UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Análise preditiva aplicada ao Bitcoin: um estudo comparativo

Felipe Boteon Calderaro

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Felipe Boteon Calderaro

Análise preditiva aplicada ao Bitcoin: um estudo comparativo

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Jó Ueyama

Versão original

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

Calderaro, Felipe Boteon
S856m Análise preditiva aplicada ao Bitcoin: um estudo comparativo / Felipe Boteon Calderaro ; orientador Jó Ueyama. – São Carlos, 2023.
59 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Bitcoin. 2. Análise Preditiva. 3. Regressão. I. Ueyama, Jó, orient. II. Título.

Felipe Boteon Calderaro

Análise preditiva aplicada ao Bitcoin: um estudo comparativo

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Jó Ueyama

Original version

São Carlos

2023

Dedico este trabalho à memória do meu cachorro Juca,
fiel companheiro que esteve presente
em grande parte dessa jornada de conhecimento.

RESUMO

Calderaro, F. B. **Análise preditiva aplicada ao Bitcoin: um estudo comparativo**. 2023. 59p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Este trabalho compara diferentes abordagens preditivas para os retornos e volatilidade do Bitcoin utilizando modelos de regressão. Para a modelagem do problema são utilizados dados históricos de preços de diversos ativos financeiros, como índices S&P 500 e Nasdaq, taxa de câmbio EUR/USD, títulos de 10 anos do governo americano, ouro e petróleo, combinados com diferentes índices de sentimento. Durante o processo, são testados diferentes métodos de seleção de variáveis, bem como diferentes modelos preditivos: Linear Regression, K-Neighbors, Decision Tree, Random Forest e Gradient Boosting. Dente os modelos testados, os resultados foram melhores para a predição da volatilidade e o Random Forest foi o modelo que apresentou o melhor desempenho na predição de ambas as variáveis. Todos os modelos de predição de volatilidade foram beneficiados com a inclusão dos índices de sentimento, o que não ocorreu para os modelos de predição dos retornos.

Palavras-chave: Bitcoin. Inteligência Artificial. Regressão.

ABSTRACT

Calderaro, F. B. **Predictive analytics applied to Bitcoin: a comparative study**. 2023. 59p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

This work compares different predictive approaches for Bitcoin returns and volatility using regression models. Historical price data of several financial assets are used, such as S&P 500 and Nasdaq indices, EUR/USD exchange rate, 10-year US government bonds, gold and oil, combined with different sentiment indices. During the process, different variable selection methods are tested, as well as different predictive models: Linear Regression, K-Neighbors Regressor, Decision Tree Regressor, Random Forest Regressor and Gradient Boosting Regressor. Among the tested models, the results were better for the prediction of volatility and the Random Forest Regressor was the one that presented the best performance in the prediction of both dependent variables. All volatility prediction models benefited from the inclusion of sentiment indices, which did not occur for return prediction models.

Keywords: Bitcoin. Artificial intelligence.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DNN	*Deep Neural Network*
DQN	*Deep Q-Network*
GB	*Gradient Boosting*
HAR	*Heterogeneous Autoregressive*
JNN	*Jordan Neural Network*
LR	*Linear Regression*
LSTM	*Long Short-Term Memory*
ML	*Machine Learning*
MLP	*Multilayer Perceptron*
MLR	*Multiple Linear Regression*
NN	*Neural Network*
RF	*Random Forest*
RL	*Reinforcement Learning*
RNN	*Recurrent Neural Network*
SLR	*Simple Linear Regression*
SVM	*Support Vector Machine*
VADER	*Valence Aware Dictionary and sEntiment Reasoner*
VAR	*Vector Autoregression*
SETAR	*Self Exciting Threshold Autoregressive*
NAR	*Non-Linear Autoregressive*

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

Nos últimos anos, temos testemunhado um aumento significativo no uso e na adoção das criptomoedas como parte integrante da economia mundial. O crescente interesse e aceitação das criptomoedas, como o Bitcoin, Ethereum e outras, têm sido impulsionados por uma combinação de fatores, incluindo avanços tecnológicos, facilidade de transferência internacional de fundos, potencial de valorização e a crescente desconfiança em relação às moedas tradicionais em face da instabilidade econômica global. Esse fenômeno tem despertado a atenção de investidores, instituições financeiras e governos, que buscam compreender e regular o impacto das criptomoedas nos mercados financeiros e nas políticas monetárias. Embora o uso generalizado das criptomoedas também tenha levantado preocupações sobre segurança, regulamentação e seu potencial uso em atividades ilegais, não se pode negar que elas estão desempenhando um papel cada vez mais relevante no cenário econômico global. (MOUGAYAR, 2016)

Com a popularização desses ativos digitais, plataformas de negociação eletrônica de criptomoedas têm surgido em todo o mundo, permitindo que investidores possam comprar e vender criptomoedas de forma rápida e fácil, 24 horas por dia, 7 dias por semana. No entanto, o mercado de criptomoedas ainda é altamente volátil e arriscado, o que requer cuidado e atenção por parte dos investidores (OKSANEN *et al.*, 2022). Definir estratégias de negociação de criptomoedas pode ser um desafio complexo, pois existem muitas variáveis em jogo. Alguns fatores a serem considerados incluem o tipo de criptomoeda sendo negociada, a volatilidade do mercado, o nível de risco que o negociante está disposto a assumir e a experiência do negociante em usar uma determinada estratégia. Cada estratégia tem seus próprios prós e contras, e pode ser mais adequada para diferentes situações e estilos de negociação.

A proposta deste trabalho é comparar diferentes modelos preditivos aplicados ao Bitcoin, que hoje detém a maior capitalização no mercado das criptomoedas[1]. Serão utilizados algoritmos de regressão para analisar dados históricos de preços e outras variáveis relevantes, com o objetivo de identificar padrões e tendências que possam indicar movimentos futuros de preços. Além disso, serão incluídos dados de fontes externas, como notícias e eventos macroeconômicos, com o objetivo de melhorar sua precisão e eficácia.

[1] fonte: `https://coinmarketcap.com/` - vista em 30/07/2023

1.2 Justificativa e Motivação

O interesse na negociação de criptomoedas tem crescido significativamente nos últimos anos e o volume de negociação tem sido visto como um sinal de maturidade do mercado, impulsionado pela crescente aceitação das criptomoedas pelos investidores institucionais (TAPSCOTT; TAPSCOTT, 2018), além da criação das moedas digitais que vem sendo adotadas pelos bancos centrais ao redor do globo[2]. O Bitcoin é a criptomoeda mais conhecida e negociada, representando a parte mais significativa do mercado total de criptomoedas[3].

Neste contexto, este trabalho se propõe a investigar se é possível criar um modelo preditivo para o Bitcoin utilizando-se de modelos de regressão e a combinação de dados de mercado e análise de sentimento.

1.3 Questão de Pesquisa e Objetivos

Nesta pesquisa, será investigada a aplicação de diversas técnicas de inteligência artificial à negociação de criptomoedas. A principal questão de pesquisa é:

Q1 *É possível prever de maneira assertiva os retornos ou a volatilidade do Bitcoin utilizando técnicas de inteligência artificial?*

Diante desta questão de pesquisa, é definido o seguinte objetivo para o desenvolvimento deste trabalho:

- Comparar a performance de diferentes modelos de regressão para os retornos e volatilidade do Bitcoin, a partir de dados de mercado e índices de sentimento baseados em dados de redes sociais e notícias.

Os resultados deste trabalho podem ser úteis para investidores em criptomoedas, bem como para empresas que desenvolvem soluções de negociação eletrônica baseadas em criptomoedas. Além disso, esta abordagem pode ser aplicada a outras áreas de investimento para ajudar os investidores a tomar decisões mais bem fundamentadas.

1.4 Organização do texto

No capítulo seguinte, será realizada uma revisão básica sobre os principais conceitos e ferramentas utilizadas neste trabalho, como o Bitcoin, os algoritmos de aprendizado supervisionado, a análise de sentimentos e ferramentas para comparar o desempenho das diversas estratégias.

[2] fonte: `https://www.atlanticcouncil.org/cbdctracker/` - vista em 30/07/2023
[3] fonte: `https://coinmarketcap.com/` - vista em 30/07/2023

No capítulo 3 serão apresentados os principais trabalhos relacionados. Nos capítulos 4 e 5 reside a principal parte do conteúdo apresentado neste trabalho. Neles, será apresentada a proposta da pesquisa e será descrita toda a avaliação experimental, com apresentação dos dados e parâmetros utilizados nos modelos, bem como a apresentação dos resultados obtidos.

No último capítulo será apresentada uma breve conclusão desta pesquisa, além das oportunidades de evolução mapeadas ao longo de seu desenvolvimento.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Bitcoin

Bitcoin é uma criptomoeda descentralizada que foi criada em 2008 por um indivíduo ou grupo de indivíduos sob o pseudônimo de Satoshi Nakamoto (NAKAMOTO, 2008). Desde então, o Bitcoin tem ganhado cada vez mais atenção de investidores e *traders* em todo o mundo devido ao seu potencial de lucro e à sua natureza descentralizada.

Diferentemente das moedas tradicionais emitidas por governos, o Bitcoin não é controlado por nenhuma autoridade central, como um banco central. Em vez disso, ele funciona em uma rede de computadores *peer-to-peer*, onde os usuários podem realizar transações diretamente entre si, sem intermediários.

Uma das principais vantagens do Bitcoin é a segurança que ele oferece. Todas as transações são registradas em um livro contábil digital, chamado de *blockchain*, que é distribuído em vários computadores ao redor do mundo. Isso significa que é muito difícil para qualquer pessoa ou entidade adulterar as transações, tornando o Bitcoin resistente a fraudes e hackeamentos. Além disso, o Bitcoin também oferece privacidade aos seus usuários, pois as transações são realizadas de forma pseudônima, ou seja, sem a necessidade de identificação pessoal. Isso permite que as pessoas mantenham sua privacidade financeira e protejam suas informações pessoais.

No atual cenário macroeconômico, o Bitcoin também é visto como um ativo de refúgio seguro contra a inflação e a desvalorização da moeda. Como o Bitcoin não é controlado por nenhum governo ou autoridade central, ele não está sujeito às políticas monetárias que podem levar à inflação ou à desvalorização da moeda fiduciária. Vale ressaltar que o Bitcoin também é considerado por muitos como uma reserva de valor, semelhante ao ouro, devido à sua oferta limitada e à crescente demanda (NARAYANAN *et al.*, 2016).

Embora o Bitcoin ainda seja considerado uma moeda volátil e de alto risco, sua popularidade e adoção continuam altas em todo o mundo, especialmente entre investidores e entusiastas de tecnologia. O principal motivo da utilização do Bitcoin neste trabalho é a sua representatividade no mercado das criptomoedas. Considerando o período entre 16/08/2022 e 14/08/2023, o Bitcoin teve um volume médio diário de negociação de mais 24 bilhões de dólares e um valor de mercado médio de mais de 462 bilhões de dólares[1], sendo a moeda mais negociada e com maior valor de mercado no período.

[1] fonte: `https://coinmarketcap.com/` - vista em 30/07/2023

2.2 Algoritmos de Aprendizado Supervisionado

Os algoritmos de aprendizado de máquina supervisionados são uma classe de algoritmos que usam um conjunto de dados de treinamento rotulado para aprender a prever saídas corretas a partir de novos dados não rotulados. Em outras palavras, um modelo supervisionado é alimentado com um conjunto de dados de treinamento, onde cada exemplo inclui um conjunto de características (também conhecidas como variáveis independentes) e um rótulo (também conhecido como variável dependente), e o objetivo é aprender uma função que possa prever o rótulo correto para novos exemplos com base em suas características. Alguns exemplos comuns de algoritmos de aprendizado de máquina supervisionados incluem árvores de decisão, redes neurais, regressão linear, regressão logística, entre outras. Os algoritmos supervisionados são amplamente utilizados em uma grande variedade de aplicações, incluindo classificação, regressão, detecção de anomalias, processamento de fala e imagem, entre outros (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

2.2.1 Regressão Linear Múltipla

A regressão linear é uma técnica estatística utilizada para identificar e modelar o relacionamento entre duas variáveis contínuas, sendo uma delas a variável independente e a outra a variável dependente (MONTGOMERY; PECK; VINING, 2012). No mundo das finanças, a regressão linear tem diversas aplicações, como na análise de risco e retorno de investimentos, previsão de preços de ações e análise de correlação entre variáveis financeiras, como taxa de juros, inflação e desempenho do mercado financeiro. Através da análise de regressão, é possível obter insights valiosos para tomada de decisões financeiras, ajudando a identificar tendências e padrões em dados históricos e projetar cenários futuros com maior confiança.

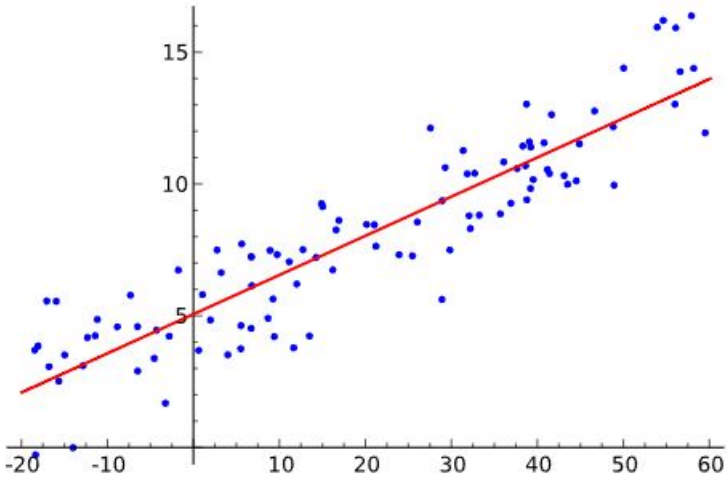

Figura 1 – Regressão linear simples

A regressão linear múltipla é uma técnica estatística utilizada para analisar a relação entre várias variáveis independentes e uma variável dependente contínua. É uma extensão da regressão linear simples, que envolve apenas uma variável independente.

Na regressão linear múltipla, a relação entre as variáveis é modelada por meio de uma equação linear que pode ser expressa como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

Onde y é a variável dependente, $x_1, x_2, ..., x_n$ são as variáveis independentes, $\beta_0, \beta_1, \beta_2, ..., \beta_n$ são os coeficientes de regressão que medem o impacto de cada variável independente em y, e ϵ é o erro aleatório que representa as influências não-modeladas e não medidas nas variáveis independentes (GUJARATI; PORTER, 2009).

A regressão linear múltipla é amplamente utilizada em várias áreas, como economia, finanças, ciência social e ciência médica, para entender como as várias variáveis independentes afetam a variável dependente e para fazer previsões. A análise de regressão também pode ser usada para identificar a importância relativa das diferentes variáveis independentes e para testar a significância estatística de seus efeitos sobre a variável dependente.

2.2.2 K-Nearest Neighbors

O algoritmo de aprendizagem supervisionada K-Nearest Neighbors é uma técnica simples, intuitiva e bastante difundida tanto para problemas de classificação quanto regressão em análise de dados. Ele é baseado no princípio de que pontos de dados semelhantes tendem a ter rótulos ou valores semelhantes. Os passos incluem escolher um valor para K (número de vizinhos), calcular a distância entre pontos, selecionar os K vizinhos mais próximos, votar na classe mais comum ou calcular a média para prever o valor. O desempenho é avaliado com métricas apropriadas, assim como nas outras classes de algoritmos, como por exemplo acurácia, precisão, recall para classificação ou erro médio para regressão, que é foco deste estudo. Adiante, abordaremos mais detalhes sobre as medidas de desempenho adotadas neste trabalho. Embora simples, o KNN, de acordo com os parâmetros utilizados, pode ser exigente em termos computacionais (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

2.2.3 Decision Tree

Decision Tree, ou Árvore de Decisão, é um algoritmo usualmente utilizado para problemas de classificação. Este método constrói uma estrutura de árvore onde cada nó interno representa uma decisão baseada em uma característica específica dos dados. Cada ramo (aresta) representa uma ramificação da decisão com base no valor dessa característica.

Os nós folha da árvore contêm as previsões finais ou valores alvo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

No caso do Decision Tree Regressor, a construção da árvore é adaptada para realizar regressão, e neste caso modela relações não-lineares entre variáveis de entrada e variáveis de saída numérica, criando uma estrutura em forma de árvore que toma decisões com base nas características dos dados para fazer previsões de valores contínuos.

A profundidade da árvore e outros hiperparâmetros influenciam o seu tamanho e a sua complexidade. Árvores mais profundas podem se ajustar mais aos dados de treinamento, mas podem resultar em *overfitting*, ou seja, uma modelagem excessivamente adaptada aos dados de treinamento, o que prejudica o desempenho em novos dados não vistos.

Na figura 2 é possível observar um exemplo de uma árvore de decisão binária, onde cada nó dá origem a dois ramos.

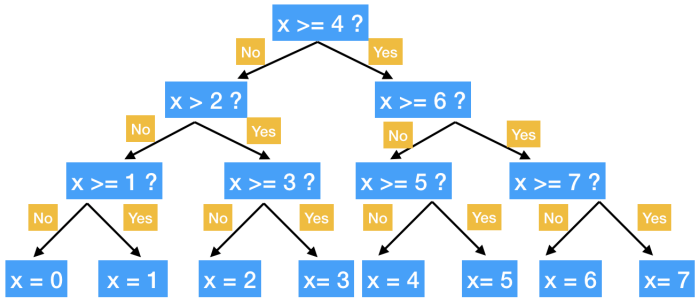

Figura 2 – Representação de árvore de decisão binária

2.2.4 Random Forest

O Random Forest é um algoritmo de aprendizado de máquina que pertence à categoria de métodos de conjunto (*emsemble methods*), onde são combinados vários modelos de aprendizado para produzir um modelo final mais robusto e preciso. A ideia principal por trás do Random Forest é criar múltiplas árvores de decisão durante o treinamento e combiná-las para fazer previsões mais confiáveis e precisas. Cada árvore de decisão é treinada em uma amostra aleatória dos dados de treinamento e utiliza um processo chamado *bagging* para criar a diversidade entre as árvores. O algoritmo reduz a variância e o overfitting, já que a média das previsões das árvores minimiza erros individuais. Além disso, ele fornece uma medida de importância das características, permitindo identificar quais influenciam mais nas previsões. (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

2.2.5 Histogram-Based Gradient Boosting

Gradient Boosting é um algoritmo que cria um modelo preditivo forte combinando múltiplos estimadores fracos, geralmente árvores de decisão rasas. Ele funciona construindo

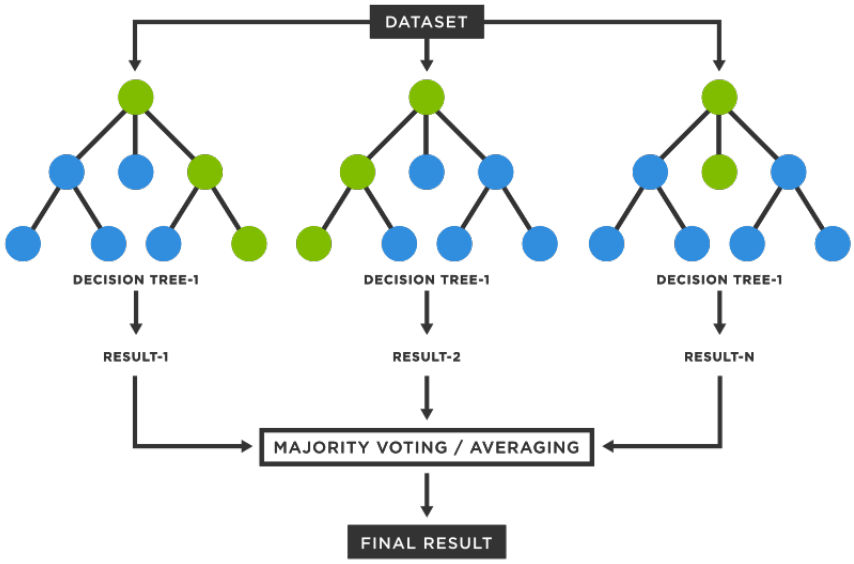

Figura 3 – Representação simplificada do algoritmo de Random Forest

cada nova árvore para corrigir os erros cometidos pelas previsões anteriores, tornando o modelo final mais preciso. No Gradient Boosting clássico, os resíduos (diferença entre as saídas reais e as previsões) são ajustados em cada iteração, o que pode ser computacionalmente intensivo em conjuntos de dados grandes.

Já o Histogram Gradient Boosting é uma otimização do Gradient Boosting que busca melhorar a eficiência. Ele utiliza histogramas para representar os resíduos, em vez de trabalhar diretamente com os valores brutos dos resíduos. Isso acelera o treinamento das árvores, pois os histogramas reduzem a complexidade computacional ao lidar com resíduos discretos. Essa abordagem melhora significativamente o desempenho em conjuntos de dados volumosos, permitindo construir modelos mais precisos em menos tempo (CHEN *et al.*, 2015).

2.3 Análise de sentimento

A análise de sentimento é uma técnica da área de processamento de linguagem natural que visa identificar e extrair informações subjetivas dos textos, como opiniões, emoções e atitudes expressas pelos autores. Seu objetivo principal é compreender a polaridade do sentimento, ou seja, determinar se o texto expressa uma visão positiva, negativa ou neutra em relação a um determinado tópico (PANG; LEE, 2008)

Essa técnica utiliza algoritmos de aprendizado de máquina e processamento de linguagem natural para analisar e classificar textos automaticamente. A análise de sentimento é amplamente aplicada em diversas áreas, incluindo marketing, ciências sociais, atendimento ao cliente, política, saúde e muitas outras. Algumas das principais aplicações

da análise de sentimento atualmente são:

Monitoramento de mídias sociais: A análise de sentimento é utilizada para monitorar a opinião pública em relação a marcas, produtos, eventos ou qualquer assunto discutido nas redes sociais. Empresas podem utilizar essa técnica para acompanhar a reputação da marca, identificar problemas de satisfação do cliente e realizar ações de marketing direcionadas.

Como exemplos da utilização da análise de sentimento, podemos citar a análise de feedback de clientes, análise de pesquisas e enquetes, análise de tendências de mercado e detecção de emoções em conteúdo de mídia, entre outros.

2.4 Métodos de seleção de variáveis

Nesta seção serão listados os diferentes métodos de seleção de variáveis utilizados neste trabalho. Métodos de seleção de variáveis são importantes nos algoritmos de aprendizado supervisionado porque permitem identificar as variáveis que são mais relevantes para a construção do modelo. Isso pode levar a uma melhora no desempenho do modelo, pois ele será treinado com um conjunto de dados mais focado e relevante. Além disso, métodos de seleção de variáveis podem ajudar a reduzir o tempo de treinamento do modelo e a melhorar a sua interpretabilidade (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

2.4.1 Select K-Best

O método Select K-Best é usado para selecionar as melhores K features (variáveis) de um conjunto de dados com base em algum critério de pontuação. Geralmente, ele utiliza testes estatísticos como ANOVA, Qui-quadrado ou funções de pontuação mutual information para avaliar a relevância de cada feature em relação à variável alvo. O usuário precisa especificar o valor de K, ou seja, quantas features deseja selecionar.

2.4.2 Variance Inflation Factor (VIF):

VIF é uma métrica utilizada para identificar a multicolinearidade entre as variáveis independentes em um modelo de regressão. Ele calcula quanto a variância de um coeficiente de regressão é aumentada devido à multicolinearidade. Valores altos de VIF indicam alta correlação entre as variáveis independentes, o que pode levar a estimativas imprecisas e instáveis do modelo.

2.4.3 Recursive Feature Elimination (RFE)

O RFE é um método de seleção de features que opera de maneira recursiva. Ele começa treinando um modelo usando todas as features e, em seguida, elimina a feature menos importante (ou menos relevante) com base em alguma métrica, como coeficiente

de regressão, importância em árvores de decisão, etc. Esse processo é repetido até que o número desejado de features seja atingido.

2.4.4 Recursive Feature Elimination with Cross-Validation (RFECV)

O RFECV é uma extensão do RFE que incorpora validação cruzada. Ele realiza a eliminação recursiva das features enquanto avalia o desempenho do modelo usando validação cruzada. Isso ajuda a evitar a seleção de features que podem ser úteis apenas para um conjunto específico de dados de treinamento. O RFECV ajuda a encontrar um subconjunto de features que oferece um bom desempenho médio em várias divisões dos dados.

2.5 Métricas de avaliação

Nesta seção apresentaremos as métricas selecionadas para avaliação dos modelos propostos neste trabalho. Nas notações abaixo, y e $\hat{y}$ representam os valores reais e previstos, respectivamente, $\bar{y}$ representa a média dos valores reais e n representa o número de amostras (PEDREGOSA *et al.*, 2011).

2.5.1 Coeficiente de Determinação (R^2)

O R^2, também conhecido como coeficiente de determinação, é uma métrica que mede a proporção da variabilidade na variável dependente que é explicada pela variável independente em um modelo estatístico, e será a principal métrica de avaliação a ser considerada neste estudo. Ele varia de 0 a 1, onde 1 indica um ajuste perfeito do modelo aos dados e 0 indica que o modelo não explica nenhuma variação nos dados. No contexto de regressão, o R^2 é usado para avaliar o quão bem o modelo se ajusta aos dados observados. O R^2 é calculado da seguinte forma:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

2.5.2 Erro Médio Absoluto (MAE)

O MAE é a média das diferenças absolutas entre as previsões de um modelo e os valores reais. Ele mede a magnitude média dos erros entre as previsões e os valores reais, independentemente de eles serem positivos ou negativos. Quanto menor o valor do MAE, melhor é o desempenho do modelo em fazer previsões precisas. O MAE é calculado da seguinte forma:

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}|y_i - \hat{y}_i|$$

2.5.3 Erro Percentual Absoluto Médio (MAPE)

O MAPE é uma métrica que calcula a média das porcentagens absolutas das diferenças entre as previsões e os valores reais. Ele é expresso como uma porcentagem e fornece uma ideia de quão precisas são as previsões em termos de magnitude percentual. No entanto, ele pode ser sensível a valores pequenos e pode não ser adequado quando há valores nulos.

$$MAPE(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} \frac{|y_i - \hat{y}_i|}{max(\epsilon, |y_i|)}$$

2.5.4 MSE (Erro Quadrático Médio)

O MSE é a média dos quadrados das diferenças entre as previsões e os valores reais. Ele penaliza erros maiores de forma mais significativa do que erros menores, tornando-se sensível a outliers. O MSE é frequentemente usado em problemas de regressão para avaliar a qualidade das previsões.

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2$$

2.5.5 Raiz do Erro Quadrático Médio (RMSE)

O RMSE é a raiz quadrada do MSE. Ele tem a mesma unidade que a variável de destino e é uma métrica de avaliação comumente usada para medir a dispersão dos erros em relação aos valores reais. O RMSE fornece uma ideia da magnitude média dos erros, sendo especialmente útil quando desejamos que os erros sejam penalizados de forma mais equilibrada.

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2}$$

2.6 Conclusão

Neste capítulo foram apresentadas as definições e características dos principais conceitos e ferramentas envolvidas neste trabalho. Na etapa seguinte, serão abordados os trabalhos relacionados e, posteriormente, a proposta e avaliação experimental serão detalhadas, onde será possível observar a aplicação prática dos itens apresentados nesta seção.

3 TRABALHOS RELACIONADOS

Nesta seção, serão apontados alguns trabalhos relacionados ao tema que serviram como base referencial para esta pesquisa. Há diversos trabalhos que comparam diferentes abordagens no espectro da inteligência artificial para a predição de preços, volatilidade, tendências e análises relacionadas. No campo da análise de sentimento também há vasta bibliografia disponível, utilizando diferentes abordagens desta técnica para propor modelos preditivos de tendência de preço e volatilidade para ativos financeiros variados. No entanto, não há consenso claro entre os trabalhos analisados.

Fang *et al.* (2022) foi um bom ponto de partida e apresenta uma excelente revisão bibliográfica sobre sobre a pesquisa de negociação de criptomoedas, cobrindo 146 trabalhos de que abordam diferentes tópicos, como sistemas de negociação de criptomoedas, bolhas e condições extremas, previsão de volatilidade e retorno, construção de portfólio, negociação técnica e outros

Shakri (2021) apresentou um trabalho interessante, onde utilizou 5 diferentes métodos para prever os retornos do Bitcoin: *Alternating Model Tree*, RF, MLR, MLP e *M5 Trees*. Este estudo usou vários preditores para prever retornos do Bitcoin, incluindo incerteza de política econômica, índice de volatilidade do mercado de ações, retornos de S&P, Taxas de câmbio do EUR/USD, preços do petróleo e do ouro, volatilidades e retornos. Na conclusão do trabalho, observou-se que o modelo de RF teve desempenho superior aos demais. Neste trabalho, utilizaremos uma abordagem parecida, porém com a utilização de algumas séries históricas diferentes, bem como a utilização de outros modelos.

Dias, Fernando and Fernando (2022) investiga hipóteses relacionadas ao efeito do sentimento dos investidores na previsão dos retornos e volatilidade do Bitcoin utilizando regressão quantílica, no período de 2017 a 2021. As descobertas demonstram que o interesse e as emoções dos investidores são preditores significativos dos retornos e volatilidade do Bitcoin, enquanto o VIX e o fórum Bitcointalk.org são os preditores mais adequados para representar as emoções e o interesse dos investidores, respectivamente. As descobertas também indicam uma relação não linear entre o sentimento dos investidores e os retornos e volatilidade do Bitcoin, com o poder de previsão variando com base nas condições de mercado.

Virk (2017) faz uma comparação de RF, SVM, GB e LR. Os resultados mostraram que o SVM alcançou a maior precisão (0,62) entre os algoritmos de aprendizado de máquina de classificação binomial e, na categoria de regressão, o GB conseguiu o maior R-quadrado: 0,99. Mallqui and Fernandes (2018) analisou o comportamento de ANN, SVM e algoritmos Ensemble (baseados em Redes Neurais Recorrentes e método de agrupamento k-Means)

para predição de direção de preços. Os resultados mostraram que os atributos selecionados e o melhor modelo de aprendizado de máquina alcançaram uma melhoria de mais de 10%, na precisão, para as previsões de direção de preço, em relação aos artigos utilizados como referência, usando o mesmo período de informação. Além disso, foi possível obter erros percentuais médios absolutos entre 1% e 2%. Em Ji, Kim and Im (2019) foram comparados vários métodos de aprendizado profundo, como uma DNN, uma rede LSTM, uma rede neural convolucional, uma rede residual profunda, e suas combinações para previsão de preço do Bitcoin. Os resultados experimentais mostraram que, embora os modelos de previsão baseados em LSTM tenham superado ligeiramente os outros modelos de previsão para previsão de preço do Bitcoin (regressão), os modelos baseados em DNN tiveram o melhor desempenho para previsão de preços altos e baixos (classificação).

Em Uras *et al.* (2020), alguns métodos foram estudados: SLR para previsão de séries univariadas usando apenas preços de fechamento e o modelo de MLR para séries multivariadas usando preços e dados de volume, MLP e LSTM, tendo sua aplicação nas criptomoedas comparadas aos trabalhos realizados no mercado de ações. Neste trabalho, os algoritmos de LSTM e de regressão tiveram um bom desempenho, e algumas conclusões interessantes foram relatadas, destacando uma grande melhora nos modelos regressivos ao segmentar o histórico de dados em partições que apresentassem uma certa tendência, distanciando-se do chamado "random walk". M. *et al.* (2020) comparou LR e SVM usando uma série temporal composta por preços de fechamento diários da criptomoeda Ether com diferentes comprimentos de janela e filtros com diferentes coeficientes de peso. A conclusão deste trabalho foi que o método SVM conseguiu maior acurácia (96,06%) do que o método LR (85,46%).

D'Amato, Levantesi and Piscopo (2022) baseia-se em técnicas de aprendizado profundo, especificamente, Jordan Neural Network (JNN), Non-Linear Autoregressive Neural Network (NLANN) e Self Exciting Threshold Autoregressive (SETAR) para analisar a volatilidade do Bitcoin.

Felizardo *et al.* (2022) estende a comparação entre a abordagem supervisionada e o aprendizado por reforço. Utilizando a arquitetura ResNet no ator ResNet-LSTM (RSLSTM-A), realizou-se a comparação com técnicas de aprendizado por reforço clássicas e recentes, como aprendizado por reforço recorrente, rede Q profunda e crítico de ator de vantagem considerando as criptomoedas Bitcoin, Litecoin, Ethereum, Monero, Nxt e Dash. O estudo mostra que a abordagem proposta alcança melhor desempenho geral, confirmando a hipótese que o aprendizado supervisionado pode superar o aprendizado por reforço para negociação de ativos.

Giaglis *et al.* (2015) estudou a relação do Bitcoin com variáveis financeiras, fatores tecnológicos e também a influência do sentimento coletivo extraído de dados do Twitter, com a utilização de SVMs, concluindo através de regressões de curto prazo que há correlação

positiva entre o preço do Bitcoin e o sentimento nas redes sociais.

Colianni, Rosales and Signorotti (2015) utilizou algoritmos de aprendizado supervisionado, como regressão logística, Naive Bayes e SVM, para analisar dados obtidos no Twitter e obteve uma precisão superior a 90% para previsão da direção dos preços hora a hora ou diários do Bitcoin.

Valencia, Gómez-Espinosa and Valdes (2019) comparou a utilização de redes neurais, SVM e RF usando elementos do Twitter e dados de mercado como recursos de entrada. Os resultados mostram que é possível prever mercados de criptomoedas usando aprendizado de máquina e análise de sentimento, onde os dados do Twitter por si só podem ser usados para prever certas criptomoedas e que NN superam os outros modelos. Garcia and Schweitzer (2015) apresenta um modelo combinando o volume de operações, valência emocional e polarização de opinião, conforme expresso em tweets relacionados ao Bitcoin por mais de 3 anos. Foi observado que os aumentos na polarização de opinião e no volume transacional precedem o aumento dos preços do Bitcoin e que a valência emocional precede a polarização de opinião e o aumento dos volumes transacionais. Este modelo foi aplicado a uma estratégia de negociação que apresentou ótimos resultados em menos de um ano.

Oikonomopoulos *et al.* (2022) usou a análise de sentimentos do Twitter para prever as flutuações de preço das criptomoedas usando a biblioteca VADER. A previsibilidade dos retornos de preços é examinada com Vector Autoregression (VAR) e previsões altamente precisas para duas das sete criptomoedas foram alcançadas. Mais especificamente, as previsões de preços de Ethereum e Polkadot atingiram 99,67% e 99,17% de precisão, respectivamente. Em Abraham *et al.* (2018), foram analisados o sentimento e o volume de buscas (Google Trends) e de tweets em redes sociais (Twitter), e concluiu-se que o volume de buscas e tweets foi a variável que melhor conseguiu prever a direção do preço do Bitcoin e do Ethereum.

Pichl and Kaizoji (2017) estudou a volatilidade do Bitcoin ao longo dos últimos cinco anos, realizando uma análise em várias escalas, desde os dados instantâneos até as escalas de 5 minutos, 1 hora e 1 dia, e utiliza um modelo autorregressivo para a volatilidade realizada.

Na Tabela 1, é possível observar um comparativo dos trabalhos analisados.

Estudo	Variável Analisada	Modelos Preditivos	Utiliza Índices de Sentimento?
Colianni, Rosales and Signorotti (2015)	Direção	LGR, Naive Bayes e SVM	Sim
Garcia and Schweitzer (2015)	Direção	VAR	Sim
Giaglis *et al.* (2015)	Preço	SMV	Sim
Lamon, Nielsen and Redondo (2017)	Direção	LGR, Naive Bayes e SVM	Sim
Pichl and Kaizoji (2017)	Volatilidade	HAR	Não
Virk (2017)	Preço	RF, SVM, GB e LR	Não
Mallqui and Fernandes (2018)	Direção	ANN, SVM, RNN e K-Means	Não
Ji, Kim and Im (2019)	Preço	DNN, LSTM, CNN, ResNet	Não
Abraham *et al.* (2018)	Direção	MLR	Sim
Valencia, Gómez-Espinosa and Valdes (2019)	Preço	NN, SVM e RF	Sim
Uras *et al.* (2020)	Preço	SLR, MLR, MLP e LSTM	Não
M. *et al.* (2020)	Preço	LR e SVM	Não
Shakri (2021)	Retorno	AMT, RF, MLR, MLP e M5 Trees	Sim
Dias, Fernando and Fernando (2022)	Volatilidade	QR	Sim
D'Amato, Levantesi and Piscopo (2022)	Volatilidade	JNN, SETAR, NAR NN	Não
Felizardo *et al.* (2022)	Preço	ResNet, RL e DQN	Não
Oikonomopoulos *et al.* (2022)	Retorno	VAR	Sim

Tabela 1 – Estudos sobre modelos preditivos em criptomoedas

4 METODOLOGIA

Neste capítulo será detalhada a metodologia deste trabalho, que tem como objetivo comparar a previsão dos retornos e volatilidade do Bitcoin através de utilização de modelos de aprendizado supervisionado, a partir de um conjunto de dados históricos de ativos financeiros e índices de sentimento selecionados.

4.1 Ativos financeiros

Na Tabela 2 é possível observar o símbolo utilizado para a representação do ativo e descrição resumida de cada componente presente no modelo.

Símbolo	**Descrição**
BTC	Bitcoin
SPX	Índice S&P500
NDQ	Índice Nasdaq
EUR	Taxa de câmbio EUR/USD
GLD	Ouro
TYX	Título Treasury 10-Year
OIL	Crude OIL

Tabela 2 – Símbolos e descrições dos ativos financeiros utilizados.

A seguir, serão abordados alguns detalhes a respeito de cada um dos ativos mencionados:

1. Bitcoin: representa a variável de interesse neste trabalho. Além das variáveis que este trabalho se propõe a estudar e que serão abordadas na sequência, o preço do Bitcoin também é determinado pela oferta e demanda no mercado, e pode ser influenciado por vários fatores, como adoção global, regulamentação governamental, eventos geopolíticos e sentimento do mercado.

2. Índice S&P 500: O S&P 500 é um índice ponderado de mercado que acompanha o desempenho de 500 grandes empresas negociadas nas bolsas de valores dos Estados Unidos. Ele é frequentemente usado como um indicador amplo do desempenho do mercado de ações nos EUA. O preço do S&P 500 é influenciado pelo desempenho das empresas incluídas no índice, bem como por fatores econômicos, políticos e eventos globais que afetam o mercado de ações. S&P faz referência a Standard & Poor's, uma empresa de consultoria financeira.

3. Nasdaq: Nasdaq é uma bolsa de valores eletrônica nos Estados Unidos, onde são negociadas várias empresas de tecnologia e do setor de crescimento. Além disso, o

termo "Nasdaq"também é frequentemente usado para se referir ao índice Nasdaq Composite, que acompanha o desempenho de mais de 3.000 ações listadas no Nasdaq. O índice Nasdaq é especialmente conhecido por incluir empresas de tecnologia, e no passado recente é possível observar uma forte correlação entre os movimentos nos preços do índice e do Bitcoin.

4. Taxa de câmbio EUR/USD: A taxa de câmbio EUR/USD representa a taxa de câmbio entre o euro (EUR) e o dólar americano (USD). Essa taxa indica quantos dólares americanos são necessários para comprar um euro. Ela é influenciada por diversos fatores, incluindo políticas monetárias dos bancos centrais, indicadores econômicos, eventos políticos e o sentimento do mercado em relação ao euro e ao dólar americano.

5. Ouro: O ouro é um metal precioso amplamente reconhecido como reserva de valor. Historicamente, o ouro tem sido utilizado como uma proteção contra a inflação e a volatilidade dos mercados financeiros. O preço do ouro é influenciado por fatores como oferta e demanda, taxas de juros, valor do dólar americano, instabilidade geopolítica e incerteza econômica.

6. Títulos do governo americano (10 anos): Os títulos do governo americano, como os títulos do Tesouro dos Estados Unidos, são considerados investimentos seguros e de baixo risco. Eles representam empréstimos feitos pelos investidores ao governo dos Estados Unidos em troca de juros. Os preços dos títulos do governo são afetados por fatores como a taxa de juros, inflação, oferta e demanda por esses títulos, bem como pela confiança dos investidores na solvência do governo. Esses títulos são conhecidos por sua liquidez e são amplamente utilizados como uma referência para as taxas de juros de longo prazo nos mercados financeiros.

7. Crude Oil: refere-se ao petróleo bruto, que é uma das commodities mais negociadas no mercado financeiro. Existem vários mercados onde o petróleo bruto é negociado, como a New York Mercantile Exchange (NYMEX) nos Estados Unidos e a Intercontinental Exchange (ICE) em Londres. Os contratos futuros de petróleo bruto são uma forma comum de investimento nesse ativo financeiro. O preço do petróleo bruto é influenciado por uma série de fatores, como oferta e demanda globais, eventos geopolíticos, políticas governamentais e condições econômicas. Alterações na oferta, como descobertas de novas reservas ou interrupções na produção, e mudanças na demanda, impulsionadas pelo crescimento econômico global ou pela adoção de fontes de energia alternativas, podem afetar significativamente o preço do petróleo bruto.

4.2 Análise de Sentimento

Além do histórico de preços, retornos e volatilidades históricas, mencionados na seção anterior, foram incluídos também alguns indicadores de sentimento, conforme indicado na Tabela 3:

Símbolo	Descrição
VIX	CBOE Volatility Index
FEAR	Bitcoin Fear & Greed Index
NEWS	FRBSF Daily News Sentiment Index
TWITTER	Twitter-based Economic Uncertainty Index
WIKI	Número de visitas à pagina da Wikipedia

Tabela 3 – Símbolos e descrições dos índices de sentimento utilizados.

1. VIX: esta é a abreviação do CBOE Volatility Index da Chicago Board Options Exchange, e representa uma medida financeira que reflete a expectativa de volatilidade dos preços no mercado de ações. Também é conhecido como o "índice de medo", pois tende a aumentar quando os investidores estão preocupados com a possibilidade de quedas acentuadas nos preços das ações. O VIX é calculado com base nas opções de compra e venda de curto prazo de um índice amplo, como o S&P 500. Quando os mercados estão mais voláteis, os preços das opções aumentam, levando a um aumento no valor do VIX. Em resumo, o VIX serve como um indicador da confiança e do sentimento dos investidores em relação à estabilidade futura do mercado de ações (WHALEY, 2009).

2. Bitcoin Fear & Greed Index: indicador desenvolvido para medir o sentimento predominante no mercado de Bitcoin com base em uma variedade de fatores e métricas. Ele oferece uma representação simplificada do estado emocional dos participantes do mercado, indicando se eles estão mais inclinados ao medo ou à ganância em relação ao preço e às perspectivas do Bitcoin (BITCOIN..., 2018). O índice é calculado usando uma combinação de informações, incluindo:

    - Volatilidade de Preço: variação de preço do Bitcoin. Uma volatilidade maior pode indicar sentimentos extremos.
    - Volume de Negociação: quantidade de Bitcoin sendo comprada e vendida. Um volume elevado pode indicar movimentos significativos no mercado.
    - Atividade em Mídias Sociais: sentimento expresso nas redes sociais e fóruns relacionados ao Bitcoin.
    - Pesquisas de Mercado: pesquisas e enquetes que avaliam a opinião dos investidores sobre o mercado.

- Dominância de Mercado: participação do Bitcoin em relação a outras criptomoedas no mercado.
- Índices de Preço Atuais e Históricos: como o preço atual do Bitcoin se compara a seu histórico recente.

Com base nesses dados, o índice é calculado e apresentado como um número entre 0 e 100. Pontuações mais baixas, próximas a 0, tendem a indicar um mercado dominado pelo medo, o que pode sugerir oportunidades de compra para investidores que buscam preços mais baixos. Pontuações mais altas, próximas a 100, indicam um mercado dominado pela ganância, o que pode sinalizar a possibilidade de uma correção de preços em breve.

3. FRBSF Daily News Sentiment Index: este é um índice desenvolvido pelo Federal Reserve Bank of San Francisco para acompanhar o sentimento nas notícias diárias com o objetivo de avaliar o impacto das notícias sobre a economia (BUCKMAN *et al.*, 2020).

   O índice é construído através da análise de artigos de notícias e outros textos para determinar o tom emocional presente neles, ou seja, se as notícias têm um sentimento positivo, negativo ou neutro. Essa análise é feita usando técnicas de processamento de linguagem natural e aprendizado de máquina.

   O FRBSF Daily News Sentiment Index pode ser usado como uma ferramenta para entender como o sentimento nas notícias pode afetar as decisões econômicas, como investimentos e gastos dos consumidores. Por exemplo, um aumento no sentimento positivo nas notícias pode indicar otimismo em relação à economia, o que poderia influenciar as decisões de investimento e gastos das pessoas e empresas.

4. Twitter-based Economic Uncertainty Index: métrica desenvolvida com base em dados do Twitter para avaliar a incerteza econômica. Ele faz parte de um esforço mais amplo para medir a incerteza econômica usando informações disponíveis online, como notícias, pesquisas de palavras-chave e, no caso desse índice, mensagens do Twitter (BAKER *et al.*, 2021).

   O índice é calculado analisando os tweets que mencionam temas relacionados à economia e à incerteza econômica. Isso pode incluir discussões sobre política econômica, indicadores econômicos, mudanças nas condições de mercado e outros tópicos afins. O índice utiliza algoritmos de processamento de linguagem natural para identificar e classificar os tweets de acordo com a intensidade da incerteza econômica expressa.

   A ideia por trás do Twitter-based Economic Uncertainty Index é que as redes sociais, como o Twitter, podem refletir o sentimento e a percepção do público em relação à economia. As pessoas muitas vezes compartilham suas opiniões e preocupações sobre

questões econômicas em plataformas de mídia social, o que pode fornecer insights sobre os níveis de incerteza.

5. Visualizações da Wikipedia: a Wikipedia é uma enciclopédia online colaborativa e de acesso gratuito, que abrange uma ampla gama de tópicos e conhecimentos. A plataforma utiliza um modelo wiki, onde qualquer pessoa pode editar as páginas existentes ou criar novos artigos, resultando em uma base de dados vasta e diversificada. Com milhões de artigos em diversos idiomas, a Wikipedia se tornou uma das fontes de informação mais acessadas globalmente, embora sua natureza aberta também tenha gerado discussões sobre precisão e confiabilidade em alguns casos. Portanto, as visualizações diárias da página sobre Bitcoin no site são coletadas para quantificar o sentimento dos investidores. Este índice de sentimento visa representar o nível de interesse do público sobre o assunto através do número de acessos ao referido website[1].

4.3 Conclusão

Neste capítulo foram apresentados detalhes sobre os ativos e índices de sentimento escolhidos para compor a análise preditiva, além de uma breve explicação sobre todos os modelos testados no desenvolvimento do trabalho. A seguir, serão expostos os detalhes da avaliação experimental, como a obtenção das séries de retorno e volatilidade, as características das séries de dados e apresentação dos resultados obtidos.

[1] `https://en.wikipedia.org/wiki/Bitcoin`

5 AVALIAÇÃO EXPERIMENTAL

5.1 Conjuntos de Dados

Para o desenvolvimento deste trabalho[1] utilizamos a versão 3.11 do Pyhton. As séries históricas preço dos ativos financeiros foram obtidos com auxílio da biblioteca *yfinance*. Esta biblioteca é uma ferramenta popular para acessar e obter dados financeiros diretamente do site Yahoo Finance. Ela permite aos desenvolvedores e analistas financeiros acessar uma ampla gama de informações, como histórico de preços de ações, dados de mercado, estatísticas financeiras e muito mais. Para este trabalho, estamos utilizando o preço de fechamento ajustado de cada um dos ativos financeiros descritos na seção anterior. Utilizamos dados do histórico de preços de 17/01/2017 até 30/06/2023.

Após a obtenção das séries históricas de preço (figura 4), foram calculadas as série de retorno r (equação 5.1, figura 5) e, a partir dos retornos, a volatilidade histórica anualizada σ (equação 5.2, figura 6) de cada ativo. As variáveis p_t e r_t representam o preço e o retorno do ativo no instante de tempo t. A variável $\hat{r}$ representa a média dos retornos no período, representado por N, cujo valor utilizado para o cálculo das volatilidades foi de 21 dias. Nas figuras 8, 9, 10 são apresentadas, respectivamente, as matrizes de correlação das séries históricas de preços, retornos e volatilidades dos ativos financeiros selecionados.

$$r_t = \ln(\frac{p_t}{p_{t-1}}) \tag{5.1}$$

$$\sigma_t = \frac{1}{N} \sum_{t-N}^{t} (r_t - \hat{r})^2 * \sqrt{252} \tag{5.2}$$

No caso dos índices de sentimento, além da biblioteca *yfinance*, as séries históricas também foram obtidas através de websites especializados em cada indicador[2]. Os históricos dos índices de sentimento estão representados na figura 7 e a matriz de correlação na figura 11.

---

[1] O código-fonte pode ser acessado em `https://github.com/fbcalderaro/MBA_IABGD`

[2] Policy Uncertainty: `https://www.policyuncertainty.com/`
FRBSF Daily News Sentiment: `https://www.frbsf.org/`. em cada abordagem.
Bitcoin Fear & Greed Index: `https://alternative.me/crypto/fear-and-greed-index/`
Visitas à Wikipedia: `https://pageviews.toolforge.org/`

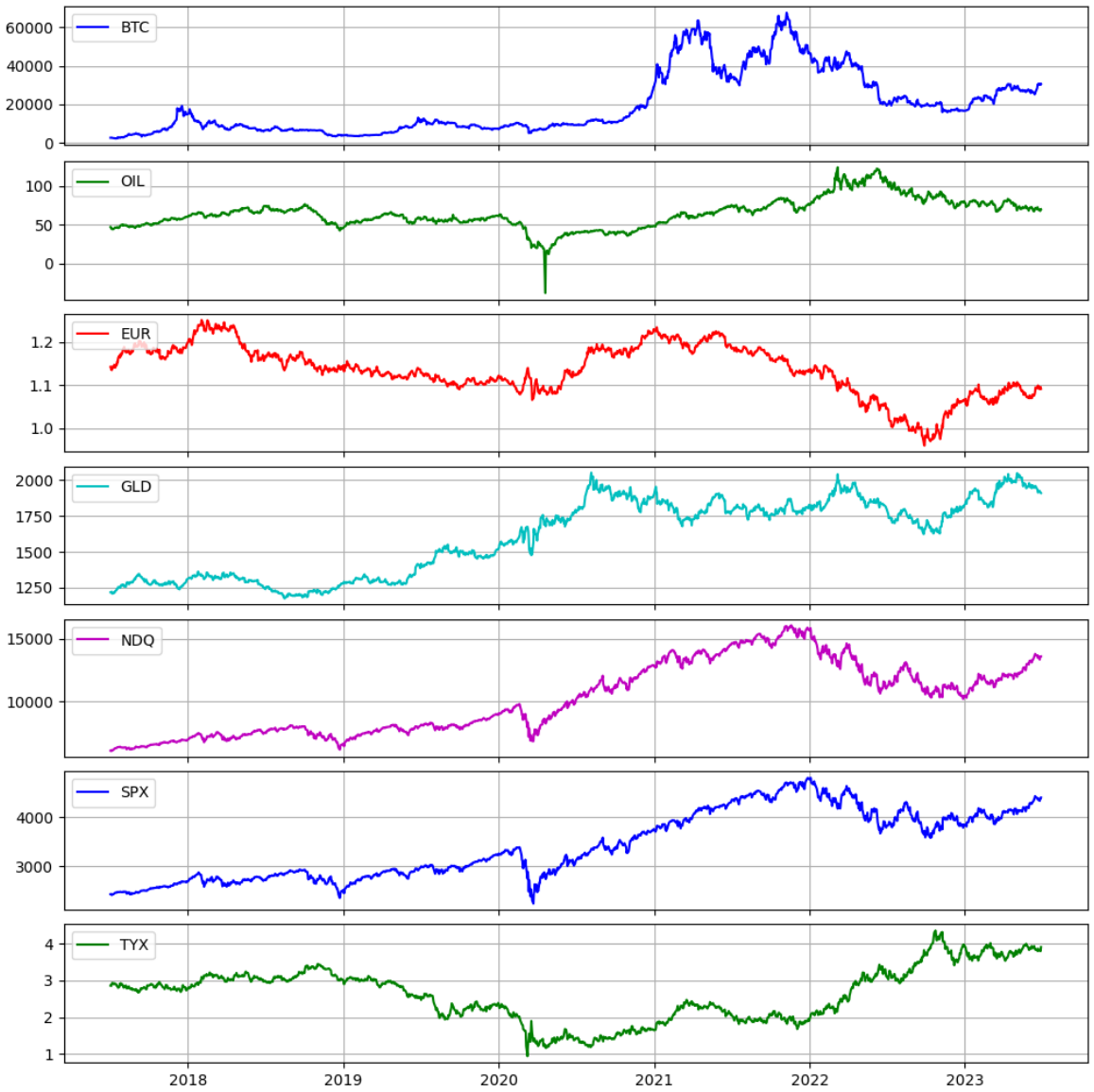

Figura 4 – Séries históricas de preços

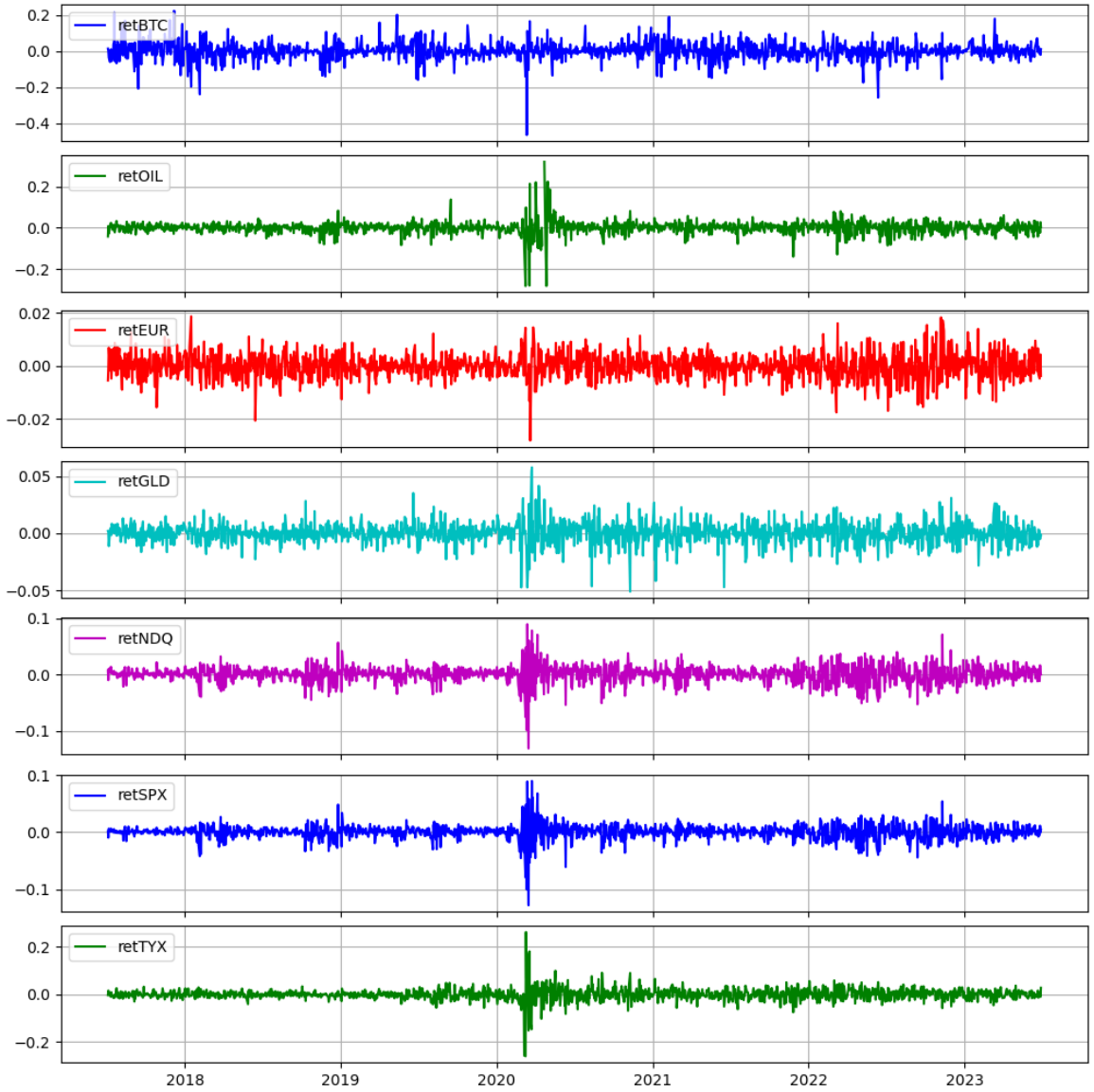

Figura 5 – Séries históricas de retornos

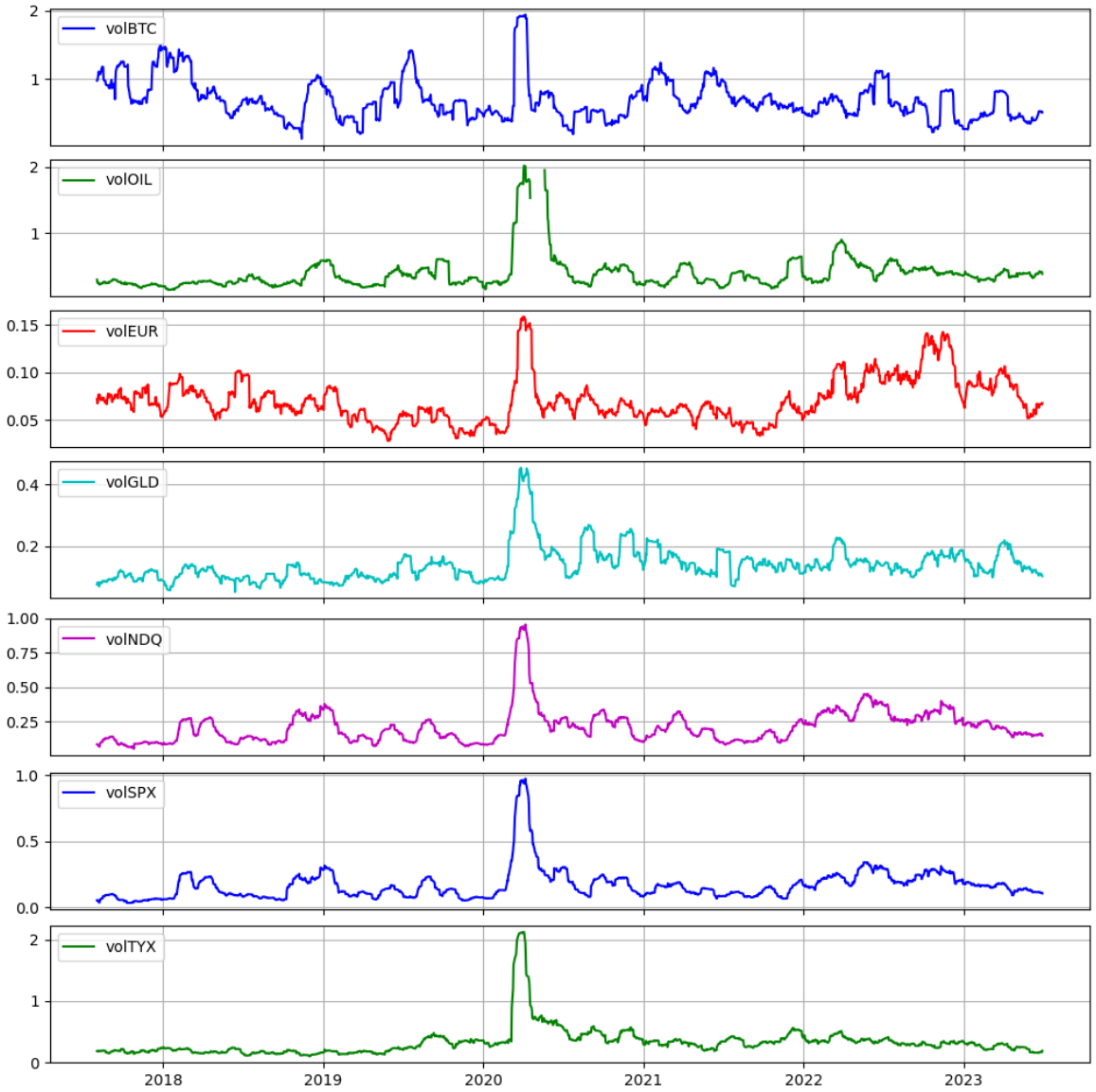

Figura 6 – Séries históricas de volatilidades

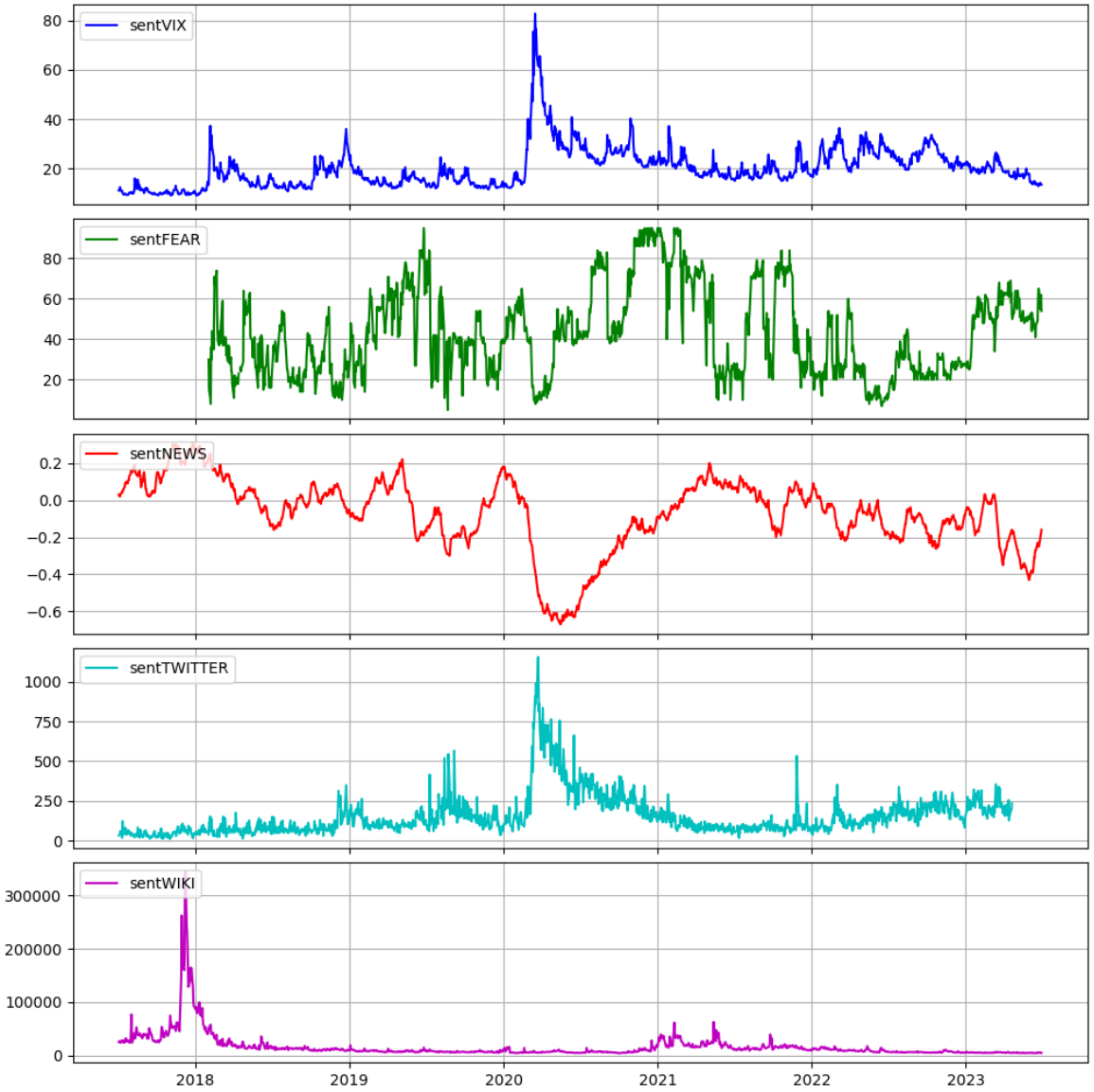

Figura 7 – Séries históricas dos índices de sentimento

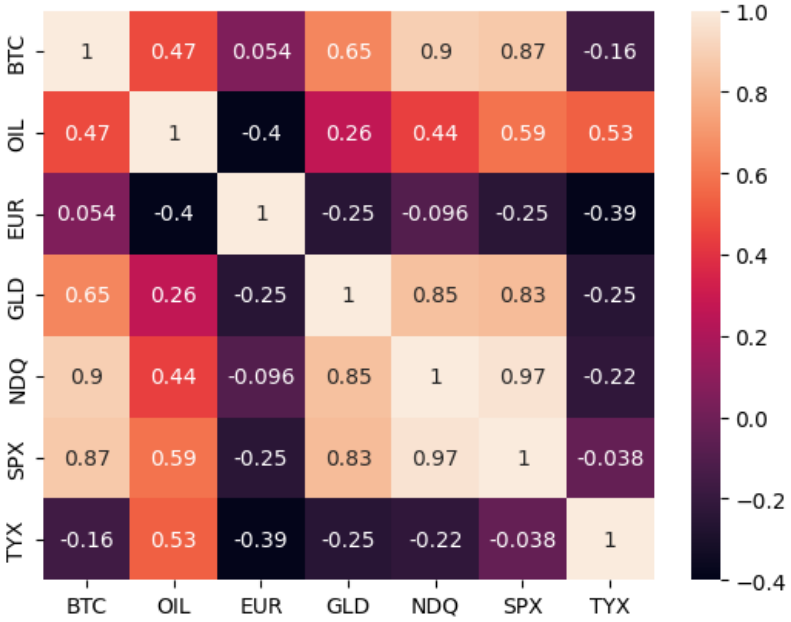

Figura 8 – Matriz de correlação das séries de preço

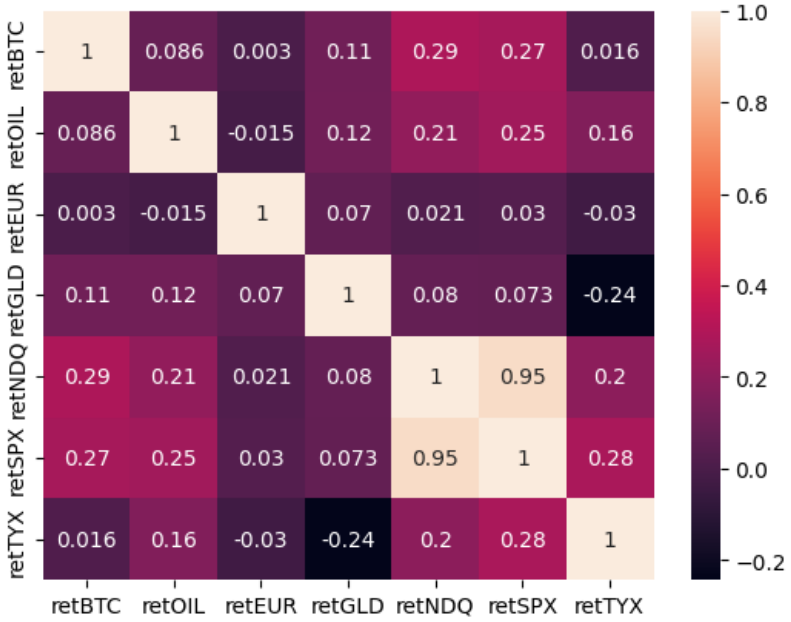

Figura 9 – Matriz de correlação das séries de retorno

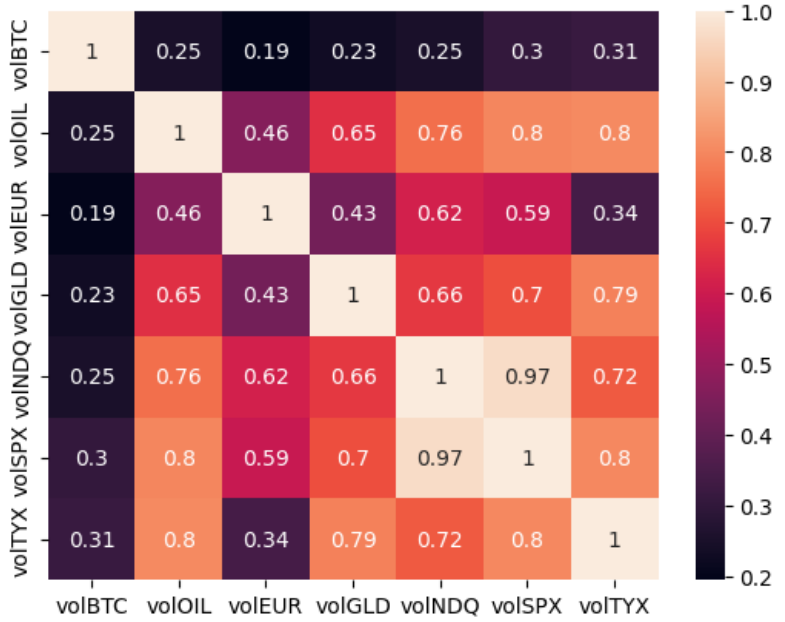

Figura 10 – Matriz de correlação das séries de volatilidade

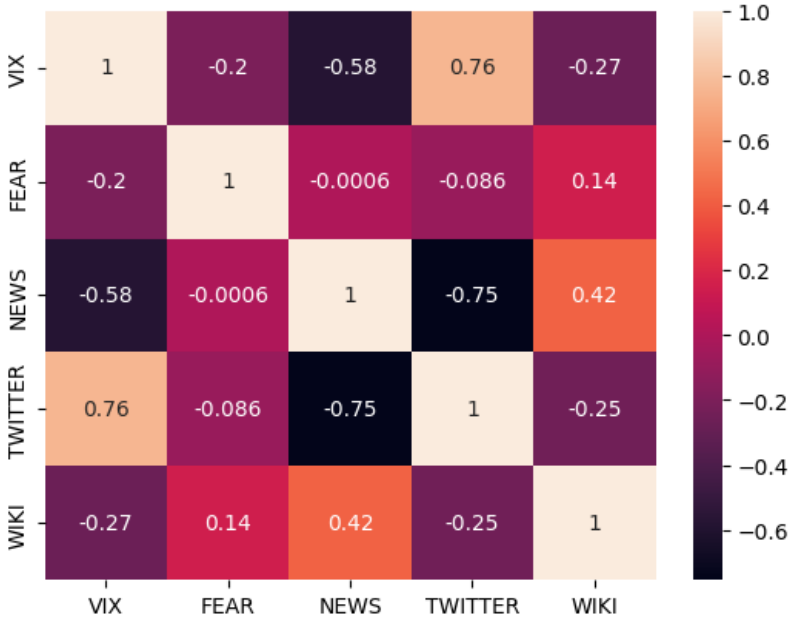

Figura 11 – Matriz de correlação das séries de índices de sentimento

5.2 Configuração Experimental

Em praticamente todas as etapas do experimento foi utilizada a biblioteca *sklearn*. A biblioteca sklearn, também conhecida como scikit-learn, é uma das bibliotecas de aprendizado de máquina mais populares e amplamente utilizadas em Python. Ela fornece uma ampla gama de algoritmos de aprendizado de máquina e ferramentas para pré-processamento de dados, validação de modelos, seleção de modelos e avaliação de desempenho (PEDREGOSA *et al.*, 2011).

5.2.1 Pré-processamento

A etapa de pré-processamento consiste na junção de todas as séries de dados obtidas e limpeza do conjunto, excluindo-se as informações indesejadas para cada modelo e os registros com valores ausentes. Foram criados dois conjuntos de dados de entrada distintos para o nosso modelo, o primeiro contendo apenas as séries históricas dos retornos e das volatilidades, e o segundo sendo o primeiro acrescido das séries históricas dos índices de sentimento, para posterior análise comparativa de desempenho dos modelos.

5.2.2 Seleção de Variáveis

Neste trabalho foram testados quatro diferentes métodos de seleção de variáveis, descritos na Tabela 4:

Símbolo	**Descrição**
SKB	Select K-Best
VIF	Variance Inflation Factor
RFE	Recursive Feature Elimination
RFECV	RFE with Cross-Validation

Tabela 4 – Métodos de seleção de variáveis

Nas Tabelas 5 e 6 é possível observar quais variáveis foram selecionadas por cada método considerando os diferentes conjuntos de entrada (com ou sem os índices de sentimento, representado na coluna "Sent?").

Método	Sent?	Variáveis Selecionadas
SKB	Não	retOIL, retGLD, retSPX
SKB	Sim	FEAR, retGLD, retSPX
RFE	Não	retEUR, retGLD, retNDQ
RFE	Sim	retEUR, retGLD, retNDQ
VIF	Não	retOIL, retEUR, retGLD, retTYX, volOIL, volEUR, volGLD
VIF	Sim	retOIL, retEUR, retGLD, retTYX, volOIL, volEUR, volGLD, FEAR, NEWS, WIKI
RFECV	Não	retNDQ
RFECV	Sim	retEUR, retGLD, retNDQ, retSPX, retTYX, volEUR

Tabela 5 – Seleção de variáveis na previsão dos retornos.

Método	Sent?	Variáveis Selecionadas
SKB	Não	volNDQ, volTYX, volSPX
SKB	Sim	volGLD, volTYX, volNDQ
VIF	Não	retOIL, retEUR, retGLD, retTYX, volOIL, volEUR, volGLD
VIF	Sim	retOIL, retEUR, retGLD, retTYX, volOIL, volEUR, volGLD, FEAR, NEWS, WIKI
RFE	Não	retEUR, volEUR, volSPX
RFE	Sim	retEUR, volEUR, volSPX
RFECV	Não	retEUR, volEUR
RFECV	Sim	retOIL, retEUR, retSPX, volEUR, volGLD, volNDQ, volSPX, NEWS

Tabela 6 – Seleção de variáveis na previsão da volatilidade.

5.2.3 Modelos

Para executar os modelos, separamos os dados em conjuntos de treino e teste para treinar nosso modelo de regressão. Na Tabela 7 estão listados os modelos utilizados neste trabaho. Foram utilizados os 70% iniciais dos dados para o conjunto de treinamento e os 30% finais para o conjunto de testes.

Símbolo	Descrição
LR	Linear Regression
KNN	K-Nearest Neighbors
DT	Decision Tree
RF	Random Forest
HGB	Histogram-Based Gradient Boosting

Tabela 7 – Modelos de regressão utilizados.

5.3 Resultados e Discussões

As métricas escolhidas para a comparação do desempenho dos modelos foram: R^2, Erro Absoluto Médio (MAE), Erro Percentual Médio Absoluto (MAPE), Erro Quadrático Médio (MSE) e Raiz do Erro Quadrático Médio (RMSE). Nas Tabelas 8 e 9 são apresentados os comparativos das métricas de avaliação propostas para cada modelo preditivo, bem como o conjunto de dados de entrada e o método de seleção de variáveis utilizado. Os resultados estão ordenados pelo R^2 de maneira decrescente.

No caso dos retornos, é possível observar que o modelo com melhor desempenho foi o Random Forest Regressor, apresentando o maior R^2 (70,28%) e os menores erros quando comparado aos demais modelos. A adição dos índices de sentimento no conjunto de dados de entrada não foi vantajosa para este problema, representando uma piora na performance de quase todos os modelos, quando comparados com o mesmo modelo usando o conjunto de dados apenas com retornos e volatilidades na entrada. Neste caso, em todos os cenários o método de seleção de variáveis escolhido foi o RFE.

Modelo	**Sent?**	**FS**	R^2	**MAE**	**MAPE**	**MSE**	**RMSE**
RF	Não	RFE	0.7028	0.0167	1.7948	0.0007	0.0258
HGB	Não	RFE	0.6905	0.0165	2.8161	0.0007	0.0264
RF	Sim	RFE	0.6756	0.0163	1.7483	0.0007	0.0259
HGB	Sim	RFE	0.6743	0.0148	1.7162	0.0007	0.0260
DT	Não	RFE	0.4958	0.0134	1.2256	0.0011	0.0337
DT	Não	RFE	0.3893	0.0145	1.6373	0.0013	0.0355
KNN	Sim	RFE	0.2052	0.0285	2.6814	0.0016	0.0405
KNN	Não	RFE	0.1713	0.0306	5.1121	0.0019	0.0432
LR	Sim	RFE	0.1071	0.0298	3.5540	0.0018	0.0430
LR	Não	RFE	0.092	0.0304	3.1668	0.002	0.0452

Tabela 8 – Performance dos modelos na predição dos retornos ordenados pelo R^2

No caso da volatilidade, o modelo com melhor desempenho também foi o Random Forest Regressor, apresentando o maior R^2 (78,40%) e os menores erros quando comparado aos demais modelos. Neste cenário, a adição dos índices de sentimento no conjunto de dados de entrada foi bastante vantajosa, representando uma melhora considerável em todos os modelos, quando comparados com o mesmo modelo usando o conjunto de dados apenas com retornos e volatilidades na entrada. Neste caso, utilizando o conjunto de dados enriquecido com os índices de sentimento, o método de seleção de variáveis escolhido foi o RFECV, e utilizando o conjunto de dados apenas com retornos e volatilidades, os métodos de seleção de variáveis foram descartados e todas as variáveis originais foram utilizadas pelo modelo.

Observando no detalhe apenas a métrica R^2, é possível comparar a performance dos modelos considerando os contendo os índices de sentimento ou apenas com as séries de retornos e volatilidades nas figuras 12 e 13.

Modelo	Sent?	FS	R^2	MAE	MAPE	MSE	RMSE
RF	Sim	RFECV	0.7840	0.0723	0.1341	0.0174	0.1321
KNN	Sim	RFECV	0.7501	0.0964	0.1795	0.0202	0.1421
HGB	Sim	RFECV	0.7302	0.0751	0.1387	0.0218	0.1476
RF	Não	-	0.5831	0.1027	0.2037	0.0372	0.1929
HGB	Não	-	0.5768	0.1005	0.2012	0.0378	0.1943
DT	Não	-	0.5629	0.0858	0.1715	0.0353	0.1879
DT	Sim	RFECV	0.5534	0.0899	0.1748	0.0399	0.1996
LR	Sim	RFECV	0.4280	0.1659	0.2987	0.0462	0.2149
KNN	Não	-	0.2155	0.1834	0.3614	0.0700	0.2646
LR	Não	-	0.0047	0.2412	0.4537	0.0888	0.2980

Tabela 9 – Performance dos modelos na predição da volatilidade ordenados pelo R^2

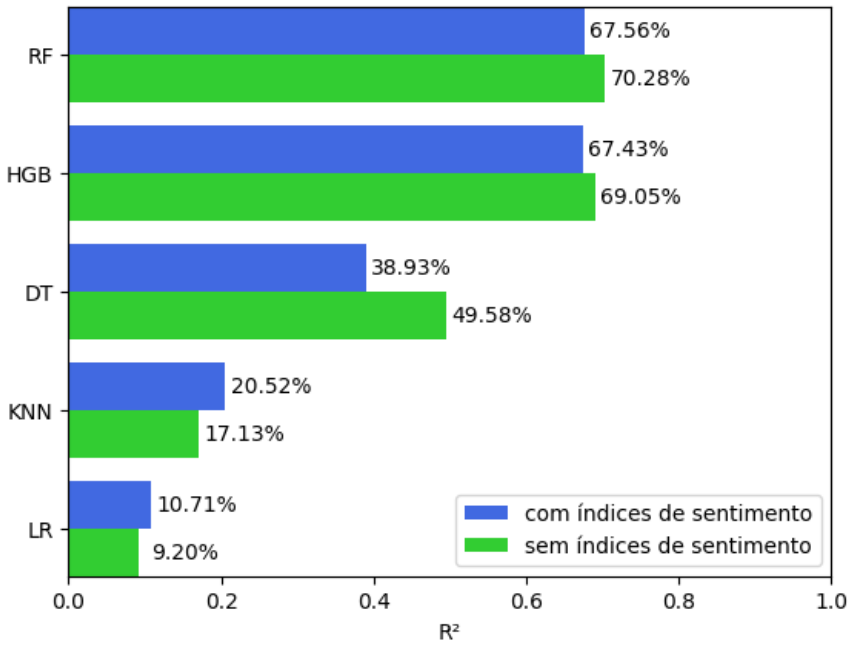

Figura 12 – Comparação do R^2 para os modelos na predição dos retornos do Bitcoin.

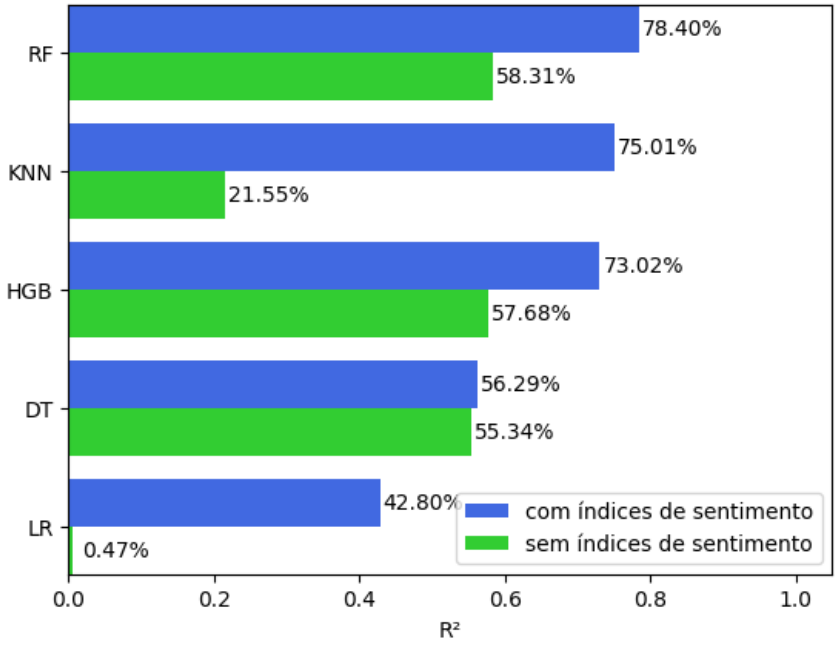

Figura 13 – Comparação do R^2 para os modelos na predição da volatilidade do Bitcoin.

6 CONCLUSÕES

Neste trabalho foi proposta uma comparação entre diversas técnicas para a predição dos retornos e volatilidade do Bitcoin. O propósito subjacente é fornecer apoio à tomada de decisão no contexto de estratégias de negociação eletrônica de criptomoedas. Para tal, foi considerada uma combinação de dados históricos dos ativos financeiros selecionados, combinados com índices de sentimento sobre o próprio Bitcoin, economia, redes sociais e notícias.

Concluímos que, dentre os modelos aplicados, tivemos melhor desempenho na predição da volatilidade, quando comparada aos retornos do Bitcoin. O Random Forest foi o modelo que apresentou melhor desempenho em ambos os cenários, apresentando um coeficiente de determinação (R^2) de 78,40% quando utilizado para prever a volatilidade. Observou-se ainda que, para a predição da volatilidade do Bitcoin, a adição dos índices de sentimento representou melhora na performance em todos os modelos, porém esta mesma melhora não foi observada na predição dos retornos.

Para trabalhos futuros, pretende-se avaliar técnicas de *tuning* para cada modelo escolhido, além de executar e documentar uma estratégia de negociação de volatilidade baseada nas predições do modelo com melhor performance. Adicionalmente, conforme mencionado em seções anteriores, a motivação para a escolha do Bitcoin como variável dependente neste trabalho deu-se principalmente pela sua predominância no mercado das criptomoedas, porém também é possível aplicar a metodologia proposta a outras criptomoedas.

REFERÊNCIAS

ABRAHAM, J. *et al.* Cryptocurrency price prediction using tweet volumes and sentiment analysis. *In*: . [*S.l.: s.n.*], 2018.

BAKER, S. R. *et al.* Twitter-derived measures of economic uncertainty. *In*: . [*S.l.: s.n.*], 2021. Available at: <https://api.semanticscholar.org/CorpusID:235399702>.

BITCOIN Fear Greed Index. 2018. <https://alternative.me/crypto/fear-and-greed-index/>. Accessed: 2023-07-31.

BUCKMAN, S. R. *et al.* News Sentiment in the Time of COVID-19. **FRBSF Economic Letter**, v. 2020, n. 08, p. 1–05, April 2020. Available at: <https://ideas.repec.org/a/fip/fedfel/87710.html>.

CHEN, T. *et al.* Xgboost: extreme gradient boosting. **R package version 0.4-2**, v. 1, n. 4, p. 1–4, 2015.

COLIANNI, S. G.; ROSALES, S. M.; SIGNOROTTI, M. Algorithmic trading of cryptocurrency based on twitter sentiment analysis. *In*: . [*S.l.: s.n.*], 2015.

DIAS, I. K.; FERNANDO, J. R.; FERNANDO, P. N. D. Does investor sentiment predict bitcoin return and volatility? a quantile regression approach. **International Review of Financial Analysis**, v. 84, p. 102383, 2022. ISSN 1057-5219. Available at: <https://www.sciencedirect.com/science/article/pii/S1057521922003337>.

D'AMATO, V.; LEVANTESI, S.; PISCOPO, G. Deep learning in predicting cryptocurrency volatility. **Physica A: Statistical Mechanics and its Applications**, v. 596, p. 127158, 2022. ISSN 0378-4371. Available at: <https://www.sciencedirect.com/science/article/pii/S0378437122001704>.

FANG, F. *et al.* Cryptocurrency trading: a comprehensive survey. **Financial Innovation**, v. 8, n. 1, p. 13, Feb 2022. ISSN 2199-4730. Available at: <https://doi.org/10.1186/s40854-021-00321-6>.

FELIZARDO, L. K. *et al.* Outperforming algorithmic trading reinforcement learning systems: a supervised approach to the cryptocurrency market. **Expert Systems with Applications**, 2022.

GARCIA, D.; SCHWEITZER, F. Social signals and algorithmic trading of bitcoin. **Royal Society Open Science**, v. 2, n. 9, p. 150288, 2015. Available at: <https://royalsocietypublishing.org/doi/abs/10.1098/rsos.150288>.

GIAGLIS, G. *et al.* Using time-series and sentiment analysis to detect the determinants of bitcoin prices. *In*: . [*S.l.: s.n.*], 2015.

GUJARATI, D. N.; PORTER, D. C. **Basic Econometrics**. 5. ed. [*S.l.: s.n.*]: McGraw-Hill/Irwin, 2009. ISBN 978-0-07-337577-9.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning Data Mining, Inference, and Prediction**. [*S.l.: s.n.*], 2009.

JI, S.; KIM, J.; IM, H. A comparative study of bitcoin price prediction using deep learning. **Mathematics**, v. 7, n. 10, 2019. ISSN 2227-7390. Available at: <https://www.mdpi.com/2227-7390/7/10/898>.

LAMON, C.; NIELSEN, E.; REDONDO, E. Cryptocurrency price prediction using news and social media sentiment. *In*: . [*S.l.: s.n.*], 2017.

M., P. *et al.* Prediction of the price of ethereum blockchain cryptocurrency in an industrial finance system. **Computers Electrical Engineering**, v. 81, p. 106527, 2020. ISSN 0045-7906. Available at: <https://www.sciencedirect.com/science/article/pii/S0045790618331343>.

MALLQUI, D.; FERNANDES, R. Predicting the direction, maximum, minimum and closing prices of daily bitcoin exchange rate using machine learning techniques. **Applied Soft Computing**, v. 75, 12 2018.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis**. 5th. ed. [*S.l.: s.n.*]: John Wiley Sons, Inc, 2012.

MOUGAYAR, W. **The business blockchain: Promise, practice, and application of the next internet technology**. [*S.l.: s.n.*]: John Wiley amp; Sons, 2016.

NAKAMOTO, S. **Bitcoin: A Peer-to-Peer Electronic Cash System**. 2008. Available at: <www.bitcoin.org>.

NARAYANAN, A. *et al.* **Bitcoin and Cryptocurrency Technologies**. 2016.

OIKONOMOPOULOS, S. *et al.* Cryptocurrency price prediction using social media sentiment analysis. *In*: **2022 13th International Conference on Information, Intelligence, Systems Applications (IISA)**. [*S.l.: s.n.*], 2022. p. 1–8.

OKSANEN, A. *et al.* Gambling and online trading: emerging risks of real-time stock and cryptocurrency trading platforms. **Public Health**, v. 205, p. 72–78, 2022. ISSN 0033-3506. Available at: <https://www.sciencedirect.com/science/article/pii/S0033350622000348>.

PANG, B.; LEE, L. **Opinion mining and sentiment analysis**. 2008. 1-135 p.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PICHL, L.; KAIZOJI, T. Volatility analysis of bitcoin. **Quantitative Finance and Economics**, v. 1, n. 4, p. 474–485, 2017.

SHAKRI, I. Time series prediction using machine learning: a case of bitcoin returns. **Studies in Economics and Finance**, ahead-of-print, 11 2021.

TAPSCOTT, D.; TAPSCOTT, A. **Blockchain Revolution: How the technology behind Bitcoin is changing money, business and the world**. [*S.l.: s.n.*]: Portfolio Penguin, 2018.

URAS, N. *et al.* **Forecasting Bitcoin closing price series using linear regression and neural networks models**. 2020.

VALENCIA, F.; GóMEZ-ESPINOSA, A.; VALDES, B. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. **Entropy**, v. 21, p. 1–12, 06 2019.

VIRK, D. S. Prediction of bitcoin price using data mining. *In*: . [*S.l.: s.n.*], 2017.

WHALEY, R. Understanding the vix. **The Journal of Portfolio Management**, v. 35, p. 98–105, 02 2009.