



INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Classificação de tendências e previsão de preço do Índice S&P 500: trading recorrendo a deep learning e XGBoost

Marta Sofia Morgado das Neves

Mestrado em Ciência de Dados

Orientadora:

Dra. Diana Mendes, Professora

ISCTE - Instituto Universitário de Lisboa

Co-Orientadora:

Dra. Anabela Costa, Professora

ISCTE - Instituto Universitário de Lisboa

Setembro, 2023

Departamento de Métodos Quantitativos para Gestão e Economia

Departamento de Ciências e Tecnologia da Informação

Classificação de tendências e previsão de preço do Índice S&P 500: trading recorrendo a deep learning e XGBoost

Marta Sofia Morgado das Neves

Mestrado em Ciência de Dados

Orientadora:

Dra. Diana Mendes, Professora

ISCTE - Instituto Universitário de Lisboa

Co-Orientadora:

Dra. Anabela Costa, Professora

ISCTE - Instituto Universitário de Lisboa

Setembro, 2023

Agradecimentos

Ao redigir estas palavras, um sentimento de nostalgia percorre-me, pois, esta página simboliza não apenas a conclusão de uma etapa importante da minha jornada, mas também a oportunidade de contemplar a imensa sorte que tenho. Encontro-me cercada por pessoas incríveis, dispostas a tudo para verem-me feliz e apoiarem-me na conquista dos meus objetivos. Assim, posso afirmar com convicção que a realização desta dissertação é o fruto do contributo positivo de diversas pessoas. Desejo expressar a minha profunda gratidão a todos aqueles cujo ânimo e apoio foram decisivos para tornar este feito possível.

Desejo manifestar, com elevado apreço, os meus sentimentos de gratidão às minhas orientadoras, a Professora Doutora Diana Mendes e a Professora Doutora Anabela Costa, cujas orientações conferiram uma qualidade indiscutível a esta dissertação. A sua confiança nas minhas capacidades incentivou-me a superar obstáculos e a alcançar um resultado que estou orgulhosa de apresentar. Obrigada por terem acreditado em mim!

Aos meus pais, a quem jamais poderei retribuir devidamente e a quem qualquer reconhecimento nunca será suficiente. Nenhuma etapa do meu percurso teria sido possível se não fosse graças à sua dedicação, por acreditarem mais em mim do que qualquer pessoa e por terem abdicado de tanto em prol da minha felicidade!

Agradeço particularmente à minha mãe, por ser o maior exemplo de garra que eu conheço, por me ter dado asas para voar, por acreditar em mim, por me ter ensinado a ler e, principalmente, a prontidão constante para me ajudar em tudo na vida!

Ao meu namorado, agradeço por todo o amor, incentivo, apoio e compreensão. A realização desta dissertação foi sem dúvida uma tarefa menos hercúlea por o ter ao meu lado.

Ao resto da minha família, agradeço por estarem ao meu lado, por apoiarem as minhas ambições e por serem um pilar de força e incentivo.

Por fim, mas não menos importante, a todos os meus colegas com os quais tive o privilégio de trabalhar, até o momento, pois esta jornada não teria sido possível sem a colaboração e apoio de cada um deles.

Resumo

Esta dissertação visa otimizar as estratégias de investimento no índice S&P 500 através do desenvolvimento de dois modelos: previsão dos seus movimentos, discernindo se o preço subirá ou declinará no dia subsequente, além de antecipar o respetivo valor de fecho. Os algoritmos escolhidos foram o *eXtreme Gradient Boosting* (XGBoost) e o *Long Short-Term Memory* (LSTM), devido aos seus desempenhos comprovados empiricamente.

Para aprimorar o desempenho dos modelos, foram incorporadas variáveis exógenas, a saber: preço do petróleo, Índice VIX e Índice Dow Jones, com o objetivo de capturar eventos de maior imprevisibilidade que possam influenciar o comportamento do referido Índice.

No problema de classificação, o modelo atingiu uma *accuracy* de 94%, sugerindo a capacidade de discernir com confiabilidade se o preço subirá ou descerá no dia subsequente.

No que concerne ao problema de previsão do preço de fecho, o modelo registou um *Root Mean Square Error* (RMSE) de 8.94, indicando que, em média, as previsões distanciam-se 8.94 dólares do valor real. Este resultado mostra que o modelo possui uma boa capacidade de previsão, apesar da margem de erro presente.

A inclusão das variáveis exógenas permitiu aumentar a *accuracy* do problema de classificação, em 20 pontos percentuais, desempenhando, assim, um papel significativo na melhoria das previsões sobre o movimento do índice S&P 500. Porém, no problema de regressão, esta inclusão, não contribuiu de forma positiva para o desempenho do modelo.

Realça-se que, qualquer aplicação prática dos modelos propostos requer análises adicionais e ponderações devido à volatilidade e complexidade dos mercados financeiros.

Palavras-chave: Previsão; Classificação; S&P 500; Variáveis Exógenas; XGBoost; LSTM

Abstract

This dissertation aimed to optimize investment strategies in the S&P 500 index by developing two models: predicting its movements, discerning whether the price will rise or decline the following day and anticipating its closing value. Due to their empirically proven performances, the algorithms chosen were eXtreme Gradient Boosting (XGBoost) and Long Short-Term Memory (LSTM).

To improve the performance of the models, exogenous variables were incorporated, namely: oil price, VIX Index and Dow Jones Index, to capture events of greater unpredictability that may influence the behaviour of the Index.

In the classification problem, the model achieved an *accuracy* of 94%, suggesting the ability to reliably discern whether the price will rise or fall the following day.

Regarding the problem of predicting the closing price, the model recorded a *Root Mean Square Error* (RMSE) of 8.94, indicating that, on average, the predictions are 8.94 dollars away from the actual value. This result shows that the model has a good predictive capacity, despite the margin of error present.

Including the exogenous variables increased the *accuracy* of the classification problem by 20 percentage points, thus playing a significant role in improving predictions about the movement of the S&P 500 index. However, in the regression problem, this inclusion did not contribute positively to the model's performance.

It should be noted that any practical application of the proposed models requires additional analysis and weighting due to the volatility and complexity of financial markets.

Keywords: Forecasting; Classification; S&P 500; Exogenous Variables; XGBoost; LSTM

Índice

Agradecimentos	i
Resumo	iii
Abstract	v
CAPÍTULO 1. Introdução	1
CAPÍTULO 2. Revisão da Literatura	5
2.1. Estudo e Análise dos Artigos Seleccionados	5
2.1.1 Métodos estudados na classificação da tendência	7
2.1.2 Métodos estudados na previsão de preço	9
2.1.3 Métodos Híbridos e Ensemble em Machine Learning	10
2.2 Conclusões do Estudo Realizado	11
CAPÍTULO 3. Metodologia	15
3.1 Enquadramento Económico	15
3.2 Variáveis Exógenas	20
3.3 Pré-Processamento dos Dados	24
3.4 Incorporação de Variáveis na Modelação	25
3.4.1 Exploração e Seleção de Variáveis para o Problema de Classificação	26
3.4.2 Seleção de Variáveis para o Problema de Regressão	30
3.5 Modelação	31
3.5.1 Construção do Modelo para o Problema de Classificação	32
3.5.1.1 Experiência 1	35
3.5.1.2 Experiência 2	35
3.5.1.3 Experiência 3	36
3.5.1.4 Experiência 4	37
3.5.1.5 Experiência 5	38
3.5.1.6 Experiência 6	39
3.5.2 Construção do Modelo para o Problema de Regressão	42
3.5.2.1 XGBoost	44
3.5.2.2 LSTM	45
CAPÍTULO 5. Conclusões	51
Referências Bibliográficas	55

ANEXOS	58
Anexo A – Variáveis Originais	58
1. Índice S&P 500	58
2. Índice VIX	58
3. Petróleo	59
4. Índice DOW	59
Anexo B – Lista de Variáveis	60

Índice de Figuras

Figura 1 - Resultados da query no Scopus com “trend” e “stock markets”, entre 1955 e 2023.	6
Figura 2 - Frequência de utilização dos modelos mais populares nos artigos analisados.	7
Figura 3 - Frequência de utilização das métricas nos artigos analisados.	7
Figura 4 - Metodologia CRISP - DM.	15
Figura 5 - Evolução do preço de fecho do petróleo, desde o ano 2000 até a atualidade.	17
Figura 6 - Evolução do preço de fecho do índice S&P 500, de 1990 até à atualidade.	18
Figura 7 - S&P 500 Retorno Anual Histórico.	19
Figura 8 - Relação entre o índice VIX e o índice S&P 500, no ano de 2022.	22
Figura 9 - Pesos de cada setor do Índice S&P 500.	22
Figura 10 – Matriz de correlação das variáveis mais relevantes.	28
Figura 11 - Features selecionadas para o problema de classificação.	29
Figura 12 - Features selecionadas para o problema de regressão.	31
Figura 13 - Metodologia adotada para o desenvolvimento do problema de classificação.	33
Figura 14 - Resultados obtidos com as experiências desenvolvidas no problema de classificação.	41
Figura 15 - Autocorrelação da variável que mede a magnitude as oscilações do índice S&P 500.	43
Figura 16 - Autocorrelação Parcial da variável que mede a magnitude as oscilações do S&P 500.	43
Figura 17 - Autocorrelação Parcial da variável que mede a magnitude as oscilações do S&P 500.	43
Figura 18 - Autocorrelação Parcial da variável que mede a magnitude as oscilações do S&P 500.	43
Figura 19 - Arquitetura LSTM otimizada.	46
Figura 20 - Previsões do preço de fecho do Índice S&P 500.	47
Figura 21 - Previsão do preço de fecho do Índice S&P 500 - Modelo Final	48

Índice de Tabelas

Tabela 1 – Experiência n.º 1: XGBoost sem otimização de parâmetros.	35
Tabela 2 - Experiência n.º 2: XGBoost com otimização de parâmetros.	36
Tabela 3 - Experiência n.º 3: XGBoost com otimização de parâmetros, utilizando novo conjunto de features e ajuste de parâmetros.	37
Tabela 4 - Experiência n.º 4: XGBoost com otimização de parâmetros, utilizando novo conjunto de features e ajuste de parâmetros.	38
Tabela 5 - Experiência n.º 5: XGBoost com otimização de parâmetros, utilizando novo conjunto de features e ajuste de parâmetros.	39
Tabela 6 - Experiência n.º 6: XGBoost com otimização de parâmetros, utilizando novo conjunto de features e ajuste de parâmetros.	40
Tabela 7 - Experiência n.º 1: XGBoost Regressor.	44
Tabela 8 - Experiências: Resultados obtidos com o uso da rede LSTM para prever o preço de fecho do Índice S&P 500.	46
Tabela 9 - Resultados obtidos com a última experiência na previsão do preço de fecho.	47

Introdução

Nos últimos anos, o Índice S&P 500 tem despertado um interesse considerável entre investidores e *traders* devido à sua inegável importância como um indicador fundamental do desempenho do mercado de ações nos Estados Unidos. Constituído por 500 empresas de elevado valor de mercado, o Índice representa uma relevância inestimável na avaliação da saúde económica do país, graças à sua abrangência de setores económicos. A sua composição diversificada, incluindo setores como energia, tecnologia, saúde, finanças e muito mais, confere-lhe uma relevância singular e proporciona *insights* essenciais na compreensão e interpretação do cenário económico vivenciado.

À luz do estudo intitulado *Markets will be markets: An analysis of long-term returns from S&P 500* realizado pela empresa de renome McKinsey & Company, o Índice S&P 500 tem sido um dos principais pontos de referência para investimento, graças à sua longa história de retornos consistentes e à sua abrangência representativa que oferece. É certo que, atrair retornos elevados e consistentes é, de facto, um objetivo comum para muitos *traders* e investidores. O investimento no Índice S&P 500 pode proporcionar oportunidades de crescimento substanciais em comparação com investimentos mais conservadores. No entanto, é importante destacar que investimentos com retornos mais elevados, geralmente, têm intrínseco um nível de risco também mais elevado.

Diante da inerente volatilidade dos mercados financeiros, os *traders* têm-se dedicado a utilizar diversas abordagens analíticas com o intuito de mitigar os riscos associados aos seus investimentos no Índice S&P 500. Essas abordagens analíticas envolvem, entre outras técnicas, a análise técnica, análise fundamental e o desenvolvimento de modelos preditivos, visando identificar padrões de comportamento, investigar fatores macroeconómicos relevantes e antecipar os movimentos de mercado, com o propósito de alcançar uma maior eficiência na alocação de recursos financeiros.

O prognóstico das tendências e dos valores do Índice S&P 500, tem sido objeto de uma minuciosa análise e constitui uma área de pesquisa extensivamente estudada. A capacidade de realizar tais previsões com precisão tem um valor inestimável, conferindo vantagens significativas a investidores e *traders*.

Primeiramente, é importante destacar que a antecipação com precisão dos movimentos do Índice S&P 500 confere aos investidores a capacidade de tomar decisões conscientes em relação às suas estratégias de investimento. Através dessa antecipação, obterão oportunidade de identificar momentos oportunos para comprar ou vender ações, aumentando significativamente as suas hipóteses de obter lucros substanciais. Revela-se igualmente imprescindível, a precisão da previsão, visto que, tal habilidade acarretará uma vantagem crucial: a mitigação de riscos. Conscientes de que os mercados financeiros são inerentemente voláteis e sujeitos a incertezas, a capacidade de antecipar tais movimentos possibilita que os investidores adotem medidas de proteção para as suas carteiras. Por seu turno, preparar-se-ão de forma adequada para enfrentar períodos de volatilidade e, até mesmo, evitar possíveis quedas abruptas.

Na sequência dos factos supramencionados, tornou-se imprescindível a modelização de um modelo de elevada robustez, destinado a determinar com precisão as tendências e preços do Índice S&P 500. Com o intento de obter benefícios tangíveis, a presente dissertação procura seguir uma abordagem metodológica, procurando a utilização de técnicas de *machine learning* e inteligência artificial. Dessa confrontação, espera-se que os resultados deste projeto possam oferecer *insights* valiosos de modo a aprimorar a tomada de decisão e a gestão de risco.

Nesse contexto, a presente dissertação adota uma abordagem que utiliza dois algoritmos de *machine learning* amplamente estudados e reconhecidos na literatura científica: *eXtreme Gradient Boosting* (XGBoost) e *Long Short-Term Memory* (LSTM). A seleção destes algoritmos foi fundamentada com base nas suas eficiências, empiricamente comprovadas, e na capacidade de lidar com a complexidade inerente aos dados financeiros. Por meio da aplicação de algoritmos de *machine learning*, o objetivo primordial é extrair *insights* que possibilitem uma avaliação do Índice S&P 500, proporcionando, assim, previsões alicerçadas em fundamentos sólidos e confiáveis.

As vantagens dos algoritmos XGBoost e LSTM no contexto da previsão do Índice S&P 500 são inquestionáveis e revestem-se de extrema importância. Ambos se destacam pela sua capacidade de abordar de forma eficaz conjuntos de dados financeiros, caracterizados pela presença de múltiplas variáveis, e não-linearidade inerente ao ambiente dinâmico do mercado de ações. O XGBoost, em virtude das suas técnicas avançadas de *gradient boosting*, evidencia-se como uma poderosa ferramenta ao capturar relações não triviais entre as variáveis, conferindo-lhe a capacidade de realizar previsões de elevada *performance*. Por sua vez, a LSTM destaca-se pela sua capacidade de apreender dependências de longo e curto prazo presentes nos dados, permitindo a exploração e compreensão de padrões temporais intrínsecos ao comportamento do Índice S&P 500. Essas características tornam os algoritmos indicados para este projeto, que tem como foco a previsão do preço de fecho e a classificação da tendência, do referido Índice, contribuindo, assim, de forma significativa para uma melhor compreensão e antecipação dos movimentos do mercado financeiro.

Realce-se, no entanto que, não obstante a reconhecida eficácia e capacidade dos modelos XGBoost e LSTM na análise de dados financeiros complexos, é possível que não sejam totalmente capazes de capturar adequadamente, eventos macroeconômicos, como crises financeiras e flutuações de preço, e outras ocorrências de escala global, como políticas monetárias e fenômenos geopolíticos, que exerçam influência significativa sobre o comportamento do Índice S&P 500, devido à sua intrínseca imprevisibilidade. Ora, por essa razão, considerou-se igualmente pertinente realizar uma análise criteriosa e exploração aprofundada de determinadas variáveis exógenas que possam exercer um impacto relevante na previsão do Índice S&P 500, com o intuito de avaliar a sua contribuição efetiva na melhoria do desempenho preditivo dos modelos utilizados.

A dissertação de que ora nos ocupamos encontra-se estruturado em quatro Capítulos, sendo que o primeiro Capítulo se consubstancia na presente introdução.

O Capítulo dois, conduzirá a uma revisão de literatura, abordando os métodos mais utilizados para enfrentar os desafios da classificação da tendência e previsão de preços do Índice S&P 500. Nesse contexto, explorar-se-á diversas abordagens de *machine learning*, dando destaque particularmente aos algoritmos XGBoost e LSTM. Esta revisão proporcionará uma visão compreensiva das metodologias mais utilizadas e dos resultados alcançados, fornecendo conhecimento sólido para o desenvolvimento dos modelos propostos neste trabalho.

No terceiro Capítulo, dedicado inteiramente à descrição de toda a metodologia adotada, recorrendo ao CRISP-DM, desde a etapa de *Data Understanding* até o desenvolvimento de modelos e análise dos resultados obtidos. Adicionalmente, tal Capítulo ambicionará detalhar todas as experiências realizadas com a intenção de aprimorar os resultados alcançados. Numa fase inicial, o *Data Understanding* abrangerá a exploração criteriosa dos diversos conjuntos de dados utilizados, bem como a identificação das variáveis mais pertinentes para a previsão do Índice S&P 500. Com efeito, serão utilizadas técnicas de seleção de variáveis, visando obter uma percepção detalhada da importância das diversas características incorporadas. Num segundo momento, adotar-se-á técnicas de *Data Preparation*, com o intento de tratar os dados ausentes e outras transformações essenciais de modo a garantir a qualidade dos dados utilizados na construção dos modelos. Por fim, na etapa de *Model Development*, proceder-se-á à implementação dos algoritmos XGBoost e LSTM, com o ajuste de parâmetros adequados para cada método. Com a implementação rigorosa destes algoritmos e a correta calibração dos parâmetros, procurar-se-á obter previsões mais precisas e confiáveis para o comportamento do Índice S&P 500, contribuindo para uma melhor compreensão do cenário financeiro em questão.

A encerrar a presente dissertação, encontrar-se-ão as principais conclusões alcançadas no desenvolvimento da mesma. Com efeito, realizar-se-á uma análise crítica ponderada nos resultados logrados

CAPÍTULO 2

Revisão da Literatura

Neste Capítulo, será realizado uma análise da literatura científica recente referente à previsão da tendência e flutuações do mercado de ações. O estudo baseou-se numa seleção criteriosa de artigos científicos provenientes da plataforma *Scopus*, publicados a partir do ano de 2019. O objetivo principal desta análise consiste em investigar os modelos de previsão e as métricas de desempenho frequentemente usados na literatura, com o propósito de identificar os mais predominantes e aqueles que apresentam resultados competitivos.

Uma tendência distinta foi observada durante a análise dos artigos selecionados. Os modelos LSTM e XGBoost emergem como os mais frequentemente utilizados na literatura. Adicionalmente, constatou-se que a métrica de avaliação mais comum para esses modelos é o RMSE. Ao examinar a primeira consulta, verificou-se um aumento gradual no número de publicações sobre o tema nos últimos seis anos, em comparação com o período anterior, refletindo um crescimento de aproximadamente 70%.

Portanto, este Capítulo abordará de forma minuciosa todos os aspetos supramencionados, fornecendo uma análise detalhada dos modelos de previsão mais predominantes, das métricas de desempenho adotadas e do crescimento da literatura científica acerca da previsão de tendências e flutuações do mercado de ações ao longo dos últimos anos.

2.1. Estudo e Análise dos Artigos Selecionados

Os mercados de ações têm se tornado uma componente intrínseca da economia global num passado recente. Qualquer oscilação nesses mercados exerce influência tanto sobre a esfera financeira pessoal e corporativa, quanto sobre a saúde económica de uma nação. Conforme destacado por [Price \(2022\)](#), o mercado de ações sempre deteve posição proeminente entre as opções de investimento, devido aos elevados retornos que oferece. Entretanto, investir no mercado de ações é inerentemente associado a certos riscos, decorrentes do seu comportamento imprevisível. Portanto, um modelo preditivo inteligente para a antecipação dos movimentos do mercado de ações seria amplamente desejável e objeto de grande interesse ([Hassan et al., 2005](#)).

O mercado de ações é um sistema complexo e muito volátil, e por esse motivo, a previsão da tendência do mercado de ações é caracterizada pela intensidade dos dados, do ruído existente, pelos dados serem ou não estacionários, alto grau de incerteza e relações ocultas ([Yu et al., 2008](#)).

Considerando que o mercado de ações constitui um sistema complexo e dinâmico, além de representar um dos investimentos mais populares e rentáveis, optou-se por conduzir uma investigação com vista a analisar os algoritmos adotados para abordar essa problemática e identificar aqueles que têm obtido os melhores resultados. O propósito deste trabalho é desenvolver um algoritmo que demonstre um desempenho superior em relação aos trabalhos anteriores, uma vez que ainda se identifica espaço para inovação neste domínio específico.

Portanto, a primeira tarefa consistiu na reunião de literatura que fosse relevante para o domínio que está a ser analisado para construir um corpo abrangente de conhecimento. O *Scopus* é um dos três motores de pesquisa mais utilizados para consultar artigos académicos ([Meho et al., 2007](#)). Por esse motivo, foi o eleito para a coleção e consulta dos diversos artigos presentes na revisão de literatura.

A previsão da tendência das ações constitui um trunfo fundamental para qualquer investidor. Tal vantagem pode levar a uma antecipação da compra ou venda de determinada ação, permitindo assim reduzir o risco de investimento ([Yoo et al., 2005](#)).

Portanto, o interesse em construir um modelo com alta precisão preditiva, para que se possa ter uma boa compreensão da volatilidade do mercado de ações, tem sido vasta ao longo dos anos ([Wang et al., 2019](#)).

Estudos e livros publicados sobre previsão da tendência de ações podem ser encontrados desde 1955. No entanto, uma pesquisa efetuada na base de dados *Scopus* para resultados incluindo as palavras-chave “trend” e “stock markets” revela um enorme aumento no número de publicações a partir da década de 2000, conforme visível na Figura 1.

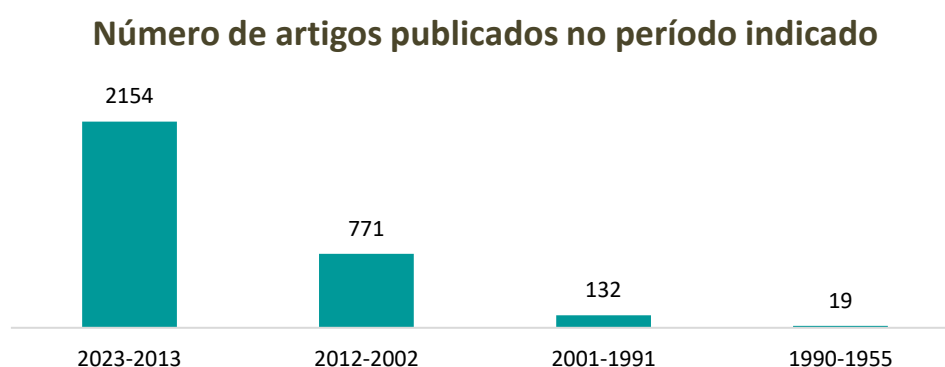


Figura 1 - Resultados da query no Scopus com “trend” e “stock markets”, entre 1955 e 2023.

Além disso, nos últimos dez anos foram publicados mais do dobro dos artigos que nos 30 anos anteriores, revelando uma preocupação crescente relativamente a esta temática.

Após a filtragem e seleção dos artigos, analisou-se os modelos de previsão da tendência que cada artigo abordava no sentido de perceber quais os modelos mais utilizados pelos investigadores.

Posto isto, verificou-se que os algoritmos mais utilizados são o método LSTM e XGBoost, seguindo-se o *Support Vector Machine* (SVM) e *Random Forest*, conforme consta na Figura 2.

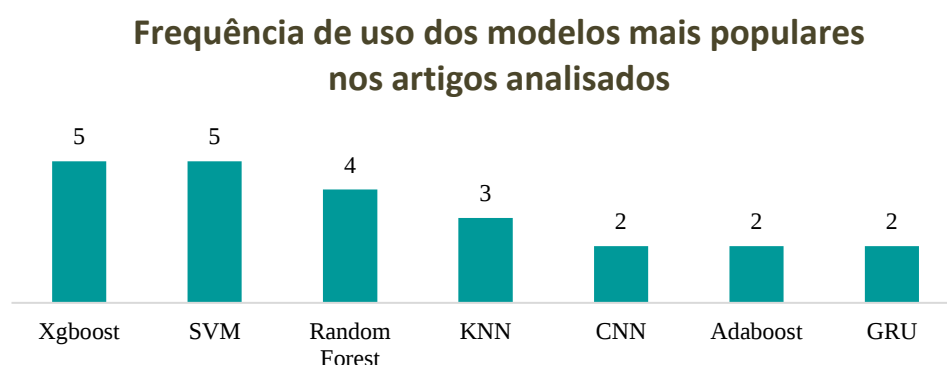


Figura 2 - Frequência de utilização dos modelos mais populares nos artigos analisados.

Em relação às métricas de *performance*, observou-se que a métrica mais utilizada é a *accuracy* para classificação, uma vez que mede número de previsões corretas face ao total de previsões. E para problemas de regressão, destaca-se o RMSE, como sendo a medida mais popular, dado que esta analisa quão afastados estão, em média, as previsões dos valores reais dos preços das ações (Figura 3).

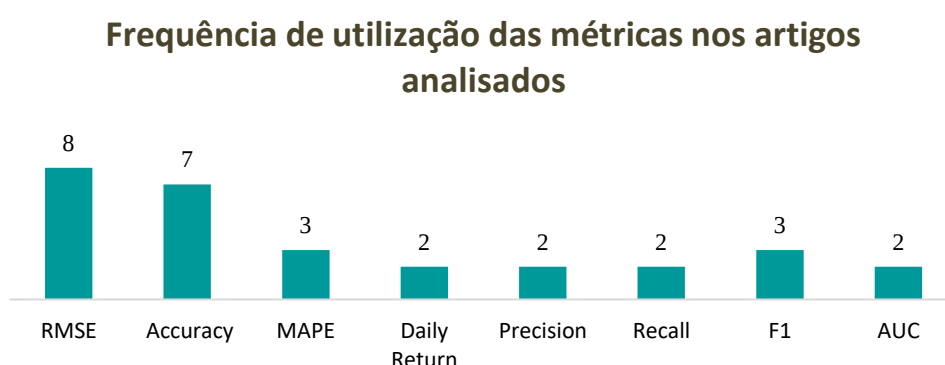


Figura 3 - Frequência de utilização das métricas nos artigos analisados.

Para a revisão de literatura, decidiu-se dividir entre problemas de classificação e de regressão. Nos parágrafos abaixo, estão descritos os modelos implementados assim como o respetivo desempenho para a classificação das flutuações no mercado de ações.

2.1.1 Métodos estudados na classificação da tendência

[Ghosh et al. \(2022\)](#) estudaram o LSTM e *Random Forest* para prever os movimentos de ações constituintes do S&P 500 de janeiro de 1993 a dezembro de 2018, para negociação diária. Em cada dia de negociação, compravam as 10 ações com maior probabilidade de subir, com o intuito de comprar barato e ir vender no mercado a um preço superior. Além disso, ainda realizaram *short-selling* nas 10

ações com maior probabilidade de descer, com o objetivo de pedirem uma ação emprestada a uma entidade, vender no mercado ao preço atual e depois se o preço descer comprariam novamente. Com este sistema, obtiveram retornos diários superiores com o modelo LSTM, com um valor de 0.41%. Ou seja, se investir hoje 12 mil euros ao fim de um ano este valor aumenta para 53 mil euros, obtendo um lucro de 41 mil euros. No futuro, tencionam analisar esta abordagem aplicada a criptomoedas.

O modelo XGBoost tem evidenciado ser dos melhores algoritmos para problemas de classificação encontrados na literatura, sendo que supera todos os outros existentes e é capaz de prever a longo prazo ([Dey et al., 2016](#)). No que segue, analisou-se a performance de algoritmos de *boosting* face a outros modelos.

[Ampomah et al. \(2020\)](#) compararam a eficácia dos modelos de *machine learning* baseados em árvores na previsão da direção do movimento do preço de ações de três bolsas diferentes (NYSE, NASDAQ e NSE), de 2005 a 2019. Os resultados mostram que em termos de *accuracy* o *Adaboost* teve um melhor desempenho face aos outros modelos, seguindo o XGBoost com valores de 0.8606 e 0.8538, respetivamente. Estes autores, tencionam no futuro incorporar modelos de *machine learning* que envolvam processos *Gaussianos*, uma técnica de regularização e ainda técnicas baseadas em *kernel*.

Da mesma forma, [Park et al. \(2019\)](#) implementaram a previsão do movimento de ações no Índice S&P 500 e os resultados indicam que o classificador *Boosting* com *Adaboost* real prevê os preços com um melhor desempenho, tendo obtido uma *accuracy* de 0.6840. No entanto, os autores anteriores obtiveram um poder preditivo mais elevado.

Recentemente, [Wang et al. \(2023\)](#) abordaram a complexa tarefa de prever as tendências do mercado de ações, incorporando informações oportunas de notícias e redes sociais em modelos preditivos. A abordagem proposta combina indicadores técnicos e sentimentos de textos de redes sociais como fatores influentes em diferentes situações. A pesquisa apresenta um modelo híbrido de três fases que combina métodos tradicionais de *machine learning* e *deep learning* como principais preditores. A inclusão de indicadores técnicos e fatores externos, como o sentimento de notícias, demonstrou melhorar a precisão geral da previsão em comparação com abordagens que consideram apenas um desses aspetos. Além disso, concluíram que a seleção criteriosa de indicadores técnicos e a ampliação da quantidade e variedade de fontes de dados de sentimentos podem contribuir para aprimorar ainda mais os resultados. Ao comparar modelos de *machine learning* e *deep learning*, não observaram uma vantagem significativa dos algoritmos de *deep learning*. Os modelos tradicionais de *machine learning* podem alcançar desempenhos competitivos para tarefas relativamente simples. O modelo CNN (*Convolutional Neural Network*) destaca-se como o melhor para prever tendências de ações, seguido pelo modelo XGBoost, com uma ligeira redução de 0.72 pontos percentuais na *accuracy* em relação ao CNN.

2.1.2 Métodos estudados na previsão de preço

Nesta secção, são apresentados diversos trabalhos relacionados com a previsão de preços de ações ou dos seus retornos.

LSTM está entre os modelos de *deep learning* mais populares usados atualmente. Também está a ser aplicado em previsão de séries temporais, que é um problema particularmente difícil de resolver devido à presença de tendências de longo prazo, flutuações sazonais e cíclicas, e ruído. O desempenho do LSTM é altamente dependente da escolha dos valores de vários hiperparâmetros que caracterizam o modelo, a fim de obter os melhores resultados ([Yadav et al., 2020](#)). Os estudos apresentados abaixo, ilustram o alto desempenho do LSTM face a outros modelos.

[Pokhrel et al. \(2022\)](#) realizou uma análise comparativa de três modelos de *deep learning*, nomeadamente, LSTM, *Gated Recurrent Unit* (GRU) e CNN, na previsão do preço de fecho do dia seguinte da Bolsa de Valores do Nepal (Índice NEPSE), de julho de 2016 a janeiro de 2021. Os resultados experimentais mostram que a arquitetura do modelo LSTM fornece um desempenho superior com alta precisão de previsão, com um RMSE de 10.466. Para atingirem este desempenho, utilizaram dados financeiros, macroeconómicos, indicadores técnicos e ainda análise de sentimento de notícias financeiras. Além disso, o algoritmo de otimização com o melhor desempenho foi o Adam. Como forma de melhoria do modelo, sugere-se a utilização de modelos híbridos combinando outra rede neuronal, e além das notícias financeiras, outras análises de sentimentos poderiam ser incluídas, como por exemplo dos *influencers* nas redes sociais.

[Bhandari et al. \(2022\)](#) testou o LSTM de camada única e multicamada para prever o preço de fecho do Índice S&P 500, no dia seguinte. Os resultados experimentais mostram que o modelo LSTM de camada única fornece um ajuste superior em comparação com os modelos de LSTM multicamadas. Neste estudo, foi recolhido os dados referentes a 15 anos, de 2006 a 2020, e além disso, o segundo conjunto de variáveis são variáveis macroeconómicas que influenciam significativamente o desempenho do mercado das ações, explicando potenciais informações na previsão. Como [Pokhrel et al. \(2022\)](#), este artigo também evidenciou melhores resultados com o otimizador Adam, apresentando um RMSE de 28.65. Como sugestão de melhoria, referem que poderiam no futuro implementar algoritmos híbridos.

[Gumelar et al. \(2020\)](#) estudaram a previsão do preço de fecho de ações de 25 empresas, na Bolsa de Valores da Indonésia (IDX), entre 2000 e 2019. Os dois algoritmos utilizados foram o LSTM e o XGBoost, e ambos demonstraram uma excelente *accuracy* no resultado da previsão, com 99.995% e 99%, respetivamente.

Ao contrário de [Pokhrel et al. \(2022\)](#), [Medarhri et al. \(2022\)](#), avaliou seis técnicas de *machine learning* para o Índice S&P 500, para uma janela temporal de 10 anos, sendo o preço de fecho a variável *target* para análise. Os resultados deste estudo indicam que o *Multilayer perceptron* (MLP) e as técnicas de LSTM foram os modelos mais precisos em comparação com outros. O próximo trabalho poderá concentrar-se em novas técnicas de *machine learning*, técnicas de otimização e explorar novas variáveis, tais como indicadores financeiros ou económicos.

[Sethia et al. \(2019\)](#), implementaram um modelo para previsão do preço das ações para o quinto dia seguinte e, consequentemente, fornecer uma estratégia diária de compra/venda para o Índice S&P 500. Os dados de treino e previsão abrangeram mais de 12 anos, de 2000 a 2017. O modelo LSTM provou superar os outros modelos, tais como GRU, SVM e ANN, com um RMSE de 0.0428.

[Lanbouri et al. \(2020\)](#) apresentam um artigo baseado em indicadores técnicos com LSTM para prever o preço do Índice S&P 500 de um minuto, cinco minutos e dez minutos à frente. Tal como os autores acima mencionados, para além do histórico de preços, adicionaram indicadores de análise técnica e estratégia, por esse motivo, o *output* deste modelo mostra que o método proposto é eficaz e adequado para previsão alguns minutos antes, sendo que o RMSE mais baixo foi obtido para os 10 minutos posteriores.

Num estudo recente, conduzido pelos autores [Grupta et al. \(2023\)](#), foram investigados os desafios da previsão precisa do mercado de ações, empregando métodos de inteligência artificial. Estes recorreram a *Multilayer Linear-Regression*, CNN, e LSTM com o objetivo de analisar as tendências de preços ao longo de vários períodos e prever o preço de fecho de cinco empresas de diferentes setores. Para essa análise, foram utilizados dados diários de preços obtidos entre 3 de novembro de 2012 e 3 de novembro de 2022, sendo as variáveis consideradas apenas os preços históricos. Após avaliação, constatou-se que o modelo LSTM demonstrou ser o mais preciso na previsão do preço de fecho das empresas em diferentes comprimentos de sequência.

2.1.3 Métodos Híbridos e Ensemble em Machine Learning

Existem diversos estudos analisando e comparando a performance dos modelos híbridos ou *ensemble* face aos modelos individuais, os resultados analíticos têm evidenciado que estas técnicas combinando dois ou mais modelos têm obtido o melhor desempenho na previsão ([Chou et al., 2013](#)). A diferença entre estes métodos é que, enquanto os métodos *ensemble* trabalham de forma independente para alcançar um resultado, os híbridos trabalham juntos para prever um único resultado. Os modelos que se seguem, confirmam a *performance* mais elevada face aos outros modelos.

[Yun et al. \(2021\)](#) recorreram a um modelo de previsão *ensemble Genetic Algorithm-XGBoost*, utilizando para o efeito o Índice de preços de ações da Coreia do Sul de 2000 a 2016, para prever a direção do movimento para o dia seguinte. Compararam o desempenho do algoritmo elaborado, com outros modelos como *Decision Tree*, *Random Forest*, *SVM* e *KNearest Neighbor*, e concluíram que o modelo híbrido foi o melhor, apresentando uma *accuracy* de 93.28%.

Tal como os autores referidos anteriormente, [Kamara et al. \(2022\)](#) também elaboraram um modelo *ensemble*, mas, neste caso combinando LSTM, com CNN e ainda mecanismos de atenção com o intuito de capturar informações dos preços de ações de longo prazo. Neste estudo, utilizaram o Índice NYSE e o otimizador Adam, tendo obtido um RMSE de 0.029 sendo este inferior a todos os outros modelos utilizados como comparação.

[Worasuchee et al. \(2022\)](#) também propuseram um modelo *ensemble* baseado em XGBoost e LightGBM, para prever a tendência do preço de fecho do segundo dia posterior utilizando para o efeito, preços históricos diários de 20 ações da Bolsa da Tailândia.

[Lin et al. \(2021\)](#) utilizaram uma combinação de modelos utilizando para o efeito LSTM com *Complete Ensemble Empirical Mode Decomposition with Adaptive Noise* (CEEMDAN) para prever o preço do Índice S&P 500, com 10 anos de histórico. Além disso, o artigo utiliza modelos para comparação de resultados, tais como SVM, *Backward Propagation* (BP), rede Elman e *Wavelet Neural Networks* (WAV). Os resultados obtidos evidenciam que o modelo adotado tem boa precisão de previsão e robustez, tendo obtido um RMSE de 20.88, o que pode efetivamente prever a tendência de mudanças de preços e efetivamente evitar perdas de investimento.

Por fim, [Liwei et al. \(2021\)](#) elaboraram um modelo híbrido contendo LSTM e XGBoost baseado em otimização *bayesiana* e, este estudo evidencia que o modelo híbrido tem melhor capacidade de aproximação e generalização na previsão, obtendo uma *accuracy* de 60%. Apesar das conclusões obtidas por estes autores, comparando a performance com outros artigos, este apresenta uma exatidão mais baixa, como por exemplo, relativamente ao [Yun et al. \(2021\)](#).

2.2 Conclusões do Estudo Realizado

Durante a análise dos artigos referentes ao problema da classificação da tendência, constatou-se que alguns autores conseguiram alcançar resultados significativamente superiores em termos de *accuracy* ao incorporar variáveis exógenas nas suas abordagens. De entre esses autores, destacam-se [Ampomah et al \(2020\)](#) e [Wang et al \(2023\)](#), cujos estudos se destacaram por apresentarem um desempenho mais robusto e consistente. Esta abordagem, ao considerar variáveis exógenas no conjunto de dados em análise, revelou-se uma estratégia eficaz na detecção precisa da classificação da tendência. Estes

resultados obtidos, neste sentido, suscitam grande interesse, abrindo caminho para novas investigações.

Já no que concerne ao problema de regressão, com base nos resultados obtidos por [Pokhrel et al. \(2022\)](#), é possível inferir que a inclusão de variáveis exógenas na previsão de preços pode não representar a escolha mais apropriada. A respeito da incorporação de dados macroeconômicos, indicadores técnicos e análise de sentimento, o modelo analisado revelou o segundo maior valor de RMSE. Diante desta constatação, emerge, então, a necessidade de conduzir um estudo suplementar acerca da inclusão de variáveis exógenas na previsão de preços, com o intuito de investigar se, efetivamente, elas contribuem para aprimorar o RMSE.

Diante do evento supramencionado, um dos objetivos da dissertação será realizar uma análise criteriosa das variáveis exógenas e investigar o comportamento dessas variáveis nos problemas de classificação e regressão. Nesse contexto, pretende-se identificar fatores distintos, contextos temporais ou características específicas dos dados nos quais as variáveis exógenas possam desempenhar um papel relevante na melhoria das previsões. Por meio dessa abordagem, visa-se adquirir uma compreensão substancial e fundamentada acerca do valor e do impacto das variáveis exógenas nos problemas de classificação e regressão.

Comparando os modelos estudados pelos investigadores, concluiu-se que o LSTM e XGBoost são os modelos que apresentam melhor performance para previsão e classificação, respetivamente. Além disso, concluiu-se que os modelos híbridos, comparativamente com outros modelos estudados, geraram uma melhor capacidade de aproximação e de generalização. Muitos investigadores acreditam que combinar diferentes modelos ou utilizar modelos híbridos pode ser uma solução eficaz para melhorar a performance e superar algumas limitações ([Khashei et al., 2019](#)).

Além disso, os resultados analisados também revelaram a eficácia dos modelos *ensemble*. O modelo desenvolvido por [Yun et al. \(2021\)](#), demonstra que estes apresentam uma performance mais elevada comparativamente a modelos individuais, tais como os modelos estudados e analisados por [Ampomah et al. \(2020\)](#) e [Park et al. \(2019\)](#). Adicionalmente, [Kamara et al. \(2022\)](#) gerou o segundo valor de RMSE mais baixo face a todos os artigos analisados, sendo que o terceiro RMSE inferior observado ([Lanbouri et al., 2020](#)) é cerca de 15 vezes superior a este.

Atendendo ao exposto, nesta tese serão abordados exclusivamente modelos individuais XGBoost e LSTM, tendo em conta a complexidade inerente e o maior tempo de processamento associados aos modelos híbridos e *ensemble*. O objetivo principal dessa abordagem consiste em avaliar rigorosamente o desempenho desses modelos individuais e analisar o impacto das variáveis exógenas nos dois problemas em estudo.

Além disso, é crucial ressaltar que a seleção do Índice S&P 500 como objeto de análise baseia-se na sua extensa cobertura e reconhecimento na literatura especializada. Dado que este Índice é amplamente estudado, optou-se por incluí-lo na pesquisa com o intuito de viabilizar uma comparação mais abrangente e significativa em relação a estudos anteriores. Dessa maneira, ao empregar o S&P 500 em conjunto com os modelos individuais, procura-se obter perspectivas esclarecedoras sobre a influência das variáveis exógenas nos problemas específicos em análise.

Metodologia

Neste projeto, optou-se por adotar a metodologia CRISP-DM (*Cross – Industry Standard Process for Data Mining*), o que denota a escolha de uma abordagem estruturada e iterativa, como pode ser observado na Figura 4, caracterizando-se como uma prática altamente valorizada no desenvolvimento de projetos de ciência de dados.

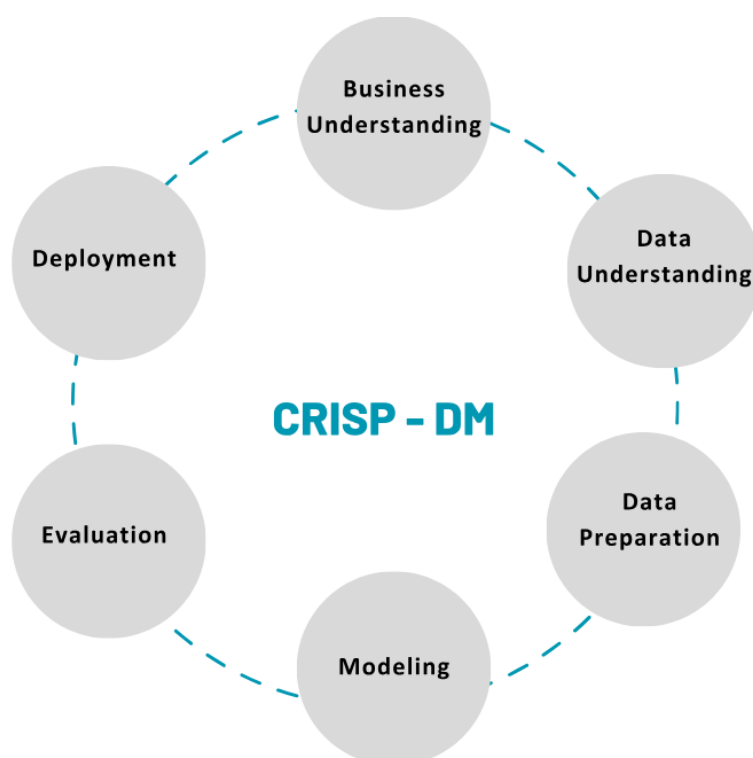


Figura 4 - Metodologia CRISP - DM.

Desse modo, a adesão a esta metodologia proporciona um sólido alicerce para a condução do projeto, enquanto favorece a obtenção de soluções bem-sucedidas e passíveis de aprimoramento, através de iterações coerentes e cuidadosamente orientadas.

3.1 Enquadramento Económico

A prática de *day trading* no mercado financeiro tem-se assumido como uma prática comum há muitos anos. Veja-se que, o seu impacto sobre os mercados financeiros tem sido significativo e, com o avanço tecnológico, o *day trading* tem-se tornado cada vez mais acessível aos investidores individuais, resultando num aumento de adesão a esta abordagem comparativamente ao passado.

Hodiernamente, a procura pelo *day trading* surge como forma de obter ganhos no mercado financeiro. Porém, é importante ressaltar que o *day trading* é uma estratégia em que os clientes devem estar cientes dos riscos envolvidos e estar preparados para lidar com a volatilidade do mercado financeiro. Motivo pelo qual, se impõe, a esta dissertação, dedicar atenção à construção de ferramentas adequadas de modo a gerir o risco financeiro e, conseqüentemente, evitar perdas.

Tecidas estas considerações, cumpre afirmar que a história do mercado financeiro é repleta de eventos que deixaram uma marca indelével na trajetória económica global. Entre esses eventos extremos, destacam-se a bolha da internet, a crise financeira de 2008, a queda dos preços do petróleo e a pandemia global. Cada um desses acontecimentos desafiou os mercados, alterou perspectivas e testou a resiliência das sociedades, com repercussões significativas no panorama económico mundial.

No final do século XX, a bolha da internet trouxe consigo um período de especulação frenética sem precedentes. O surgimento da era digital abriu novas fronteiras, repletas de promessas de inovação e transformação. Investidores foram envolvidos num clima de otimismo exacerbado, impulsionados pela convicção de que empresas de tecnologia e internet seriam a chave para um futuro próspero. No entanto, essa euforia desmedida resultou em avaliações inflacionadas e um crescimento insustentável. Quando a realidade começou a impor-se, a bolha estourou, acarretando uma desvalorização abrupta das ações das empresas de internet e um colapso nos mercados de ações. Tal cenário culminou numa recessão económica que procurou uma reavaliação das perspectivas e uma procura por bases sólidas e sustentáveis para o crescimento económico.

No que concerne à crise financeira de 2008, esta marcou um ponto de virada na história financeira moderna. Originada pela crise das hipotecas *subprime* nos Estados Unidos, essa crise deflagrou uma reação em cadeia global que abalou o sistema financeiro mundial. Instituições financeiras de renome enfrentaram dificuldades, a confiança nos mercados foi quebrada e uma crise de liquidez instaurou-se. O S&P 500, um dos principais indicadores do mercado de ações, sofreu uma queda significativa, atingindo um valor mínimo face ao pânico generalizado e à incerteza predominante. As repercussões dessa crise financeira foram abrangentes, desencadeando uma recessão global e influenciando mudanças regulatórias substanciais visando mitigar riscos sistémicos no futuro.

A queda dos preços do petróleo também desempenhou um papel relevante no panorama económico global. A volatilidade nos mercados de petróleo não apenas afeta a indústria energética, mas também possui implicações significativas para a estabilidade económica global. Entre 2014 e 2016, os preços do petróleo sofreram uma queda acentuada devido a diversos fatores, como o excesso de oferta e a desaceleração da procura global. Essa dinâmica pode ser claramente observada ao analisarmos a Figura 5, apresentada abaixo, onde se evidencia a tendência descendente dos preços do petróleo nesse intervalo de tempo. Tal cenário teve impactos negativos nas empresas do setor

energético, incluindo aquelas listadas no S&P 500, resultando num declínio do Índice e desafios para toda a indústria petrolífera.

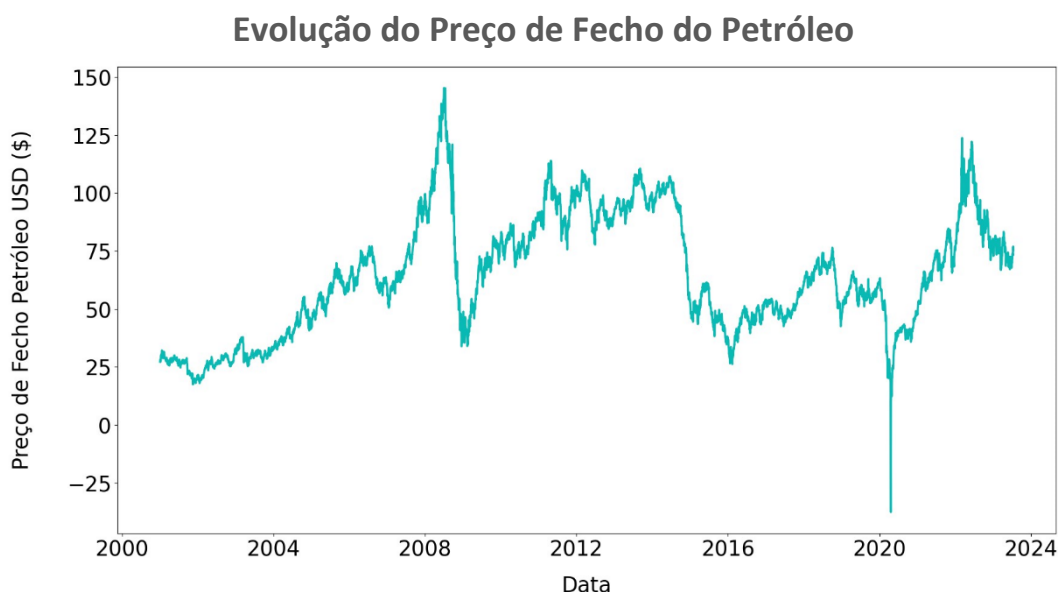


Figura 5 - Evolução do preço de fecho do petróleo, desde o ano 2000 até a atualidade.

A pandemia global, que teve início em 2019 e intensificou-se em 2020, provocou um impacto sem precedentes na economia e nos mercados financeiros. O surgimento do novo coronavírus e a disseminação da doença conhecida como COVID-19 tiveram repercussões significativas em escala global. Restrições à mobilidade, encerramento de empresas e interrupções na cadeia de suprimentos foram adotados como medidas de contenção, afetando gravemente a atividade económica. O S&P 500, assim como outros Índices de mercado, sofreu uma queda acentuada devido à aversão ao risco e à incerteza generalizada. A pandemia desencadeou uma série de consequências económicas, como a contração do PIB, o aumento do desemprego e a volatilidade dos mercados financeiros.

Por fim, além dos eventos mencionados anteriormente, outro fator importante que tem afetado a economia global e o Índice S&P 500 nos últimos tempos é o período de inflação alta que tem sido observado. Cumpre, assim, constatar que, a inflação alta tem consequências significativas para os investidores e empresas listadas no Índice S&P 500, o que gera preocupações e impacta os mercados financeiros, incluindo o Índice S&P 500.

Considerando essas circunstâncias, por meio da análise visual da Figura 6, é possível discernir a magnitude desses eventos históricos e a sua influência abrangente no mercado financeiro. A representação gráfica captura os extremos ascendentes e descendentes do Índice S&P 500 ao longo do tempo, enfatizando os momentos críticos em que esses eventos mencionados ocorreram. Esta ilustração oferece uma perspectiva das oscilações significativas que esses eventos históricos provocaram no panorama financeiro. As fases de expansão económica são vividamente retratadas por

picos ascendentes, enquanto as crises e recessões são graficamente representadas por quedas abruptas e acentuadas.

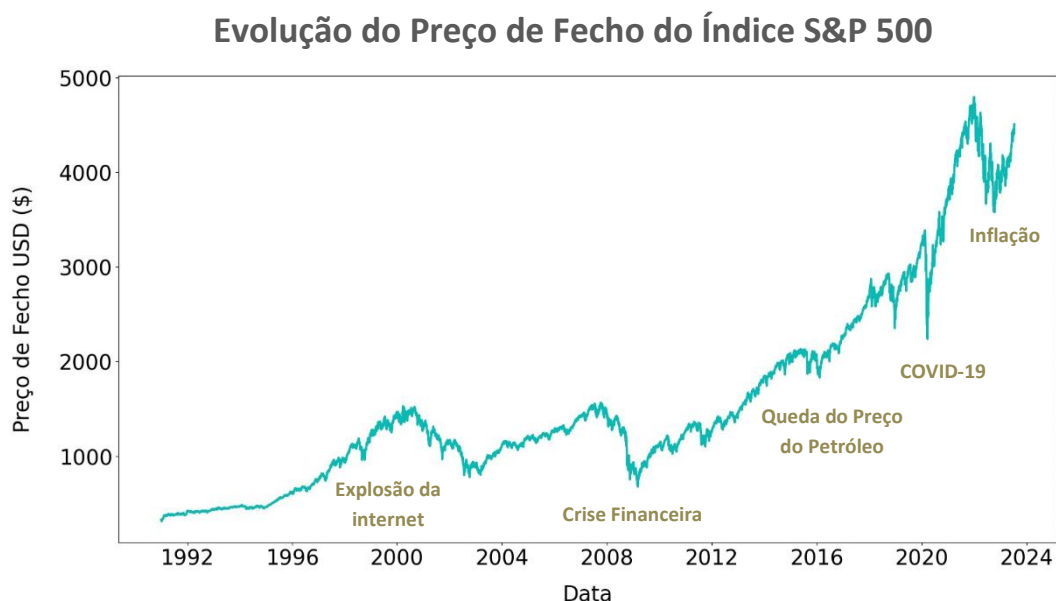


Figura 6 - Evolução do preço de fecho do índice S&P 500, de 1990 até à atualidade.

Mediante esta visualização gráfica, emerge uma compreensão mais profunda da inerente volatilidade que permeia os mercados financeiros e, por conseguinte, ressalta a necessidade premente de estar devidamente preparado para enfrentar os desafios que se manifestam ao longo dessa jornada económica.

Portanto, a Figura 6, proporciona uma perspetiva tangível dos eventos históricos que moldaram o cenário do mercado financeiro. Mais do que um simples esboço gráfico, instiga a adotar uma abordagem informada e prudente ao realizar investimentos, conscientes de que, apesar das incertezas inerentes, os mercados podem-se recuperar e oferecer oportunidades para aqueles que se dispõem a compreender e gerir os riscos inerentes com discernimento.

Além disso, diante das características mencionadas do Índice S&P 500, é importante ressaltar que este apresenta diversas vantagens para investidores que procurem lucros de curto prazo. Sendo certo que o Índice S&P 500 é composto por 500 das maiores empresas listadas na bolsa de valores dos EUA, significa que este é altamente líquido permitindo, deste modo, aos *traders* entrar e sair das suas posições facilmente. Por outro lado, considerando a alta volatilidade deste Índice e o acesso à alavancagem, poderá, então, surgir oportunidades de lucro significativas caso se consiga antecipar as mudanças de preço do mercado.

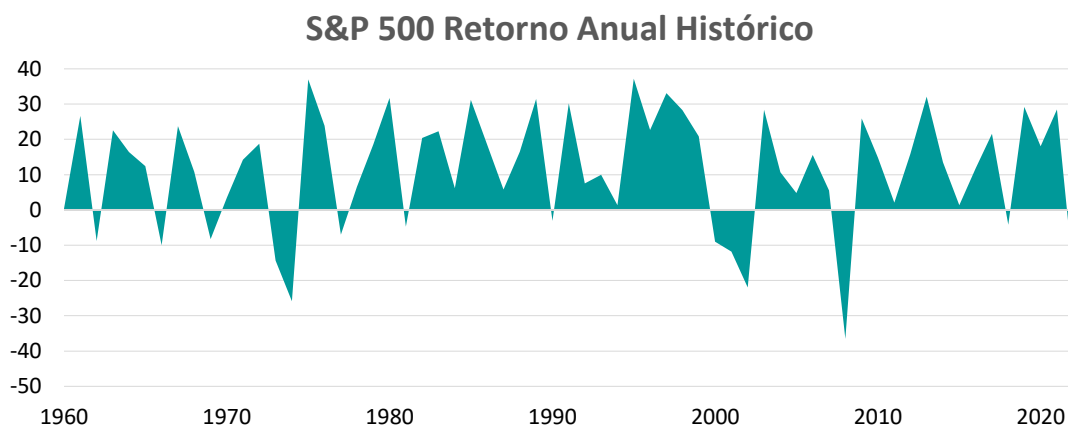


Figura 7 - S&P 500 Retorno Anual Histórico.
Fonte: Aswath Damodaran, NYU Stern School of Business

No que concerne ao Índice S&P 500, cumpre ainda notar que, ao longo do último século, o retorno anual médio (que engloba dividendos) foi de aproximadamente 10%, sem considerar a inflação. De facto, este resultado positivo anual pode ser uma vantagem para os *traders* diários pois potencialmente propicia ganhos a longo prazo. Por exemplo, um *day trader* que alcance uma média diária de lucro de 1% pode gerar um retorno anual superior a 250% se negociar em todos os dias úteis do ano, *ceteris paribus*.

Entende-se que o desempenho histórico do Índice S&P 500 não é uma garantia de desempenho futuro. Investir no Índice S&P 500 é inerentemente sujeito a riscos e incertezas, e é de extrema importância realizar uma análise cuidadosa dos riscos e oportunidades de investimento.

Reconhecendo esta problemática, torna-se imprescindível a implementação de algoritmos que auxiliem no processo de tomada de decisão, possibilitando que os investidores possam reagir com rapidez às mudanças no mercado. Essa necessidade é ainda mais relevante no caso do Índice S&P 500, dada a sua alta sensibilidade às variações nas condições económicas e políticas.

Nesse contexto, o objetivo do presente trabalho visa analisar essa problemática e auxiliar os *day traders*. A proposta consiste em desenvolver um modelo capaz de classificar a tendência do Índice S&P 500, bem como determinar o preço de fecho. Esta abordagem pode proporcionar alguns benefícios aos investidores.

No que concerne ao problema da tendência, o objetivo reside em determinar a direção diária do Índice S&P 500, para o dia seguinte, através da realização de previsões que estimem, de forma precisa, a ocorrência de uma tendência ascendente ou descendente. Especificamente, procura-se identificar se o preço de fecho do referido Índice registou uma variação positiva ou negativa em relação ao dia anterior. Tal abordagem proporciona aos investidores uma perspetiva sobre a evolução do mercado, subsidiando as suas decisões de negociação. Por outro lado, no contexto da regressão, almeja-se

estimar o valor absoluto do preço de fecho do Índice S&P 500 ao término do dia de negociação seguinte, conferindo-lhe, dessa maneira, uma medida precisa e quantitativa do desempenho do Índice nesse intervalo temporal.

A implementação bem-sucedida desse modelo acarretaria diversas vantagens para os *day traders* e demais investidores do S&P 500. Primeiramente, eles teriam acesso a informações preditivas sobre a tendência do Índice, possibilitando antecipar movimentos de mercado e ajustar as suas estratégias de investimento de forma fundamentada. Além disso, a obtenção do preço de fecho do Índice S&P 500 também é uma informação crucial para os investidores. Esse valor é utilizado como referência para calcular o desempenho diário do mercado, avaliar ganhos ou perdas e elaborar estratégias de investimento. Como resultado, seriam potencialmente alcançados melhores retornos financeiros e redução de riscos associados.

À luz da premissa supramencionada, foi determinado que uma análise cuidadosa de variáveis exógenas é crucial para compreender as flutuações do Índice S&P 500. A inclusão dessas variáveis pode, portanto, propiciar uma melhor compreensão das variações do Índice S&P 500, bem como uma previsão mais precisa do seu desempenho futuro. Em decorrência dessa premissa, a análise de variáveis exógenas, tais como o Índice VIX, o preço do petróleo e o Índice Dow Jones *Industrial Average*, foi motivada pelo reconhecimento de que tais variáveis possuem potencial para desempenhar um papel relevante na explicação das flutuações do referido Índice.

3.2 Variáveis Exógenas

Nesta secção, serão tratadas e analisadas as variáveis exógenas que se revelaram relevantes para o Índice S&P 500. Tais variáveis são consideradas fatores externos ao mercado financeiro capazes de exercer influência sobre os movimentos do referido Índice, o que possibilita uma compreensão mais precisa do seu comportamento.

Com a introdução do Índice VIX, conhecido como o "Índice do medo", surge um potencial indicador útil para a volatilidade implícita do mercado de ações dos Estados Unidos nos próximos 30 dias. Valores elevados do VIX pode indicar que os investidores estão preocupados com possíveis declínios no mercado de ações, levando-os a adquirir opções de venda como uma forma de proteção. Nesse sentido, o aumento do VIX pode servir como um alerta para os investidores, possivelmente associado a quedas subsequentes no Índice S&P 500.

É importante ressaltar que o VIX é apenas um indicador e não deve ser considerado uma previsão infalível. A volatilidade do mercado de ações é influenciada por diversos fatores e pode ocorrer em resposta a eventos imprevistos. Portanto, embora o aumento do VIX possa estar associado a quedas subsequentes do Índice S&P 500, é essencial que os investidores realizem uma análise abrangente e considerem outras informações relevantes antes de tomar decisões de investimento. Por esse motivo, foram analisadas outras variáveis exógenas que poderiam exercer influência no Índice S&P 500. Ao reconhecer que a volatilidade do mercado de ações é determinada por uma ampla gama de fatores, é fundamental considerar diversas variáveis para uma análise abrangente e precisa.

A análise da Figura 8 revela a existência de uma correlação negativa entre o Índice VIX e o Índice S&P 500, o que pode ser corroborado através da Figura 10, localizada na página 28. Essa correlação negativa indica que, de modo geral, quando o Índice VIX apresenta picos significativos, indicando maior volatilidade implícita e ansiedade entre os investidores, é mais provável que ocorram subsequentes quedas no Índice S&P 500.

A Figura 8, que consiste numa representação gráfica adicional, proporciona uma visão mais detalhada da relação entre o Índice VIX e o Índice S&P 500. Ao observarmos atentamente essa figura, podemos identificar períodos nos quais o Índice VIX atinge os seus valores mais elevados, ao passo que o Índice S&P 500 apresenta um declínio correspondente. Tal consistência na direção oposta dos movimentos sugere uma correlação negativa entre essas duas variáveis.

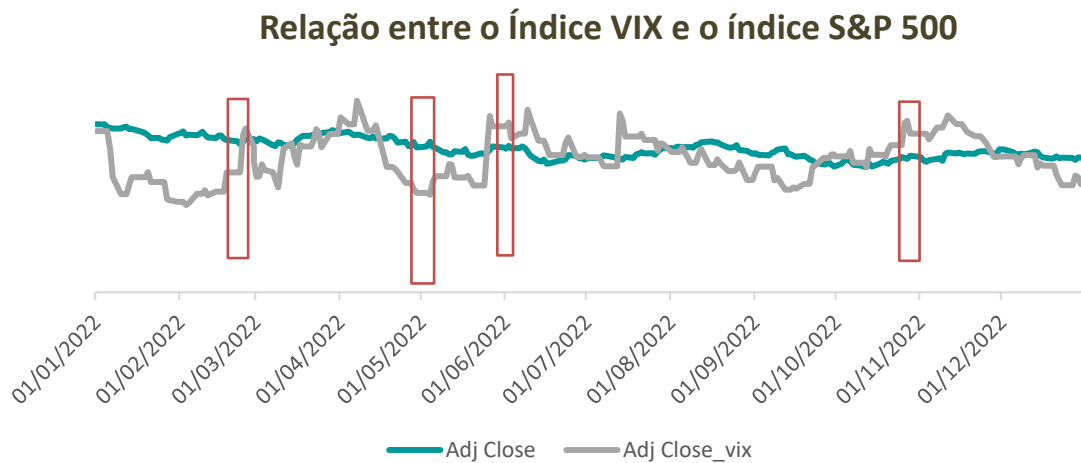


Figura 8 - Relação entre o índice VIX e o índice S&P 500, no ano de 2022.

Torna-se imperativo para uma compreensão aprofundada das oscilações do Índice S&P 500 realizar uma análise da sua composição empresarial, considerando que o Índice é composto por 11 setores distintos, tal como é possível observar pela Figura 9. É importante considerar a influência do preço do petróleo sobre as variações do Índice, tendo em vista que o setor de energia é diretamente afetado pelo preço do petróleo, o que pode impactar os custos de produção, as margens de lucro e, por conseguinte, o valor das ações.

Pesos dos diferentes Setores no Índice S&P 500

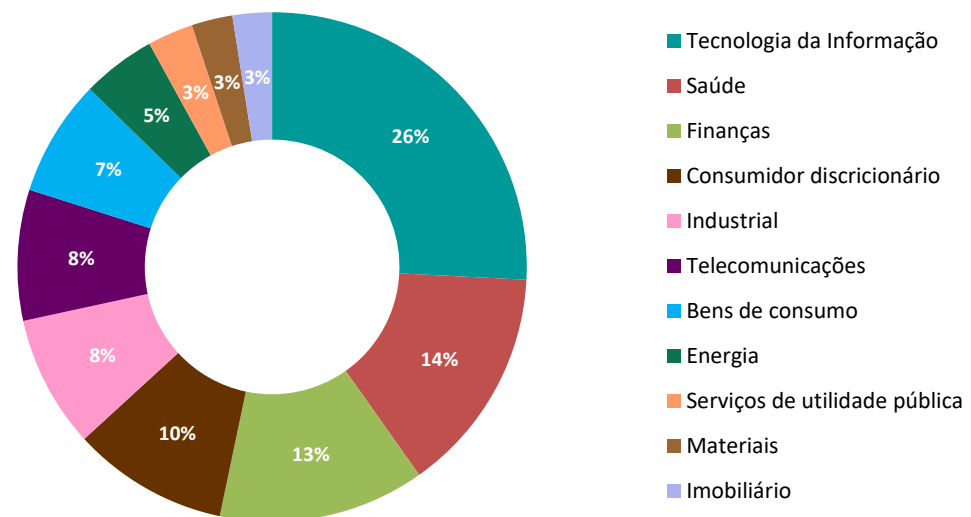


Figura 9 - Pesos de cada setor do Índice S&P 500.

Cumprе ressaltar que, embora o setor energético represente apenas 5% do peso total do Índice, empresas de diversos setores também podem ser indiretamente afetadas pelo preço do petróleo. Isso decorre do aumento dos custos de produção causado pelo aumento de preço do petróleo, o que pode reduzir as margens de lucro dessas empresas, assim como o preço dos seus produtos. Como resultado, tal situação pode ter um efeito negativo nas suas ações e, dessa forma, afetar a performance geral do Índice S&P 500.

Igualmente passível de impactar o desempenho do Índice S&P 500, afigura-se a introdução do Índice Dow Jones, uma vez que este último é frequentemente considerado um reflexo da saúde em geral do mercado de ações dos Estados Unidos. O Dow Jones é composto pelas 30 empresas líderes em diversos setores da economia dos Estados Unidos, o que o torna um indicador importante para avaliar o desempenho do mercado de ações. É digno de notar que, a movimentação desses dois Índices tende a ocorrer em direções similares na maioria das vezes, o que aumenta a relevância do Dow Jones como um importante indicador do mercado de ações e fornecer *insights* sobre possíveis mudanças no S&P 500.

Por derradeiro, em complemento, faz-se salutar mencionar que a análise técnica figura como uma das principais ferramentas empregues pelos investidores no que se refere à tomada de decisões de investimento e à previsão do comportamento dos ativos financeiros. De entre as diversas técnicas utilizadas, a inserção de variáveis como o RSI (*Relative Strength Index*) e o MACD (*Moving Average Convergence-Divergence*) pode ser considerada uma estratégia eficiente para antever os movimentos do Índice S&P 500 e, conseqüentemente, tomar decisões informadas.

Após a identificação das fontes de dados relevantes e análise do seu conteúdo, é necessário dar início à etapa de preparação dos dados, visando transformá-los e adequá-los ao problema em questão.

Tendo como premissa a intenção de fundamentar o presente estudo com base em séries temporais, almeja-se explorar e analisar o comportamento de um conjunto de variáveis ao longo do tempo. A seleção desta abordagem fundamenta-se no facto, de que as séries temporais apresentam a habilidade de proporcionar informações relevantes acerca do comportamento dos dados, permitindo previsões e decisões com maior grau de precisão. Cumprе destacar que a série temporal objeto do presente estudo está intrinsecamente relacionada ao contexto financeiro, sendo que a mesma se encontra disponível na plataforma *Yahoo Finance*. Nesse sentido, a decisão de utilizar a *Yahoo Finance* como fonte de dados configura-se como uma escolha de alta qualidade e confiabilidade, garantindo que as análises e previsões derivadas da mesma sejam fundamentadas em informações precisas e atualizadas.

No âmbito da definição do conjunto de dados a ser analisado, é imprescindível mencionar a criação de múltiplas variáveis derivadas das variáveis originais. Essas variáveis incluem, por exemplo, as variações de preço em relação ao dia anterior, bem como variáveis históricas que capturam o valor observado há quatro e dez dias atrás. A inclusão dessas variáveis adicionais é justificada pela razão de que, em séries temporais financeiras, o uso de informações históricas e outras informações relevantes pode ser uma estratégia eficiente para melhorar a precisão do desempenho do Índice S&P 500 e, assim, maximizar o valor do investimento.

Neste contexto, a introdução de variáveis que contêm informações sobre o comportamento passado pode ser útil na identificação de padrões e tendências relevantes que possam influenciar o desempenho futuro. Além disso, a utilização de variáveis que refletem as variações de preço em relação ao dia anterior pode contribuir para capturar a dinâmica de curto prazo do mercado, bem como a resposta dos investidores a eventos e notícias recentes. Esse aspecto é especialmente relevante em momentos de alta volatilidade, quando a capacidade de reagir rapidamente aos acontecimentos pode ser fundamental para tomar decisões de investimento informadas.

3.3 Pré-Processamento dos Dados

Na metodologia CRISP-DM é fundamental garantir a integridade e consistência dos dados utilizados, bem como a sua adequação aos objetivos e hipóteses em questão. Para tanto, é necessário realizar a limpeza e pré-processamento dos dados, a fim de remover inconsistências, preencher valores omissos e tratar outras possíveis irregularidades.

No âmbito do conjunto de dados do Índice VIX, foi identificada a existência de valores omissos decorrentes da falta de dados ou incompatibilidades nas datas após a realização de uma transformação da data para obter a volatilidade esperada para o respectivo dia. Para obter a volatilidade prevista para um determinado dia, a data foi modificada antecipando 30 dias e, em seguida, aplicou-se o método de imputação *forward fill* para preencher os valores ausentes com os valores anteriores disponíveis na mesma coluna. Tal decisão foi tomada com base na suposição de que os valores de volatilidade esperada para um determinado dia estão diretamente relacionados aos valores do dia anterior e que a aplicação do método *forward fill* não comprometeria a integridade dos dados. Ademais, o uso deste método garante a inexistência de lacunas nos dados, o que pode ser de grande importância para análises posteriores.

Com o intuito de viabilizar a análise comparativa e integrada dos dados, efetuou-se uma operação de redimensionamento no *dataset* final, o qual contém a junção dos quatro conjuntos de dados: Índice S&P 500, preço do petróleo, VIX e *Índice* Dow Jones. Este procedimento consistiu em ajustar todos os conjuntos de dados para que se iniciassem a partir de setembro de 2000, momento em que os registros do preço do petróleo passaram a ser disponibilizados. Por meio deste alinhamento temporal, tornou-se possível realizar análises integradas e coerentes, englobando, assim, informações relevantes e disponíveis a partir da mesma data de referência.

Finda a etapa anterior, a qual consistiu no redimensionamento do conjunto de dados, resultou num conjunto final contendo 5757 linhas e 83 variáveis, cobrindo o período que se estende desde setembro de 2000 até abril de 2023. É importante destacar que os dados possuem uma frequência diária, ou seja, as informações são registradas diariamente ao longo desse intervalo temporal. Este conjunto de dados é adequado para realizar uma análise ampla e detalhada das variáveis que afetam a volatilidade do Índice S&P 500. Com o número de variáveis disponíveis, é possível investigar as principais variáveis que afetam as oscilações do Índice, permitindo uma melhor compreensão dos fatores subjacentes que impulsionam os movimentos do mercado financeiro.

3.4 Incorporação de Variáveis na Modelação

Para fins de redução de dimensionalidade no conjunto de dados, foram utilizados dois algoritmos para selecionar as variáveis mais relevantes: ANOVA e LASSO¹. O objetivo deste passo foi minimizar o impacto de características irrelevantes ou ruidosas, assegurando, assim, a inclusão apenas de características informativas na análise. Para tanto, foram realizados dois processos distintos de seleção de características, tendo em vista a existência de dois problemas diferentes no estudo em questão, um de classificação e outro de regressão.

¹ Tanto a análise da variância (ANOVA) como o método LASSO são amplamente adotados como abordagens populares para a seleção de variáveis em tarefas de previsão e são implementados como funções na biblioteca scikit-learn. Para obter informações mais detalhadas sobre estes métodos, recomenda-se a leitura dos seguintes documentos:

https://scikitlearn.org/stable/modules/feature_selection.html

https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html

As Figuras 11 e 12, apresentadas nas páginas 29 e 31, respectivamente, ilustram o processo de seleção de variáveis, em que as variáveis selecionadas foram identificadas como importantes em ambos os métodos utilizados. Tal abordagem de seleção de recursos consiste em considerar as características destacadas pelo método LASSO e pelo método ANOVA, com o objetivo de identificar aquelas que mais contribuem para o desempenho do modelo em análise. Esta estratégia de seleção de variáveis permite obter uma seleção mais refinada de características relevantes para o modelo, tendo em vista que ambos os métodos adotam abordagens complementares para a identificação de características importantes. Com a combinação dos métodos, é possível selecionar apenas as variáveis que apresentam maior impacto no desempenho do modelo, o que contribui para uma melhor interpretação dos resultados obtidos e para uma maior confiabilidade do modelo em relação à sua capacidade de generalização.

3.4.1 Exploração e Seleção de Variáveis para o Problema de Classificação

No tocante ao processo de seleção de *features* para o problema de classificação, a variável "variation_dow" foi selecionada como uma das características mais relevantes para prever a tendência do Índice S&P 500, tal como é possível observar pela Figura 11. Esta variável representa a variação percentual do preço do Índice Dow Jones no dia atual em relação ao dia anterior.

O Índice Dow Jones é amplamente considerado um indicador líder do mercado de ações americano e, portanto, a variação percentual do preço desse Índice pode fornecer informações úteis sobre as expectativas do mercado de ações e, por conseguinte, ajudar a prever a tendência do Índice S&P 500. Ademais, dado que os mercados financeiros são altamente interligados, a variação do Índice Dow Jones pode influenciar a tendência do Índice S&P 500.

A relação existente entre o preço do petróleo e a previsão da tendência do Índice S&P 500 está profundamente conectada a diversos fatores macroeconômicos e geopolíticos que afetam tanto o mercado de petróleo quanto o mercado de ações.

Em primeiro lugar, é imprescindível enfatizar que o petróleo é um recurso energético crucial para a economia global, e o seu preço tem um impacto substancial nos mercados financeiros. O petróleo é utilizado na produção de combustíveis para veículos, no aquecimento de residências e fábricas, além de ser uma importante fonte de matéria-prima para a produção de diversos produtos químicos e plásticos.

O aumento do preço do petróleo pode ter um efeito negativo sobre a economia global, uma vez que eleva os custos de produção e transporte de bens e serviços, o que pode levar a um aumento da inflação e a uma desaceleração do crescimento económico. Por outro lado, a queda do preço do petróleo pode ser positiva para a economia, uma vez que reduz os custos de produção e transporte, deixando mais recursos disponíveis para outras atividades económicas.

No que se refere ao mercado de ações, as empresas que produzem e vendem petróleo e gás natural, denominadas "empresas de energia", possuem uma forte influência no desempenho do Índice S&P 500. Grandes multinacionais, como a ExxonMobil e Chevron fazem parte desse setor. Quando o preço do petróleo está elevado, as empresas de energia geralmente apresentam um bom desempenho no mercado de ações, uma vez que os seus lucros e receitas aumentam. Por outro lado, quando o preço do petróleo está baixo, essas empresas tendem a ter um desempenho inferior, uma vez que os seus lucros e receitas são impactados.

Desse modo, o preço do petróleo pode ser considerado uma das variáveis mais relevantes para a previsão da tendência do Índice S&P 500, uma vez que reflete tanto a oferta e a procura de petróleo quando as condições macroeconómicas e geopolíticas afetam a economia global. Quando o preço do petróleo está elevado, pode ser um sinal de que a economia está em crescimento e que as empresas de energia terão um bom desempenho no mercado de ações. No entanto, se o preço do petróleo estiver excessivamente elevado, pode indicar uma pressão inflacionária que pode prejudicar a economia global e, consequentemente, o desempenho do Índice S&P 500.

Veja-se que, apesar da análise por meio da matriz de correlações não revelar uma correlação significativa entre a variável alvo e o preço do petróleo, os métodos utilizados para selecionar as variáveis mais relevantes na classificação da tendência indicaram que o preço do petróleo apresenta um dos maiores poderes explicativos. Nesse contexto, a seleção de variáveis com base em métodos sofisticados forneceu uma visão mais completa e precisa sobre as variáveis que têm maior influência na previsão da variável alvo, possibilitando a identificação de padrões mais complexos e relevantes para o problema em questão.

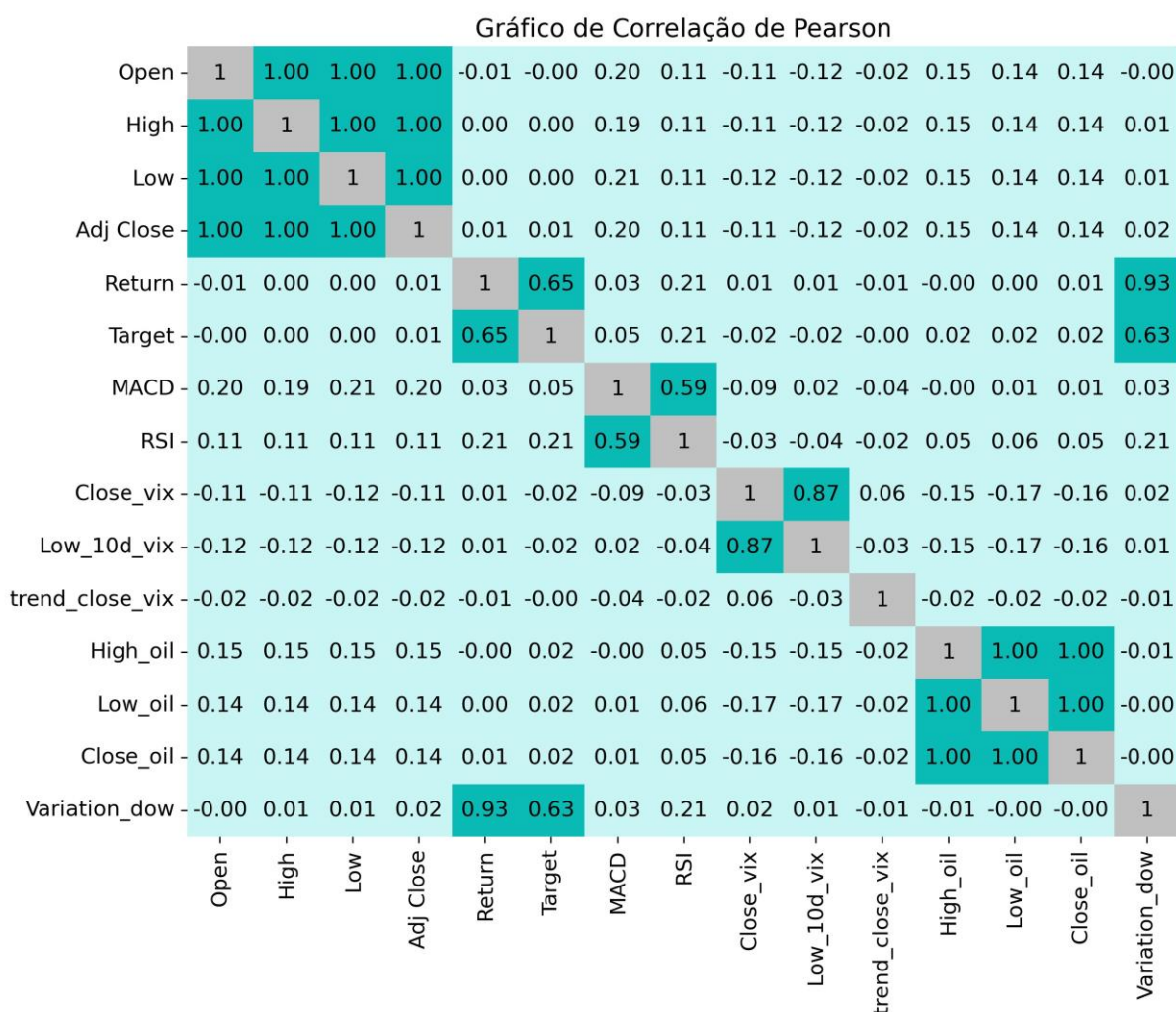


Figura 10 – Matriz de correlação das variáveis mais relevantes.

No que concerne a esta análise, cumpre notar que a variável "Return" é também amplamente reconhecida como uma das características mais pertinentes na previsão da tendência do Índice S&P 500. Em primeiro lugar, o cálculo da diferença percentual entre os preços de fecho e de abertura do S&P 500 proporciona um indicador direto do desempenho diário do mercado de ações dos Estados Unidos. Posto isto, o retorno é frequentemente utilizado como um indicativo da atividade económica geral e das tendências do mercado. Ademais, investidores e analistas monitorizam de perto o retorno do Índice, uma vez que ele reflete a confiança e as expectativas dos participantes do mercado.

Por seu turno, o retorno do S&P 500 pode ser influenciado por fatores macroeconómicos que impactam a confiança dos investidores. Por exemplo, notícias positivas sobre empregabilidade ou um aumento nos lucros corporativos podem impulsionar o retorno do Índice, indicando otimismo e confiança na economia. Por outro lado, eventos negativos, como incertezas políticas ou preocupações com uma possível recessão, podem resultar em retornos negativos, refletindo falta de confiança e aversão ao risco por parte dos investidores.

Features Selecionadas - Tendência

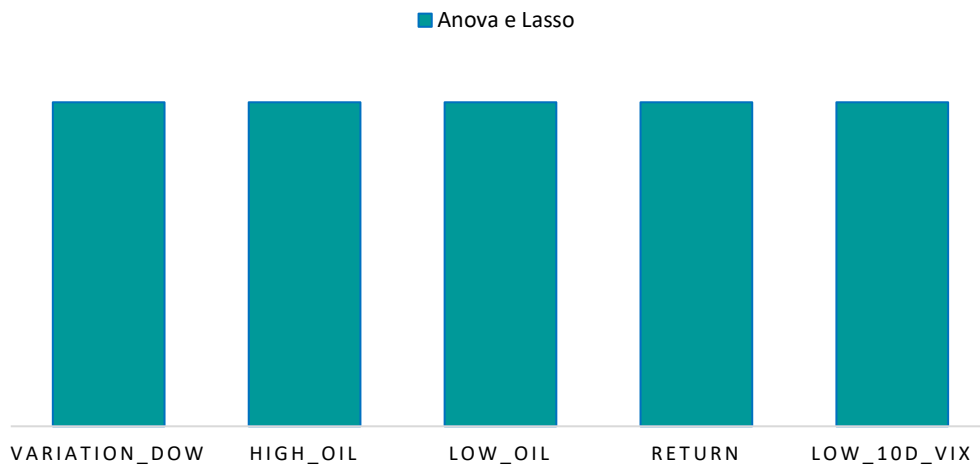


Figura 11 - Features selecionadas para o problema de classificação.

Daqui se extrai que, o retorno do S&P 500 fornece informações valiosas sobre as tendências de curto prazo do mercado. Investidores e *traders* utilizam essas informações para identificar padrões e sinais de compra ou venda. Por exemplo, um retorno positivo substancial pode indicar uma tendência de alta, enquanto um retorno negativo pode sugerir uma tendência de baixa. Compreender as tendências passadas auxilia na tomada de decisões informadas sobre investimentos futuros.

Por fim, não se pode deixar de notar que, o retorno do S&P 500 desempenha um papel relevante na gestão de risco. Ao calcular a variação diária dos preços de abertura e fecho, é possível obter informações sobre a volatilidade do mercado. Essas informações podem ser usadas para estabelecer níveis adequados de *stop loss*, que são pontos de preço pré-determinados nos quais um investidor decide vender um ativo para limitar as perdas. Portanto, o retorno do Índice é um componente essencial na definição de estratégias de *stop loss*.

Ao concluir esta seção que aborda as variáveis consideradas relevantes para determinar a tendência do Índice S&P 500, é importante salientar que a última *feature* considerada significativa para explicar essa tendência é a variável que avalia o preço mínimo do Índice VIX nos últimos dez dias.

A volatilidade implícita, representada pelo Índice VIX, é comumente reconhecida como um indicador de temor ou incerteza nos mercados financeiros. Quando o VIX apresenta um valor elevado, isso indica que os investidores estão mais preocupados com a volatilidade e tendem a adotar uma postura defensiva, reduzindo o seu apetite por riscos. A relação entre a variável "low_10d_vix" e a tendência do Índice S&P 500 pode ser explicada da seguinte maneira: se o preço mínimo do Índice VIX nos últimos dez dias for elevado, indicando um maior nível de volatilidade e incerteza, isso pode afetar negativamente a confiança dos investidores no mercado. Essa falta de confiança pode levar a uma redução na procura por ações, resultando em quedas no Índice S&P 500. Por outro lado, se o preço mínimo do Índice VIX nos últimos dez dias for relativamente baixo, indicando menor volatilidade e maior estabilidade no mercado, os investidores podem sentir-se mais confiantes e propensos a assumir riscos. Isso pode impulsionar a procura por ações e, potencialmente, resultar numa tendência de alta no Índice S&P 500.

3.4.2 Seleção de Variáveis para o Problema de Regressão

Ao abordar o problema de regressão para prever o preço do Índice S&P 500, realizou-se uma análise para determinar as variáveis relevantes. Ao contrário do problema de classificação, neste caso, as variáveis endógenas, ou seja, aquelas diretamente relacionadas ao Índice em si, são consideradas as mais importantes. Essa relevância decorre do facto de que o preço do Índice é influenciado pelas flutuações e interações internas dos ativos que o compõem. Portanto, as informações sobre componentes do Índice, como os preços das ações individuais, volumes de negociação, volatilidade histórica e outros indicadores correlacionados, fornecem uma base sólida para estimar o preço futuro do Índice S&P 500.

É importante ressaltar que a relevância das variáveis exógenas e endógenas pode variar dependendo do objetivo específico da análise financeira. Por exemplo, em determinadas situações, pode ser necessário considerar tanto as variáveis exógenas quanto as endógenas para obter uma visão abrangente do mercado e prever com precisão o preço do Índice. A seleção adequada das variáveis, considerando o contexto e o objetivo da análise, é fundamental para desenvolver modelos de previsão mais eficazes e relevantes para o Índice S&P 500.

Features Seleccionadas - Preço

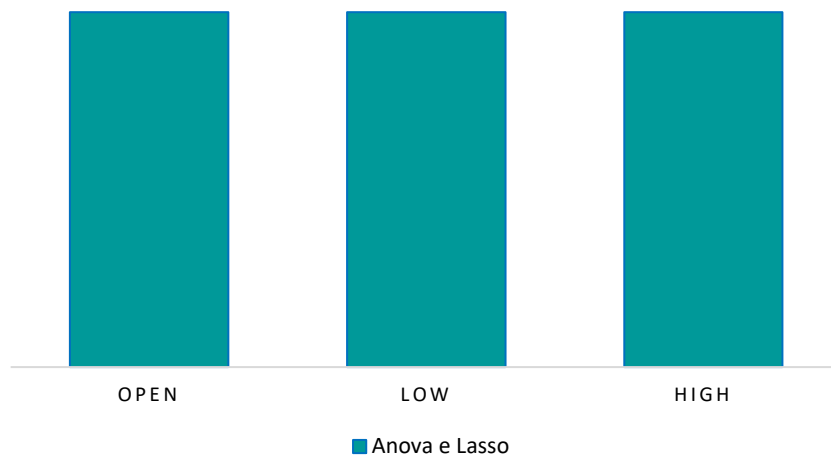


Figura 12 - Features seleccionadas para o problema de regressão.

Ao concluir esta secção que aborda as variáveis consideradas relevantes para determinar o preço de fecho do Índice S&P 500, após minuciosa análise estatística, constatou-se que as variáveis endógenas emergiram como as mais importantes para explicar a variabilidade dos dados.

3.5 Modelação

Antes de iniciar a etapa de modelação, é pertinente abordar algumas considerações acerca do não tratamento de valores extremos. Os dados financeiros, nomeadamente os relacionados aos Índices de mercado como o Índice S&P 500, são conhecidos pela sua inerente volatilidade e pela presença de movimentos abruptos e eventos extremos.

Ora, tais movimentos bruscos podem ser desencadeados por eventos económicos, políticos ou corporativos de grande relevância, tais como divulgações de resultados financeiros, alterações nas políticas governamentais, crises financeiras ou eventos geopolíticos. Esses eventos têm o potencial de gerar alterações súbitas nos preços das ações e, consequentemente, no Índice S&P 500.

Nesse contexto, os valores discrepantes nos dados financeiros, como pontos extremos nos preços do Índice S&P 500, são considerados parte intrínseca da variabilidade normal do mercado. Tais pontos podem representar momentos de extrema volatilidade ou eventos excepcionais que afetam de maneira significativa o comportamento do referido Índice.

Dessa forma, ao realizar análises e previsões relacionadas com o Índice S&P 500, é esperado que a presença de valores extremos seja observada, os quais não precisam necessariamente de ser tratados como um problema a ser corrigido. Em vez disso, esses pontos extremos podem conter informações de alta relevância acerca de eventos significativos que impactam o mercado de ações.

É certo que, antes de avançar para a fase de modelação dos dados, é fundamental estabelecer a adoção de um procedimento de divisão do conjunto de dados em treino, teste e validação. No caso do problema de classificação, a divisão foi realizada entre conjuntos de treino e teste, seguindo uma proporção fixa de 80% para treino e 20% para teste. No entanto, para o problema de regressão, optou-se por particionar o conjunto de dados em conjuntos de treino, teste e validação, com uma divisão proporcional de 80%, 10% e 10%, respetivamente. Ao dividir os dados dessa maneira, procura-se garantir uma avaliação robusta do modelo num conjunto de dados não utilizado durante o treino, proporcionando uma estimativa realista da sua capacidade de generalização e desempenho em cenários práticos.

3.5.1 Construção do Modelo para o Problema de Classificação

No âmbito da resolução do problema de classificação da tendência do Índice S&P 500, foi realizado um estudo aprofundado com a finalidade de otimizar o desempenho do modelo XGBoost implementado. É importante destacar que XGBoost é um modelo *ensemble*, que combina várias árvores de decisão para realizar as previsões de classificação.

A variável alvo determinada para o problema de classificação é uma variável binária, assumindo o valor um quando o preço de fecho do Índice S&P 500 aumenta em relação ao dia anterior e o valor zero quando o preço diminui face ao dia anterior.

Com o propósito de alcançar a meta de otimização do modelo XGBoost, foram empregues estratégias sofisticadas, como o ajuste de hiperparâmetros, para encontrar a configuração ideal do modelo. Além disso, foram realizadas experiências sistemáticas, que envolveram a inclusão e exclusão de variáveis, bem como a análise do impacto dessas alterações no desempenho preditivo do modelo. Todos esses esforços foram direcionados para obter um modelo XGBoost altamente eficiente e preciso na classificação das tendências do Índice S&P 500. A abordagem metodológica seguida foi detalhadamente delineada na Figura 13.

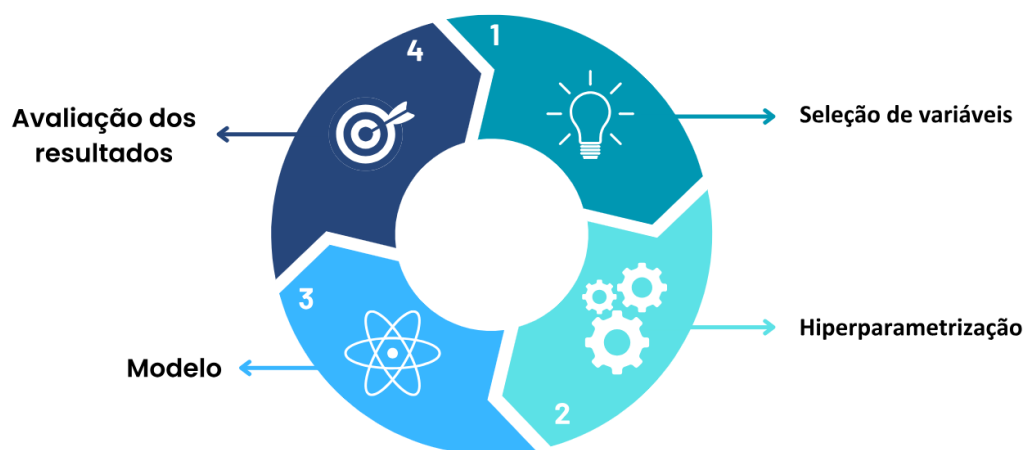


Figura 13 - Metodologia adotada para o desenvolvimento do problema de classificação.

No que se refere às técnicas de Hiperparametrização, inicialmente recorreu-se ao método *grid search*, cujo propósito é identificar a combinação ótima de hiperparâmetros² por meio de uma avaliação exaustiva de todas as possíveis combinações. Entretanto, ao introduzir o parâmetro de regularização contra *overfitting* denominado *reg_alpha*, constatou-se que tal inclusão provocou uma significativa degradação no desempenho computacional do processo. Diante dessa limitação, optou-se por prosseguir com o método *random search*. Esta estratégia permitiu explorar de forma mais eficiente o espaço de hiperparâmetros, selecionando combinações aleatórias para avaliação. Esta mudança possibilitou prosseguir com procura pela configuração ótima dos hiperparâmetros, enquanto se manteve um tempo de execução razoável. Veja-se que Bergstra e Bengio, em 2012, comprovam a igual eficiência do *random search* em comparação com o *grid search*, demonstrando empiricamente que o *random search* pode alcançar resultados semelhantes ou superiores, enquanto ainda reduz o custo computacional.

Revela-se, então, imprescindível levar em consideração múltiplos hiperparâmetros a fim de otimizar o desempenho do algoritmo XGBoost. De entre desses hiperparâmetros, alguns destacam-se devido à sua relevância para o processo de ajuste.

² De modo a obter uma listagem minuciosa de todos os hiperparâmetros do XGBoost Classifier, solicita-se que faça uma leitura ao seguinte documento: <https://xgboost.readthedocs.io/en/stable/parameter.html>. Adicionalmente, caso deseje obter informações mais detalhadas sobre as funções de otimização de parâmetros Random Search e Grid Search da biblioteca Scikit-Learn, recomenda-se a leitura dos seguintes documentos: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html Com o intuito de aprofundar o entendimento acerca de qualquer conceito teórico relacionado ao campo de *deep learning*, é cordialmente sugerida a consulta da obra intitulada “Deep Learning with Python”, de autoria de François Chollet.

Ora, um desses hiperparâmetros é a *Learning Rate*, que governa a velocidade com que o modelo adquire conhecimento durante o processo de treino. Uma taxa de aprendizagem reduzida resulta numa aprendizagem mais lenta, proporcionando um refinamento mais cauteloso dos pesos relativos aos exemplos de treino. Por outro lado, uma taxa de aprendizagem mais elevada pode conduzir a uma aprendizagem mais rápida, embora com maior propensão a saltos excessivos nos pesos.

Outro hiperparâmetro crucial é o *Min Child Weight*, o qual regula a complexidade do modelo. Tal parâmetro estabelece o peso mínimo necessário para a criação de um novo nó na árvore. Valores mais elevados desse hiperparâmetro incentivam a formação de nós contendo um número maior de amostras, o que pode reduzir a complexidade do modelo e prevenir o fenómeno de *overfitting*.

O hiperparâmetro *Gamma* desempenha um papel significativo na poda das árvores durante o processo de treino. Este indica a redução mínima exigida na perda de treino para que uma nova divisão num nó seja considerada. Valores mais altos de *Gamma* conduzem a um maior número de podas na árvore, tornando o modelo mais conservador na sua estrutura.

Adicionalmente, o *Max Depth* é um hiperparâmetro que limita o número de camadas presentes numa árvore de decisão. Ao controlar a profundidade da árvore, é possível administrar a complexidade do modelo e mitigar a tendência ao *overfitting*. Contudo, estabelecer uma profundidade excessivamente baixa pode conduzir a *underfitting*, enquanto uma profundidade excessivamente alta pode propiciar o surgimento do *overfitting*.

O Número de Estimadores é um outro hiperparâmetro relevante, indicando a quantidade de estimadores (árvores) a serem criados durante o processo de treino. Cada estimador contribui para a decisão final do modelo. Um maior número de estimadores amplia a capacidade do modelo, porém também pode aumentar o tempo de treino e o risco de *overfitting*.

Por fim, o *Reg Alpha* é um hiperparâmetro de regularização que governa a penalização L1 (regularização L1) aplicada aos pesos dos nós da árvore. Essa penalização auxilia na prevenção do *overfitting*, ao encorajar a obtenção de pesos inferiores e a simplificação do modelo.

Tecidas as considerações supramencionadas, revela-se imprescindível ajustá-los adequadamente de modo a obter um modelo XGBoost com o melhor desempenho possível, considerando as considerações intrínsecas do conjunto de dados e o objetivo da tarefa de classificação. Nesta senda, experimentou-se diferentes combinações de valores para esses hiperparâmetros com o intuito de encontrar a configuração mais adequada para o problema em questão.

3.5.1.1 Experiência 1

Na primeira experiência, foi adotada a utilização de um modelo XGBoost sem otimização de parâmetros, com o propósito de estabelecer uma abordagem preliminar fundamentada em características mais simples, visando obter um modelo de menor complexidade e com variáveis de fácil compreensão, tais como:

- Adj_Close;
- Volume;
- Open.

Consequentemente, na Tabela 1 são apresentados os resultados obtidos nas diversas métricas avaliadas. Os resultados obtidos não foram satisfatórios, o que motivou a realização de novas iterações para aprimorar o modelo.

Tabela 1 – Experiência n.º 1: XGBoost sem otimização de parâmetros.

XGBoost		
Hiperparâmetro	Default	Métricas
Learning_rate	0.3	Accuracy: 78%
Min_child_weight	1	
Gamma	0	
Max_depth	6	Precision: 77%
N_estimators	100	
		Recall: 84%

3.5.1.2 Experiência 2

De forma a aprimorar o desempenho do modelo, procedeu-se à otimização dos parâmetros, com a intenção de identificar uma combinação ótima de valores que pudesse gerar resultados mais satisfatórios. Esta abordagem envolve a exploração sistemática de diversas configurações do modelo XGBoost, através da adaptação dos parâmetros, visando maximizar sua capacidade preditiva. Ao otimizar os parâmetros, desejou-se encontrar uma configuração que fosse capaz de reproduzir resultados mais favoráveis, em conformidade com as métricas de avaliação estabelecidas. Tal procedimento permitiu explorar plenamente o potencial do modelo e aprimorar a sua eficácia preditiva, tal como é possível observar, na Tabela 2, através do incremento dos valores das métricas de avaliação.

Tabela 2 - Experiência n. 92: XGBoost com otimização de parâmetros.

XGBoost			
Hiperparâmetro	Range	Value	Métricas
Learning_rate	[0.0005,0.005,0.1]	0.1	▪ Accuracy: 81% ▪ Precision: 80% ▪ Recall: 85%
Min_child_weight	[0,1,5,10]	1	
Gamma	[0,0.1]	0	
Max_depth	[2,3,5,7]	3	
N_estimators	[100,200,300,400,500,600,700]	300	

3.5.1.3 Experiência 3

De um ponto de vista puramente experimental, optou-se, na terceira experiência, por realizar a exclusão de determinadas variáveis e a inclusão de outras, com o propósito de investigar a capacidade de classificação da tendência do Índice S&P 500 utilizando o novo conjunto de variáveis. Esta abordagem visa analisar se as variáveis selecionadas apresentam um poder preditivo mais significativo em relação à tendência do S&P 500, permitindo uma compreensão e classificação mais precisa desse indicador. A remoção das variáveis anteriores tem como objetivo direcionar o enfoque para as características mais relevantes e explorar a habilidade do modelo em classificar a tendência do S&P 500 de maneira mais precisa.

As variáveis selecionadas foram as seguintes:

- Adj_Close;
- Adj_Close_vix;
- Adj_Close_oil;
- MACD;
- Open;
- trend_close_vix
- RSI.
- Variation_oil;
- Variation_dow.

Como pode ser observado na Tabela 3, este procedimento permitiu alcançar um acréscimo de aproximadamente 10 pontos percentuais nas três métricas de avaliação do modelo, em comparação com a experiência anterior. Estes resultados evidenciam uma melhoria significativa no desempenho do modelo, indicando que a substituição das variáveis resultou numa capacidade de classificação mais precisa da tendência do S&P 500. Essa diferença positiva nas métricas de avaliação reforça a relevância das variáveis selecionadas na modelação do comportamento desse Índice.

Tabela 3 - Experiência n.º 3: XGBoost com otimização de parâmetros, utilizando novo conjunto de features e ajuste de parâmetros.

XGBoost			
Hiperparâmetro	Range	Value	Métricas
Learning_rate	[0.0005,0.005,0.1,0.15,0.2]	0.005	▪ Accuracy: 89% ▪ Precision: 89% ▪ Recall: 90%
Min_child_weight	[1,5,10]	10	
Gamma	[0,0.1]	0.2	
Max_depth	[3,5,7,9]	3	
N_estimators	[300,400,500,600,700,1000]	1000	

No entanto, é importante ressaltar que, na última experiência mencionada, embora as métricas tenham alcançado valores altamente significativos e satisfatórios, o número de variáveis explicativas utilizado permaneceu substancialmente elevado. Diante dessa circunstância, com o propósito de reduzir a complexidade do modelo através da diminuição do número de variáveis, procedeu-se, na quarta experiência, à exclusão de três variáveis com o intuito de examinar o efeito dessa remoção no desempenho do modelo. Esta abordagem tem por objetivo avaliar a importância dessas variáveis na capacidade de classificação da tendência do S&P 500, enquanto visa simplificar e aprimorar a interpretação do modelo, promovendo uma maior parcimónia na seleção das variáveis utilizadas.

3.5.1.4 Experiência 4

Após a remoção das variáveis Adj_Close_vix, Adj_Close_oil e Variation_oil, foi observado que, em comparação com a experiência anterior, os valores das métricas permaneceram constantes, com exceção do *recall*, que registou um aumento de um ponto percentual. A estabilidade das métricas sugere que a exclusão das três variáveis não teve um impacto substancial no desempenho geral do modelo, o que por sua vez contribui para aumentar a simplicidade do modelo. O aumento no *recall* indica que o modelo está a capturar de forma mais eficaz os casos em que o Índice S&P 500 apresenta uma tendência crescente, melhorando a sua capacidade de identificar corretamente esses casos.

Tabela 4 - Experiência n.º 4: XGBoost com otimização de parâmetros, utilizando novo conjunto de features e ajuste de parâmetros.

XGBoost			
Hiperparâmetro	Range	Value	Métricas
Learning_rate	[0.0005,0.005,0.1,0.15,0.2]	0.005	▪ Accuracy: 89% ▪ Precision: 89% ▪ Recall: 91%
Min_child_weight	[1,5,10]	10	
Gamma	[0,0.1]	0.2	
Max_depth	[3,5,7,9]	3	
N_estimators	[300,400,500,600,700,1000]	100	

3.5.1.5 Experiência 5

Na penúltima experiência, foram realizadas duas importantes modificações no modelo. Primeiramente, foi adicionado um novo hiperparâmetro denominado *reg_alpha*, com o propósito de controlar o *overfitting*, ou seja, reduzir a tendência do modelo em ajustar-se excessivamente aos dados de treino. Além disso, houve uma modificação no conjunto de variáveis de entrada, resultando numa redução do número de variáveis utilizadas, com o intuito de simplificar o modelo. Esta redução das variáveis foi fundamenta nos métodos estatísticos ANOVA e LASSO, os quais identificaram as variáveis que possuíam maior capacidade de influenciar a tendência do Índice S&P 500, de acordo com suas respectivas metodologias. Esta abordagem teve como objetivo selecionar as variáveis mais relevantes e informativas, a fim de melhorar a capacidade preditiva do modelo.

Com esta experiência, foi possível observar um leve aumento nas métricas de avaliação do desempenho do modelo. Estes resultados sugerem que as modificações realizadas, como a introdução do hiperparâmetro *reg_alpha* e a seleção de variáveis com base nos métodos ANOVA e LASSO, contribuíram para uma melhoria marginal no desempenho do modelo. Embora o incremento nas métricas tenha sido modesto, essas melhorias indicam um progresso na capacidade de classificação da tendência do Índice S&P 500.

As variáveis selecionadas foram as seguintes:

- Variation_dow,
- Low_10d_vix,
- Low_oil

Tabela 5 - Experiência n.º 5: XGBoost com otimização de parâmetros, utilizando novo conjunto de features e ajuste de parâmetros.

XGBoost			
Hiperparâmetro	Range	Value	Métricas
Learning_rate	[0.0005,0.005,0.1,0.15,0.2]	0.005	▪ Accuracy: 90% ▪ Precision: 92% ▪ Recall: 90%
Min_child_weight	[1,5,10]	5	
Gamma	[0,0.1,0.2]	0.2	
Max_depth	[3,5,7,9]	9	
N_estimators	[300,400,500,600,700,1000]	300	
Reg_alpha	[0.0001,0.001,0.01]	0.0001	

Por fim, mesmo com resultados considerados satisfatórios até então, foi realizada uma última experiência visando obter melhorias ainda mais significativas.

3.5.1.6 Experiência 6

Esta última experiência consistiu, então, exclusivamente na substituição da variável "variation_dow" pela variável "return". Com essa modificação, como pode ser observado na Tabela 6, observou-se um aumento de quatro pontos percentuais na *accuracy* em relação à experiência anterior, um aumento de dois pontos percentuais na *precision* e um aumento de cinco pontos percentuais no *recall*. Estes resultados indicam que a substituição da variável teve um impacto positivo no desempenho do modelo, aumentando sua capacidade de classificação e previsão da tendência do Índice S&P 500.

- Return
- Low_10d_vix,
- Low_oil

Após a realização da última experiência, observou-se um incremento de quatro pontos percentuais na *accuracy* em relação à experiência anterior, alcançando um valor de 94%. Isso significa que o modelo apresenta uma taxa elevada de previsões corretas em relação à tendência do Índice S&P 500, o que é muito relevante para a tomada de decisões de investimento. A *accuracy* elevada indica que o modelo possui um desempenho satisfatório na classificação dos casos.

Tabela 6 - Experiência n.º 6: XGBoost com otimização de parâmetros, utilizando novo conjunto de features e ajuste de parâmetros.

XGBoost			
Hiperparâmetro	Range	Value	Métricas
Learning_rate	[0.0005,0.005,0.1,0.15,0.2]	0.005	▪ Accuracy: 94% ▪ Precision: 94% ▪ Recall: 95%
Min_child_weight	[1,5,10]	10	
Gamma	[0,0.1,0.2]	0.1	
Max_depth	[3,5,7,9]	3	
N_estimators	[300,400,500,600,700,1000]	400	
Reg_alpha	[0.0001,0.001,0.01]	0.01	

Já a *precision* de 94% indica a habilidade do modelo em prever corretamente 94% dos casos em que a tendência do Índice é crescente. Esta métrica é crucial para evitar falsos positivos, ou seja, previsões incorretas de aumento da tendência. Com uma *precision* elevada, os investidores têm uma maior confiança de que as previsões de aumento da tendência são precisas e podem tomar decisões de investimento com maior segurança. É, ainda, essencial salientar que os falsos positivos têm implicações financeiras negativas. Quando o modelo prevê uma tendência crescente que não se concretiza, isto é, na realidade, o Índice S&P 500 apresenta uma tendência decrescente, os investidores podem incorrer em perdas financeiras. Tais perdas ocorrem porque os investidores podem abrir posições de compra com base na previsão incorreta e, posteriormente, ter de acionar o *stop loss* para limitar as suas perdas. Portanto, uma *precision* elevada é fundamental para minimizar os falsos positivos e evitar perdas financeiras desnecessárias.

Com um *recall* de 95%, o modelo demonstra a capacidade de identificar corretamente 95% dos casos em que o Índice S&P 500 apresenta tendência crescente. Esta métrica reflete a eficácia do modelo em evitar falsos negativos, ou seja, a ocorrência de casos em que a tendência crescente não é identificada corretamente. Os falsos positivos, por sua vez, correspondem aos casos em que o modelo previu erroneamente uma tendência decrescente que na verdade se transformou numa tendência positiva. Nesses casos, os investidores podem ser levados a vender as suas posições devido ao temor de uma queda no Índice, quando na verdade o Índice acaba por subir. Evitar falsos positivos é fundamental para evitar perdas desnecessárias e permitir que os investidores tomem decisões de venda de forma mais precisa e fundamentada.

Em suma, na conclusão desta secção, pode-se afirmar que o aprimoramento da inclusão de variáveis exógenas no modelo de classificação, resultou num avanço significativo na sua *performance* preditiva. O modelo final exibiu uma elevada *accuracy* de 94%, denotando habilidade em discernir com precisão se o preço de fecho irá ascender ou declinar, face ao dia anterior. Neste contexto, em outras palavras, esta *accuracy* de 94% reflete que, num conjunto de 100 dias, o modelo previu corretamente o movimento do preço em 94 ocasiões.

Comparativamente ao modelo *baseline* inicial, o progresso obtido é significativo, aumentando em cerca de 20 pontos percentuais o desempenho preditivo. Este aprimoramento, deve-se, em grande parte, às técnicas utilizadas para a seleção de *features*. A seleção criteriosa das variáveis exógenas, determinadas através de técnicas avançadas de seleção de *features*, desempenhou um papel crucial no aumento do valor das métricas de performance. Estas técnicas permitiram identificar e incluir apenas as variáveis mais relevantes para o processo de classificação, evitando, assim, o impacto negativo de variáveis irrelevantes ou redundantes.

Na Figura 14, estão resumidos os valores das métricas obtidos para cada uma das seis experiências realizadas.

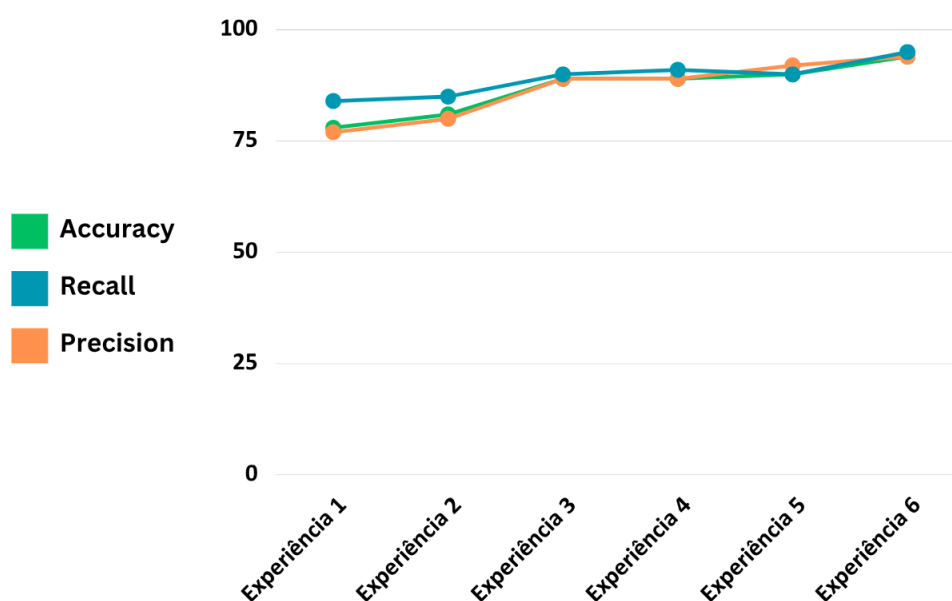


Figura 14 - Resultados obtidos com as seis experiências desenvolvidas no problema de classificação.

Ademais, é imprescindível destacar que a otimização dos hiperparâmetros também desempenhou um papel preponderante na obtenção de resultados aprimorados. A meticulosa procura e a criteriosa seleção dos hiperparâmetros contribuíram significativamente para aprimorar a eficácia do modelo e maximizar a sua capacidade de previsão.

Em virtude do nível de *accuracy*, atingindo a marca de 94%, torna-se inquestionável que o desenvolvimento deste modelo se revelou vantajoso. A sua precisão consistente na previsão oferece uma vantagem competitiva valiosa, capacitando tomadas de decisão informadas e confiáveis num cenário caracterizado pela volatilidade. Nesse contexto, este modelo permite reduzir substancialmente o risco associado à alavancagem, uma vez que as previsões corretas, em grande parte dos dias, minimizam a probabilidade de ocorrência de perdas significativas. Isto proporciona maior confiança aos investidores que utilizam alavancagem ao tomar decisões de investimento, permitindo-lhes capitalizar de forma mais segura nas oportunidades de mercado.

Com base na comparação com a revisão de literatura, notadamente considerando o artigo de [Ampomah et al. \(2020\)](#), o modelo desenvolvido ostentou uma superioridade, apresentando uma *accuracy* de oito pontos percentuais maior do que o modelo XGBoost abordado neste estudo. Veja-se que, o estudo conduzido por [Ampomah et al. \(2020\)](#) revelou a *accuracy* mais elevada para o XGBoost entre os artigos analisados. Adicionalmente, ao observar a revisão de literatura, constata-se que o segundo melhor modelo da revisão da literatura para o XGBoost obteve uma *accuracy* de 74%. Deste modo, é evidente que o modelo desenvolvido neste projeto apresentou um notável aumento de 20 pontos percentuais em comparação com o referencial apresentado na literatura. Consequentemente, esta significativa melhoria de desempenho demonstra a eficácia e superioridade do modelo proposto face aos artigos anteriormente analisados.

3.5.2 Construção do Modelo para o Problema de Regressão

Após a conclusão da construção do modelo destinado a identificar a tendência do Índice S&P 500, procedeu-se ao desenvolvimento de um modelo LSTM com o intuito de efetuar previsões do preço de fecho. Tal escolha baseou-se na capacidade das redes neurais recorrentes do tipo LSTM capturarem dependências temporais presentes em séries temporais financeiras a curto e a longo prazo.

É pertinente salientar que um dos objetivos consistia em estimar o valor diário do impacto decorrente das variações do Índice S&P 500, a fim de determinar a magnitude dessas oscilações e avaliar o impacto resultante. No entanto, constatou-se que a variável criada com esse propósito não exibiu sinais de autocorrelação (ou dependência ao longo do tempo), conforme evidenciado na Figura 15, onde a função de autocorrelação mostra um padrão típico de um processo de ruído branco. Tal ausência de autocorrelação implica a inviabilidade de previsão dessa variável.

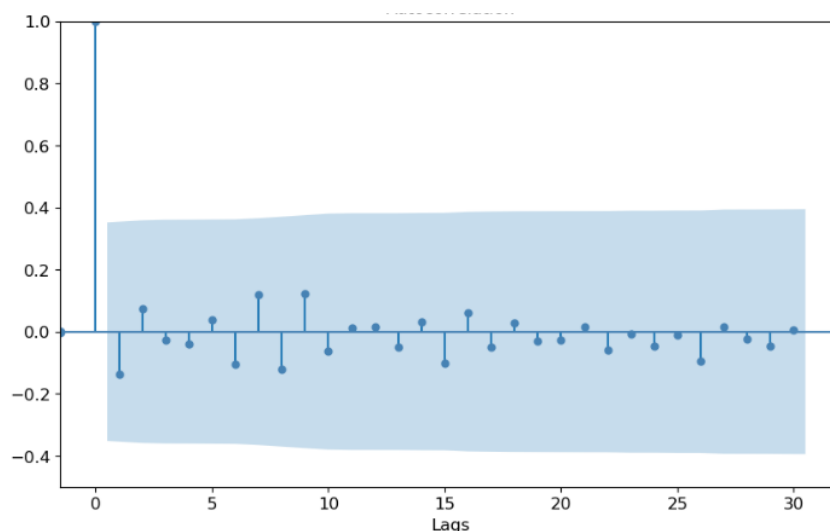


Figura 15 - Autocorrelação da variável que mede a magnitude das oscilações do Índice S&P 500.

Mais se pode concluir que a ausência de autocorrelação nessa variável reforça a hipótese da teoria da eficiência dos mercados financeiros. Ora, a teoria da eficiência dos mercados eficientes, formulada por [Eugene Fama](#) na década de 1960, postula que os preços dos ativos financeiros refletem de forma imediata todas as informações pertinentes e disponíveis. De acordo com essa teoria, os mercados são considerados eficientes na medida em que os investidores não conseguem consistentemente superar o desempenho médio do mercado por meio da aplicação de estratégias baseadas em análise técnica ou em informações históricas.

Nesta senda, de acordo com a perspectiva da teoria da eficiência dos mercados financeiros, os retornos diários das ações não podem ser preditos de maneira consistente, uma vez que os preços já contemplam todas as informações disponíveis. A ausência de autocorrelação na variável em estudo reforça essa ideia, indicando que os dados se comportam como ruído branco ou *random walk*, onde não há padrões ou dependência serial.

Em consequência, prosseguiu-se com a tarefa de prever de forma precisa o preço de fecho diário do Índice S&P 500. O objetivo subjacente consiste em obter o valor do preço de fecho ao término do dia seguinte de negociação, empregando-se informações históricas referentes aos dias anteriores. Essas informações incluem valores passados e padrões observados nas transações realizadas com o Índice durante o período definido.

3.5.2.1 XGBoost

Com o intuito de obter resultados comparativos, realizou-se uma experiência inicial utilizando o modelo XGBoost *Regressor*³, preservando rigorosamente a mesma estrutura adotada na última experiência executada com o XGBoost *Classifier* no problema previamente analisado, bem como as variáveis selecionadas, tais como:

- Return
- Low_10d_vix,
- Low_oil

Além disso, salienta-se que esta arquitetura do XGBoost *Regressor* foi utilizada com os parâmetros otimizados, que foram obtidos através de um cuidadoso processo de procura e ajuste de hiperparâmetros. Essa procura de parâmetros foi conduzida na secção 3.5.1 desta dissertação, onde foram exploradas diversas combinações de valores para os hiperparâmetros do XGBoost.

Perante esta abordagem adotada, veja-se, todavia, que, os resultados obtidos não se revelaram satisfatórios, dado o erro médio quadrático evidenciar uma média de erro de aproximadamente 757 dólares em relação aos valores reais, conforme apresentado na Tabela 7. Com esta diferença considerável entre as previsões e o valor real médio do variável *target*, que é de 1921\$, é possível inferir que este modelo poderia resultar em perdas financeiras significativas. Além disso, o MAPE (*Mean absolute percentage error*) de 28.83% indica que, em média, as previsões diferem em quase 29% do valor real. Esta taxa de erro percentual médio é bastante significativa e pode impactar negativamente a tomada de decisões financeiras.

Tabela 7 - Experiência n.º 1: XGBoost Regressor.

XGBoost			
Hiperparâmetro	Range	Value	Métricas
Learning_rate	[0.0005,0.005,0.1,0.15,0.2]	0.15	• RMSE: 757 • MAPE: 28.83% • $\bar{X}$ TARGET: 1921
Min_child_weight	[1,5,10]	10	
Gamma	[0,0.1,0.2]	0.1	
Max_depth	[3,5,7,9]	9	
N_estimators	[100,300,400,500,600,700,1000]	100	
Reg_alpha	[0.0001,0.001,0.01]	0.01	

³ O XGBoost Regressor é um algoritmo de *machine learning* baseado em árvores de decisão, amplamente utilizado para tarefas de regressão devido à sua eficiência e precisão dos resultados. Para mais informações sobre os parâmetros específicos do XGBoost Regressor, consulte o seguinte documento: <https://xgboost.readthedocs.io/en/stable/parameter.html>

Desta forma, é crucial rever e aprimorar o modelo de previsão e considerar abordagens que possam proporcionar maior precisão nas previsões. Reduzir a diferença entre as previsões e os valores reais é fundamental para evitar perdas financeiras e garantir que as decisões sejam fundamentadas em informações mais confiáveis.

3.5.2.2 LSTM

Nas experiências subsequentes, desenvolveu-se uma rede LSTM utilizando técnicas de otimização de hiperparâmetros, a fim de determinar os valores mais adequados para alcançar o melhor desempenho possível. A arquitetura resultante é composta por duas camadas ocultas de tipo LSTM, em que o número de unidades é selecionado entre 25, 50 ou 75 por meio do processo de otimização dos parâmetros. A função de ativação pode ser escolhida entre "relu" ou "tanh" para ambas as camadas LSTM. Após as camadas LSTM, são adicionadas duas camadas densas, com a primeira contendo 25 unidades e a última com uma única unidade de saída. A rede é compilada utilizando os algoritmos de otimização "adam" ou "rmsprop", e a função de perda utilizada é o *Mean Squared Error* (erro médio quadrático). Essa arquitetura possibilita que a rede LSTM capture informações temporais em sequências de dados e realize previsões visando otimizar o seu desempenho.

Para avaliar o desempenho da arquitetura desenvolvida, conduziram-se três experiências, nas quais apenas as variáveis de entrada foram modificadas. Nestas três primeiras experiências, a escolha dos melhores hiperparâmetros permaneceu consistente, indicando que um número ótimo de unidades na rede LSTM é 75, o algoritmo de otimização adotado é o "adam" e a função de ativação empregue é o "tanh", o esquema da arquitetura LSTM encontra-se na Figura 19. Mediante a Tabela 8, é possível visualizar que a terceira experiência resultou num desempenho superior, superando os resultados previamente obtidos. Destaca-se que as características utilizadas nessa experiência foram selecionadas por meio dos métodos de seleção de variáveis LASSO e ANOVA.

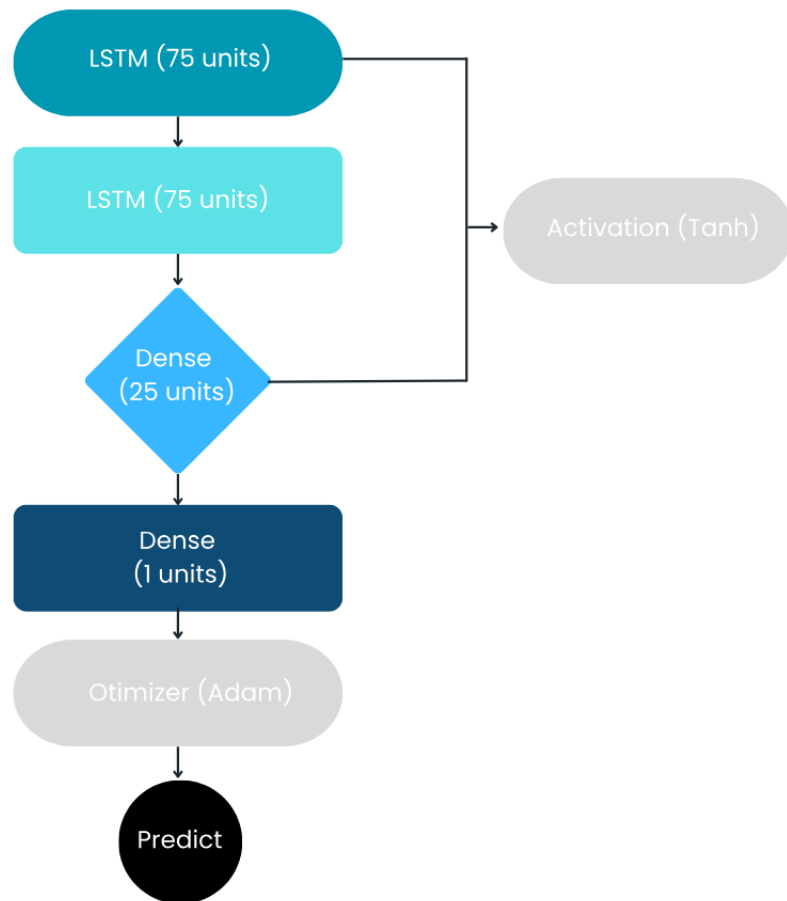


Figura 19 - Arquitetura LSTM otimizada.

Observa-se que a terceira experiência, com as variáveis Open, Low e High selecionadas pelo método LASSO e ANOVA, obteve até ao presente momento o melhor desempenho em termos das métricas de avaliação consideradas. Estes resultados corroboram a eficácia da combinação dos hiperparâmetros otimizados e da seleção adequada das variáveis de entrada, evidenciando a relevância das técnicas empregues na construção da arquitetura da rede LSTM. Com base nos dados apresentados na Tabela 8 e visualizados de forma gráfica na Figura 20, é possível perceber os resultados obtidos com as três experiências realizadas, bem como o desvio da previsão em relação aos valores reais.

Tabela 8 - Experiências: Resultados obtidos com o uso da rede LSTM para prever o preço de fecho do Índice S&P 500.

LSTM				
Experiência	Features	RMSE	MAPE	$\bar{X}$ TARGET
1	Return / Low_10d_vix / Low_oil	996	44%	1 921
2	Return/Variation_oil/Variation_dow/Adj_close_previous	22	2.23%	
3	Open / Low / High	16	1.90%	

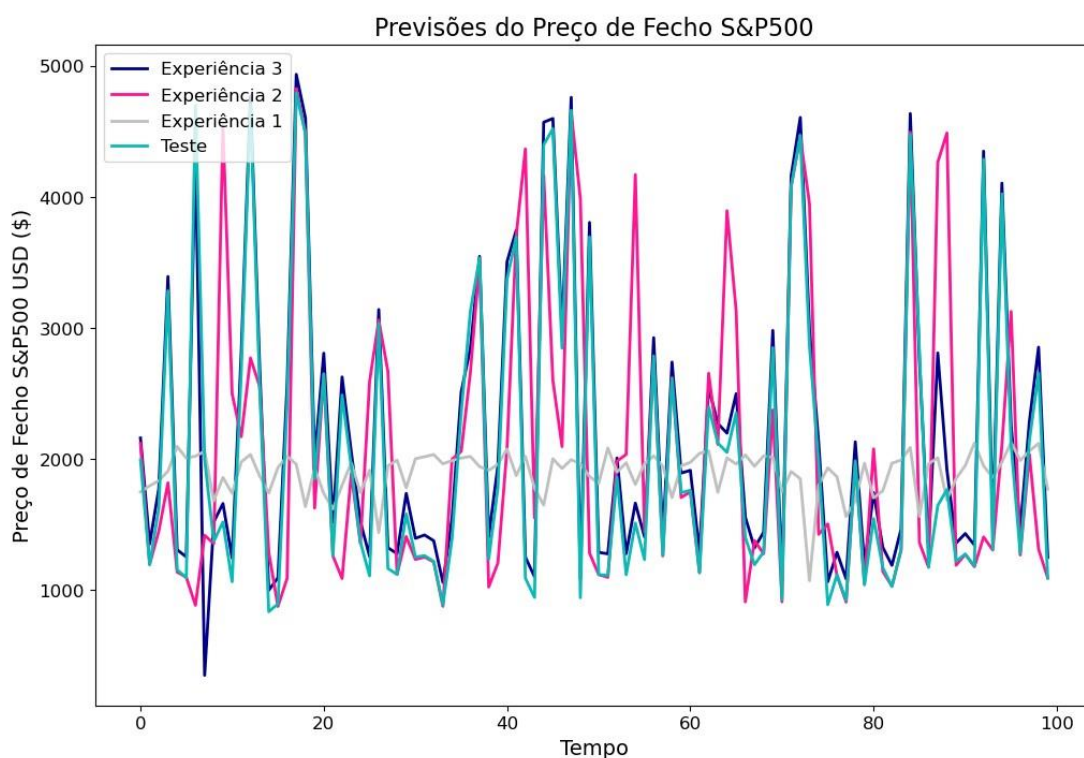


Figura 20 - Previsões do preço de fechamento do Índice S&P 500.

Verifica-se que, até o presente momento, as variáveis endógenas destacaram-se como as que obtiveram o melhor desempenho na previsão do preço. Com o intuito de estudar o comportamento das métricas, decidiu-se realizar uma quarta e última experiência, utilizando apenas o histórico do preço de fechamento do Índice S&P 500, dos últimos 20 dias. Surpreendentemente, esta abordagem resultou num desempenho superior, como pode ser corroborado pelos resultados apresentados na Tabela 9 e visualizados de forma gráfica na Figura 21, onde se evidencia a comparação entre os valores reais e os valores preditos. Este último modelo empregado na previsão do preço do Índice S&P 500 revelou um RMSE de 8.94, indicando que, em média, as previsões exibem um erro de meros 8.94\$ em relação aos valores reais. Além disso, o MAPE obtido indica que o modelo determina um erro médio percentual de apenas 1.71% em relação aos valores reais, representando cerca de 32.8\$ em termos absolutos. Vale salientar que, em cenários onde a simplicidade é prevalente, o modelo de regressão obteve uma eficácia exímia ao realizar as previsões quando comparado com o valor médio do Índice. Por conseguinte, podemos concluir que, ao desenvolver modelos mais simples e de compreensão mais acessível, foi possível obter os melhores resultados na previsão do preço.

Tabela 9 - Resultados obtidos com a última experiência na previsão do preço de fechamento.

LSTM				
Experiência	Features	RMSE	MAPE	$\bar{X}$ TARGET
4	Adj_Close	8.94	1.71%	1 921

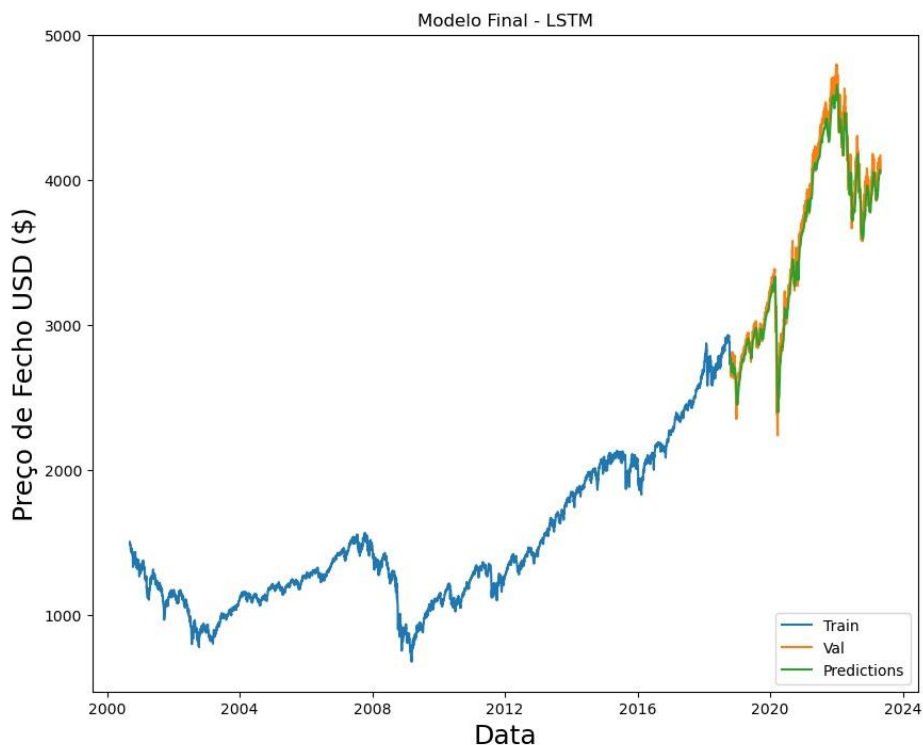


Figura 21 - Previsão do preço de fechamento do Índice S&P 500 - Modelo Final

Tais métricas apresentam uma importância de relevo no âmbito do *day trading* e da gestão financeira, pois podem exercer um impacto direto na rentabilidade e na mitigação de riscos. No contexto do *day trading*, onde as operações se desenvolvem em prazos curtos e baseiam-se em previsões de curto prazo, a precisão das previsões de preço reveste-se de caráter crucial. Um modelo com valores reduzidos de RMSE e MAPE sugere que as previsões se encontram próximas dos valores reais, o que pode auxiliar os *day traders* na tomada de decisões informadas e oportunas. Isso pode resultar numa maior probabilidade de realizar *trades* bem-sucedidas, o que, por sua vez, pode levar a ganhos financeiros significativos.

Para este problema, observou-se, então, que a introdução de variáveis exógenas, como por exemplo o preço do petróleo ou o valor do Índice VIX, apenas adicionava ruído ao modelo de previsão de preço. No entanto, é importante ressaltar que, ao considerar a classificação da tendência, a inclusão de variáveis exógenas revelou-se crucial. A previsão do preço do Índice S&P 500 apresenta um desafio complexo devido à sua natureza volátil e à influência de múltiplos fatores no mercado financeiro. Embora a inclusão de variáveis exógenas possa, a princípio, parecer uma abordagem promissora para aprimorar a precisão das previsões de preço, é importante reconhecer que tais variáveis podem, em determinados casos, introduzir ruído indesejável e não conferir um benefício tangível à precisão global do modelo. Contudo, quando se trata da classificação da tendência do Índice S&P 500, as variáveis exógenas desempenham um papel crucial. Essas variáveis externas possuem a capacidade de fornecer informações complementares sobre as condições do mercado, sinalizando possíveis tendências ascendentes ou descendentes.

De acordo com a análise realizada considerando as características específicas do modelo de previsão desenvolvido neste projeto, notadamente a sua arquitetura LSTM e o objetivo de prever o preço de fecho do Índice S&P 500 em base diária, é digno de notar que o modelo desenvolvido nesta secção, destaca-se como o terceiro melhor de entre os estudos literários analisados, levando em consideração tais características.

Cumprе ressaltar que os artigos que alcançaram resultados mais satisfatórios na previsão do Índice S&P 500 adotaram uma abordagem abrangente, incorporando diversas fontes de informação complementares e estratégicas. Tais fontes incluem dados do *Google Trends*, uma ampla gama de indicadores económicos e, ainda, análise de sentimento. Outro aspeto relevante, e que contribui para ostentar a diferença existente nas *performances*, é a incorporação de diferentes períodos temporais de histórico de dados.

CAPÍTULO 5

Conclusões

Neste Capítulo, são apresentadas as principais conclusões que se lograram alcançar com a presente dissertação, relativamente à eficiência dos algoritmos XGBoost e LSTM e, bem como, sobre a sua aplicabilidade na previsão de dados financeiros, marcados por alta volatilidade e complexidade.

Todos os conjuntos de dados utilizados no estudo são de natureza diária e foram recolhidos através da plataforma *Yahoo Finance*, abrangendo os períodos de setembro de 2000 até abril de 2023. Adicionalmente, ao conjunto de dados do Índice S&P 500, foram acrescentados outros três conjuntos contendo variáveis exógenas, com o intuito de incrementar o desempenho dos modelos. Ademais, também foram inseridas duas variáveis representativas de indicadores técnicos, a saber, o RSI e o MACD, visando auxiliar no processo de previsão. Após a construção das diversas variáveis, realizou-se o tratamento de dados necessário para mitigar a presença de valores omissos e assegurar a integridade dos dados.

Após a conclusão da etapa supramencionada, foi realizada a seleção de variáveis mediante a aplicação de duas técnicas: LASSO e ANOVA. O objetivo foi obter as variáveis mais explicativas para cada variável *target*.

No âmbito da problemática da classificação, enfocada na tendência do Índice S&P 500, a incorporação do Índice VIX e do preço do petróleo demonstrou um incremento significativo nas métricas de desempenho, refletindo, assim, um reforço na análise das oscilações do preço de fecho do Índice S&P 500. Consequentemente, pode-se concluir que a inclusão do Índice VIX, como variável explicativa, permitiu a captura e reflexão das alterações no sentimento e percepção dos investidores relativamente ao risco vivenciado nos mercados, contribuindo, deste modo, para o aprimoramento das métricas de *performance* do modelo desenvolvido para prever as direções do Índice S&P 500. Por sua vez, o preço do petróleo exerceu uma influência significativa sobre o Índice S&P 500, uma vez que as flutuações nos preços do petróleo, afetam diversos setores económicos e as perspetivas de crescimento, impactando, consequentemente, o desempenho geral do Índice. Veja-se que, o aumento nos preços do petróleo acarreta um consequente aumento nos custos operacionais das empresas, o que pode, por sua vez, exercer pressão negativa sobre os lucros e o desempenho das empresas incorporadas no Índice S&P 500.

Acrescenta-se, ainda, que inicialmente, um modelo *baseline* foi desenvolvido, contendo unicamente variáveis endógenas, e com a adição estratégica de determinadas variáveis exógenas, alcançou-se um aumento de cerca de 20 pontos percentuais na *accuracy*. O modelo final revelou, portanto, uma *accuracy* de 94%, evidenciando a sua capacidade em discernir com precisão o movimento do preço de fecho do Índice S&P 500 em relação ao dia anterior. Este valor de *accuracy* enfatiza a vantagem estratégica decorrente do referido modelo, pois tal valor de *accuracy* implica que ao longo de 365 dias, o modelo prevê erroneamente apenas 22 ocasiões.

Como estratégia adicional, é pertinente mencionar que o segundo problema a ser abordado consistia na previsão do impacto das variações do Índice S&P 500, a fim de avaliar a sua magnitude. No entanto, ao analisar a variável criada com essa finalidade, constatou-se que a mesma não apresentava autocorrelação, sugerindo que os preços refletem eficientemente todas as informações disponíveis no mercado financeiro. Ora, esse resultado está em consonância com a teoria da eficiência dos mercados financeiros, que postula que os preços dos ativos financeiros já incorporam todas as informações relevantes. Deste modo, optou-se, portanto, por prosseguir com o foco na previsão do preço de fecho do Índice S&P 500 para o dia seguinte.

Assim, neste aspeto, no problema de regressão, cujo objetivo é prever o preço de fecho do Índice S&P 500 para o dia seguinte, foram conduzidas várias experiências, com o intuito de aprimorar as métricas de desempenho. Ao longo do desenvolvimento do modelo, constatou-se que a inclusão de variáveis exógenas resultava apenas no acréscimo de ruído do modelo, não contribuindo significativamente para a melhoria das previsões. Consequentemente, o melhor modelo foi obtido através da utilização do histórico de 20 dias do preço de fecho do referido Índice como a única variável preditiva.

Após a conclusão do processo de desenvolvimento do último modelo, foi alcançado um valor da métrica RMSE de apenas 8.94. Tal valor de RMSE, revela a amplitude dos valores que o modelo é capaz de prever. Esta estreita faixa de variação de preços, ficará situada entre 1912.06\$ e 1929.04\$, visto que o preço médio do Índice S&P 500 é de 1291 dólares. Esta discrepância sugere que o modelo desenvolvido possui uma margem de erro relativamente pequena em relação ao intervalo de valores possíveis. Para investidores e *traders*, possuir um modelo com uma baixa margem de erro representa um ativo inestimável, uma vez que possibilita a tomada de decisões fundamentadas e precisas, resultando, deste modo, na redução de riscos associados às operações financeiras.

A respeito de todas as contribuições advindas dos Capítulos precedentes, é importante ressaltar que esta dissertação se encontra sujeita a diversas limitações. De entre elas, destaca-se a constatação de que, embora os modelos desenvolvidos tenham apresentado métricas de desempenho satisfatórias, ainda persiste uma pequena, porém relevante, margem de erro. É imprescindível ter plena consciência de que essa margem de erro, embora reduzida, pode acarretar perdas significativas.

A sensibilidade do contexto que este trabalho se insere, especialmente no âmbito dos mercados financeiros, configura-se como um elemento de suma importância a ser ponderado. Os referidos mercados caracterizam-se pela sua intrínseca volatilidade, bem como pela natureza dinâmica e complexa, o que implica a necessidade de uma abordagem cautelosa. Consequentemente, a aplicação de uma estratégia eficaz em momentos propícios pode, em contrapartida, revelar-se desfavorável durante períodos de adversidade, culminando em prejuízos consideráveis.

Ora, é, portanto, essencial reconhecer que, mesmo diante dos progressos alcançados ao longo desta dissertação, a plena compreensão da imprevisibilidade da potencial instabilidade inerente aos mercados financeiros exige uma postura prudente e cautelosa na sua aplicação. A procura pela excelência e aperfeiçoamento dos modelos desenvolvidos deve ser acompanhada por uma consciência contínua das possíveis incertezas que permeiam esse contexto, com o propósito de mitigar quaisquer impactos adversos que possam decorrer dos seus investimentos.

Em acréscimo, em virtude da inerente volatilidade dos mercados financeiros, outra limitação que merece ser destacada é a exclusividade do uso de dados de preços históricos neste estudo. Dessa forma, a incorporação de análise de tendências, análise de sentimento ou, até mesmo, a inclusão de informações advindas de notícias pertinentes poderia evidenciar um elemento aliado na procura por resultados mais confiáveis. Nesse sentido, para futuros estudos, torna-se imperativo considerar a inclusão de tais dados complementares, dado que estes poderiam figurar como um aliado na procura por resultados de maior confiança e solidez.

Por fim, é importante considerar a possibilidade de resolver os problemas mencionados através do uso de modelos híbridos. Conforme destacado na revisão de literatura realizada, tais modelos têm demonstrado *performances* notáveis, apesar da sua complexidade. Muitos investigadores acreditam que a combinação de diferentes modelos ou a adoção de abordagens híbridas pode constituir uma solução eficaz para aprimorar o desempenho e superar as limitações existentes ([Khashei et al., 2019](#)).

Referências Bibliográficas

- Ampomah, E. K., Qin, Z., & Nyame, G. (2020). Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. *Information*, 11(6), 332.
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).
- Bhandari, H. N., Rimal, B., Pokhrel, N. R., Rimal, R., Dahal, K. R., & Khatri, R. K. (2022). Predicting stock market index using LSTM. *Machine Learning with Applications*, 100320.
- Binder, C., & Kamdar, R. (2022). Expected and realized inflation in historical perspective. *Journal of Economic Perspectives*, 36(3), 131-56.
- Chou, J. S., & Pham, A. D. (2013). Enhanced artificial intelligence for ensemble approach to predicting high performance concrete compressive strength. *Construction and Building Materials*, 49, 554-563.
- Dey, S., Kumar, Y., Saha, S., & Basak, S. (2016). Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting. PESIT South Campus.
- Fahmy, Y. A., & Kandil, M. (2003). The Fisher effect: new evidence and implications. *International Review of Economics & Finance*, 12(4), 451-465.
- Fama, E. F. (1960). Efficient market hypothesis. Diss. PhD Thesis, Ph. D. dissertation.
- Farmer, J. D., Gallegati, M., Hommes, C., Kirman, A., Ormerod, P., Cincotti, S., ... & Helbing, D. (2012). A complex systems approach to constructing better models for managing financial markets and the economy. *The European Physical Journal Special Topics*, 214(1), 295-324.
- Ghosh, P., Neufeld, A., & Sahoo, J. K. (2022). Forecasting directional movements of stock prices for intraday trading using LSTM and random forests. *Finance Research Letters*, 46, 102280.
- Gumelar, A. B., Setyorini, H., Adi, D. P., Nilowardono, S., Widodo, A., Wibowo, A. T., ... & Christine, E. (2020, September). Boosting the Accuracy of Stock Market Prediction using XGBoost and Long Short-Term Memory. In *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)* (pp. 609-613). IEEE.
- Gupta, A., Joshi, K., Patel, M., & Pratap, M. V. (2023, March). Stock Prices Prediction Using Machine Learning. In *2023 2nd International Conference for Innovation in Technology (INOCON)* (pp. 1-7). IEEE.

- Hassan, M. R., & Nath, B. (2005, September). Stock market forecasting using hidden Markov model: a new approach. In 5th International Conference on Intelligent Systems Design and Applications (ISDA'05) (pp. 192-196). IEEE.
- Kamara, A. F., Chen, E., & Pan, Z. (2022). An ensemble of a boosted hybrid of deep learning models and technical analysis for forecasting stock prices. *Information Sciences*, 594, 1-19.
- Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*.
- Khashei, M., & Hajirahimi, Z. (2019). A comparative study of series arima/mlp hybrid models for stock price forecasting. *Communications in Statistics-Simulation and Computation*, 48(9), 2625-2640.
- Kumar, G., Jain, S., & Singh, U. P. (2021). Stock market forecasting using computational intelligence: A survey. *Archives of Computational Methods in Engineering*, 28(3), 1069-1101.
- Lam, M. (2004). Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. *Decision support systems*, 37(4), 567-581.
- Lanbouri, Z., & Achchab, S. (2020). Stock market prediction on high frequency data using long-short term memory. *Procedia Computer Science*, 175, 603-608.
- Lin, Y., Yan, Y., Xu, J., Liao, Y., & Ma, F. (2021). Forecasting stock index price using the CEEMDAN-LSTM model. *The North American Journal of Economics and Finance*, 57, 101421.
- Liwei, T., Li, F., Yu, S., & Yuankai, G. (2021). Forecast of LSTM-XGBoost in Stock Price Based on Bayesian Optimization. *INTELLIGENT AUTOMATION AND SOFT COMPUTING*, 29(3), 855-868.
- Medarhri, I., Hosni, M., Nouisser, N., Chakroun, F., & Najib, K. (2022, October). Predicting Stock Market Price Movement using Machine Learning Techniques. In 2022 8th International Conference on Optimization and Applications (ICOA) (pp. 1-5). IEEE.
- Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the american society for information science and technology*, 58(13), 2105-2125.
- Park, M., Lee, M. L., & Lee, J. (2019). Predicting stock market indices using classification tools. *Asian Economic and Financial Review*, 9(2), 243-256.
- Pokhrel, N. R., Dahal, K. R., Rimal, R., Bhandari, H. N., Khatri, R. K., Rimal, B., & Hahn, W. E. (2022). Predicting nepse index price using deep learning models. *Machine Learning with Applications*, 9, 100385.
- Price, M. (2022). Average Stock Market Return. Retrieved 11 December 2022, from <https://www.fool.com/investing/how-to-invest/stocks/average-stock-market-return/>
- S&P Dow Jones Indices – A Division of S&P 500 (2023, junho). S&P 500® - S&P Dow Jones Indices. <https://www.spglobal.com/spdji/en/indices/equity/sp-500/#data>

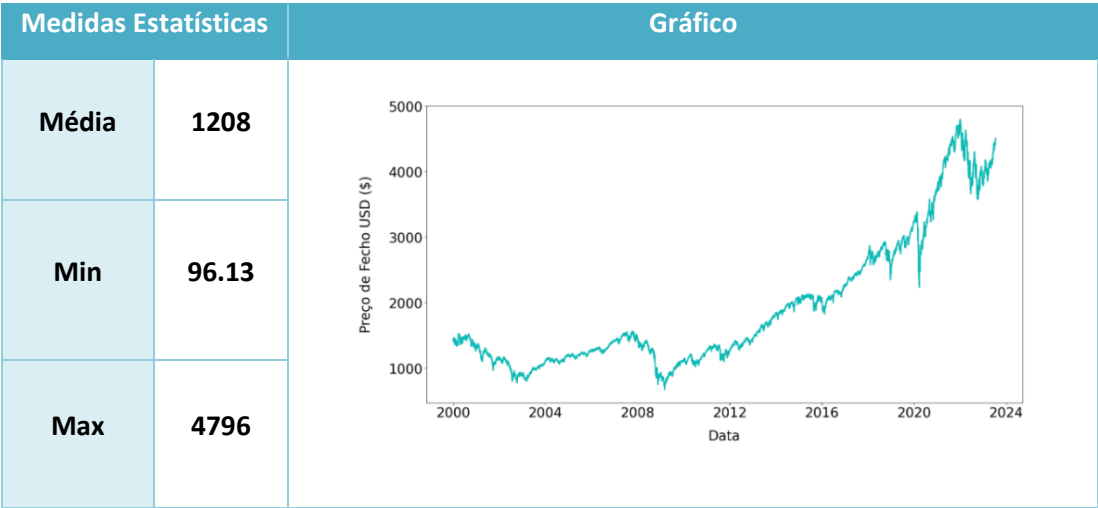
- Salisu, A. A., Ndako, U. B., & Akanni, L. O. (2020). New evidence for the inflation hedging potential of US stock returns. *Finance Research Letters*, 37, 101384.
- Salisu, A. A., Raheem, I. D., & Ndako, U. B. (2020). The inflation hedging properties of gold, stocks and real estate: A comparative analysis. *Resources Policy*, 66, 101605.
- Sethia, A., & Raut, P. (2019). Application of LSTM, GRU and ICA for stock price prediction. In *Information and communication technology for intelligent systems* (pp. 479-487). Springer, Singapore.
- Wang, Y., Liu, H., Guo, Q., Xie, S., & Zhang, X. (2019). Stock volatility prediction by hybrid neural network. *IEEE Access*, 7, 154524-154534.
- Wang, Z., Hu, Z., Li, F., Ho, S. B., & Cambria, E. (2023). Learning-based stock trending prediction by incorporating technical indicators and social media sentiment. *Cognitive Computation*, 15(3), 1092-1102.
- Worasuchep, C. (2022). Ensemble Classifier for Stock Trading Recommendation. *Applied Artificial Intelligence*, 36(1), 2001178.
- Yadav, A., Jha, C. K., & Sharan, A. (2020). Optimizing LSTM for time series prediction in Indian stock market. *Procedia Computer Science*, 167, 2091-2100.
- Yoo, P. D., Kim, M. H., & Jan, T. (2005, November). Machine learning techniques and use of event information for stock market prediction: A survey and evaluation. In *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)* (Vol. 2, pp. 835-841). IEEE.
- Yu, L., Chen, H., Wang, S., & Lai, K. K. (2008). Evolving least squares support vector machines for stock market trend mining. *IEEE Transactions on evolutionary computation*, 13(1), 87-102.
- Yun, K. K., Yoon, S. W., & Won, D. (2021). Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process. *Expert Systems with Applications*, 186, 115716.

ANEXOS

Anexo A – Variáveis Originais

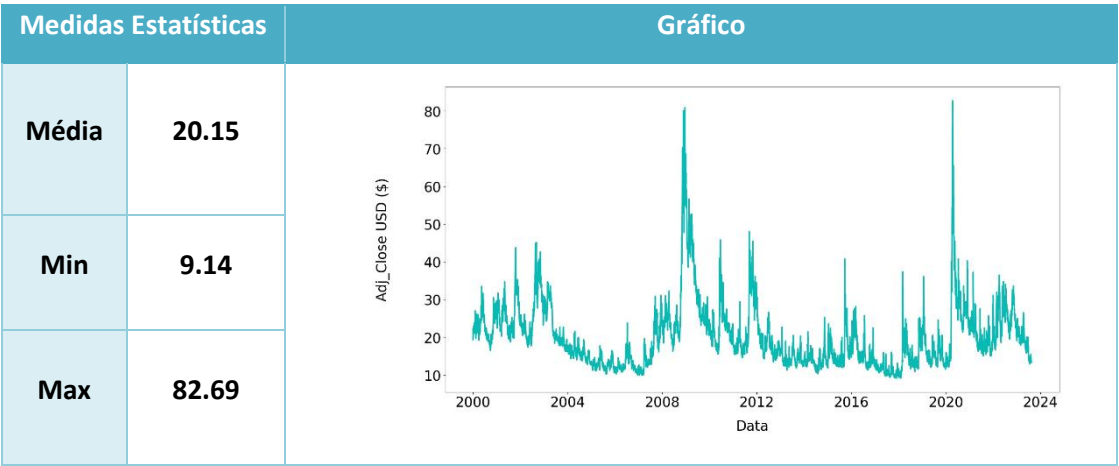
1. Índice S&P 500

Adj_Close

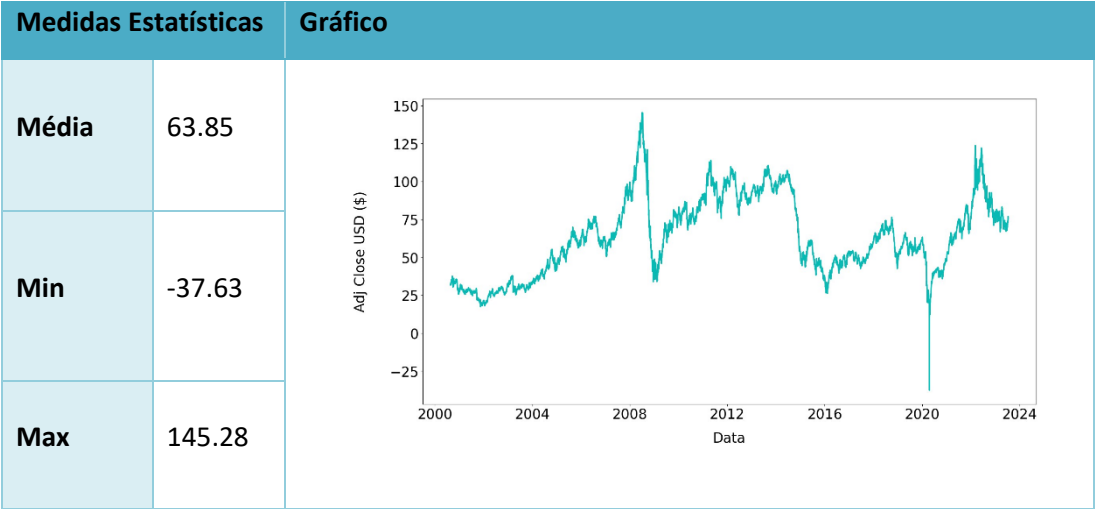


2. Índice VIX

Adj_Close



3. **Petróleo**
Adj_Close



4. **Índice DOW**
Adj_Close



Anexo B – Lista de Variáveis

- Open_sp500
- High_sp500
- Low_sp500
- Close_sp500
- Adj Close_sp500
- Volume_sp500
- Open_10d_sp500
- Open_4d_sp500
- High_10d_sp500
- High_4d_sp500
- Low_10d_sp500
- Low_4d_sp500
- Close_10d_sp500
- Close_4d_sp500
- Adj Close_10d_sp500
- Adj Close_4d_sp500
- Volume_10d_sp500
- Volume_4d_sp500
- Return_sp500
- Previous_Close_sp500
- Impact_value_sp500
- Target_trend
- MACD_sp500
- RSI_sp500
- Open_vix
- High_vix
- Low_vix
- Close_vix
- Adj Close_vix
- Open_10d_vix
- Open_4d_vix
- High_10d_vix

- High_4d_vix
- Low_10d_vix
- Low_4d_vix
- Close_10d_vix
- Close_4d_vix
- Adj Close_10d_vix
- Adj Close_4d_vix
- Volume_10d_vix
- Volume_4d_vix
- diff_close_vix
- trend_close_vix
- Open_oil
- High_oil
- Low_oil
- Close_oil
- Adj Close_oil
- Volume_oil
- Open_10d_oil
- Open_4d_oil
- High_10d_oil
- High_4d_oil
- Low_10d_oil
- Low_4d_oil
- Close_10d_oil
- Close_4d_oil
- Adj Close_10d_oil
- Adj Close_4d_oil
- Volume_10d_oil
- Volume_4d_oil
- Previous_Close_oil
- Variation_oil
- Open_dow
- High_dow

- Low_dow
- Close_dow
- Adj Close_dow
- Volume_dow
- Open_10d_dow
- Open_4d_dow
- High_10d_dow
- High_4d_dow
- Low_10d_dow
- Low_4d_dow
- Close_10d_dow
- Close_4d_dow
- Adj Close_10d_dow
- Adj Close_4d_dow
- Volume_10d_dow
- Volume_4d_dow
- Variation_dow