso, essa abordagem permite avaliar a robustez e estabilidade do modelo ao longo do tempo, especialmente em cenários onde o comportamento da série pode mudar devido a fatores econômicos, estruturais ou sazonais.

No contexto do mercado financeiro, a validação *walk-forward* é amplamente recomendada, pois reflete de maneira mais fiel a aplicação prática dos modelos preditivos, nos quais decisões são tomadas com base apenas nas informações disponíveis até aquele momento. Dessa forma, os resultados obtidos por meio dessa técnica tendem a ser mais realistas e confiáveis quando comparados a métodos tradicionais de divisão de dados.

2.4. Árvores de Decisão

As Árvores de Decisão são algoritmos de aprendizado supervisionado amplamente utilizados em problemas de classificação e regressão. Sua principal característica é a estrutura hierárquica em forma de árvore, onde cada nó interno representa uma condição de teste em um atributo, cada ramo representa o resultado desse teste, e cada nó folha representa uma saída predita.

No contexto de regressão, as Árvores de Decisão funcionam dividindo o espaço de entrada em regiões que minimizam o erro quadrático entre os valores reais e previstos. Essa divisão é feita de forma recursiva, selecionando os pontos de corte (*thresholds*) que melhor separam os dados, com base em métricas como o erro quadrático médio (MSE).

Entre suas principais vantagens estão a interpretabilidade e a capacidade de capturar relações não lineares entre variáveis. No entanto, modelos de árvore única tendem a sofrer com o *overfitting* (sobreajuste), especialmente em conjuntos de dados complexos. Por isso, técnicas baseadas em múltiplas árvores, como Random Forest e XGBoost, são normalmente utilizadas para melhorar o desempenho preditivo, combinando diversas árvores em uma abordagem única.

2.5. Indicadores Técnicos

Indicadores técnicos são ferramentas matemáticas aplicadas sobre séries de preços e volume de negociação, com o objetivo de extrair informações relevantes sobre tendência, força do movimento, volatilidade e possíveis pontos de reversão do mercado. Ao transformar dados brutos em variáveis derivadas, esses indicadores auxiliam modelos de aprendizado de máquina a capturar padrões temporais e comportamentos não lineares, comuns em séries financeiras.

O Relative Strength Index (RSI) é um oscilador que varia entre 0 e 100 e mede a intensidade dos movimentos de preço em um determinado período, geralmente 14 dias. Valores elevados indicam condições de sobrecompra, enquanto valores baixos sugerem sobrevenda, fornecendo sinais importantes sobre possíveis reversões de curto prazo. Já o Moving Average Convergence Divergence (MACD) baseia-se na diferença entre duas médias móveis exponenciais de períodos distintos, permitindo identificar mudanças na direção e na força da tendência predominante do ativo.

As médias móveis simples (SMA) e exponenciais (EMA) são utilizadas para suavizar as oscilações do preço, facilitando a identificação de tendências de médio e longo prazo. Enquanto a SMA atribui pesos iguais aos dados históricos, a EMA dá maior importância aos valores mais recentes, tornando-a mais sensível a mudanças rápidas no mercado. Essas características são particularmente relevantes para modelos preditivos, pois ajudam a diferenciar movimentos estruturais de flutuações aleatórias.

A volatilidade histórica mede o grau de variação dos preços ao longo do tempo, fornecendo uma estimativa do risco associado ao ativo. Em conjunto, o retorno diário, calculado a partir da variação percentual do preço de fechamento, permite que os modelos aprendam não apenas os níveis de preço, mas também a dinâmica das variações, o que é essencial para capturar movimentos abruptos e mudanças de regime.

Indicadores de *momentum*, como o Momentum e o Rate of Change (ROC), avaliam a velocidade e a intensidade das mudanças de preço, auxiliando na identificação de

acelerações ou desacelerações do movimento do mercado. Esses indicadores são úteis para prever continuidades ou reversões de tendência, especialmente em janelas de curto prazo.

O Average True Range (ATR) é um indicador de volatilidade que considera *gaps* e oscilações intra diárias, sendo eficaz para capturar períodos de maior instabilidade no mercado. Complementarmente, as Bandas de Bollinger, compostas por uma média móvel central e dois limites baseados no desvio padrão, permitem identificar momentos em que o preço se encontra relativamente elevado ou deprimido em relação ao seu comportamento histórico. A largura das bandas, por sua vez, indica o nível de volatilidade do mercado em determinado período.

Por fim, o marcador de picos de preço (*spike flag*) é utilizado para sinalizar movimentos extremos ou atípicos na série temporal, funcionando como um mecanismo auxiliar para destacar eventos de alta intensidade que podem influenciar o processo de aprendizado dos modelos.

De forma geral, a utilização conjunta desses indicadores amplia a representação do comportamento do mercado, permitindo que os algoritmos de aprendizado de máquina capturem relações complexas entre tendência, volatilidade e dinâmica de preços. Essa abordagem contribui para um treinamento mais robusto e informativo, reduzindo a dependência exclusiva dos preços brutos e favorecendo a generalização dos modelos preditivos.

2.6. Métricas de Avaliação

A avaliação dos modelos preditivos foi realizada por meio de métricas específicas para problemas de regressão:

- MAE (*Mean Absolute Error*): mede o erro médio absoluto entre as previsões e os valores reais.

- RMSE (*Root Mean Squared Error*): raiz quadrada da média dos erros ao quadrado; penaliza erros maiores.

- MAPE (*Mean Absolute Percentage Error*): erro percentual médio; útil para comparar ativos com escalas diferentes.

- R^2 (Coeficiente de Determinação): mede a proporção da variabilidade dos dados reais explicada pelo modelo. Valores próximos de 1 indicam que o modelo consegue explicar bem os movimentos da série, enquanto valores baixos indicam baixa capacidade explicativa. O R^2 também ajuda a identificar possíveis problemas de *overfit* ou *underfit*, quando comparado com outras métricas de erro.

O uso de múltiplas métricas é essencial para uma avaliação completa, pois cada métrica evidencia diferentes aspectos do desempenho do modelo. Enquanto MAE e RMSE informam sobre a magnitude dos erros, MAPE permite comparações relativas entre ativos e R^2 mostra a capacidade explicativa global do modelo. Essa combinação garante que a análise do desempenho dos algoritmos seja robusta e consistente, considerando tanto a precisão quanto a aderência aos padrões da série temporal

3. Trabalhos Relacionados

Nesta seção, são apresentados e analisados trabalhos relacionados que abordam a predição de preços de ações utilizando técnicas de aprendizado de máquina. O objetivo é compreender quais métodos, algoritmos e métricas têm maior eficiência na predição do preço de ações, bem como os pontos fortes e fracos de cada algoritmo. Para isso, foram selecionados quatro estudos que exploram diferentes métodos preditivos aplicados ao mercado financeiro, com foco na comparação entre modelos tradicionais e avançados.

O artigo (CARVALHO; BARBOZA; PASCHOARELLI, 2022), analisou o mercado de ações brasileiro utilizando algoritmos de aprendizado de máquina aplicados às ações VALE3, PETR4 e ITUB4, no período de 11/02/2015 a 10/02/2020. Para melhorar o modelo, o estudo utilizou indicadores técnicos que são utilizados no mercado financeiro, como RSI, Oscilador Estocástico (%K), MACD, PROC, OBV e Williams %R, além disso, foram aplicados pesos que aumentam a importância dos dados mais recentes. Os algoritmos analisados foram Redes Neurais Artificiais (RNA), Random Forest (RF) e Support Vector Machine (SVM), testando a previsão para 30, 60 e 90 dias. Para analisar os modelos foram observadas as métricas de acurácia, precisão, recall e especificidade, além do uso de curvas ROC para avaliação gráfica do desempenho. Os resultados indicaram que os modelos Random Forest e SVM apresentaram desempenho superior, com acurácia entre 91% e 97%, enquanto as RNAs ficaram entre 71% e 89%. O kernel SVM-GHI, geralmente usado para reconhecimento de imagens, também obteve bom desempenho. O estudo organizou os dados de forma efetiva e utilizou bons indicadores para enriquecer o treinamento dos algoritmos, com isso, a análise do estudo conseguiu ser eficiente utilizando as métricas e gráficos para entender os pontos fortes e fracos de cada algoritmo.

Em comparação com o presente trabalho, observa-se como semelhança o uso de algoritmos tradicionais de aprendizado de máquina, como Random Forest e SVM/SVR, bem como a aplicação de indicadores técnicos para enriquecer o conjunto de dados e melhorar a capacidade preditiva dos modelos. No entanto, este estudo se diferencia principalmente pelo foco em um problema de regressão, voltado à previsão dos preços das ações CMIG4 e TAEE11, enquanto o trabalho de Carvalho et al. (2022) adota uma abordagem de classificação, avaliando o desempenho por meio de métricas como acurácia e curvas ROC. Além disso, este trabalho incorpora o método de validação *walk-forward*, visando reduzir atrasos na identificação de novos ciclos de preços e mitigar problemas de *overfitting*, aspecto que não é explorado explicitamente no estudo analisado. Dessa forma, este trabalho contribui ao aprofundar a análise temporal das previsões e ao comparar diferentes modelos sob a perspectiva da adaptação dinâmica aos ciclos de mercado.

Outro trabalho que foi selecionado é o (MATTOS; ASSIS; AUGUSTO; ALMEIDA; 2022), o qual escolheu cinco ações da B3 sendo elas: Petrobras (PETR4), Vale (VALE3), Banco Itaú (ITUB3), Banco Bradesco (BBDC3) e Ambev (ABEV3) para treinamento de algoritmos de machine learning na predição do preço dessas. Para apoiar a modelagem, foram utilizados indicadores técnicos os quais são RSI e o Estocástico, que fornecem informações sobre tendências, reversões e movimentos

neutros de preço. O estudo analisou os algoritmos Regressão Linear, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Árvores de Decisão e Random Forest, visando prever os preços nos próximos 30 dias. Os resultados indicaram que os classificadores SVM e KNN apresentaram maior acurácia em comparação aos demais, embora a Regressão Linear tenha mostrado valores preditivos próximos aos preços reais, o qual pode ser útil na definição de alvos de compra e venda. O artigo mostra que o desempenho varia conforme a ação analisada, além disso, a combinação de modelos, como a Regressão Linear com um classificador, pode aumentar a eficácia nas previsões, mostrando a diversidade dos algoritmos e parâmetros envolvidos.

Em relação a este presente trabalho, observa-se como principal semelhança o uso de algoritmos amplamente empregados no mercado financeiro, como SVM e Random Forest, bem como a utilização de indicadores técnicos para capturar padrões de tendência e comportamento dos preços. Contudo, este estudo se diferencia ao adotar uma abordagem estritamente voltada à regressão de preços, avaliando o desempenho dos modelos por meio de métricas contínuas como RMSE, MAE, MAPE e R^2, enquanto o trabalho de Mattos et al. (2022) prioriza métricas de acurácia típicas de classificação.

Já o artigo (RAMOS, 2023) utilizou algoritmos de aprendizado de máquina aplicados a diferentes ativos financeiros, tanto do mercado indiano (como Tata Motors, Axis Bank, Maruti e HCL) quanto do mercado brasileiro (Ambev, Itaú Unibanco e Vale). Foram testados diversos modelos, incluindo Support Vector Machine (SVM), Support Vector Regression (SVR), Árvore de Decisão, XGBoost, modelos de séries temporais, LSTM e Redes Neurais Convolucionais, com a avaliação baseada nas métricas: RMSE, MAE e MAPE. O estudo seguiu a metodologia proposta por Hiransha et al. (2018) e concluiu que algoritmos mais avançados, como LSTM e XGBoost, apresentam desempenho superior em relação a modelos mais tradicionais como o SVR, especialmente em função da natureza não-linear e volátil dos dados do mercado financeiro. Porém, o SVR ainda se mostrou capaz de prever a tendência geral dos preços. A pesquisa reforça a importância de utilizar modelos robustos e múltiplas métricas para avaliação preditiva, por causa da complexidade da previsão de preços de ações.

Comparando com o presente trabalho, observa-se uma forte semelhança na escolha dos algoritmos analisados, uma vez que modelos como LSTM, XGBoost, SVR e árvores de decisão também são avaliados nesta pesquisa. Além disso, ambos os estudos utilizam métricas de erro contínuo, como RMSE, MAE e MAPE, reforçando uma abordagem consistente para problemas de regressão em séries temporais financeiras. Entretanto, este trabalho se diferencia ao aplicar o método de validação *walk-forward* para ajustar dinamicamente as janelas de treinamento aos ciclos de preços das ações analisadas, além de incluir o coeficiente de determinação (R^2) como métrica adicional para avaliar o grau de explicação da variabilidade dos dados. Outra diferença relevante está no foco da análise, que busca identificar o melhor modelo para cada ação individualmente, enquanto o estudo de Ramos (2023) enfatiza uma comparação geral entre algoritmos e mercados distintos. Dessa forma, o presente trabalho complementa os achados da literatura ao aprofundar a análise do desempenho dos modelos em ações específicas do mercado brasileiro, considerando a influência temporal e os ciclos de preços.

Por fim, o estudo (HUPSEL; CUNHA; BARRETO, 2022) realizou um comparativo

entre Redes Neurais e Regressão Linear na predição de valores de ações da bolsa brasileira, foram utilizados dados do período entre 2020 e 2022, incluindo informações de preços de abertura, fechamento, máximas, mínimas e volume de negociações. O conjunto de dados foi dividido em 75% para treino, 24% para teste e 1% para validação. Ambos os modelos foram avaliados com base nas métricas RMSE, MAE e MSE, fazendo previsões para os 30 dias seguintes. Os resultados mostraram que, embora a Regressão Linear tenha tido um desempenho satisfatório, a Rede Neural obteve resultados melhores, demonstrando maior precisão na previsão dos preços. O estudo destaca a capacidade preditiva das redes neurais em cenários de alta variabilidade, mostrando que seu uso é mais eficaz para aplicações futuras no mercado financeiro.

Comparando com o presente trabalho, nota-se que ambos os estudos utilizam Redes Neurais e métodos de regressão para previsão de preços de ações, reforçando a relevância desses algoritmos para séries temporais financeiras. Entretanto, o presente trabalho se diferencia ao analisar quatro modelos distintos (XGBoost, LSTM, SVR e Random Forest) aplicados a duas ações específicas, além de implementar a validação *walk-forward* para ajustar dinamicamente a janela de treinamento aos ciclos de preços das ações. Outra diferença é o uso do coeficiente de determinação (R^2) e do erro percentual absoluto (MAPE).

4. Metodologia

Esta seção apresenta os procedimentos metodológicos adotados para a construção e avaliação dos modelos preditivos aplicados à previsão de preços de ações brasileiras. A metodologia foi estruturada em duas etapas principais: o tratamento dos dados e o desenvolvimento dos modelos de regressão. A primeira etapa compreende desde a descrição do conjunto de dados até os processos de coleta, limpeza, transformação e normalização. Já a segunda etapa apresenta os quatro algoritmos utilizados no estudo: XGBoost, LSTM, Random Forest e SVR.

Todo o processo de implementação, incluindo o pré-processamento dos dados, o treinamento dos modelos e a avaliação dos resultados, foi desenvolvido em ambiente Python e está disponibilizado em um repositório público no GitHub: https://github.com/JoaoVitorA7/tccJoaoAraujoCosta.git, com o objetivo de garantir transparência, reprodutibilidade e validação externa dos experimentos realizados.

A metodologia aqui apresentada tem como objetivo garantir que todos os modelos sejam comparáveis entre si, com as mesmas entradas e ambiente de avaliação, permitindo uma análise consistente dos seus desempenhos na tarefa de previsão de preços das ações TAEE11 e CMIG4.

4.1. Conjunto de Dados

O conjunto de dados utilizado neste trabalho foi construído com o objetivo de prever os preços de ações brasileiras, considerando tanto variáveis financeiras específicas dos ativos quanto fatores macroeconômicos externos. Foram selecionadas duas ações do setor de energia elétrica negociadas na B3: Taesa (TAEE11) e Cemig (CMIG4). Para cada uma dessas ações, foram coletados dados históricos diários no período de

01/01/2010 a 31/12/2019, contendo as seguintes variáveis: data, preço de abertura, preço de fechamento, preço máximo (alta), preço mínimo (baixa) e volume negociado. Esses dados são essenciais para representar o comportamento de mercado ao longo do tempo e identificar padrões relevantes.

Além dos dados das ações, foram incorporadas ao conjunto variáveis macroeconômicas que influenciam diretamente os preços dos ativos. Entre elas estão:

- Taxa de Juros (SELIC): impacta o custo do capital e o valor presente dos fluxos de caixa futuros das empresas;

- IPCA (Índice Nacional de Preços ao Consumidor Amplo): reflete a inflação do período, afetando o poder de compra e os custos operacionais;

- Câmbio (USD/BRL): importante para empresas com exposição ao comércio internacional ou com custos atrelados à moeda estrangeira;

- IEE (Índice de Energia Elétrica): representa o desempenho do setor ao qual pertencem as ações analisadas.

Cada uma dessas variáveis foi associada à sua respectiva data de referência, permitindo a sincronização temporal com os preços das ações no período compreendido entre 01/01/2010 e 31/12/2019. A escolha desse intervalo temporal foi motivada por dois fatores principais.

Primeiramente, trata-se de um período suficientemente longo para capturar diferentes ciclos econômicos, contemplando fases de crescimento, instabilidade e recuperação da economia brasileira, o que é fundamental para modelos de séries temporais aplicados ao mercado financeiro. Essa diversidade de cenários possibilita que os algoritmos aprendam padrões associados tanto a períodos de maior volatilidade quanto a fases mais estáveis dos preços das ações.

Em segundo lugar, o intervalo selecionado garante consistência e disponibilidade conjunta de todas as variáveis macroeconômicas utilizadas — SELIC, IPCA, câmbio (USD/BRL) e IEE — evitando lacunas de dados que poderiam comprometer o processo de treinamento dos modelos. Além disso, ao limitar a análise até o final de 2019, evita-se a inclusão de eventos extraordinários, como a pandemia da COVID-19, que introduziram choques atípicos capazes de distorcer os padrões históricos e prejudicar a capacidade de generalização dos modelos.

A escolha das ações CMIG4 e TAEE11 também se relaciona às características do setor elétrico brasileiro. Empresas desse setor tendem a apresentar maior previsibilidade de receitas quando comparadas a companhias de outros segmentos da economia, uma vez que grande parte de suas atividades é baseada em contratos de longo prazo e em regulação estatal. No caso das empresas de transmissão de energia, como a TAEE11, a remuneração ocorre por meio de receitas anuais permitidas definidas em contratos de concessão, o que reduz a exposição direta às oscilações de mercado. Já empresas integradas ou com atuação em geração e distribuição, como a Cemig (CMIG4), podem apresentar maior sensibilidade a fatores de mercado, como variações no consumo de energia e condições hidrológicas. Dessa forma, a análise dessas duas ações permite

observar o comportamento de modelos de previsão em ativos que pertencem ao mesmo setor econômico, mas com estruturas operacionais distintas.

Além disso, foram calculados diversos indicadores técnicos baseados no histórico de preços e volumes, com o objetivo de enriquecer o conjunto de entrada dos modelos. Esses indicadores auxiliam na identificação de tendências, volatilidade, momentos de reversão e força do movimento de preços. Entre os indicadores utilizados estão: RSI (Índice de Força Relativa), MACD (*Moving Average Convergence Divergence*), médias móveis (EMA e MA), *Bollinger Bands*, volatilidade histórica, *momentum*, *rate of change* (ROC), *average true range* (ATR), retorno diário e um indicador de pico de preço (*spike_flag*).

A integração entre variáveis de preço, contexto macroeconômico e indicadores técnicos busca fornecer uma base rica e diversificada de informações, permitindo que os modelos de regressão aprendam padrões mais complexos e melhorem sua capacidade preditiva.

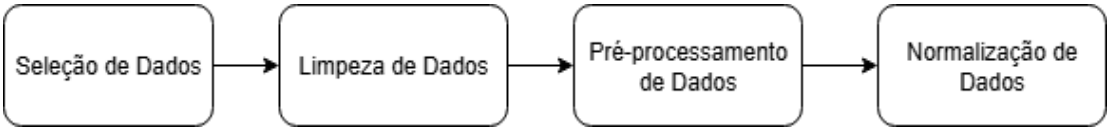

Figura 3: Fluxograma dos dados
Fonte: Elaborado pelo autor.

4.1.1. *Seleção dos Dados*

A coleta de dados para este trabalho foi realizada a partir de múltiplas fontes confiáveis e públicas, com o objetivo de reunir informações relevantes tanto do mercado financeiro quanto de indicadores macroeconômicos. Os dados foram obtidos no formato CSV e processados em Python para posterior integração e uso nos modelos de aprendizado de máquina.

Nome	Tipo	Descrição
Data	Data	Data de referência do registro
Ticker	String	Código de identificação da ação
Open	Float	Preço de abertura da ação no dia
Close	Float	Preço de fechamento da ação no dia
High	Float	Maior preço negociado no dia
Low	Float	Menor preço negociado no dia
Volume	Integer	Volume negociado no dia
AberturaIEE	Float	Valor de abertura do índice de energia elétrica
Selic	Float	Taxa de juros
usd_brl	Float	Taxa de câmbio

Quadro 1: Metadados utilizados e o que representam

Os dados históricos das ações TAEE11 (Taesa) e CMIG4 (Cemig), assim como os

valores de câmbio (USD/BRL), foram obtidos através da plataforma Kaggle, que disponibiliza séries temporais atualizadas e organizadas para fins acadêmicos e de pesquisa. Já os dados referentes ao IEE (Índice de Energia Elétrica) foram extraídos do site Investing.com, sendo esse índice importante para representar o desempenho setorial das empresas analisadas.

As variáveis taxa de juros (SELIC) e IPCA (Índice Nacional de Preços ao Consumidor Amplo) foram coletadas diretamente do site do Banco Central do Brasil, utilizando arquivos CSV disponíveis na seção de séries temporais econômicas.

4.1.2. *Limpeza de Dados*

Após a coleta, foi necessário realizar a etapa de limpeza dos dados, com o objetivo de garantir a consistência e a integridade do conjunto que seria utilizado para o treinamento dos modelos. Inicialmente, foram selecionadas apenas as colunas relevantes de cada arquivo CSV, descartando informações irrelevantes ou redundantes que acompanhavam os dados brutos provenientes das fontes originais. Essa filtragem permitiu manter apenas os atributos essenciais para a predição, como preços das ações, indicadores técnicos e variáveis macroeconômicas.

Em seguida, os diferentes conjuntos de dados foram integrados em um único arquivo, por meio de scripts em Python. A unificação foi feita utilizando a coluna data como chave principal para sincronizar as diferentes fontes. Após a junção dos dados, foram removidos todos os registros que apresentavam valores nulos (*missing values*), a fim de evitar inconsistências no treinamento dos algoritmos e garantir que todos os modelos recebessem entradas completas e compatíveis.

4.1.3. *Pré Processamento de Dados*

A etapa de transformação dos dados tem o objetivo de preparar o conjunto para o treinamento dos modelos de regressão. Primeiramente, todos os valores numéricos extraídos dos arquivos CSV foram convertidos para o tipo float, garantindo a compatibilidade com os algoritmos de aprendizado de máquina.

Além disso, foram calculados diversos indicadores técnicos diretamente em Python, por meio de funções matemáticas aplicadas sobre os dados de preço e volume das ações. Entre os indicadores adicionados ao conjunto, estão: RSI, MACD, médias móveis exponenciais e simples, volatilidade histórica, retorno diário, *momentum*, ROC, ATR, largura das bandas de Bollinger e um marcador de picos de preço (*spike flag*). Esses indicadores foram adicionados como novas colunas, enriquecendo a base de dados com informações derivadas que auxiliam os modelos na detecção de padrões temporais e tendências de mercado.

Outro ponto essencial foi a sincronização temporal das séries. Os arquivos originais apresentavam datas desordenadas ou com lacunas, além de frequências diferentes entre as variáveis analisadas. Enquanto os preços das ações estavam disponíveis em frequência diária, alguns indicadores macroeconômicos, como o IPCA, eram divulgados em frequência mensal. Para garantir a consistência temporal dos dados, inicialmente foi realizada a ordenação cronológica das séries no intervalo de 01/01/2010 a 31/12/2019. Em seguida, foi aplicado um processo de alinhamento em que cada registro diário do dataset principal, contendo os preços das ações, recebeu os

valores correspondentes das demais variáveis.

4.1.4. *Normalização de Dados*

A normalização dos dados é uma etapa fundamental no pré-processamento, especialmente em algoritmos de aprendizado de máquina sensíveis à escala das variáveis, como redes neurais e máquinas de vetor de suporte. No presente trabalho, foi aplicada a técnica de normalização *StandardScaler*, que transforma os dados de modo que eles tenham uma média de 0 e um desvio padrão de 1.

Para isso, foram utilizadas as funções *StandardScaler*() da biblioteca sklearn.preprocessing. A normalização foi aplicada separadamente nas variáveis independentes (*features*). O vetor de entrada X foi composto pelas variáveis selecionadas para o treinamento, enquanto o vetor Y representou a variável alvo a ser prevista.

Essa abordagem assegurou que todas as variáveis contribuíssem de maneira equilibrada durante o treinamento dos modelos, evitando que atributos com escalas maiores dominassem o processo de aprendizado. A aplicação consistente da normalização também facilitou a convergência dos algoritmos e melhorou a estabilidade numérica das redes neurais.

4.2. Modelos

Após o tratamento e preparação dos dados, foram selecionados quatro algoritmos de aprendizado de máquina para a tarefa de previsão dos preços das ações: XGBoost, LSTM, Random Forest e SVR. A escolha desses modelos levou em consideração sua ampla utilização na literatura para problemas de séries temporais e regressão, bem como sua capacidade de capturar diferentes padrões nos dados.

Cada modelo foi treinado utilizando os mesmos conjuntos de dados e indicadores, de modo a garantir uma base de comparação justa entre as abordagens. O processo de treinamento e avaliação seguiu uma estratégia de validação *walk-forward*, adequada a problemas de séries temporais, na qual os modelos são treinados com dados históricos e avaliados em janelas futuras, respeitando a ordem temporal das observações.

Durante a fase de treinamento, os dados foram organizados em janelas sequenciais de treino e teste, e as previsões geradas ao longo do processo foram divididas em 30 e 60 dias para avaliação global do desempenho. As previsões obtidas foram avaliadas por meio de métricas apropriadas para problemas de regressão, como RMSE, MAE, MAPE e R^2.

Nos subtópicos a seguir, são apresentadas as características e configurações adotadas para cada modelo, destacando suas principais particularidades, os parâmetros utilizados e o motivo da escolha de cada abordagem no contexto deste trabalho.

4.2.1. *Estratégia de Validação Walk-Forward*

A avaliação dos modelos preditivos desenvolvidos neste trabalho foi realizada por meio da validação *walk-forward*, uma abordagem específica para séries temporais, adotada de forma padronizada em todos os modelos analisados. Diferentemente de métodos tradicionais de validação que realizam divisões aleatórias dos dados, a

validação *walk-forward* preserva a ordem temporal das observações, garantindo que o modelo seja treinado apenas com informações passadas e avaliado em dados futuros. Essa estratégia é essencial em problemas financeiros, nos quais o uso de informações futuras durante o treinamento caracterizaria vazamento de dados (*data leakage*), comprometendo a validade dos resultados.

A série temporal foi dividida em janelas sequenciais de treinamento e teste, com tamanho fixo. Para cada iteração do processo, o modelo foi treinado utilizando um intervalo histórico de dados e, em seguida, testado no intervalo imediatamente posterior. Após a etapa de teste, a janela foi deslocada para frente no tempo, repetindo-se o procedimento até o final da série.

Formalmente, o processo pode ser descrito da seguinte forma, além disso, os valores adotados serão apresentados no tópico dos seus respectivos modelos:

- Uma janela de tamanho W foi utilizada para treinamento do modelo;
- Um intervalo subsequente de tamanho T foi reservado para teste;
- Após a avaliação, a janela foi deslocada em T observações;
- O procedimento foi repetido até que não houvesse dados suficientes para compor uma nova janela de teste.

Neste trabalho, foi adotada uma janela deslizante, na qual o tamanho do conjunto de treinamento permaneceu constante ao longo das iterações, refletindo um cenário realista de re-treinamento periódico dos modelos.

Em cada iteração da validação *walk-forward*, o modelo foi treinado exclusivamente com os dados disponíveis na janela de treinamento, após isso, foram geradas previsões para todo o intervalo de teste associado àquela janela.

As previsões obtidas em cada iteração foram armazenadas e concatenadas, formando uma série contínua de valores previstos ao longo do tempo. Dessa forma, os valores reais correspondentes foram armazenados, permitindo a comparação direta entre valores observados e previstos ao final do processo.

4.2.2. *XGBoost*

O algoritmo XGBoost (*Extreme Gradient Boosting*) foi utilizado neste trabalho por sua eficiência e desempenho em tarefas de regressão, especialmente em conjuntos de dados tabulares com múltiplas variáveis. Para a implementação, utilizou-se a biblioteca $xgboost$, por meio da classe $XGBRegressor$, que permite personalizar os hiperparâmetros do modelo de forma prática.

O modelo foi configurado com os seguintes parâmetros: $objective = 'reg{:}\,squarederror'$, que define o problema como uma regressão com função de perda baseada no erro quadrático; $n\_estimators = 300$, indicando o número de árvores que compõem o modelo; $learning\_rate = 0.1$, que controla a taxa de atualização a cada iteração; $max\_depth = 4$, limitando a profundidade de cada árvore para evitar *overfitting*; $subsample = 1.0$, que define a fração de amostras

utilizadas por árvore; $colsample\_bytree = 0.8$, especificando a fração de variáveis consideradas por árvore; e $random\_state = 30$, garantindo reprodutibilidade dos resultados.

O conjunto de dados foi dividido em 80% para treinamento e 20% para teste, com as variáveis previamente normalizadas. A variável alvo utilizada foi o preço de fechamento da ação. O modelo foi treinado com os dados históricos e indicadores técnicos previamente preparados, permitindo ao XGBoost aprender padrões complexos e realizar previsões de forma eficiente.

4.2.3. *LSTM*

O modelo LSTM (Long Short-Term Memory) foi utilizado com o objetivo de capturar padrões temporais e dependências de longo prazo nos dados históricos das ações. Essa arquitetura é especialmente adequada para séries temporais, como é o caso da previsão de preços no mercado financeiro. A implementação foi realizada com a biblioteca TensorFlow, utilizando a interface Sequential do Keras.

A arquitetura do modelo contou com uma camada LSTM contendo 64 unidades, configurada com $return\_sequences = False$ por se tratar de uma tarefa de regressão univariada. Em seguida, foi adicionada uma camada de *Dropout* com taxa de 20%, com o intuito de reduzir o *overfitting*. A camada final foi uma camada densa com um único neurônio, responsável por gerar a saída contínua correspondente no preço da ação.

O modelo foi compilado com a função de perda *mean_squared_error*, e o otimizador utilizado foi o $Adam$, com taxa de aprendizado de 0.01. O treinamento foi conduzido durante até 80 épocas, com *batch size* de 16 e divisão de 20% dos dados de treino para validação. Para evitar o *overfitting*, foi aplicado o mecanismo de *early stopping*, que interrompe o treinamento quando não há melhora na perda de validação após 10 épocas consecutivas.

A LSTM foi aplicada tanto aos dados da TAEE11 quanto da CMIG4, utilizando os dados normalizados e estruturados em janelas de tempo, conforme exigido por redes neurais recorrentes.

4.2.4. *Random Forest*

O modelo Random Forest foi utilizado neste trabalho como uma abordagem de aprendizado de máquina baseada em árvores de decisão. Esse algoritmo é conhecido por sua robustez, capacidade de lidar com dados ruidosos e por evitar *overfitting* por meio da construção de múltiplas árvores de forma aleatória. Para a implementação, foi utilizada a classe $RandomForestRegressor$ da biblioteca Scikit-learn.

O modelo foi configurado com 300 árvores ($estimators = 300$), sem limitação de profundidade ($maxdepth = None$), permitindo que cada árvore crescesse livremente conforme os dados. Os parâmetros de divisão foram definidos como $min\_samples\_split = 3$ e $min\_samples\_leaf = 1$, controlando a complexidade das árvores. O parâmetro $max\_features = \sqrt{x}$ foi utilizado para limitar o número de atributos considerados em cada divisão, promovendo diversidade entre as árvores.

A construção das árvores foi feita com *bootstrap sampling* ($bootstrap = True$), e o

modelo foi avaliado internamente com a métrica de *out-of-bag score* ($oob\_score = True$), uma técnica que utiliza as amostras não incluídas no treinamento de cada árvore para estimar o desempenho do modelo. A execução foi paralisada com $n\_jobs = -1$ para otimizar o tempo de processamento.

O modelo foi aplicado aos dados das ações TAEE11 e CMIG4, utilizando as variáveis normalizadas e os indicadores técnicos previamente calculados. A Random Forest foi utilizada para estimar diretamente os preços de fechamento, com o objetivo de capturar padrões não lineares entre as variáveis explicativas e o comportamento do ativo.

4.2.5. *SVR*

O modelo Support Vector Regression (SVR) foi utilizado com o intuito de aplicar uma abordagem baseada em margens de erro e maximização da generalização para prever os preços das ações. O SVR é uma extensão do algoritmo de Support Vector Machine (SVM) voltado para tarefas de regressão, buscando encontrar uma função que se ajuste aos dados dentro de uma margem de tolerância (*epsilon*), enquanto penaliza desvios fora dessa margem por meio de uma função de custo controlada pelo parâmetro C.

A implementação foi feita por meio da classe SVR da biblioteca *Scikit-learn*, utilizando o kernel linear ($kernel = 'linear'$) para modelar a relação entre as variáveis de forma direta e interpretável. O modelo foi configurado com $C = 20$, o que indica uma penalização menos rígida para erros fora da margem, $gamma = scale$, que controla a influência de cada ponto de dado na função de decisão, e $epsilon = 0.05$, definindo uma faixa muito estreita de tolerância ao erro.

O modelo foi aplicado sobre os dados das ações TAEE11 e CMIG4, previamente normalizados. Foram utilizadas todas as variáveis técnicas e macroeconômicas como entradas, e a saída foi definida como o preço de fechamento.

Embora o SVR tenha desempenho competitivo em certos cenários, sua aplicação em séries temporais financeiras é desafiadora devido à sensibilidade a *outliers*, à necessidade de ajuste fino dos hiperparâmetros e à limitação de escalabilidade em grandes volumes de dados. Ainda assim, sua inclusão no estudo permitiu uma comparação relevante com os demais modelos mais complexos.

4.3. Ameaças à validade e limitações

A análise de modelos de aprendizado de máquina aplicada à previsão de preços de ações envolve uma série de decisões metodológicas que podem influenciar diretamente os resultados obtidos. Nesse contexto, é fundamental discutir as ameaças à validade interna e externa do estudo, bem como suas principais limitações, a fim de fornecer uma interpretação adequada dos resultados e delimitar o escopo das conclusões apresentadas.

No que se refere à validade interna, uma das principais ameaças está relacionada às escolhas realizadas durante o processo de modelagem. A definição do tamanho da janela temporal utilizada para o treinamento dos modelos, bem como dos horizontes de previsão de 30 e 60 dias, influencia a forma como os algoritmos capturam padrões de curto e longo prazo nas séries temporais. Janelas maiores tendem a favorecer a aprendizagem de tendências mais estáveis, enquanto janelas menores podem tornar os

modelos mais sensíveis a oscilações locais e ruídos do mercado. Dessa forma, os resultados obtidos refletem diretamente essas decisões e podem variar caso outras configurações sejam adotadas. Outro fator relevante para a validade interna é a seleção do conjunto de variáveis utilizadas no treinamento. O uso de indicadores técnicos, dados macroeconômicos e informações setoriais enriquece a representação do comportamento do mercado, porém também introduz dependência em relação à qualidade e à relevância desses indicadores. A estratégia de normalização dos dados e a abordagem de previsão direta para múltiplos horizontes também impactam a estabilidade e a capacidade de generalização dos modelos.

Quanto à validade externa, os resultados deste estudo não podem ser generalizados para qualquer ativo financeiro ou contexto de mercado. A análise foi conduzida com ações específicas do setor elétrico brasileiro, que apresentam características próprias, como menor volatilidade relativa, comportamento regulado e ciclos de preços distintos de setores como tecnologia ou commodities. Assim, o desempenho dos modelos observado neste trabalho pode não se repetir em ativos com dinâmicas diferentes ou em mercados internacionais. Além disso, o período histórico analisado influencia diretamente o aprendizado dos algoritmos. As previsões foram realizadas com base em dados coletados dentro de um intervalo temporal específico, refletindo as condições econômicas, políticas e regulatórias daquele período. Mudanças estruturais no mercado, crises financeiras, choques macroeconômicos ou eventos extraordinários podem alterar significativamente o comportamento dos preços, reduzindo a capacidade dos modelos treinados em dados passados de manter o mesmo nível de desempenho em períodos futuros. Portanto, não é possível afirmar que os resultados se estendem de forma irrestrita para previsões realizadas muito além do intervalo temporal considerado no estudo.

Entre as principais limitações deste trabalho, destaca-se o fato de que as previsões se baseiam exclusivamente em dados históricos e variáveis quantitativas. Eventos exógenos inesperados, como alterações regulatórias, crises políticas ou acontecimentos de grande impacto econômico, não são explicitamente modelados, embora possam exercer influência significativa sobre os preços das ações. Além disso, o estudo se restringe à previsão em frequência diária, não abordando cenários de alta frequência ou aplicações diárias.

Por fim, este trabalho não tem como objetivo propor estratégias de investimento ou tomada de decisão financeira, tampouco garantir desempenho futuro dos modelos analisados. O foco está na avaliação comparativa da capacidade preditiva de diferentes algoritmos de regressão em séries temporais financeiras, considerando métricas estatísticas e análise gráfica. Dessa forma, os resultados devem ser interpretados como uma análise experimental dentro de um contexto específico, servindo como base para estudos futuros que explorem outros ativos, horizontes de previsão ou abordagens metodológicas.

5. Resultados

Este capítulo apresenta os resultados obtidos a partir da aplicação dos quatro modelos de regressão propostos: XGBoost, LSTM, Random Forest e SVR, com o objetivo de prever o preço das ações TAEE11 e CMIG4. Para avaliar o desempenho de cada modelo, foram utilizadas quatro métricas amplamente aplicadas em problemas de regressão: RMSE, MAE, MAPE e R^2. Essas métricas permitem quantificar o erro entre

os valores reais e os valores previstos, considerando diferentes aspectos da distribuição do erro. Além disso, todos os modelos foram configurados em ambas as ações com uma janela de treinamento *(window_size)* de 110 dias e um conjunto de teste *(test_size)* de 30 dias.

Os tópicos estão separados para análise das duas ações individualmente, no qual cada subcapítulo apresenta os resultados específicos e os gráficos dos respectivos modelos para cada ação, após isso, a comparação final será mostrada em uma tabela com todos os modelos aplicados para tal ação e suas métricas.

Essa abordagem visa não apenas comparar numericamente o desempenho dos modelos, mas também fornecer uma avaliação visual da aderência das previsões aos dados reais, permitindo uma análise mais completa da qualidade de cada técnica.

5.1. CMIG4

5.1.1. *XGBoost*

Essa abordagem de *walk-forward* permitiu que o modelo fosse treinado continuamente em períodos recentes da série temporal, ajustando-se de forma incremental às mudanças no comportamento do preço da ação. O XGBoost foi afinado para capturar variações de curto prazo. O conjunto de teste foi utilizado para avaliar o desempenho do modelo em dados não vistos, garantindo uma simulação realista da capacidade preditiva.

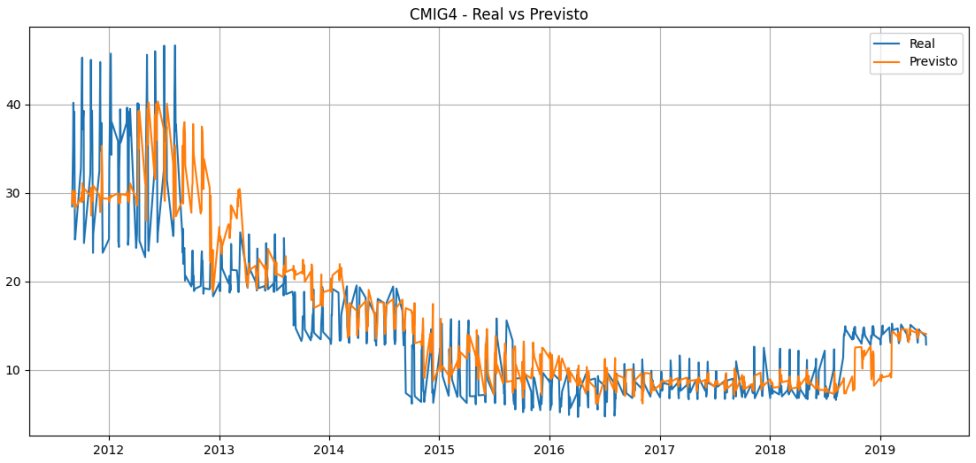

Figura 3: Gráfico da previsão de 30 dias da ação CMIG4 do algoritmo XGBoost.

Fonte: Elaborado pelo autor.

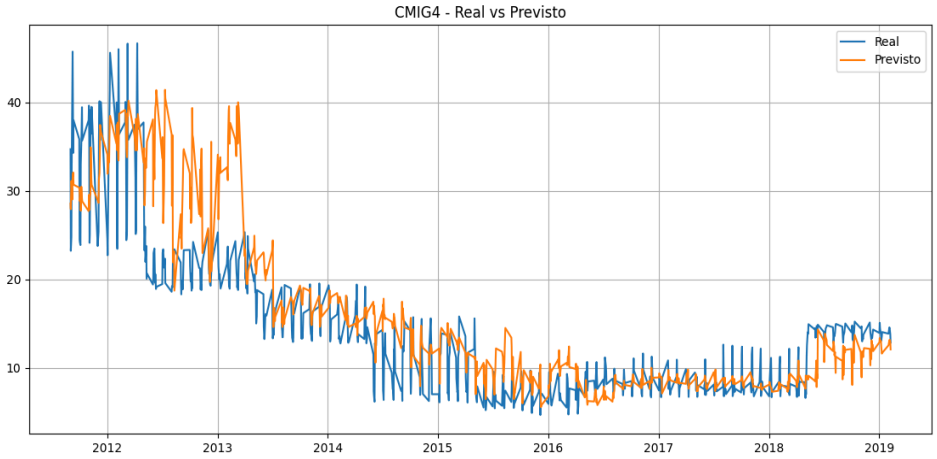

Figura 4: Gráfico da previsão de 60 dias da ação CMIG4 do algoritmo XGBoost.

Fonte: Elaborado pelo autor.

Como observado nos gráficos, o modelo reconheceu bem as variações de curto prazo, mantendo os picos e vales sem suavização excessiva na média. No entanto, apresentou um leve atraso para identificar novas tendências em mudanças mais bruscas de ciclo de preços, refletindo a dependência do XGBoost em padrões históricos recentes, contudo, a previsão de 30 dias teve um desempenho melhor em relação ao de 60 nesse quesito, o qual adaptou-se mais rápido à mudança de tendência. Apesar desse atraso momentâneo, o modelo mostrou boa sensibilidade às oscilações locais e foi capaz de capturar adequadamente os movimentos da série temporal durante o período de teste.

5.1.2. *LSTM*

Para a ação CMIG4, o modelo LSTM foi configurado com um comprimento de sequência (*seq_len*) de 5 dias. O *walk-forward* foi aplicado para que o modelo fosse treinado continuamente em blocos recentes da série temporal, ajustando-se às mudanças nos padrões de preços de forma incremental. O conjunto de teste foi utilizado para avaliar o desempenho do modelo em dados não vistos, assegurando uma medição mais realista.

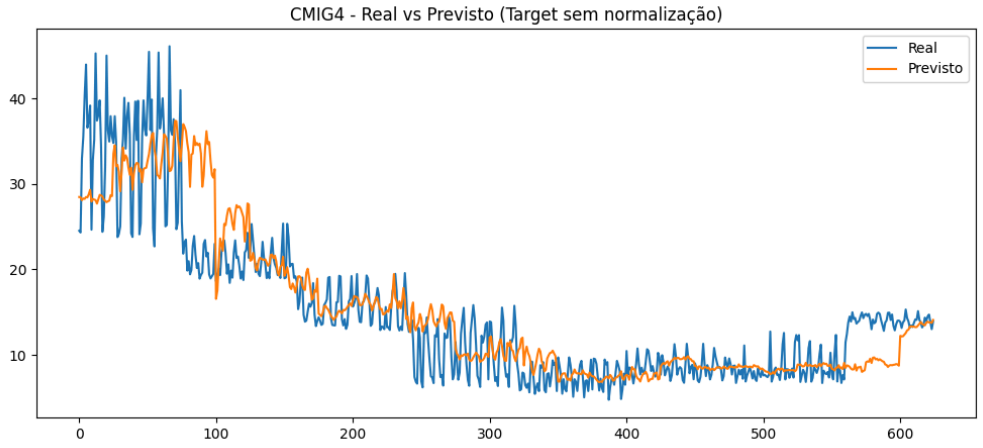

Figura 5: Gráfico da previsão de 30 dias da ação CMIG4 do algoritmo LSTM.

Fonte: Elaborado pelo autor.

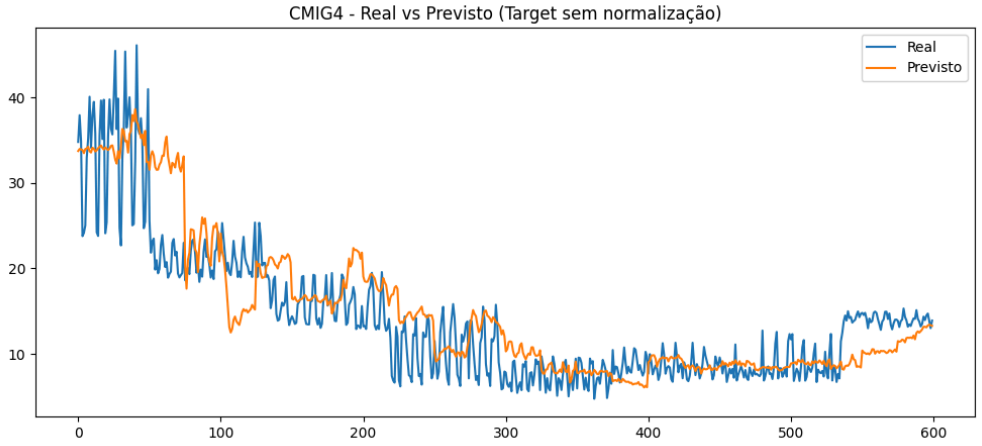

Figura 6: Gráfico da previsão de 60 dias da ação CMIG4 do algoritmo LSTM.

Fonte: Elaborado pelo autor.

O modelo LSTM apresentou suavização nos picos e vales, acompanhando com um certo atraso em mudanças mais bruscas as tendências de longo prazo da ação. Apesar dessa suavização, o modelo conseguiu seguir adequadamente os ciclos de preço, porém as previsões de 60 dias tiveram um desempenho inferior, o qual conteve alguns *outliers*, refletindo pequenas discrepâncias momentâneas na previsão em relação aos valores reais.

5.1.3. *SVR*

O método de *walk-forward* foi aplicado, permitindo que o modelo fosse treinado em blocos consecutivos da série temporal e ajustasse suas previsões conforme os padrões mais recentes dos preços. Essa abordagem garante que o SVR aprenda de forma contínua e avalie seu desempenho em dados não vistos, oferecendo uma medição realista da capacidade preditiva do modelo.

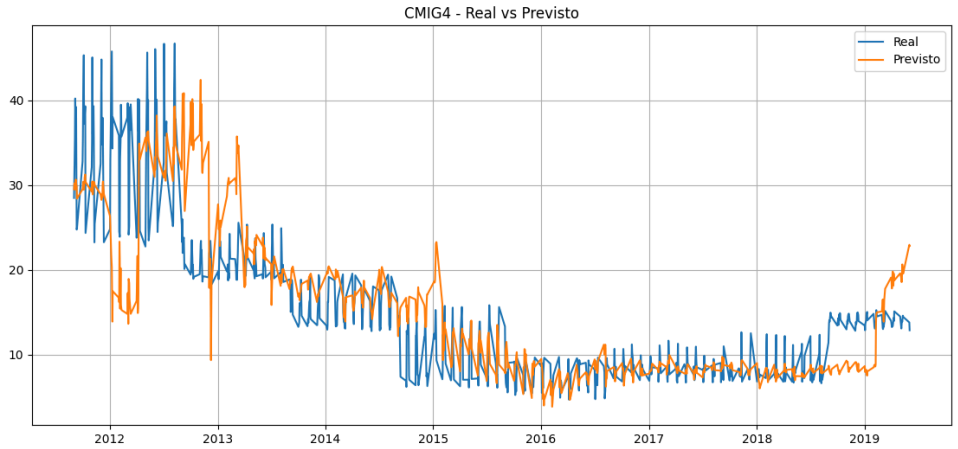

Figura 7: Gráfico da previsão de 30 dias da ação CMIG4 do algoritmo SVR.

Fonte: Elaborado pelo autor.

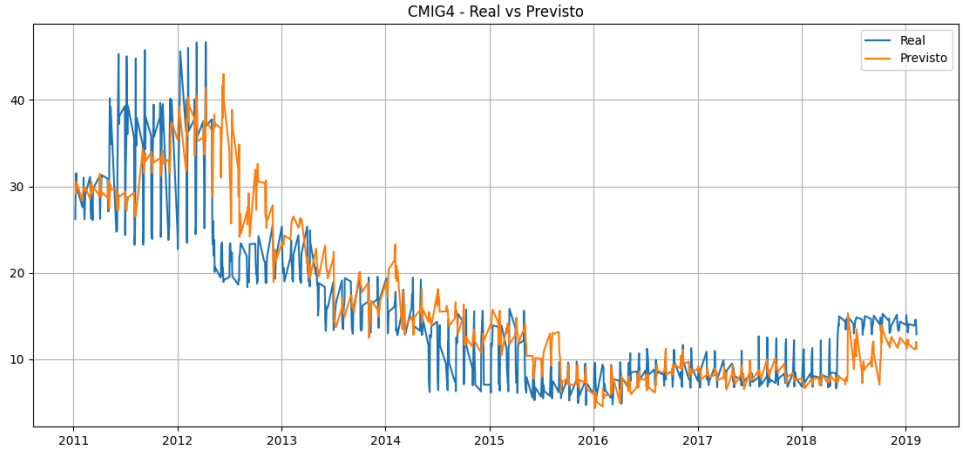

Figura 8: Gráfico da previsão de 60 dias da ação CMIG4 do algoritmo SVR.

Fonte: Elaborado pelo autor.

O desempenho do SVR apresentou comportamento distinto entre as previsões de 30 e 60 dias: nas de 30 houveram muitos erros no começo do gráfico, embora após isso o modelo tenha ajustado o ciclo, já na previsão de 60 dias esses erros não aconteceram, já que o modelo apresentou leve suavização no início, além disso, mostrou-se eficiente em acompanhar as tendências e mudanças de longo prazo, ajustando-se bem aos ciclos da ação e mantendo consistência nas previsões ao longo do tempo.

5.1.4. *Random Forest*

O desempenho do Random Forest apresentou uma suavização moderada nos picos e vales de curto prazo, com um leve atraso na identificação de tendências de longo prazo. No entanto, o modelo de 60 dias gerou *outliers* no início do gráfico por conta do atraso em reconhecer a mudança de tendência, no geral, ambos conseguiram seguir de forma estável os ciclos de preços da ação.

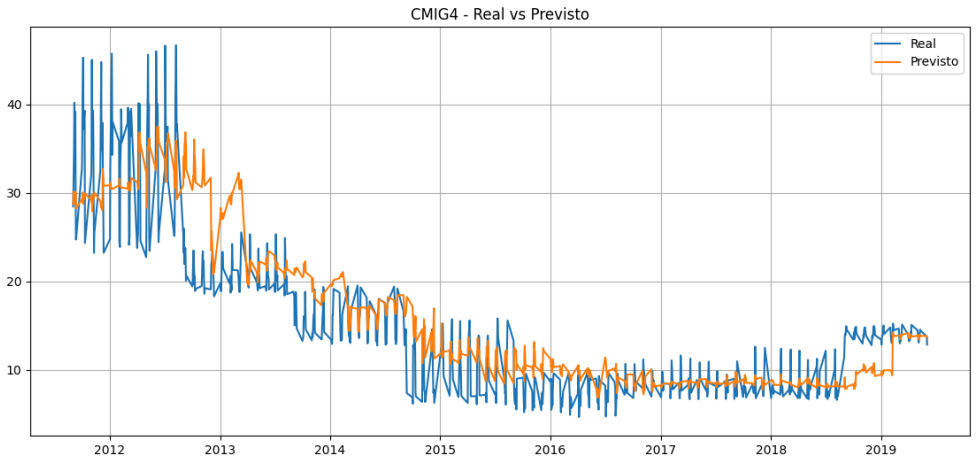

Figura 9: Gráfico da previsão de 30 dias da ação CMIG4 do algoritmo random forest.
Fonte: Elaborado pelo autor.

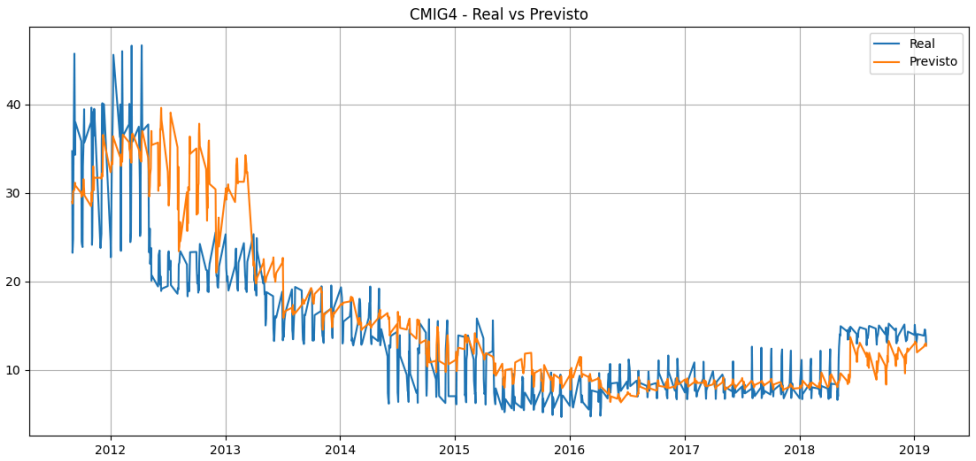

Figura 10: Gráfico da previsão de 60 dias da ação CMIG4 do algoritmo random forest.
Fonte: Elaborado pelo autor.

5.1.5. *Discussão dos Resultados*

Este tópico tem como ideia principal responder as perguntas definidas nos objetivos do trabalho, além disso, mostrar os aspectos gerais de cada modelo diante das mudanças aplicadas:

1. Qual algoritmo de aprendizado de máquina apresenta o melhor desempenho preditivo para cada ação analisada (CMIG4 e TAEE11), considerando métricas de erro e aderência às tendências de preços?
2. Como os diferentes modelos se comportam em relação à captura de variações de

curto prazo, reconhecimento de mudanças de tendência e estabilidade das previsões ao longo do tempo?

Com base nas tabelas de 30 e 60 dias para a CMIG4, podemos fazer a seguinte análise dos modelos:

Para o horizonte de 30 dias, observa-se que o LSTM apresenta o melhor desempenho geral em termos de equilíbrio entre RMSE, MAE, R^2 e MAPE, conseguindo capturar bem as tendências. O Random Forest apresenta métricas próximas, especialmente em R^2, mas o erro percentual (MAPE) é ligeiramente maior. O XGBoost e o SVR tiveram desempenhos inferiores, sendo que o SVR apresentou o maior RMSE, indicando dificuldade em acompanhar fielmente os picos e vales de curto prazo nesse período.

MODELO	RMSE	MAE	R2	MAPE
LSTM	4.5316	3.0895	0.7389	0.2163
XGB	4.7330	3.2736	0.7250	0.2361
SVR	6.0363	3.8895	0.5526	0.2574
RF	4.6569	3.3133	0.7337	0.2435

Tabela 1: Resultados das métricas dos 30 dias previstos dos modelos utilizados na CMIG4

MODELO	RMSE	MAE	R2	MAPE
LSTM	4.4274	3.2420	0.6881	0.2563
XGB	5.4442	3.5560	0.5524	0.2624
SVR	5.0133	3.4872	0.7064	0.2482
RF	5.0211	3.4249	0.6193	0.2532

Tabela 2: Resultados das métricas dos 60 dias previstos dos modelos utilizados na CMIG4

No horizonte de 60 dias, a tendência muda um pouco. O LSTM ainda mantém boa performance, mas observa-se uma leve redução em R^2 e aumento no MAPE, mostrando que quanto maior o horizonte de previsão, o modelo suaviza mais as variações e tem mais dificuldade em capturar movimentos abruptos. O SVR se destaca nesse horizonte por apresentar o maior R^2 entre os modelos para 60 dias, sugerindo que consegue acompanhar a tendência de longo prazo melhor que XGBoost e Random Forest, que tiveram aumento considerável nos erros absolutos e percentuais. O XGBoost apresenta o pior desempenho para 60 dias, indicando que seu ajuste para o

curto prazo não generaliza tão bem quando o horizonte de previsão aumenta.

5.2. TAEE11

5.2.1. *XGBoost*

A configuração foi mantida de forma consistente em relação aos demais modelos aplicados à mesma ação, garantindo comparabilidade entre os resultados. O modelo foi ajustado com os parâmetros que apresentaram melhor desempenho durante os testes, e avaliado com base nas métricas de erro e na análise gráfica das previsões em relação aos valores reais.

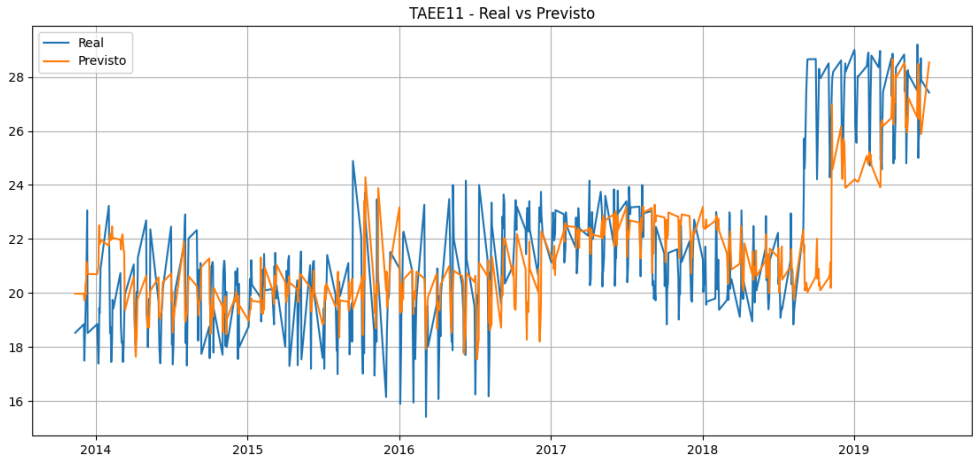

Figura 11: Gráfico da previsão de 30 dias da ação TAEE11 do algoritmo XGBoost.

Fonte: Elaborado pelo autor.

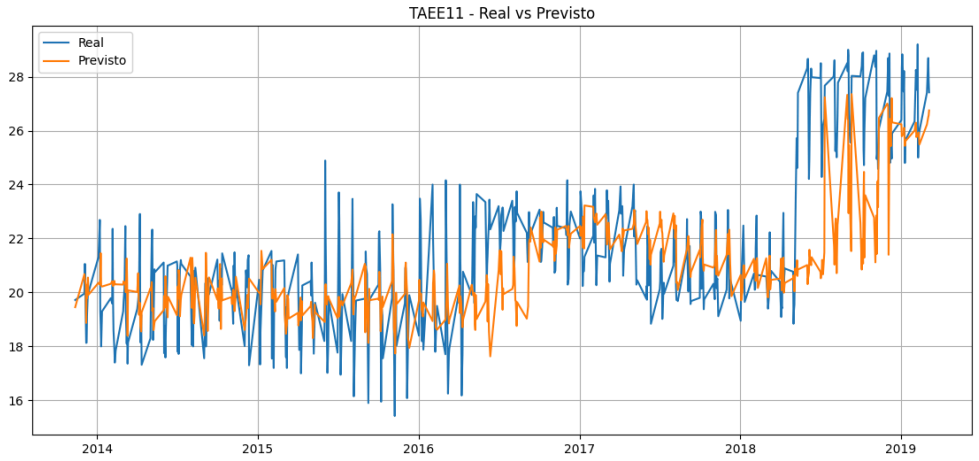

Figura 12: Gráfico da previsão de 60 dias da ação TAEE11 do algoritmo XGBoost.

Fonte: Elaborado pelo autor.

A análise dos gráficos indica que o XGBoost foi capaz de acompanhar de forma eficaz tanto a tendência de longo prazo quanto as variações de curto prazo dos preços da ação

TAEE11 em ambas previsões. O modelo conseguiu reproduzir adequadamente os picos e vales da série, demonstrando boa sensibilidade às oscilações locais. Observou-se apenas um atraso pontual na identificação de uma mudança de tendência ocorrida por volta do ano de 2019, período no qual o modelo demorou alguns instantes para se ajustar ao novo ciclo de preços. Após essa adaptação, o XGBoost voltou a apresentar previsões alinhadas ao comportamento real da série, indicando boa capacidade de generalização para esse ativo.

Em resumo, para a CMIG4: 30 dias o LSTM é o mais equilibrado, RF próximo, SVR e XGBoost ficam atrás. Já nos 60 dias o LSTM ainda performa bem, mas SVR se aproxima em R^2, enquanto XGBoost apresenta maior dificuldade. O aumento do horizonte de previsão impacta todos os modelos, com aumento de erros e diminuição do R^2, mostrando que a previsão de longo prazo é mais desafiadora.

5.2.2. *LSTM*

Para a ação TAEE11, o modelo LSTM foi treinado utilizando a estratégia *walk-forward*, com sequências temporais (*seq_len*) de 5 períodos. Essa configuração foi definida buscando equilibrar a capacidade do modelo em capturar dependências temporais com a estabilidade do treinamento. O uso de sequências permitiu ao LSTM explorar padrões temporais mais longos, enquanto a validação *walk-forward* garantiu que as previsões fossem sempre realizadas em dados futuros, respeitando a natureza temporal da série.

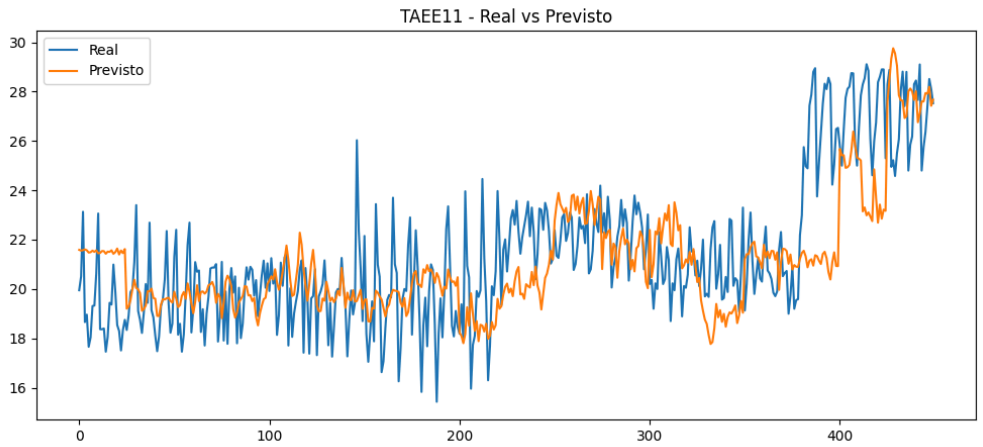

Figura 13: Gráfico da previsão de 30 dias da ação TAEE11 do algoritmo LSTM.

Fonte: Elaborado pelo autor.

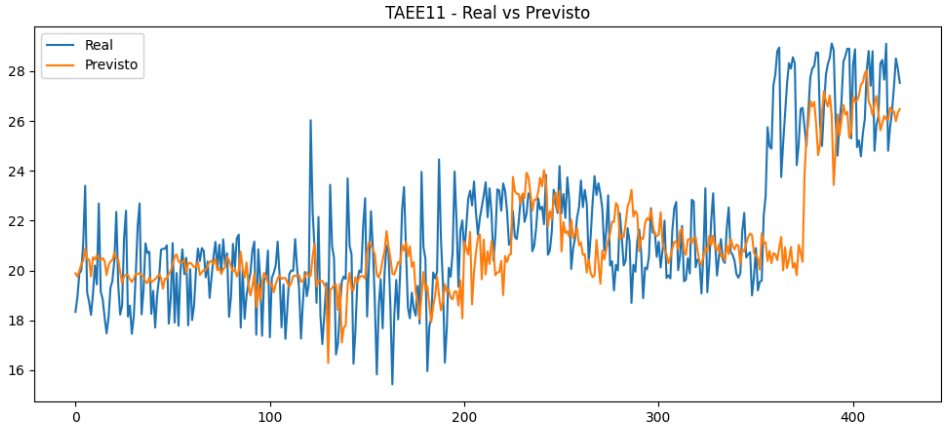

Figura 14: Gráfico da previsão de 60 dias da ação TAEE11 do algoritmo LSTM.

Fonte: Elaborado pelo autor.

A análise dos gráficos revela que o modelo apresentou um comportamento semelhante nas duas propostas, em alguns trechos, as previsões tendem a representar uma média dos movimentos, enquanto em outros momentos o modelo conseguiu acompanhar de forma mais próxima os picos e vales da série. Observou-se também a presença de erros mais evidentes em determinados ciclos de preços, especialmente em torno do dia 400, período no qual o modelo não conseguiu acompanhar adequadamente o comportamento real dos preços. Além disso, de forma semelhante ao XGBoost, o LSTM apresentou um atraso na identificação de uma mudança de tendência observada no final do gráfico, indicando dificuldade momentânea de adaptação a um novo padrão de preços.

5.2.3. *SVR*

A configuração permitiu que o modelo fosse continuamente estimado ao longo da série temporal, utilizando apenas informações passadas para realizar previsões futuras, mantendo a consistência temporal do processo de avaliação. Os parâmetros do SVR foram definidos buscando um compromisso entre capacidade de generalização e sensibilidade às variações dos dados.

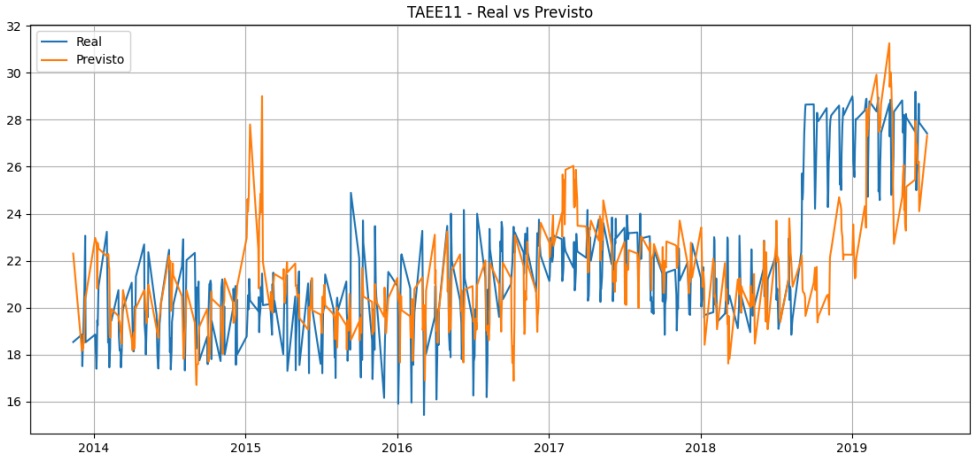

Figura 15: Gráfico da previsão de 30 dias da ação TAEE11 do algoritmo SVR.

Fonte: Elaborado pelo autor.

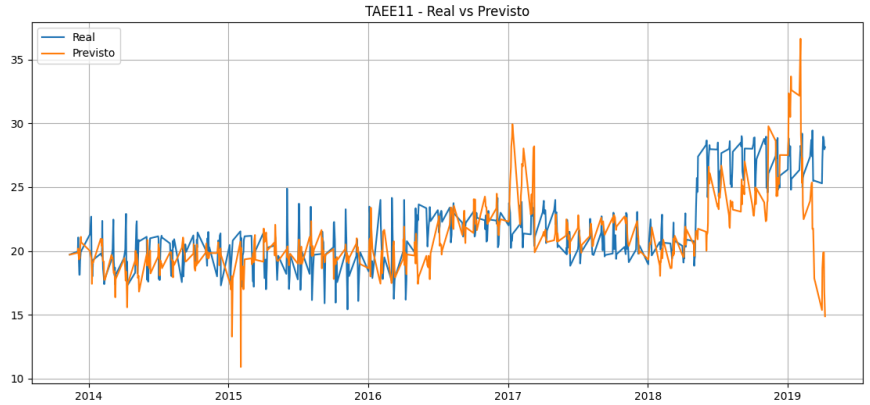

Figura 16: Gráfico da previsão de 60 dias da ação TAEE11 do algoritmo SVR.

Fonte: Elaborado pelo autor.

A partir da análise dos gráficos, observa-se que o modelo apresentou um comportamento mais agressivo em relação aos preços, tendo muitos *outliers* e ruídos em ambos, esse comportamento indica uma sensibilidade pontual do modelo a padrões específicos do conjunto de treinamento naquele período. Apesar disso, o comportamento da tendência de longo prazo foi bem capturado, acompanhando de forma consistente os ciclos principais da série.

5.2.4. *Random Forest*

A configuração possibilita que o modelo fosse continuamente estimado ao longo da série temporal, garantindo que cada previsão fosse feita utilizando apenas dados históricos disponíveis até aquele ponto, mantendo a consistência temporal da análise. Os parâmetros do modelo foram definidos buscando equilíbrio entre capacidade de generalização e sensibilidade às variações de curto e longo prazo.

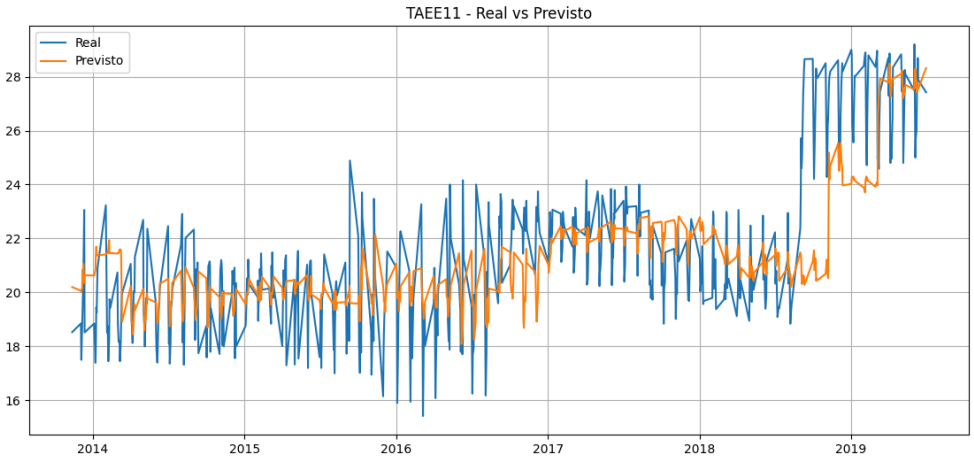

Figura 17: Gráfico da previsão de 30 dias da ação TAEE11 do algoritmo random forest.
Fonte: Elaborado pelo autor.

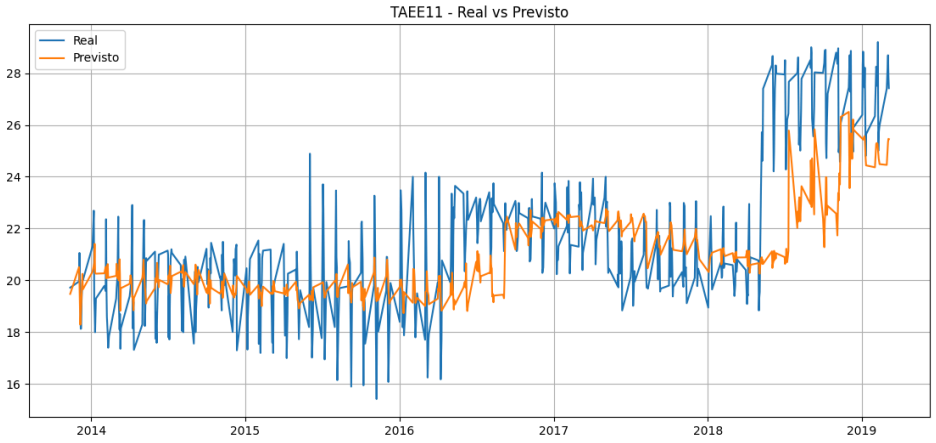

Figura 18: Gráfico da previsão de 60 dias da ação TAEE11 do algoritmo random forest.
Fonte: Elaborado pelo autor.

Na análise dos gráficos, observa-se que o modelo acompanhou bem a tendência de longo prazo, apresentando o mesmo atraso identificado no XGBoost durante a mudança de tendência em 2019, no final do gráfico. A previsão se manteve equilibrada, sem suavizar excessivamente nem exagerar os picos e vales. Entre 2015 e 2016, quando ocorreram alterações de curto prazo mais extremas, o modelo conseguiu manter a mesma variação de curto prazo observada nos dados, sem amplificar os picos ou vales, demonstrando estabilidade e consistência na captura dos ciclos do preço da ação.

5.2.5. *Discussão dos Resultados*

Este tópico tem como ideia principal responder as perguntas definidas nos objetivos do trabalho, além disso, mostrar os aspectos gerais de cada modelo diante das mudanças aplicadas:

1. Qual algoritmo de aprendizado de máquina apresenta o melhor desempenho preditivo para cada ação analisada (CMIG4 e TAEE11), considerando métricas de erro e aderência às tendências de preços?
2. Como os diferentes modelos se comportam em relação à captura de variações de curto prazo, reconhecimento de mudanças de tendência e estabilidade das previsões ao longo do tempo?

Com base nas tabelas de 30 e 60 dias para a TAEE11, podemos analisar o desempenho dos modelos da seguinte forma:

Para o horizonte de 30 dias, o Random Forest apresenta o melhor desempenho geral, com menor RMSE e MAE, além de R^2 razoavelmente alto e o menor MAPE, indicando boa aderência às tendências e menor erro percentual. O XGBoost também teve resultados competitivos, com desempenho próximo ao RF, especialmente em R^2 e MAPE, mostrando que consegue acompanhar bem o comportamento do ativo nesse curto período. O LSTM teve desempenho ligeiramente inferior ao RF e XGBoost, com R^2 menor, refletindo um pouco de suavização das variações de curto prazo. O SVR apresentou o pior desempenho, com maiores erros e menor capacidade de seguir os movimentos de preço, embora ainda tenha conseguido captar a tendência geral.

MODELO	RMSE	MAE	R2	MAPE
LSTM	2.3214	1.7557	0.3795	0.0803
XGB	2.1711	1.6169	0.4443	0.0750
SVR	2.5489	1.9286	0.2341	0.0889
RF	2.1141	1.5845	0.4731	0.0732

Tabela 3: Resultados das métricas dos 30 dias previstos dos modelos utilizados na TAEE11

MODELO	RMSE	MAE	R2	MAPE
LSTM	2.1801	1.6141	0.4581	0.0733
XGB	2.2989	1.6700	0.3816	0.0750
SVR	2.7562	1.9478	0.1598	0.0869
RF	2.2721	1.6756	0.3959	0.0750

Tabela 4: Resultados das métricas dos 60 dias previstos dos modelos utilizados na TAEE11

Já para o de 60 dias, o LSTM apresenta melhoria em R^2 e MAPE em comparação com os 30 dias, indicando que consegue generalizar melhor para previsões de médio prazo.

O RF mantém desempenho consistente, embora com aumento do RMSE e MAE, mostrando que a previsão de longo prazo gera maior incerteza, mas ainda é o modelo mais estável. O XGBoost apresentou queda em R^2, sugerindo maior dificuldade em prever além do curto prazo, enquanto o SVR teve o pior desempenho, com R^2 ainda mais baixo e erros maiores, confirmando sua limitação para horizontes mais longos.

De modo geral, considerando o equilíbrio entre aderência às tendências e resposta às variações locais, o Random Forest se mostrou o modelo mais adequado para a TAEE11, apresentando a melhor capacidade de previsão entre os algoritmos testados.

5.3 Impacto do Walk-Forward na Qualidade das Previsões

Antes da adoção do processo de validação *Walk-Forward*, os modelos eram treinados utilizando um grande volume de dados históricos de forma acumulada. Embora essa abordagem fornecesse uma visão ampla do comportamento passado dos preços, ela apresentava limitações importantes. Em especial, observou-se que os modelos frequentemente apresentavam atraso na identificação de novas tendências e, em alguns casos, não reconheciam adequadamente mudanças estruturais nos preços. Isso ocorria porque a presença de longas janelas históricas fazia com que o aprendizado fosse excessivamente influenciado por padrões antigos, resultando em previsões demasiadamente suavizadas ou pouco responsivas às transições de regime de preços.

Com a implementação do método *Walk-Forward*, esse cenário mudou de forma significativa. O processo permitiu ajustar dinamicamente o período de treinamento utilizado em cada etapa, de modo que o modelo passasse a aprender prioritariamente com informações mais recentes e relevantes. A definição de uma janela de treinamento adequada possibilita alinhar o aprendizado dos modelos aos ciclos de mercado observados, aumentando a capacidade de adaptação às mudanças no comportamento das séries.

Como resultado, os modelos passaram a apresentar:

- Menor atraso na detecção de novas tendências, respondendo de forma mais rápida a transições de preço;
- Redução da suavização excessiva dos movimentos, preservando variações importantes no curto prazo;
- Previsões mais aderentes ao comportamento real das ações ao longo do tempo.

Portanto, a aplicação do *Walk-Forward* não apenas aprimorou a dinâmica de treinamento dos modelos, como também contribuiu diretamente para a melhoria da qualidade preditiva e da robustez dos resultados.

6. Considerações Finais e Trabalhos Futuros

O presente trabalho realizou uma análise comparativa entre diferentes algoritmos de aprendizado de máquina: LSTM, XGBoost, Random Forest e SVR, aplicados à previsão dos preços das ações CMIG4 e TAEE11. A partir da utilização de dados históricos e indicadores técnicos, foi possível avaliar o desempenho de cada modelo em diferentes horizontes de previsão, considerando métricas como RMSE, MAE,

MAPE e R^2. Os resultados mostraram que a escolha do modelo ideal depende tanto do ativo analisado quanto do horizonte temporal de previsão. Para a CMIG4, o Random Forest apresentou melhor desempenho geral, oferecendo previsões mais estáveis e menor erro percentual, enquanto o LSTM se destacou para horizontes mais longos, acompanhando a tendência de médio prazo de forma satisfatória. Para a TAEE11, embora o Random Forest tenha se mantido consistente, o LSTM mostrou vantagens ao capturar a evolução do preço em períodos mais longos, evidenciando a capacidade de cada algoritmo em lidar com características específicas de cada série temporal.

Além das contribuições práticas relacionadas à previsão de preços, este estudo evidencia a importância do uso de métricas múltiplas e da validação walk-forward, que permite uma análise mais realista do desempenho dos modelos ao longo do tempo, minimizando vieses provocados por padrões históricos específicos. Também foram identificadas limitações do trabalho, como a restrição ao conjunto de ações analisadas e ao período histórico utilizado, bem como a dependência de indicadores técnicos que podem não se aplicar da mesma forma a outros ativos ou mercados. Esses fatores podem influenciar o grau de generalização dos modelos e sugerem cautela na aplicação direta das previsões para outros contextos. Contudo, os resultados obtidos indicam que os modelos propostos apresentam potencial de aplicação prática no apoio à tomada de decisão no mercado financeiro, especialmente em horizontes de curto e médio prazo. Considerando que as previsões foram realizadas para janelas de 30 e 60 dias, os modelos podem ser utilizados como ferramentas auxiliares para decisões de compra ou venda, oferecendo uma estimativa do comportamento esperado dos preços dentro desse intervalo temporal. Embora os modelos não tenham como objetivo prever valores exatos ou antecipar eventos extremos do mercado, observou-se que as previsões permaneceram, em grande parte, dentro do intervalo real de variação dos preços, acompanhando a tendência predominante dos ativos analisados. Dessa forma, os resultados podem contribuir para a identificação de momentos mais favoráveis de entrada ou saída, quando combinados com análise técnica ou fundamentalista. Ressalta-se, contudo, que o uso prático dos modelos deve considerar suas limitações, sendo mais indicado como suporte à decisão, e não como substituto exclusivo da análise humana, especialmente para horizontes superiores a 60 dias ou em cenários de elevada instabilidade do mercado.

Como perspectivas para trabalhos futuros, destaca-se a possibilidade de expandir o estudo para um conjunto mais amplo de ações, incluindo ativos de diferentes setores e características de volatilidade, o que permitiria avaliar a robustez e generalização dos modelos. Outra linha de investigação seria a incorporação de novos tipos de dados, como notícias financeiras, indicadores macroeconômicos adicionais ou séries temporais de alta frequência, que poderiam enriquecer os modelos e melhorar a acurácia das previsões. Nesse contexto, destaca-se também o potencial de utilização de *Large Language Models* (LLMs) para a análise automática de notícias e realização de análise de sentimentos do mercado, permitindo capturar informações qualitativas que influenciam o comportamento dos preços das ações. Além disso, estudos futuros podem investigar relações entre diferentes ações ou setores do mercado, explorando possíveis dependências ou correlações que possam contribuir para melhorar o desempenho dos modelos de previsão. Também é relevante explorar técnicas avançadas de otimização e ajuste de hiperparâmetros, bem como metodologias de ensemble entre diferentes algoritmos, para combinar pontos fortes de cada modelo e potencialmente reduzir erros em previsões de curto e médio prazo.

Por fim, o trabalho contribui para a compreensão da aplicabilidade de algoritmos de aprendizado de máquina em problemas de previsão de preços de ações, fornecendo *insights* sobre a escolha do modelo adequado para diferentes cenários e destacando caminhos para aprofundamento em pesquisas futuras, com foco em maior generalização, precisão e relevância prática no mercado financeiro.

7. Referências

IBM. O que é machine learning (ML)? Disponível em: https://www.ibm.com/br-pt/topics/machine-learning.

SOUZA, Alex. Seu primeiro Projeto de Machine Learning em Python (Passo a Passo). Blog do Zouza, Medium, 2021. Disponível em: https://medium.com/@alexsouza/seu-primeiro-projeto-de-machine-learning-em-python-123abc456.

ALURA. Um pouco sobre séries temporais e suas aplicações. Alura, 2022. Disponível em: https://www.alura.com.br/artigos/series-temporais-aplicacoes.

IBM. O que é uma árvore de decisão (decision tree)? Disponível em: https://www.ibm.com/br-pt/topics/decision-tree.

IBM. O que é o XGBoost? Disponível em: https://www.ibm.com/blogs/research-br/xgboost/

CARVALHO, João Cláudio Nunes. Sobre o algoritmo XGBoost. Medium, 2020. Disponível em: https://medium.com/@joaocncarvalho/sobre-o-algoritmo-xgboost-abcdef123.

JUNIOR, Jose R. F. Redes Neurais Recorrentes — LSTM. Medium, 2021. Disponível em: https://medium.com/@jrfjunior/redes-neurais-recorrentes-lstm-123abc.

IBM. O que é random forest? Disponível em: https://www.ibm.com/br-pt/topics/random-forest.

VERMA, Nandini. Uma introdução à regressão de vetores de suporte (SVR) em aprendizado de máquina. Medium, 2020. Disponível em: https://medium.com/@nandini.verma/svr-introducao.

SCIKIT-LEARN. Regressor de vetor de suporte. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html.

SANTOS, Filipe Ramos de Souza. Predição de preço de ações utilizando aprendizado de máquina. 2023. 46 f. Trabalho de Conclusão de Curso (Bacharelado em [colocar o curso aqui]) – Universidade Federal de Ouro Preto, [Cidade], 2023. Disponível em: https://monografias.ufop.br/bitstream/35400000/5713/6/MONOGRAFIA_Predi%C3%A7%C3%A3oPre%C3%A7oA%C3%A7%C3%B5es.pdf.

PAVÃO, Emerson Antonio Freire. *Inteligência artificial aplicada ao mercado financeiro*. São Paulo: Érica, 2020.

FAHAD, Istiaq Ahmed. *Understanding Walk Forward Validation in Time Series Analysis: A Practical Guide*. Medium, 2024. Disponível em: Understanding Walk Forward Validation in Time Series Analysis: A Practical Guide | by Istiaq Ahmed Fahad | Medium

WAHYUDDIN, Eko Putra et al. *Improved LSTM hyperparameters alongside sentiment walk-forward validation for time series prediction.* Journal of Open Innovation: Technology, Market, and Complexity, [s. l.], v. 11, n. 8, 100458, dez. 2024. Disponível em: Improved LSTM hyperparameters alongside sentiment walk-forward validation for time series prediction - ScienceDirect