Universidade Federal de São Carlos – UFSCar
Centro de Ciências Exatas e de Tecnologia – CCET
Departamento de Computação – DC
Trabalho de Conclusão de Curso em Engenharia de Computação

Vitor Caligaris Figueira

Aplicação de Aprendizado de Máquina na Construção de Carteiras de Ações de Longo Prazo: Uma Abordagem Comparativa com Modelos Tradicionais

São Carlos
2025

Vitor Caligaris Figueira

Aplicação de Aprendizado de Máquina na Construção de Carteiras de Ações de Longo Prazo: Uma Abordagem Comparativa com Modelos Tradicionais

Monografia apresentada como Trabalho de Conclusão de Curso em Engenharia de Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de bacharel em Engenharia de computação.

Orientador: Heloísa de Arruda Camargo.

São Carlos

2025

Agradecimentos

Dedico este trabalho em primeiro lugar aos meus pais Mauro José Figueira e Regina Paula Caligaris, que sempre se fizeram presentes e forneceram todo o suporte necessário para esta conquista. Me ensinando que o conhecimento e o estudo são os maiores ativos que se pode obter. Muito obrigado por tudo.

A minha família, em especial a minha irmã Laura Figueira que cresceu e se desenvolveu ao meu lado, desejo uma excelente trajetória no ensino superior, assim como a minha foi.

A minha namorada, Ana Carolina Vasques, que esteve ao meu lado em todo o desenvolvimento do projeto e me deu todo o suporte e incentivo necessário para sua completude.

Aos meus amigos da graduação, em especial Pedro Klesse, Felipe Lopes e Brainer Sueverti que enfrentaram com êxito todos os desafios dessa etapa de nossas vidas e me auxiliaram em diversos momentos.

A minha orientadora, Heloísa de Arruda Camargo, que sempre ofereceu toda a ajuda acadêmica necessária e se mostrou disponível para me auxiliar no trabalho.

Resumo

O investimento em ações com foco no longo prazo chama a atenção de investidores que procuram métodos eficazes e diretos para escolher os papéis que farão parte de suas carteiras, bem como identificar ativos com maior potencial de valorização no futuro. Com o passar do tempo, diversas abordagens foram criadas para esse propósito. Diante disso, este estudo teve como finalidade a aplicação de técnicas de Aprendizado de Máquina (AM) para estimar os valores futuros das ações de dessa forma montar uma carteira de ativos capaz de bater o mercado de renda variável tradicional.

Com o objetivo de escolher ativos para compor uma carteira e prever sua rentabilidade modelos de AM podem ser usados para fazer a regressão de preço das ações e recomendar investimentos altamente rentáveis. Dessa forma, formando uma coleção de investimentos que diversificam a aplicação de capital reduzindo o risco e ainda sim superando a performance do mercado.

Neste trabalho foi desenvolvida, analisada e testada uma metodologia para montagem de uma carteira de ações para o período de alocação de um ano utilizando os seguintes modelos de AM: Regressão Linear Ridge (RL), Regressão Bayesiana (RB), Árvore de Regressão, Regressão por Vetores de Suporte (SVR), Regressão por Impulso de Gradiente (GBR). Para complementar, todos os modelos foram treinados e testados sem e com o ajuste de hiperparâmetros pelo método Busca em Grade (*grid search CV*), a fim de revelar se esse fator aumenta a performance dos modelos dessa forma, recomendando carteiras ainda melhores.

As carteiras geradas pelos modelos de AM tiveram suas rentabilidades comparadas com modelos tradicionais do mercado financeiro para recomendação de ativos como a teoria do portfólio moderno e o índice de mercado S&P 500. Ao final, o resultado obtido demonstrou que na grande maioria dos testes realizados a metodologia baseada em AM proposta por esse trabalho supera os modelos fundamentados no mercado.

Palavras-chave: aprendizado de máquina, regressão, carteira de ativos, ajuste de hiperparâmetros.

Abstract

Long-term stock investing attracts the attention of investors seeking effective and straightforward methods to select the securities to compose their portfolios, as well as to identify assets with the highest potential for future appreciation. Over time, many approaches have been developed for this purpose. In this context, this study aimed to apply Machine Learning (ML) techniques to estimate future stock values and, this way, construct a portfolio capable of outperforming the traditional equity market.

With the goal of selecting assets for a portfolio and predicting their profitability, ML models can be used to perform stock price regression and recommend highly profitable investments. This approach enables the formation of a diversified investment portfolio that reduces risk while still outperforming the market.

In this study, a methodology was developed, analyzed, and tested for constructing a stock portfolio with a one-year allocation period using the following ML models: Ridge Linear Regression (LR), Bayesian Regression (BR), Regression Tree, Support Vector Regression (SVR), and Gradient Boosting Regression (GBR). Additionally, all models were trained and tested both with and without hyperparameter tuning using the Grid Search CV method to determine whether this adjustment enhances model performance, thereby recommending even better portfolios.

The portfolios generated by the ML models had their returns compared to traditional financial market models for asset recommendation, such as Modern Portfolio Theory and the S&P 500 market index. In the end, the results showed that, in the vast majority of tests conducted, the ML-based methodology proposed in this study outperformed the market-based models.

Keywords: machine learning, regression, stock portfolio, hyperparameter tuning.

Lista de Ilustrações

Lista de Tabelas

Lista de Siglas

AM/ML - Aprendizado de Máquina / *Machine Learning.*

RL/LR - Regressão Linear Ridge / *Linear Regression*.

RB/BR - Regressão Bayesiana / *Bayesian Regression.*

AR/TR - Árvore de Regressão / *Tree Regression.*

RVS/SVR - Regressão por Vetores de Suporte / *Support Vector Regression.*

RIG/GBR - Regressão por Impulso de Gradiente / *Gradient Boosting Regression.*

ROE - *Return On Equity.*

P/L - Preço sobre lucro.

P/VP - Preço sobre valor patrimonial.

IA - Inteligência Artificial.

EQM - Erro Quadrático Médio.

NYSE - *New York Stock Exchange*.

NASDAQ - *National Association of Security Dealers Automated Quotations.*

API - *Application Programming Interface.*

EBITDA - *Earnings Before Interest Taxes Depreciation and Amortization.*

Sumário

Capítulo 1
Introdução

A constante transformação do mercado de capitais ao longo do tempo tornou esse campo cada vez mais complexo e dinâmico, tanto em relação à diversidade de operações possíveis quanto ao tipo de aplicação disponível, seja de renda fixa, variável ou derivativos. Com essa variedade de ativos e operações no mercado, surge o desafio de desenvolver estratégias eficazes para alocar capital, visando maximizar os rendimentos. Em geral, essas estratégias consideram o prazo de investimento, que pode ser de curto, médio ou longo prazo. Normalmente, quanto menor o tempo da estratégia, maior é o seu potencial de retorno, mas também o risco de perda aumenta proporcionalmente (LIU et al. 2020).

No mercado financeiro, a teoria dos mercados eficientes afirma que todas as informações disponíveis aos investidores são rapidamente absorvidas pelo mercado, impedindo que eles obtenham vantagens claras para lucros excepcionais (TITAN, 2015). Este trabalho, no entanto, contrapõe essa teoria, baseando-se em estudos de Irons (2007) e Ahsan (2012), que indicam que múltiplos de mercado, descritos matematicamente, podem fornecer informações relevantes para alcançar retornos acima da média. Entre os estudos que utilizaram essa abordagem, destacam-se aqueles que empregaram aprendizado de máquina como apoio na otimização de rendimentos. Por exemplo, Chenyao (2023) utilizou o modelo Random Forest para selecionar ações com maior potencial de retorno acima da média; já Zolotareva (2021) aplicou o XGBoost para desenvolver novas métricas explicativas dos ativos, auxiliando na escolha dos mesmos; e, por fim, Yuhan (2023) explorou diversos modelos de regressão para prever o preço futuro dos ativos, buscando uma seleção mais informada e eficiente.

Ainda no contexto dos estudos relacionados, é importante destacar que esta monografia se baseia no artigo de Mario Gambim, no qual o autor emprega modelos de aprendizado de máquina e regras fuzzy para desenvolver sistemas capazes de selecionar ativos e compor carteiras de investimento que superam o mercado. No entanto, esta monografia se diferencia do trabalho de Mario em alguns aspectos.

Primeiramente, em relação aos múltiplos escolhidos, embora ambos os trabalhos utilizem o P/L como um dos critérios para a construção da base de dados, este estudo adota o ROE e a Capitalização de Mercado, em vez do P/VP e do Volume Negociado. Em segundo

lugar, com o objetivo de aprimorar o desempenho dos modelos, esta monografia emprega ajuste de hiperparâmetros e compara seus efeitos com modelos não ajustados diferentemente do trabalho anterior. Por fim, ambos os trabalhos avaliam as carteiras desenvolvidas em relação ao S&P 500, como uma forma de medir o desempenho dos modelos em comparação com um índice de referência amplamente utilizado. No entanto, este estudo amplia essa análise ao incluir também a Teoria do Portfólio Moderno, proporcionando uma referência adicional para avaliar o quanto os modelos superam métodos tradicionais(índices - S&P 500) e sofisticados do mercado financeiro (teoria do portfólio moderno).

Para dar continuidade, dentre as diversas janelas de tempo para aplicações descritas anteriormente, as aplicações de longo prazo são de grande interesse dos operadores. Nesse tipo de modelagem, investidores buscam informações de mercado relacionadas ao preço dos ativos e saúde financeira das empresas por trás deles para escolher os ativos que podem oferecer maior potencial de rentabilidade. Ao contrário das teorias de médio e curto prazo, que em sua grande parte se baseiam nos movimentos de tendência das séries temporais dos retornos, bem como se amparam em tipos de operações de derivativos que em geral envolvem maior risco para o investidor, como por exemplo, opções e contratos futuros.

Neste trabalho, foi desenvolvida e testada uma metodologia que utiliza aprendizado de máquina para a seleção de ativos em uma carteira de investimentos com foco no longo prazo (1 ano). A principal fundamentação baseia-se na combinação da teoria de análise fundamentalista com a predição do ROE proposta por Ahsan (2012), que aponta o ROE como um indicador relevante para prever o desempenho financeiro. Assim, além de oferecer ao investidor uma forma simples e prática de selecionar ações para seu portfólio, com foco em ativos subvalorizados de acordo com a análise fundamentalista, a metodologia também permite inferir o retorno esperado dos ativos escolhidos por meio da análise do ROE.

Para alcançar esses objetivos, além de utilizar modelos de aprendizado de máquina na composição da carteira, este trabalho inclui uma segunda etapa de ajuste de hiperparâmetros. Espera-se que, com essa otimização, a performance dos modelos seja aprimorada, permitindo a seleção de carteiras ainda mais rentáveis devido à maior precisão dos modelos ajustados.

Como descrito anteriormente, a estratégia será baseada em duas teorias principais. A primeira, descrita por Irons (2007), baseada na análise fundamentalista em que os múltiplos P/L e P/VP são os necessários para decidir qual ativo colocar na carteira. Ambos tem o intuito de mostrar o quanto um ativo está sendo bem ou mal precificado no mercado, esta monografia contudo, tem enfoque apenas no múltiplo P/L, dado que o outro múltiplo a ser utilizado será abordado pela segunda teoria. O cálculo do P/L é feito dividindo o preço da

ação no momento pelo lucro que cada ação representa com base no lucro da empresa como um todo, em geral quanto menor o múltiplo P/L mais atrativo o papel, seria como comprar algo por um preço baixo e ter um lucro alto no produto. A segunda teoria analisada é a de Ahsan (2012), que indica o ROE como o melhor indicador para prever o retorno dos ativos. O ROE (Retorno sobre o Patrimônio Líquido) mede a eficiência de uma empresa em gerar lucro a partir dos recursos investidos pelos acionistas. Ele é calculado dividindo o lucro líquido pelo patrimônio líquido da empresa, resultando em um valor percentual. Por exemplo, um ROE de 15% indica que a empresa gera um lucro de 15 centavos para cada 1 real investido no patrimônio líquido. Além desses dois indicadores, será incluído o valor de mercado da empresa, que reflete seu tamanho no mercado. Esse valor é obtido multiplicando o preço da ação pelo número total de ações da empresa, resultando no valor financeiro da companhia. Essas três variáveis — P/L, ROE e valor de mercado — serão utilizadas como entradas nos modelos, cuja saída será o percentual de retorno previsto para cada ação após um período de um ano.

Em sequência, modelos de regressão foram treinados nas bases descritas anteriormente com o objetivo de montar as carteiras de investimento. Todos os atributos foram coletados na abertura de cada ano, com referência ao ano anterior, e as previsões foram realizadas para o fechamento do respectivo ano. Os modelos aplicados incluem: Regressão Linear Ridge (Regressão Linear com regularização L2), Árvore de Regressão, Regressão de Vetores de Suporte (SVR), Bayesian Ridge e Regressão por Impulso de Gradiente (GBR). Após a primeira etapa de treinamento, os modelos foram submetidos a uma nova rodada de treinamento, desta vez com os hiperparâmetros ajustados. Com as previsões de desempenho geradas, as ações foram selecionadas para compor as carteiras propostas por cada método. A avaliação dos modelos foi realizada comparando os valores previstos com os valores reais do ano analisado. Em uma primeira análise, foram comparadas as carteiras geradas sem e com o ajuste de hiperparâmetros, com o objetivo de evidenciar os ganhos obtidos por meio dessa otimização. Em uma segunda análise, as carteiras formadas pelos modelos com hiperparâmetros ajustados foram comparadas com a teoria do portfólio moderno de Harry Markowitz (1952) — um método matemático e estatístico amplamente utilizado no mercado financeiro até hoje — e com o índice de mercado americano S&P 500.

Os resultados apontaram que a montagem de carteiras por modelos de regressão utilizando aprendizado de máquina para prever o retorno de ativos em uma período de um ano tem uma performance superior, na maior parte dos casos, quando comparado a teoria do portfólio moderno e ao índice americano S&P 500.

Capítulo 2

Fundamentação Teórica

Neste capítulo serão tratados e explicados os conceitos necessários para o entendimento do trabalho, abordando os temas de alocação de carteira, mercado financeiro e aprendizado de máquina. Trabalhos que abordam problemas similares ao desta monografia também serão detalhados a seguir.

2.1 Alocação de Carteira

O principal objetivo de investir no mercado financeiro é, entre as diversas opções disponíveis, selecionar ativos que maximizem os lucros e minimizem os riscos. Essa prática é conhecida como montagem de carteira. Uma carteira de ativos nada mais é do que uma coleção de investimentos escolhidos por um indivíduo para alocar seu capital. No caso específico de ações, os investidores selecionam os papéis que consideram mais promissores para compor uma carteira exclusiva composta apenas por ativos dessa classe.

Atualmente diversas teorias que buscam explicar o funcionamento do mercado já foram desenvolvidas, essas teorias explicam como os agentes econômicos e ativos se comportam com base em ações do cotidiano e dados históricos. Tendo isso em vista, investidores formulam estratégias que podem ter como base essa teoria para cumprir o papel de maximizar o retorno e minimizar o prejuízo. Duas dessas estratégias são baseadas na Hipótese dos Mercados Eficientes e na Teoria do Portfólio Moderno que serão abordadas a seguir.

2.1.1 Hipótese dos Mercados Eficientes

Em 1970, Eugene Fama propôs a Hipótese dos Mercados Eficientes, afirmando que os preços dos ativos refletem integralmente todas as informações disponíveis no mercado. Assim, o preço de um ativo específico representa exatamente as informações conhecidas sobre ele. Essas informações incluem, principalmente, características da empresa

representada por esse ativo, como governança corporativa, nível de endividamento, setor de atuação, capital investido, margens de lucro, planos de expansão e saúde financeira.

Vale destacar que essa hipótese parte do pressuposto de que todos os participantes do mercado tomam decisões de forma completamente racional e avessa a riscos excessivos. Dessa forma, investidores não seriam influenciados por emoções ou decisões sem embasamento, devendo sempre escolher ativos com a convicção de que o mercado fornece o valor justo para cada um deles.

2.1.2 Teoria do Portfólio Moderno

A Teoria Moderna do Portfólio, desenvolvida por Harry Markowitz em 1952, é um marco na área de finanças que estabelece uma abordagem quantitativa para a construção de carteiras de investimentos. Essa teoria propõe que os investidores devem selecionar ativos considerando não apenas o retorno esperado, mas também o risco associado, medido pela variância ou desvio-padrão dos retornos. O principal objetivo é encontrar uma combinação de ativos que maximize o retorno esperado para um dado nível de risco ou, alternativamente, minimize o risco para um retorno esperado específico.

Uma das inovações centrais da teoria é o conceito de diversificação. Markowitz demonstrou matematicamente que, ao combinar ativos com diferentes níveis de risco e correlação, é possível reduzir o risco total da carteira sem comprometer o retorno esperado. Isso ocorre porque os ativos que não estão perfeitamente correlacionados tendem a compensar os movimentos negativos uns dos outros, diminuindo a volatilidade geral da carteira. Esse princípio leva à criação da chamada "fronteira eficiente", que representa o conjunto de carteiras otimizadas com a melhor relação entre risco e retorno.

As carteiras mais conhecidas baseadas na Teoria Moderna do Portfólio de Markowitz são a Máxima Razão de Sharpe e a Mínima Volatilidade. A carteira de Máxima Razão de Sharpe busca otimizar o retorno ajustado ao risco, ou seja, maximizar a relação entre o retorno excedente (acima da taxa livre de risco) e a volatilidade. Essa abordagem é ideal para investidores que desejam o melhor retorno possível para cada unidade de risco assumida. Já a carteira de Mínima Volatilidade tem como objetivo reduzir ao máximo a variabilidade dos retornos, sendo mais adequada para investidores avessos ao risco que priorizam a estabilidade sobre o retorno a longo prazo.

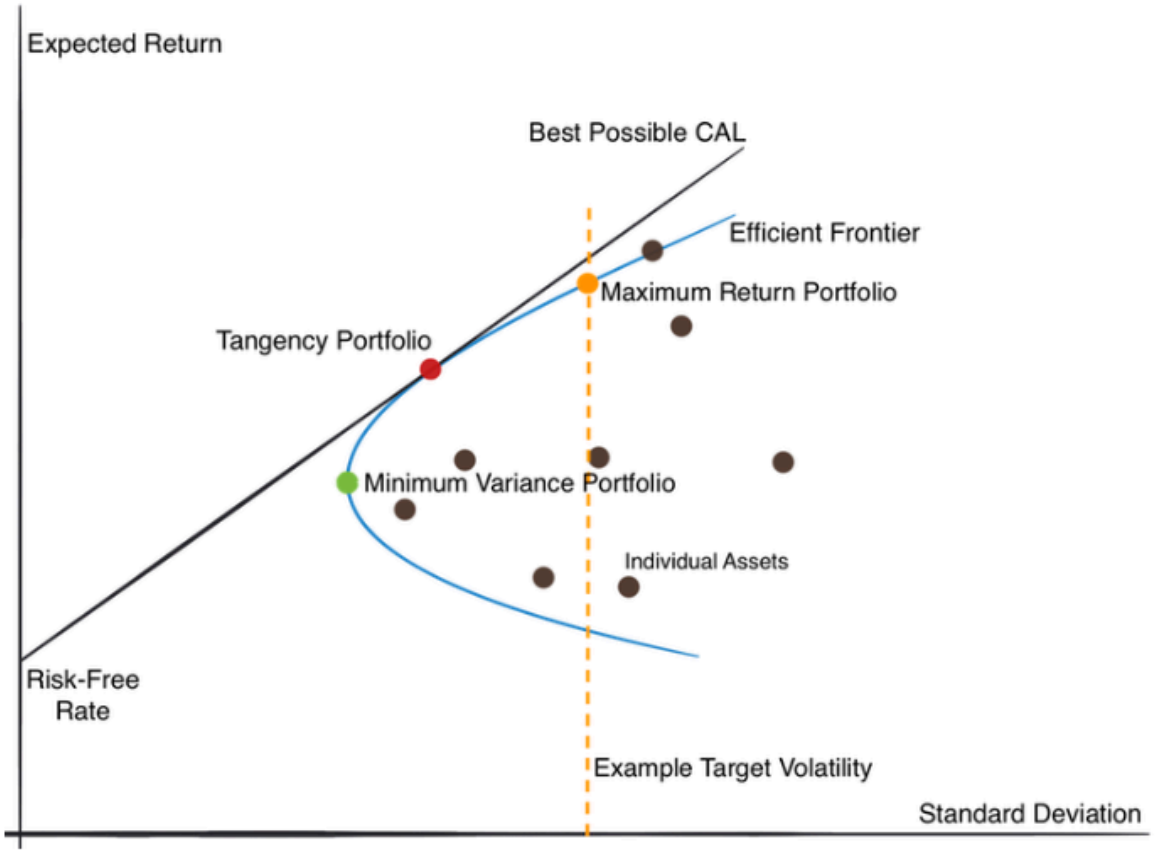

Figura 1 - Representação Gráfica da Fronteira Eficiente.
Fonte: Interactive Brokers.

2.1.3 Índice de ações

Índices no mercado de ações são indicadores que representam o desempenho de um grupo específico de ações negociadas em uma bolsa de valores. Eles são calculados com base no preço das ações das empresas que compõem o índice em tempo real, servindo como uma métrica para avaliar o desempenho geral de um setor, mercado ou economia.

Ademais, cada índice é composto por uma seleção de ações que atende a critérios específicos, como capitalização de mercado, liquidez ou setor econômico. Por exemplo, o S&P 500 mede o desempenho das 500 maiores empresas dos Estados Unidos, enquanto o Ibovespa representa as ações mais negociadas na Bolsa de Valores brasileira (B3). Os índices são amplamente utilizados por investidores como referência para analisar tendências de mercado, comparar o desempenho de carteiras de investimento e criar estratégias financeiras.

2.2 Inteligência Artificial

A inteligência artificial (IA) é um dos campos de estudo da ciência da computação, fundamentado em princípios racionais e empíricos. Essa área busca compreender como os seres humanos formulam seus pensamentos e tomam decisões, utilizando esse entendimento para modelar, de forma matemática e analítica, a estrutura de pensamento e ação. Nesse contexto, ferramentas de software são empregadas para viabilizar a modelagem e a aplicação dos modelos, enquanto avanços em hardware são utilizados para maximizar a eficiência do software, impulsionando continuamente a capacidade da inteligência artificial.

Ainda nesse tema, mais definições foram encontradas para o conceito de inteligência artificial, segundo Sheikh (2023) a IA nada mais é do que um tecnologia capaz de imitar diversas habilidades complexas de um ser humano que estão inerentemente conectadas com o pensamento inteligente, dentre essas habilidades podem ser destacadas:

- A compreensão de textos e falas para se comunicar.
- Armazenar informações de forma a estabelecer conexões lógicas entre elas.
- Ter capacidade de tirar conclusões das informações e dessa forma responder a novos questionamentos.
- Capacidade de identificar padrões e aplicá-los em situações novas.
- Através de dispositivos visuais identificar objetos e tomar decisões de movimento.

Os itens citados anteriormente e presentes na obra de Sheikh são campos de estudos da inteligência artificial já amplamente estudados atualmente.

Historicamente, o primeiro estudo científico focado em inteligência artificial foi proposto por Alan Turing em 1950, no trabalho *Computing Machinery and Intelligence*. No entanto, os pesquisadores da época enfrentavam grandes dificuldades para implementar suas teorias devido às limitações tecnológicas. Os processadores eram extremamente rudimentares e levaram décadas para se tornarem suficientemente potentes, enquanto a computação massivamente paralela ainda era inexistente. Além disso, a capacidade de armazenamento era outro obstáculo significativo: os sistemas não conseguiam lidar com grandes volumes de dados nem suportar algoritmos complexos necessários para viabilizar as técnicas de IA de forma prática.

Atualmente, ambas essas barreiras foram quebradas e a área tem experienciado um crescimento exponencial, com a habilidade de propor algoritmos e testá-los na grande maioria dos sistemas computacionais atuais. A IA está cada vez mais presente na vida

cotidiana de pessoas, pesquisadores e companhias. Um dos campos de estudo mais ativos dentro da IA é o de aprendizado de máquina, que como faz parte dessa monografia terá seus conceitos mais relevantes detalhados nas seguintes seções.

2.3 Aprendizado de Máquina

Os sistemas inteligentes devem ser capazes de utilizar dados passados como base para suas escolhas, aprimorando a tomada de decisões futuras de forma mais precisa e confiável. O aprendizado de máquina, uma das áreas de estudo dentro da inteligência artificial, permite que computadores aprendam com dados, melhorando seu desempenho sem a necessidade de programação explícita para cada tarefa. Por meio de algoritmos, esses sistemas computacionais são capazes de analisar dados, detectar padrões e, com isso, tomar decisões e realizar previsões de maneira autônoma e eficiente (ALAM, 2023).

De forma geral, os algoritmos de aprendizado de máquina geram uma solução genérica capaz de se adaptar a diversas instâncias de um mesmo problema, mantendo uma performance consistente. O processo de aprendizado das máquinas é chamado de treinamento do modelo. Após ser treinado, o modelo torna-se capaz de generalizar respostas para novos dados que não foram apresentados durante a etapa de treinamento. Com base nos dados anteriores, o modelo consegue identificar correlações com o novo conjunto de dados e, assim, realizar previsões para a nova tarefa (FRANÇA et al. 2021).

Os modelos de aprendizado de máquina podem lidar tanto com dados estruturados (em formato tabular) quanto com dados não estruturados (como textos, imagens e estímulos sonoros). Nesta monografia, destaca-se que os dados utilizados estão estruturados em formato tabular. Portanto, as seções seguintes abordarão os conceitos de aprendizado de máquina supervisionado e não supervisionado, considerando que ambos são adequados para o formato dos dados analisados.

2.3.1 Aprendizado de Máquina não supervisionado

De modo geral, trabalhos de aprendizado de máquina utilizam diversos atributos de um conjunto de dados para chegar a uma conclusão sobre o estado de um único atributo, conhecido como atributo meta. Esse atributo é responsável por definir o rótulo de cada

instância no conjunto de dados. Sendo assim, o aprendizado não supervisionado é uma técnica que dispensa dados rotulados, permitindo o treinamento do modelo sem qualquer orientação ou supervisão. Nesse contexto, o modelo classifica o conjunto de dados em várias categorias ao identificar semelhanças entre eles. Esse tipo de aprendizado opera exclusivamente com os dados de entrada, sem receber instruções prévias, desenvolvendo sua própria interpretação e organização dos dados de forma autônoma (EWUSIE et al. 2022).

Uma das principais tarefas realizadas por esse tipo de aprendizado é o agrupamento (*clustering*). Os algoritmos analisam as relações entre os dados e identificam padrões que permitam agrupar as instâncias, mesmo que grupos pré-definidos não existam. Um exemplo interessante do funcionamento do agrupamento pode ser encontrado nos estudos de Potashev (2014), em que o autor utiliza estatísticas de tolerância à contaminação do solo por hidrocarbonetos durante a fase de germinação para agrupar plantas e definir suas espécies com os grupos formados.

2.3.2 Aprendizado de Máquina supervisionado

O aprendizado supervisionado diferencia-se do aprendizado não supervisionado pela presença do atributo meta no conjunto de dados, o que os caracteriza como rotulados. Nessas bases de dados, há um ou mais atributos de entrada e um atributo de saída, conhecido como rótulo. Os modelos de aprendizado supervisionado buscam identificar funções que relacionem os atributos de entrada aos de saída, de modo a mapeá-los adequadamente (HASAN et al. 2021). A função escolhida como modelo é aquela que apresenta maior precisão ao realizar esse mapeamento. Para validar se a função selecionada é realmente a ideal, o modelo é testado com um novo conjunto de dados, diferente daquele utilizado no treinamento. Considera-se que o modelo está devidamente calibrado se as previsões para as saídas desse novo conjunto estiverem de acordo com os valores reais observados.

O aprendizado de máquina supervisionado destaca-se em dois tipos principais de problemas. O primeiro é a classificação, onde o rótulo das instâncias assume valores conhecidos, categóricos e em número limitado, como profissão, nível de escolaridade ou estado de saúde. O segundo é a regressão, em que o rótulo possui um valor numérico contínuo, como altura, peso ou volume.

Este trabalho busca prever preços de ativos, sendo assim, dados numéricos contínuos. A seguir será abordado o problema de regressão e os métodos utilizados por essa monografia para lidar com ele.

2.3.3 Hiperparâmetros

Hiperparâmetros são parâmetros que controlam o comportamento de um modelo de aprendizado de máquina, mas que não são aprendidos diretamente a partir dos dados. Ao contrário dos parâmetros do modelo (como os pesos em redes neurais), que são ajustados durante o treinamento, os hiperparâmetros precisam ser definidos antes do treinamento começar. Eles incluem configurações como a taxa de aprendizado, o número de camadas em uma rede neural, o número de árvores em uma floresta aleatória, entre outros. A escolha desses valores pode impactar significativamente a performance do modelo (ARNOLD et al. 2024).

O processo de selecionar os melhores hiperparâmetros é conhecido como *tuning* ou ajuste de hiperparâmetros. Isso pode ser feito por meio de técnicas como pesquisa em grade (*grid search*), onde várias combinações de hiperparâmetros são testadas, ou pesquisa aleatória (*random search*), que explora combinações de forma mais ampla. O objetivo é encontrar a combinação de hiperparâmetros que maximize o desempenho do modelo, de modo que ele consiga fazer previsões mais precisas e generalize bem para dados novos.

Como discutido na seção de introdução, os hiperparâmetros e o ajuste de hiperparâmetros serão utilizados com o objetivo de aprimorar a performance dos modelos de regressão, permitindo, assim, maximizar ainda mais o retorno das carteiras.

2.4 Regressão

Como já discutido anteriormente, a regressão é uma das técnicas utilizadas no aprendizado de máquina supervisionado com o objetivo de mapear dados de entrada para um dado de saída, em que a saída é um valor numérico e contínuo. Existem diversos algoritmos em AM que visam solucionar esse problema, cada um deles pode ter sua performance afetada pelo tipo, tamanho, variação e estrutura dos dados.

A seguir serão destacados os modelos de regressão utilizados neste trabalho.

2.4.1 Regressão Linear

Os modelos lineares são amplamente reconhecidos no campo da regressão. Sua principal premissa é que as saídas estão linearmente relacionadas às entradas, ou seja, espera-se que a variável alvo seja uma combinação linear dos dados de entrada. A regressão linear pode ser representada em um plano cartesiano, onde o eixo x representa as variáveis de entrada e o eixo y, as de saída. Quando há apenas uma variável de entrada relacionada diretamente a uma variável de saída, o método é classificado como regressão linear simples. Já quando há múltiplas variáveis de entrada, todas relacionadas a uma única saída, o método é denominado regressão linear múltipla.

Além disso, as regressões lineares podem ser regularizadas para evitar o *overfitting*. Duas formas amplamente utilizadas de regularização são a regressão Ridge (regularização L2) e a regressão Lasso (regularização L1). A regressão Ridge limita o tamanho dos coeficientes ao introduzir um termo de penalidade proporcional ao seu valor, enquanto a regressão Lasso reduz os coeficientes, podendo até zerar alguns deles, o que facilita a seleção de variáveis. Existem ainda outros modelos de regularização, sendo a escolha da técnica mais adequada feita geralmente por meio de validação cruzada (QU, 2024).

2.4.2 Regressão Bayesiana

Como o próprio nome salienta, a regressão Ridge Bayesiana utiliza os princípios da regularização L2 discutida anteriormente em conjunto com a inferência Bayesiana para estimar os coeficientes da regressão de forma probabilística. Em vez de ajustar os coeficientes deterministicamente, como na regressão tradicional, essa abordagem considera distribuições de probabilidade tanto para os coeficientes quanto para os hiperparâmetros de regularização.

Essa técnica é especialmente útil em cenários com poucos dados ou multicolinearidade, pois evita o *overfitting* e proporciona uma modelagem mais robusta. Além disso, a Bayesian Ridge permite analisar a incerteza dos coeficientes, fornecendo mais informação sobre o comportamento do modelo e suas previsões. É amplamente usada em áreas onde a incerteza é relevante, como finanças, medicina e modelagem preditiva de sistemas complexos, especialmente em situações de multicolinearidade ou com amostras limitadas (EFENDI, 2017).

2.4.3 Árvores de Decisão e de Regressão

A árvore de regressão é uma adaptação do modelo de árvore de decisão, amplamente utilizado para problemas de classificação, e é especialmente projetada para lidar com problemas de regressão. Esse modelo funciona criando partições do problema original de forma recursiva em subproblemas menores, seguindo uma abordagem semelhante ao método de divisão e conquista (ROKACH et al. 2005).

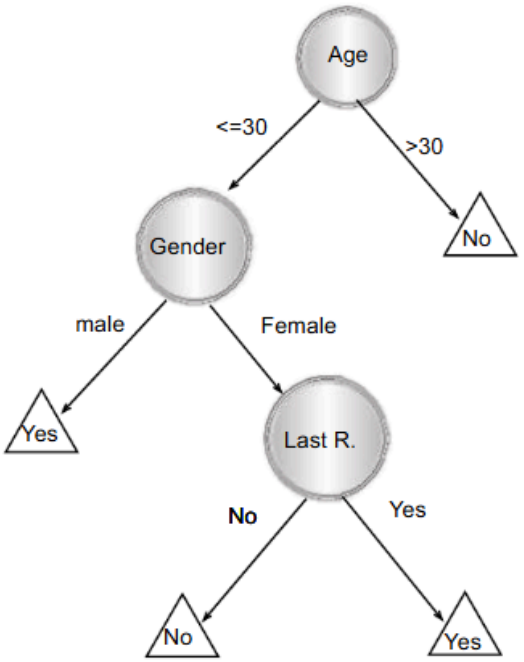

Figura 2 - Árvore de Decisão.
Fonte: (ROKACH et al. 2005).

A figura 2 destaca a estrutura de uma árvore de decisão, composta por diferentes tipos de nós. No topo, encontra-se o nó raiz, que não possui nenhum nó acima dele e é o ponto de partida da árvore, que é unicamente direcionada. Os demais nós, localizados ao longo da árvore, são chamados de nós internos, com atenção especial aos últimos nós, conhecidos como nós folha ou nós de decisão, por não possuírem filhos.

Cada nó da árvore é dividido em dois ou mais filhos, representando subespaços cada vez mais específicos do problema principal. Essas divisões são realizadas com base em uma

função discreta aplicada a um atributo de entrada. A partir desse atributo, é gerada uma regra que define as separações, permitindo que o modelo refine progressivamente a solução do problema.

Como mencionado anteriormente, a árvore de decisão trabalha com um conjunto de dados de entrada, que podem ser valores discretos ou contínuos, e retorna uma única saída localizada nos nós folha, que representam a decisão final da árvore. No contexto desta monografia, o foco está no cenário em que os nós folha fornecem uma resposta numérica, essencial para resolver problemas de regressão.

2.4.4 Regressor de Vetores de Suporte

O modelo SVR (*Support Vector Regressor*) é a versão utilizada para problemas de regressão do famoso modelo SVM (*Support Vector Machine*), amplamente empregado para tarefas de classificação. No caso do SVR, o objetivo é encontrar uma linha (ou um hiperplano, no caso de dados em mais de duas dimensões) que se ajuste aos dados de forma que a maior quantidade possível de pontos esteja próxima dessa linha, sem cometer grandes erros.

Nesse contexto, em vez de minimizar o erro de treinamento observado, o SVR busca minimizar o limite do erro de generalização, com o objetivo de alcançar uma performance mais robusta e generalizada (BASAK et al. 2007). Ou seja, ele cria uma "faixa" ao redor da linha de previsão onde erros são permitidos. Os pontos que ficam fora dessa faixa são chamados de "vetores de suporte", pois são esses pontos que ajudam a definir a posição da linha de regressão.previsão onde os erros são tolerados.

2.4.5 Regressor por Impulso de Gradiente

O regressor XGBoost, proposto por Tianqi Chen em 2014, é a forma abreviada de *Extreme Gradient Boosting Regressor*. Como o nome sugere, esse modelo constrói um conjunto de árvores de decisão de forma sequencial, onde cada nova árvore busca corrigir os erros cometidos pelas árvores anteriores. Para alcançar esse objetivo, o XGBoost utiliza a técnica de gradiente descendente para minimizar uma função de perda previamente definida.

O algoritmo de Chen começa com uma predição inicial, que geralmente corresponde à média dos valores no caso de problemas de regressão. Em seguida, árvores de decisão

baseadas nos resíduos (a diferença entre as previsões atuais e os valores reais) são criadas iterativamente. Cada nova árvore ajusta os pesos das observações com base nos gradientes calculados, corrigindo os erros das árvores anteriores e aprimorando progressivamente o modelo.

2.5 Medidas de erro e testes estatísticos

As medidas de erro dos modelos e testes estatísticos dos dados são formas encontradas para se avaliar as regressões, elas medem o quão bem o modelo se adequou aos dados fornecidos, informando assim se as predições feitas realmente estão de acordo com os dados reais. A seguir se tem uma métrica de erro apresentada, Erro Quadrático Médio (EQM), além dos testes estatísticos, t de Student e Friedman.

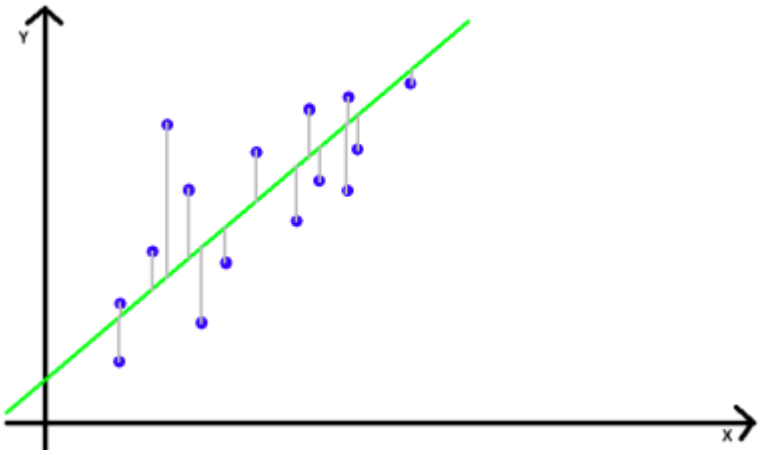

Figura 3 - Desvio dos pontos em uma regressão linear.
Fonte: (VARGAS JUNIOR, 2020).

2.5.1 Erro Quadrático Médio

Mede a média dos quadrados das diferenças entre os valores preditos pelo modelo e os valores reais, fornecendo uma indicação da precisão do modelo. Sua fórmula pode ser visualizada da seguinte forma:

$$MSE = 1/n \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

Em que:

- y_i : valor real da variável de saída no ponto i.
- $\hat{y}_i$: valor predito pelo modelo no ponto i.
- n : número total de instâncias no conjunto de dados.

Quanto menor o valor do EQM (em inglês *MSE - Mean Squared Error*) melhor o desempenho do modelo, isso indica que os valores previstos estão mais próximos dos reais. Contudo, por ser uma métrica baseada em operações quadráticas ela penaliza erros maiores mais significativamente do que erros pequenos, o que a torna útil para identificar modelos que apresentam grandes discrepâncias em algumas predições.

2.5.2 Teste de Friedman

O teste de Friedman é um método estatístico não paramétrico usado para comparar três ou mais amostras relacionadas, especialmente quando as suposições de normalidade dos dados não são atendidas. Ele é utilizado em situações onde os mesmos sujeitos ou itens são avaliados sob diferentes condições ou tratamentos, analisando se há diferenças significativas entre elas. O teste atribui ranks aos valores observados dentro de cada grupo, o que minimiza o impacto de outliers e distribuições assimétricas, tornando-o ideal para dados não paramétricos.

Os resultados do teste de Friedman indicam se há diferenças significativas entre as condições analisadas, mas não especificam quais grupos são diferentes. Caso o valor-p seja menor que o nível de significância (geralmente 0,05), rejeita-se a hipótese nula, sugerindo que pelo menos uma condição se destaca. No caso deste trabalho, o teste será utilizado para comparar os modelos de regressão e mercado entre si, utilizando como referência as três medições para cada método realizadas no cenário em que as janelas são de um ano.

2.5.3 Teste t de Student

O teste t de Student pareado é um teste estatístico utilizado para comparar duas amostras relacionadas, ou seja, quando cada observação de um grupo tem um correspondente direto no outro grupo. Ele avalia se a diferença média entre essas amostras é estatisticamente significativa, assumindo que os dados seguem uma distribuição aproximadamente normal. Esse teste é útil para medir o impacto de uma intervenção ou modificação em um experimento, pois analisa as diferenças individuais em vez de apenas comparar médias gerais.

No contexto da comparação entre um modelo de machine learning com e sem ajuste de hiperparâmetros, o teste t pareado pode ser utilizado para avaliar se a otimização dos hiperparâmetros realmente melhora o desempenho. Para isso, calcula-se a métrica de interesse (como erro médio ou acurácia) para os mesmos conjuntos de teste em ambos os experimentos e, em seguida, aplica-se o teste t pareado às diferenças dessas métricas. Se o resultado indicar uma diferença estatisticamente significativa, pode-se concluir que o ajuste de hiperparâmetros teve um impacto real na performance do modelo, e não apenas uma variação aleatória nos dados.

2.6 Trabalhos Relacionados

Primeiramente, destaca-se o trabalho de Rasekhschaffe (2019), que emprega técnicas de *machine learning* (ML) no campo das finanças quantitativas com o objetivo de prever os retornos de ativos financeiros. Este estudo não apenas explora o potencial dessas técnicas na identificação de padrões e relações complexas nos dados, mas também oferece uma importante contribuição ao abordar estratégias para lidar com desafios críticos, como o tratamento de dados financeiros e o ajuste adequado dos modelos para evitar o *overfitting*. Além disso, o autor enfatiza a importância de balancear a capacidade dos modelos de capturar relações não lineares e contextuais, sem comprometer sua generalização, proporcionando conclusões valiosas para a aplicação prática de ML em investimentos.

Já o trabalho de Wolff (2023) adota uma abordagem distinta, ao invés de buscar prever o valor numérico dos retornos dos ativos por meio de regressão. O autor utilizou dados semanais dos constituintes históricos do S&P 500, no período de 1999 a 2021, incorporando fatores de ações, fundamentos financeiros e indicadores técnicos. Os modelos desenvolvidos foram treinados para realizar uma classificação binária, prevendo se uma ação superaria ou

ficaria abaixo da mediana de retorno semanal. Essa estratégia, focada em uma classificação simples, demonstrou resultados superiores a diversos modelos complexos de regressão. Além disso, a metodologia foi replicada para o índice STOXX Europe 600, onde também apresentou resultados robustos e positivos, reforçando sua eficácia no contexto da seleção de ações.

Para finalizar, o estudo de Gambim (2022) utilizou métodos de *machine learning* para elaborar sugestões de carteiras de ativos financeiros, com destaque para a aplicação de regras fuzzy, que permitiram estabelecer relações categóricas entre as variáveis P/L, P/VP e volume negociado com o retorno dos ativos. Essa abordagem criou uma conexão clara e estruturada entre as variáveis de entrada e saída do modelo. A metodologia proposta sugere carteiras de investimento para o período de um ano, tendo como base o índice S&P500. Os resultados obtidos superaram índices tradicionais de mercado, além de métodos manuais, destacando-se como uma contribuição relevante para o uso de inteligência artificial na otimização de carteiras.

Capítulo 3
Materiais e Métodos

Neste capítulo, será detalhado o processo de formação das carteiras de ações utilizando aprendizado de máquina, baseado nos multiplicadores P/L, ROE e no valor de mercado, conforme descrito no Capítulo 1.

O objetivo principal é propor modelos de regressão para a composição de carteiras de ações, ajustar seus hiperparâmetros para alcançar um desempenho ainda mais eficiente e, por fim, comparar essa metodologia com a teoria do portfólio moderno de Harry Markowitz e com o índice de mercado S&P 500.

3.1 Fonte dos Dados

Os dados selecionados para este trabalho correspondem às ações das 300 maiores empresas listadas no índice americano S&P 500. Esse índice é composto, em sua totalidade, por 500 ações negociadas nas bolsas de valores de Nova York, a NYSE e a NASDAQ. Os ativos que integram o S&P 500 são escolhidos com base no tamanho de mercado e na liquidez. No entanto, a cada trimestre o índice sofre um rebalanceamento, no qual novas ações podem entrar ou sair do índice, portanto neste trabalho só foram analisados os ativos que permaneceram no índice durante todo o período, cobrindo o início de 2020 até o final de 2023.

Como já discutido no capítulo um, as variáveis a serem analisadas para cada ação são:

- ROE: Retorno da empresa sobre o patrimônio líquido.
- P/L: Preço de cada ação dividido pelo lucro.
- Capitalização de Mercado: valor da empresa na visão do mercado, em dólares.
- Retorno: O quanto a empresa retornou para os acionistas no período de um ano.

Todos os dados utilizados foram extraídos da API do Yahoo Finance, acessada por meio da biblioteca *yfinance* em Python. Essa biblioteca disponibiliza informações completas sobre as empresas, permitindo, a partir das fórmulas dos indicadores apresentadas anteriormente, calcular seus valores numéricos para cada ano.

3.2 Análise Exploratória

Essa etapa tem como objetivo compreender o comportamento dos dados a serem analisados, com foco principal em duas questões: o grau de dispersão dos dados e a existência de correlações significativas entre eles.

O foco desta fase foi analisar as métricas dos dados ano a ano, a fim de verificar se, nos três anos utilizados para o treinamento dos modelos, os dados apresentaram um comportamento minimamente semelhante. Primeiramente, foi realizada uma análise de correlação entre as variáveis de entrada e a variável de saída para cada ano, representada graficamente por meio de um *heatmap*. Em seguida, o nível de dispersão dos retornos dos ativos ao longo dos anos foi visualizado através de histogramas.

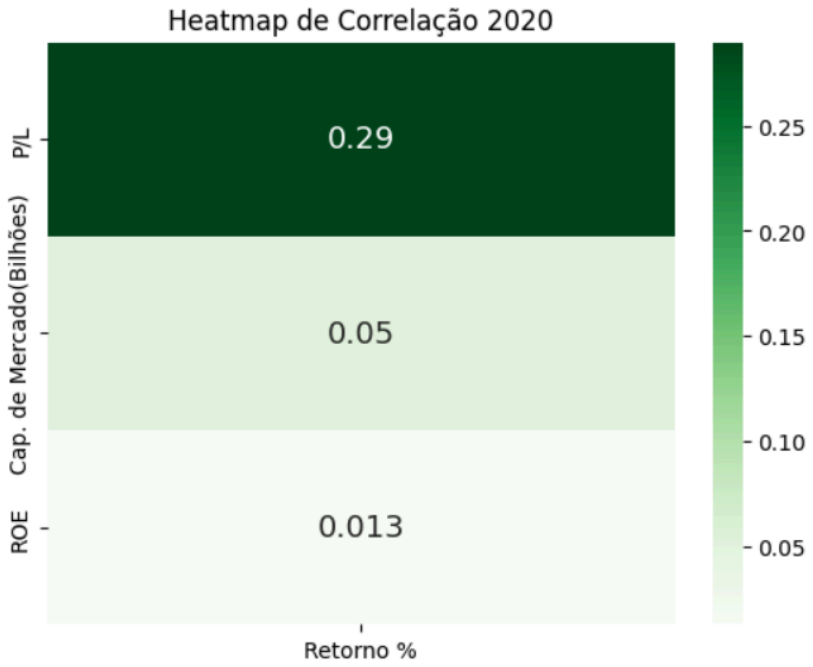

Figura 4 - Heatmap de Correlação para 2020.

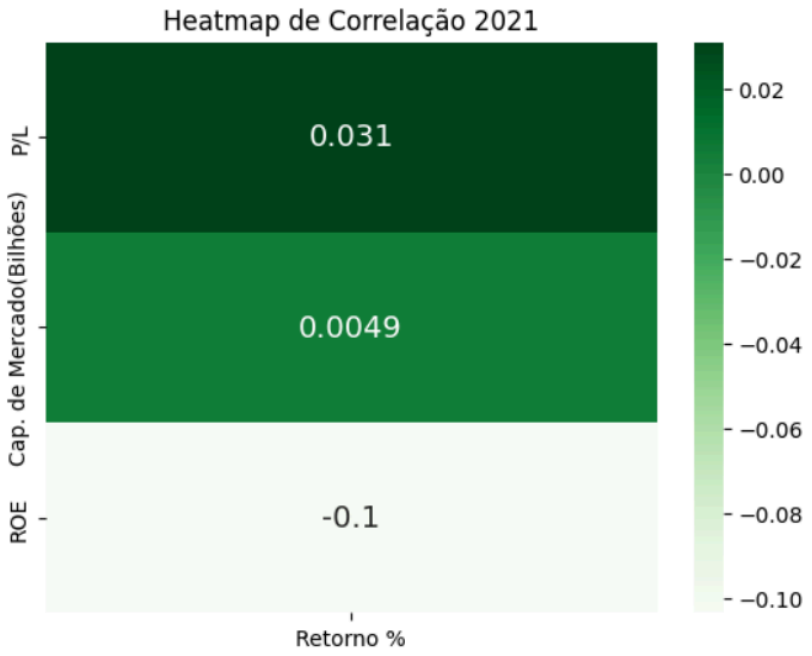

Figura 5 - Heatmap de Correlação para 2021.

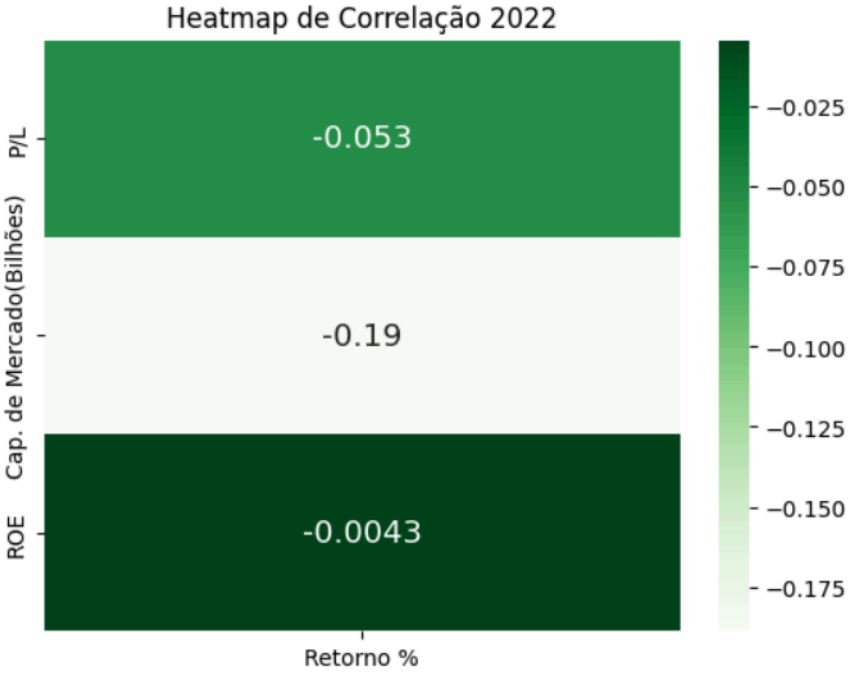

Figura 6 - Heatmap de Correlação para 2022.

As figuras acima revelam que o perfil de correlação dos anos de 2020 e 2021 apresenta certa semelhança, enquanto as correlações observadas no ano de 2022 diferem significativamente. Diante disso, é possível inferir que os dados utilizados para o treinamento em 2022 e as previsões realizadas em 2023 possam apresentar variações na performance dos modelos. A seguir, os histogramas para a análise de dispersão.

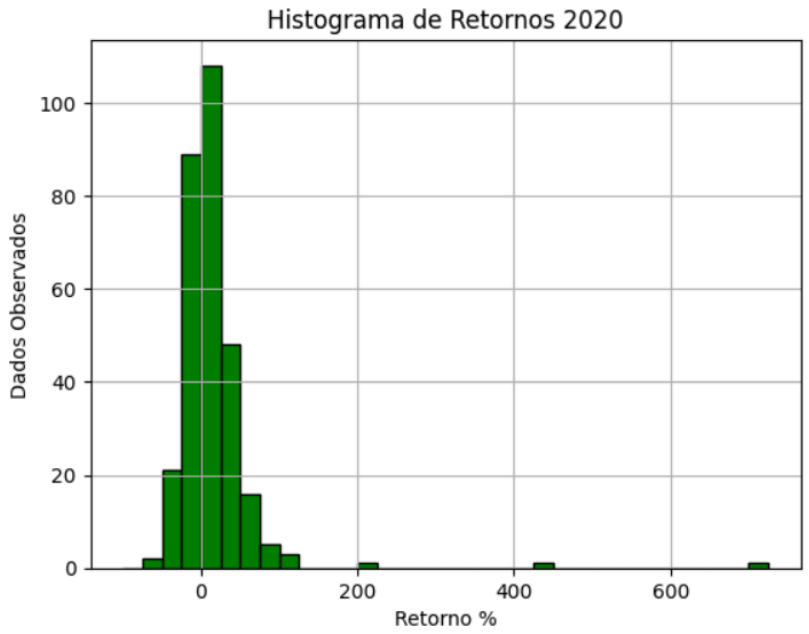

Figura 7 - Histograma de Retornos para 2020.

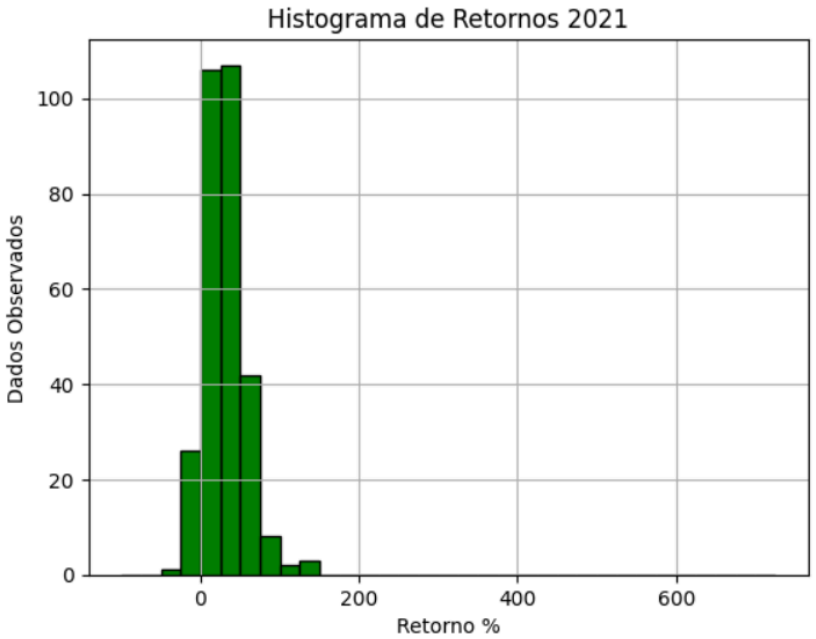

Figura 8 - Histograma de Retornos para 2021.

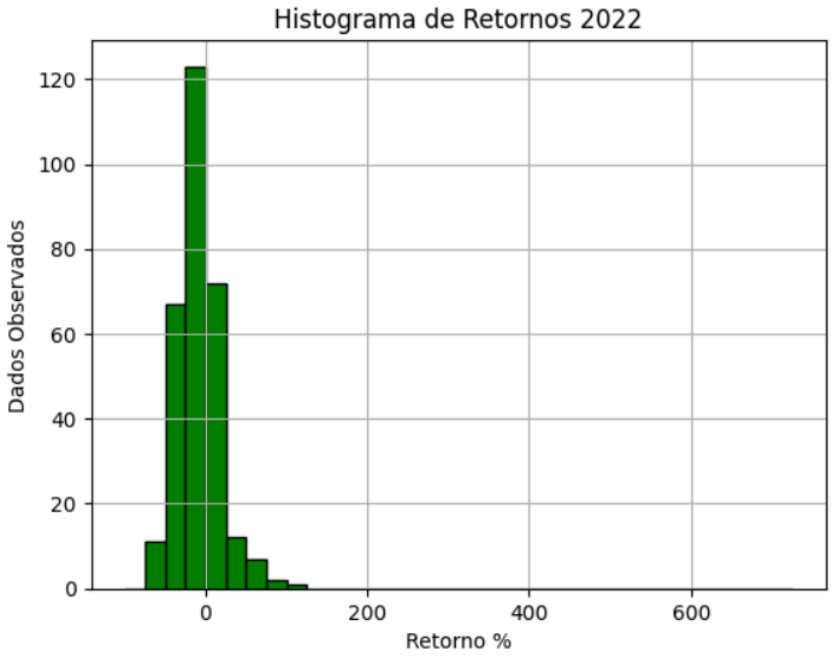

Figura 9 - Histograma de Retornos para 2022.

Em relação aos perfis anuais de dispersão dos retornos, observa-se que todos apresentam características bastante semelhantes, com a maioria das amostras concentradas entre -50% e 150%. No entanto, dois pontos relevantes merecem destaque. Primeiro, o histograma de 2020 exibe dois outliers significativos, com retornos expressivos na ordem de 400% e 700%. Segundo, o histograma de 2022 revela uma diferença marcante em relação aos outros anos: cerca de dois terços dos ativos registraram retornos negativos, indicando um declínio no preço das ações nesse período.

3.3 Limpeza e tratamento de dados

Após a obtenção dos dados por meio da API do *Yahoo Finance*, é necessário submetê-los a uma etapa de tratamento e limpeza. Essa etapa envolve a aplicação de operações nos conjuntos de dados para torná-los adequados à divisão em conjuntos de treino e teste, permitindo sua posterior utilização nos modelos.

Para começar pode ser feito o detalhamento de como cada conjunto e cada atributo foi gerado para se encaixar na base de dados. Foram gerados no total quatro conjuntos de dados referentes aos anos de 2020, 2021, 2022 e 2023, em cada conjunto os atributos presentes foram obtidos da seguinte forma:

- ROE: Lucro obtido pela empresa representante da ação durante todo o período de um ano dividido pelo valor de seu patrimônio líquido ao final do ano (dezembro).
- P/L: Dado retirado tendo como referência o início do ano em análise, obtido através da divisão do preço pelo lucro no próprio mês de janeiro.
- Capitalização de Mercado: Multiplicação entre a quantidade de cotas de ações existentes pelo preço das mesmas no mês de janeiro.
- Desempenho: Valor percentual obtido através da subtração do preço do ativo no fechamento (dezembro) pela abertura (janeiro).

A biblioteca Pandas, disponível na linguagem Python, foi a escolhida para fazer o tratamento dos conjuntos de dados.

Dessa forma, as bases foram estruturadas para cada ano no formato tabular, contendo cinco colunas: o nome do ativo e as quatro variáveis descritas anteriormente. Em seguida, os valores nulos foram removidos das bases, resultando no conjunto de dados ilustrado na figura abaixo.

	Ativo	ROE	P/L	Cap. de Mercado(Bilhões)	Retorno %
0	AAPL	0.878664	39.237197	1274.742718	76.713834
1	MSFT	0.431522	27.294739	1207.701743	38.475909
2	AMZN	0.228374	76.800701	954.699004	71.597095
3	GOOGL	0.180948	29.387895	924.162821	28.053316
4	META	0.227188	26.701189	597.663217	30.212606
...	...	...	...	...	...
290	GNRC	0.251263	40.915168	6.411390	122.929124
291	BBY	0.391977	14.263643	22.438669	14.293900
292	CHRW	0.269382	24.893426	10.452414	20.609020
293	WYNN	5.872905	-5.888533	15.492766	-21.427579
294	HPQ	-1.263465	11.390892	27.110161	18.278014

295 rows × 5 columns

Figura 10: Base de dados final 2020.

Com as bases de dados criadas, foi possível definir duas perspectivas para compor os modelos de treino e teste. A primeira perspectiva considera que os dados de treino e teste correspondem a períodos de um ano. Nesse caso, as 295 instâncias do ano de 2020 foram utilizadas para treinar o modelo, enquanto as predições foram realizadas com base nas 295 instâncias do ano de 2021.

Já em uma segunda ótica, se analisa uma base de dados maior, abrangendo um período de tempo mais longo. Nesse cenário, as bases de 2020, 2021 e 2022 foram concatenadas, formando uma única base com 885 instâncias para o treinamento. O teste, por sua vez, foi realizado utilizando os dados de um único ano, 2023. Esse segundo viés, com um recorte temporal mais extenso, busca avaliar se os modelos apresentam melhor ou pior desempenho quando treinados com um maior volume de dados e uma janela de tempo mais ampla.

3.4 Modelos Gerados sem Hiperparametrização

Como mencionado anteriormente, os modelos selecionados para a previsão dos retornos foram: Regressão Linear Ridge (Regressão Linear com regularização L2), Árvore de Regressão, Regressão de Vetores de Suporte (SVR), Bayesian Ridge e Regressão por Impulso de Gradiente (GBR). A implementação foi realizada na linguagem Python utilizando a biblioteca Scikit-learn.

Inicialmente, os modelos foram treinados com o conjunto de dados de um ano e testados com os dados do ano subsequente, no caso do conjunto concatenado de 3 anos os modelos foram aplicados da mesma forma. Nesta etapa, todos os métodos de ML foram aplicados utilizando as configurações padrão da biblioteca, sem nenhum tipo de ajuste prévio, posteriormente tiveram suas métricas retiradas.

3.5 Modelos Gerados com Hiperparametrização

Os modelos foram otimizados por meio da técnica de hiperparametrização conhecida como *grid search*. Essa abordagem realiza uma busca exaustiva por todas as combinações possíveis de um conjunto de hiperparâmetros definidos pelo usuário, avaliando o desempenho do modelo para cada combinação. A combinação que gera o melhor desempenho é selecionada como a ideal para o modelo.

Para cada ano, os modelos foram treinados e, a cada iteração, os hiperparâmetros foram ajustados conforme o conjunto de dados (ano de treinamento) e o tipo de modelo utilizado. Os hiperparâmetros e os valores selecionados para o *grid search* podem ser visualizados na tabela a seguir.

Método	Hiperparâmetro	Valores de Busca
Regressão Linear Ridge	Alpha	[0.1, 1, 10, 100, 1000, 10000]
Regressão de Vetores de Suporte (SVR)	C	[0.1, 1, 10, 100]
Regressão de Vetores de Suporte (SVR)	Epsilon	[0.01, 0.1, 0.5, 1]
Regressão de Vetores de Suporte (SVR)	Kernel	[linear, rbf]
Regressão de Vetores de Suporte (SVR)	Gamma	[scale, auto]
Árvore de Regressão	Max Depth (MD)	[3, 5, 10]
Árvore de Regressão	Min Sample Splits (MSS)	[2, 5, 10]
Árvore de Regressão	Min Samples Leafs (MSL)	[1, 2, 4]
Bayesian Ridge	Alpha 1	[0.1, 1, 10, 100]
Bayesian Ridge	Alpha 2	[0.1, 1, 10, 100]
Bayesian Ridge	Lambda 1	[0.1, 1, 10, 100]
Bayesian Ridge	Lambda 2	[0.1, 1, 10, 100]
Regressão por Impulso de Gradiente (GBR)	Number Estimators (NE)	[50, 100, 200]
Regressão por Impulso de Gradiente (GBR)	Learning Rate (LR)	[0.01, 0.05, 0.1]
Regressão por Impulso de Gradiente (GBR)	Max Depth (MD)	[3, 5, 7]

Tabela 1: Especificação dos hiperparâmetros.

Vale ressaltar que como o *grid search* é um método de ajuste de hiperparâmetro por busca exaustiva, ele pode se tornar bastante ineficiente para grades de alta dimensionalidade. A eficiência do *grid search* pode ser aproximada como o número de combinações de todos os parâmetros a serem otimizados, esse valor é obtido multiplicando a quantidade de valores

escolhidos para cada hiperparâmetro. No caso da tabela acima temos a seguinte organização de grades:

- Regressão Linear Ridge (RL): Grade de 6 entradas, 1 hiperparâmetro 6 valores, total de 6 combinações.
- Regressão de Vetores de Suporte (SVR): Grade de 12 entradas, 4 hiperparâmetros 2 com 4 valores e 2 com 2 valores, total de 64 combinações.
- Árvore de Regressão (AR): Grade de 9 entradas. 3 hiperparâmetros cada um com 3 valores, total de 27 combinações.
- Bayesian Ridge (BR): Grade de 16 entradas, 4 hiperparâmetros cada um com 4 valores, total de 256 combinações.
- Regressão por Impulso de Gradiente (GBR): Grade de 9 entradas. 3 hiperparâmetros cada um com 3 valores, total de 27 combinações.

Além disso, destaca-se que este método não encontra os hiperparâmetros ótimos para cada modelo e cada conjunto de dados, ele apenas maximiza a performance do modelo dentro dos valores de hiperparâmetro especificados para compor as grades.

A tabela 2 apresenta os resultados do *grid search* para todos os modelos e dados.

Modelo / Ano	**2020**	**2021**	**2022**	**2020 + 2021 + 2022**
RL	{ Alpha = 10000 }	{ Alpha = 10000 }	{ Alpha = 10000 }	{ Alpha = 10000 }
SVR	{ C = 10000; Epsilon = 1; Gamma = scale; Kernel = rbf }	{ C = 0.1; Epsilon = 0.1; Gamma = scale; Kernel = rbf }	{ C = 10; Epsilon = 1; Gamma = auto; Kernel = linear }	{ C = 10; Epsilon = 1; Gamma = auto; Kernel = rbf }
AR	{ MD = 3; MSS = 2; MSL = 1 }	{ MD = 3; MSS = 10; MSL = 1}	{ MD = 3; MSS = 2; MSL = 1 }	{ MD = 3; MSS = 2; MSL = 1 }
BR	{ Alpha 1 = 0.1; Alpha 2 = 100; Lambda 1 = 100; Lambda 2 = 0.1 }	{ Alpha 1 = 0.1; Alpha 2 = 100; Lambda 1 = 100; Lambda 2 = 0.1 }	{ Alpha 1 = 0.1; Alpha 2 = 100; Lambda 1 = 100; Lambda 2 = 0.1 }	{ Alpha 1 = 0.1; Alpha 2 = 100; Lambda 1 = 100; Lambda 2 = 0.1 }
GBR	{ NE = 50; LR = 0.01; MD = 7 }	{ NE = 50; LR = 0.01; MD = 7 }	{ NE = 100; LR = 0.01; MD = 5 }	{ NE = 100; LR = 0.01; MD = 3 }

Tabela 2: Hiper parâmetros utilizados nos modelos.

3.6 Seleção dos Ativos para Carteira

Com o objetivo de montar carteiras de ativos para recomendação, foram definidos os cenários descritos anteriormente. No primeiro cenário (Cenário 1), os modelos de regressão foram treinados com os dados de um ano e testados com os dados do ano subsequente, resultando na geração de três carteiras com essa configuração. Já no segundo cenário (Cenário 2), os dados de três anos consecutivos foram concatenados para o treinamento, enquanto o teste foi realizado com os dados do ano seguinte, gerando, ao final, uma única carteira.

Após a aplicação dos modelos de regressão para cada cenário, os ativos foram classificados em ordem decrescente com base nos retornos previstos no conjunto de teste. Os dez ativos com os maiores retornos esperados foram selecionados para compor a carteira de investimentos, sendo todos alocados com pesos iguais. Dessa forma, cada ativo contribui com 10% do retorno total da carteira, garantindo uma distribuição uniforme.

Capítulo 4

Apresentação e Análise dos Resultados

No Capítulo 4, serão apresentados os experimentos realizados e os resultados obtidos. A avaliação dos modelos será conduzida por meio de medidas de erro de regressão e de retorno dos ativos, comparando os resultados previstos com os valores reais. Essa abordagem permitirá identificar o modelo com melhor desempenho, analisar o impacto do ajuste de hiperparâmetros e comparar os métodos de *machine learning* com a teoria do portfólio moderno e os índices de mercado.

4.1 Cenário de janela de 1 ano

A análise a seguir detalha os modelos treinados em um ano e testados no ano seguinte, portanto um ano de janela, tendo seu erro quadrático médio, testes de Friedman e t teste de Student feitos para análise.

4.1.1 Análise de Performance de Regressão dos Modelos

Primeiramente, a análise de desempenho dos modelos de regressão será estruturada da seguinte forma: os modelos terão seus Erros Quadráticos Médios (EQMs) calculados para cada algoritmo de regressão e para cada ano de aplicação. Esses EQMs são baseados na distância entre os retornos previstos pelos algoritmos para cada ativo e os retornos reais observados no ano em que o teste foi realizado.

As seções a seguir, 4.1.1.a e 4.1.1.b, possuem uma estrutura muito semelhante. Os valores de EQM serão comparados com o objetivo de identificar os modelos com melhor desempenho em cada ano. Em seguida, será realizado o teste de Friedman nos dados para verificar se há diferença estatisticamente significativa entre os métodos. A única diferença entre as duas seções mencionadas está nos hiperparâmetros aplicados: na seção 4.1.1.a, os hiperparâmetros não foram aplicados, enquanto na seção 4.1.1.b eles estão configurados.

Por fim, a seção 4.1.1.c destaca o impacto dos hiperparâmetros na melhoria do desempenho dos modelos. Além de calcular e comparar os fatores de ganho, essa seção aplica

um teste t de Student pareado entre os modelos sem e com hiperparametrização. Assim, são realizados cinco testes com o objetivo de verificar se o ganho proporcionado pela hiperparametrização é estatisticamente significativo.

4.1.1.a Resultado dos Modelos sem hiperparâmetros

Para iniciar a demonstração dos resultados dos modelos sem ajuste de hiperparâmetros apresenta-se a tabela 3, que exibe os valores de EQM para todos os modelos treinados e os respectivos anos de teste (2021, 2022 e 2023). Os melhores resultados de EQM em cada ano estão destacados, evidenciando que o GBR apresentou o melhor desempenho em dois dos três anos analisados.

Método/Ano	**2021**	**2022**	**2023**
GBR	1103,70	1987,71	1595,96
Reg. Bayesiana	1177,17	2196,09	1396,24
Reg. Linear	1224,01	2167,56	1409,75
SVR	1185,47	2059,69	1446,66
Árvore de Regressão	1385,08	2431,04	1879,46

Tabela 3: Grade de EQM por modelo e ano sem hiperparâmetros.

Na tabela 4, é possível observar os EQMs classificados em ordem e a média de classificação na última coluna. Analisando as posições, conclui-se que, em média, o modelo *Gradient Boost* (GBR) apresenta o melhor desempenho, embora não esteja muito distante dos outros modelos. Por outro lado, o modelo que mais se destaca negativamente é a Árvore de Regressão, cuja performance é significativamente inferior em comparação aos demais.

Modelo/Ano	2021	2022	2023	Média
GBR	1	1	4	2
Bayesiana	2	4	1	2.33
SVR	3	2	3	2.66
Linear	4	3	2	3
Árvore	5	5	5	5

Tabela 4: Ranking de EQM por modelo e ano sem hiperparâmetros.

Para finalizar, como descrito anteriormente no Capítulo 2, o teste estatístico de Friedman é capaz de indicar se a diferença entre as performances de EQM é estatisticamente significativa. Conforme evidenciado pela tabela 5, embora o valor da estatística de Friedman não seja tão baixo, ele ainda é insuficiente para rejeitar a hipótese nula (p-valor > 0,05). Dessa forma, a hipótese nula de que não há evidências suficientes para considerar as diferenças como significativas é aceita.

Estatística de Friedman	6.67
p-valor	0.1546

Tabela 5: Métricas do teste de Friedman sem hiperparâmetros.

4.1.1.b Resultado dos Modelos com hiperparâmetros

O procedimento de análise para os modelos com hiperparâmetros ajustados será extremamente semelhante ao realizado anteriormente para os modelos sem ajuste de hiperparâmetros. Primeiramente, a Tabela 6 apresenta os EQMs de cada modelo implementado, e é possível notar que os valores apresentados são numericamente menores do que os observados na Seção 4.1.1.a. Além disso, a Árvore de Regressão, que anteriormente era o pior modelo em todos os anos, teve a melhor performance em 2022, superando o *Gradient Boosting* (GBR). Nos anos de 2021 e 2023, não houve mudanças em relação ao melhor modelo.

Método/Ano	2021	2022	2023
GBR	941,70	2065,32	1435,49
Reg. Bayesiana	1073,93	2194,17	1403,02
Reg. Linear	1222,25	2188,13	1409,35
SVR	1130,41	2105,29	1453,49
Árvore de Regressão	1238,28	2017,33	1538,84

Tabela 6: Grade de EQM por modelo e ano com hiperparâmetros.

A tabela 7 evidencia as posições ranqueadas e a média para cada ano de teste, e os resultados foram semelhantes aos dados sem hiperparametrização. Os maiores destaques ficam para o modelo *Gradient Boost*, que obteve a melhor média de classificações, e para a Árvore de Regressão, que apresentou um grande ganho com a hiperparametrização no ano de 2022.

	2021	2022	2023	Média
GBR	1	2	3	2
Bayesiana	2	5	1	2.66
SVR	3	3	4	3.33
Linear	4	4	2	3.33
Árvore	5	1	5	3.67

Tabela 7: Ranking de EQM por modelo e ano com hiperparâmetros.

Assim como foi feito anteriormente, o teste estatístico de Friedman foi aplicado para verificar se existe uma diferença estatisticamente significativa entre os métodos. Ao comparar a tabela 5 com a tabela 8, observa-se que o valor da estatística de Friedman diminuiu consideravelmente, o que evidencia que o ajuste de hiperparâmetros fez com que os resultados dos modelos ficassem mais próximos, em uma perspectiva numérica. Além disso, o p-valor é maior que 0,05, portanto, a hipótese nula é aceita novamente, pois não há evidências suficientes para considerar as diferenças como significativas.

Estatística de Friedman	2.13
p-valor	0.7113

Tabela 8: Métricas do teste de Friedman com hiperparâmetros.

4.1.1.c Resultado Comparativo com e sem Hiperparametrização

Com o objetivo de destacar o efeito do ajuste de hiperparâmetros nos modelos, foi analisado o ganho apresentado na performance de EQM para cada modelo e cada ano. O ganho médio foi calculado para cada modelo nos três anos e, em seguida, foi aplicado o teste t de Student, que, conforme detalhado no Capítulo 2, compara dois métodos distintos com o intuito de investigar se houve ou não uma diferença estatística relevante entre eles.

Para começar, a tabela 9 mostra os ganhos (em valor percentual) de performance para cada modelo e cada ano. Observa-se que, em alguns casos, o ajuste de hiperparâmetros não teve efeito significativo ou até resultou em uma leve piora. Contudo, na grande maioria dos casos, o efeito foi positivo, com destaque para um ganho de 22,13% na Árvore de Regressão em 2023.

Método/Ano	**2021**	**2022**	**2023**
GBR	17,20%	-3,81%	11,39%
Reg. Bayesiana	9,61%	0,09%	-0,55%
Reg. Linear	0,14%	-0,94%	0,03%
SVR	4,87%	-2,17%	-0,47%
Árvore de Regressão	11,85%	20,51%	22,13%

Tabela 9: Ganho por método e ano proporcionado pelo ajuste de hiperparâmetros.

Em seguida, o gráfico de barras retrata os ganhos médios, deixando claro que os modelos de Árvore de Regressão e *Gradient Boosting Regressor* (GBR) foram os que mais se beneficiaram do ajuste de hiperparâmetros, com ganhos de 18,16% e 8,26%, respectivamente.

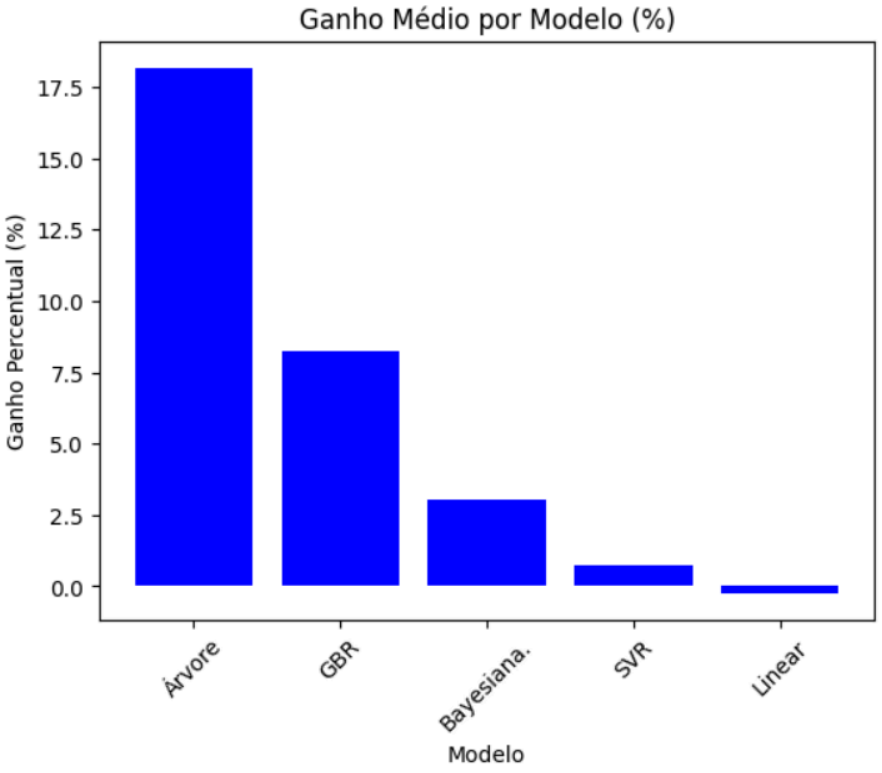

Figura 11: Ganho por método médio proporcionado pelo ajuste de hiperparâmetros.

Para finalizar, o teste t de Student foi aplicado para verificar se os ganhos obtidos com o ajuste dos hiperparâmetros foram significativamente relevantes. Para cada um dos modelos de regressão, foi gerado um p-valor utilizando as premissas do teste. Caso o p-valor fosse menor que 0,05, a hipótese nula de que não há evidências suficientes para considerar as diferenças como significativas poderia ser rejeitada. A tabela a seguir apresenta os p-valores obtidos para cada método.

Método	p-valor	Hipótese Nula
GBR	0.41	Aceita
Bayesiana	0.46	Aceita
SVR	0.48	Aceita
Linear	0.98	Aceita
Árvore	0.06	Aceita

Tabela 10: Métricas do teste t de Student para o efeito dos hiperparâmetros.

A hipótese nula foi aceita para todos os modelos, portanto, nenhum ganho pode ser considerado estatisticamente significativo.

4.1.2 Análise de mercado dos modelos

Para iniciar, a análise de mercado segue um fluxo muito similar ao da etapa de análise de regressão (4.1.1.a ou 4.1.1.b), com a principal diferença sendo a métrica analisada. Enquanto na avaliação de regressão a métrica utilizada foi o EQM, na análise de mercado será o Diferencial de Retorno (*Return Diff*) em percentual. Essa métrica é calculada pela diferença entre o retorno gerado pela melhor carteira efetiva do ano (seguindo a metodologia proposta na seção Seleção dos Ativos para Carteira – Capítulo 3) e o retorno gerado pela metodologia em análise. Assim, quanto menor o valor do Diferencial de Retorno, melhor será a performance de mercado da metodologia, ou seja, mais próxima ela estará da carteira ótima.

Como descrito anteriormente, também no Capítulo 3, as carteiras geradas pelas metodologias de *Machine Learning* (ML) serão comparadas às carteiras de máxima razão de Sharpe e mínima volatilidade propostas pela metodologia de Markowitz, além do índice de referência do mercado, o S&P 500. Em cada um dos anos analisados (2021, 2022 e 2023), a melhor carteira gerada pela metodologia de ML proposta neste trabalho foi comparada com as metodologias de mercado mencionadas. Além disso, o teste de Friedman foi aplicado nesta análise com o objetivo de verificar se a diferença no Diferencial de Retorno entre as metodologias é estatisticamente significativa.

Como primeira etapa da análise, deve-se definir os modelos de ML que tiveram a melhor performance financeira em cada ano. Como revelado pela tabela 11 abaixo, para os anos de 2021 e 2023 o modelo GBR teve o melhor Diferencial de Retorno, enquanto que no ano de 2022 a regressão por árvore de decisão apresentou resultados superiores.

Método/Ano	2021	2022	2023
GBR	58,54	45,64	77,04
Reg. Bayesiana	68,15	104,78	102,67
Reg. Linear	68,15	102,16	98,18
SVR	64,08	88,10	100,56
Árvore de Regressão	69,16	44,35	92,97

Tabela 11: Grade de Diferencial de Retorno por modelo e ano com hiperparâmetros.

Dando continuidade à análise, os três modelos de ML elegidos tiveram seus nomes substituídos nas seguintes imagens por Regressor ML (Regressor de *Machine Learning*), a fim de compará-los com as metodologias de mercado. Dessa forma, a tabela 12 foi gerada revelando que em dois dos três anos analisados (2022 e 2023) os regressores de aprendizado de máquina foram superiores aos métodos tradicionais do mercado. Contudo, em 2021 a carteira de razão de Sharpe máxima proposta por Markowitz teve melhor performance, deixando a metodologia proposta em segundo lugar.

Método/Ano	2021	2022	2023
Max Sharpe	44,95	81,37	100,77
Min Vol	86,76	77,82	111,33
Regressor ML	58,54	44,35	77,04
S&P 500	80,86	89,14	85,88

Tabela 12: Grade de Diferencial de Retorno por modelo e ano com hiperparâmetros.

Para constatar a superioridade da observada das carteiras geradas pelos modelos de ML foi gerada a tabela 13, que destaca as posições das metodologias em cada ano.

Método / Ano	2021	2022	2023	Média
Regressor ML	2	1	1	1.33
Max Sharpe	1	3	3	2.33
S&P 500	3	4	2	3
Min Vol	4	2	4	3.33

Tabela 13: Ranking de Diferencial de Retorno por modelo e ano com hiperparâmetros.

Para finalizar, o mesmo teste de Friedman realizado para a análise de EQM foi realizado para os diferenciais de retorno adotando a hipótese nula de que não há evidências suficientes para considerar as diferenças como significativas. O resultado de p-valor e estatística de Friedman podem ser visualizados a seguir na tabela 14.

Estatística de Friedman	4.20
p-valor	0.2407

Tabela 14: Métricas do teste de Friedman para Diferencial de Retorno.

Concluindo-se a análise, com o p-valor apresentado a hipótese nula é aceita.

4.2 Cenário de Janela de 3 anos de treino

Para essa etapa, os dados de 2020, 2021 e 2022 foram condensados em uma única base de dados formando a janela (3 anos), os modelos de ML foram treinados e testados no ano de 2023 com e sem ajuste de hiperparâmetros, e novamente as métricas de EQM e diferencial de retorno foram aferidas.

Primeiramente será apresentada a performance dos modelos com base na regressão, tendo em vista o EQM como principal métrica de análise. A princípio sem ajuste de

hiperparâmetros, em seguida com ajuste de hiperparâmetros, dessa forma comparando os ganhos de performance realizados pelo ajuste. E para finalizar, segue a análise econômica comparativa do diferencial de retorno em relação às metodologias tradicionais do mercado, também realizadas em uma janela de 3 anos. Vale acrescentar que como apenas uma janela conseguiu ser formada com os dados obtidos nesta etapa não contará com testes estatísticos auxiliares ao resultado, visto que todos os dados possuem uma única amostra.

4.2.1 Análise de Performance de Regressão dos Modelos (3 anos)

Como introdução à análise, pode-se destacar a performance de regressão dos modelos sem ajuste de hiperparâmetros com base no EQM como medida de erro. Os modelos de regressão linear, bayesiana e máquina de vetores de suporte foram relativamente superiores aos outros modelos como exposto pelas tabelas e gráficos abaixo.

Método	Ano	EQM
SVR	2023	954,34
Regressão Bayesiana	2023	961,70
Regressão Linear	2023	962,86
GBR	2023	1552,70
Árvore de Regressão	2023	2352,11

Tabela 15: EQM por modelo na janela de três anos sem hiperparâmetros.

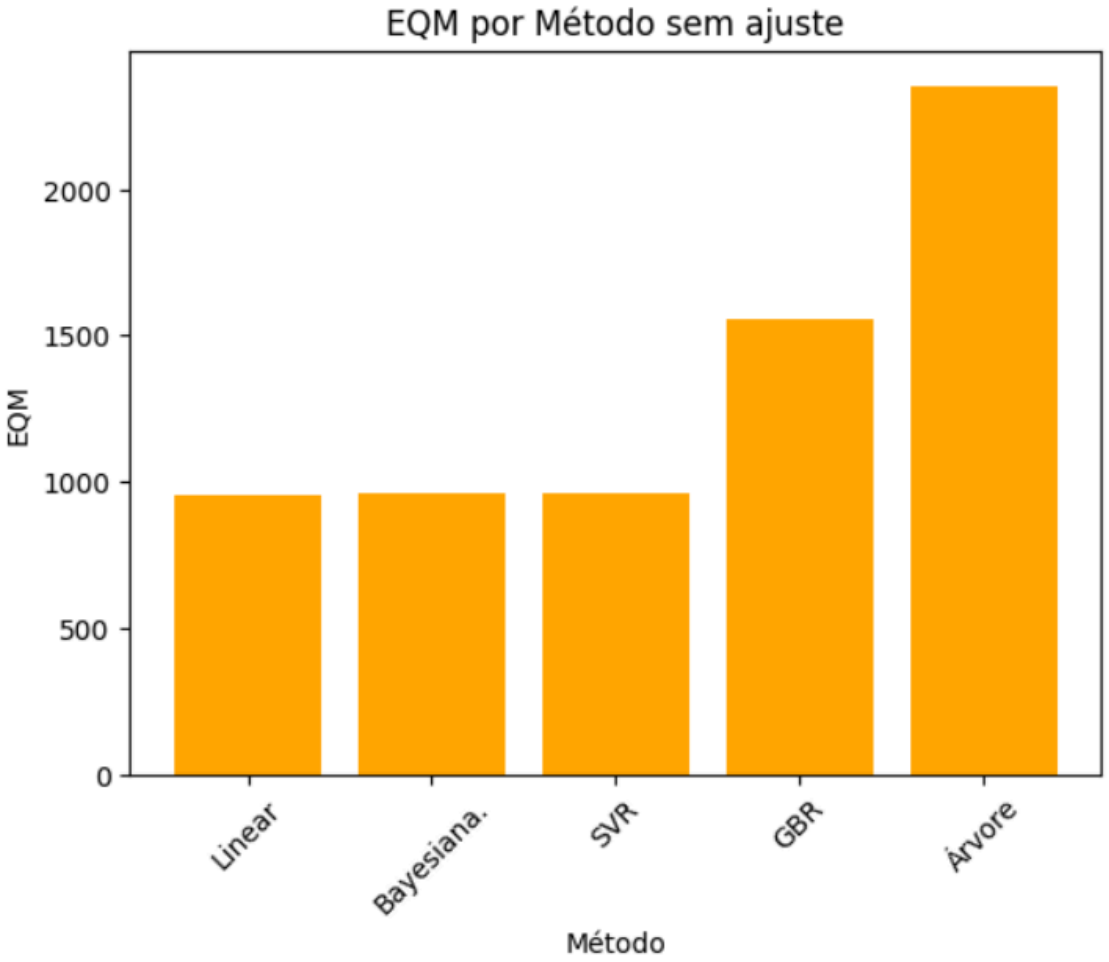

Figura 12: Gráfico de barras de EQM por modelo na janela de três anos sem hiperparâmetros.

Em sequência, os hiperparâmetros foram ajustados pelo formato descrito na seção detalhamento da busca em grade desta monografia. Os resultados obtidos foram muito parecidos demonstrando que o ajuste não teve grande influência na maioria dos modelos, exceto pelo GBR que apresentou um ganho significativo, dessa forma se igualando aos três modelos apontados no parágrafo anterior, que mantiveram seu destaque.

Abaixo podem ser visualizados os dados tanto com as referências de performance para os modelos com hiperparâmetros ajustados quanto o ganho proporcionado pelos hiperparâmetros.

Método	Ano	EQM
Regressão Bayesiana	2023	961,35
Regressão Linear	2023	962,84
SVR	2023	970,45
GBR	2023	1070,00
Árvore de Regressão	2023	2132,66

Tabela 16: EQM por modelo na janela de três anos com hiperparâmetros.

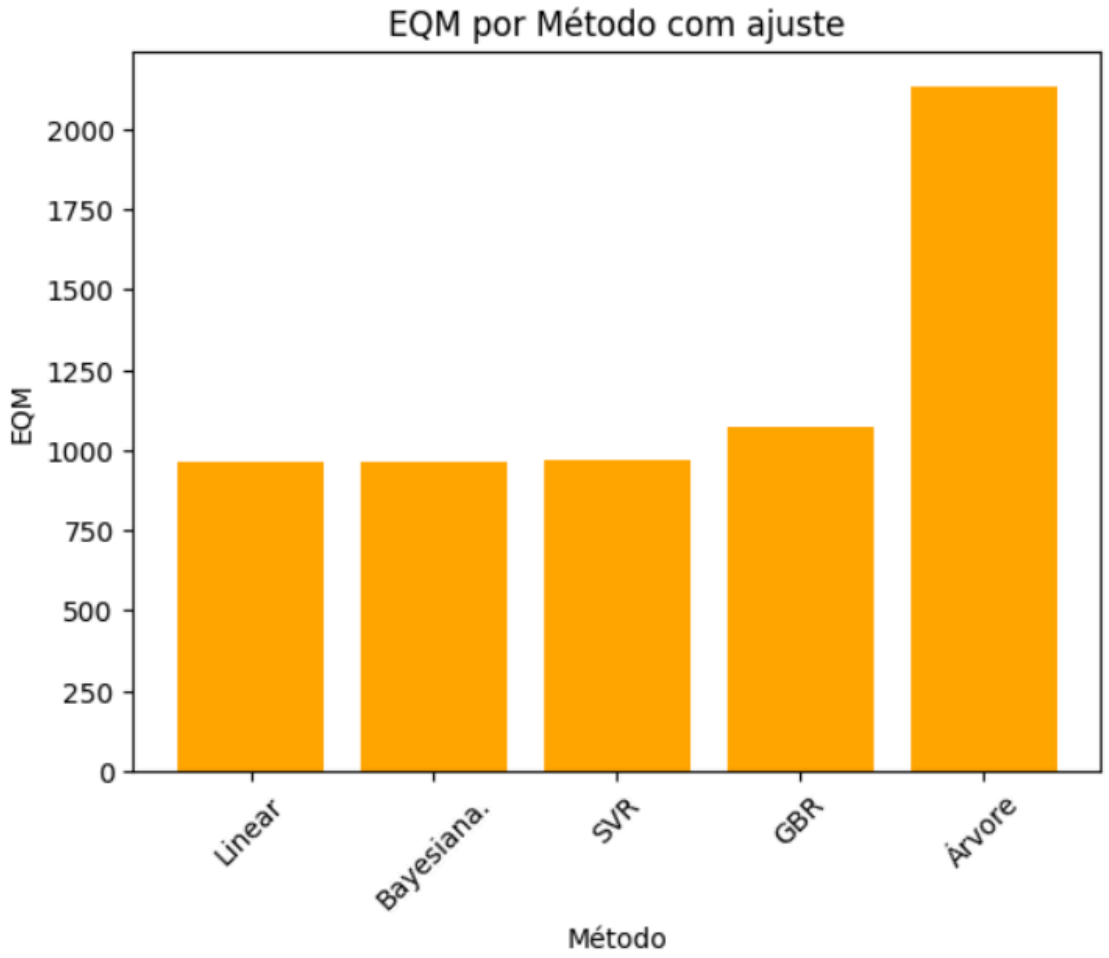

Figura 13: Gráfico de barras de EQM por modelo na janela de três anos com hiperparâmetros.

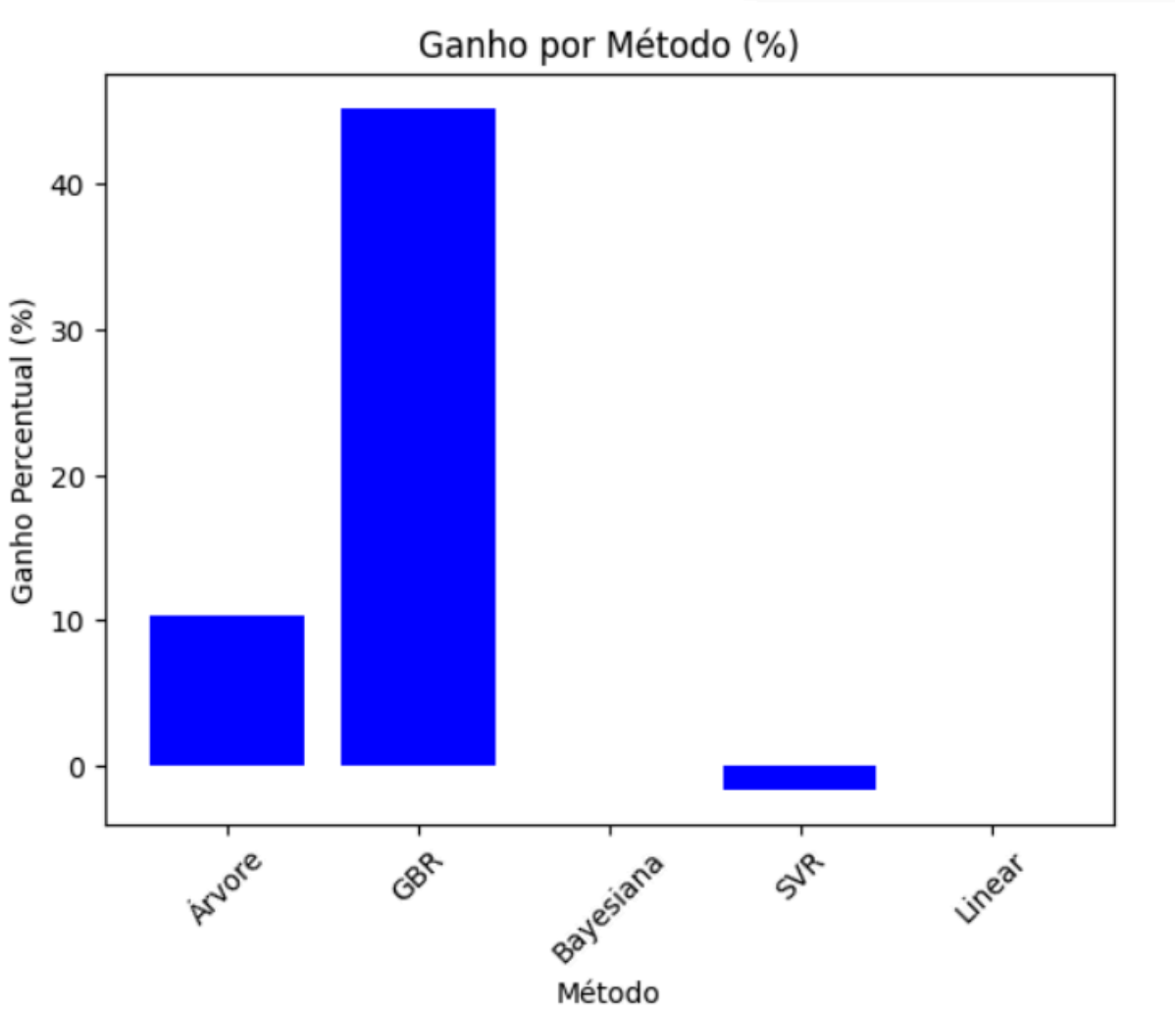

Figura 14: Gráfico de barras do ganho do ajuste de hiperparâmetros para janela de três anos.

Método	Ganho de EQM (%)
GBR	45,11%
Regressão Bayesiana	0,04%
Regressão Linear	0,00%
SVR	-1,66%
Árvore de Regressão	10,29%

Tabela 17: Ganho do ajuste de hiperparâmetros na janela de três anos.

Os dados de ganho apresentados acima revelam que o GBR e a árvore de regressão tiveram altos ganhos, contudo alguns modelos como o SVR, regressão linear e bayesiana apresentaram um ganho nulo e até piora na performance. Isso ocorre em decorrência do método de ajuste de hiperparâmetros, o *gridsearch* faz a busca e otimização dentro dos parâmetros e valores escolhidos e colocados na grade pelo programador, portanto não é garantido que o processo gerará ganhos reais.

Além disso, vale ressaltar que em relação à eficiência de tempo, os algoritmos não apresentaram longos tempos de treino, o maior gargalo da metodologia está na hiperparametrização por busca em grade, dado que um modelo é treinado para cada combinação de hiperparâmetros, fazendo com que a eficiência seja proporcional às dimensões da grade utilizada.

4.2.2 Análise de Performance de Mercado dos Modelos (3 anos)

Partindo para a análise e interpretabilidade de mercado, novamente a métrica a ser utilizada será o diferencial de retorno, dado pela diferença entre a melhor carteira possível no ano com dez ativos e a performance real da carteira sugerida por cada metodologia.

Assim como realizado com os testes da janela de um ano, o melhor modelo dentre os métodos de ML com hiperparâmetros ajustados será selecionado para ser comparado com as metodologias propostas pelo mercado financeiro e índices. Nesse caso, a regressão bayesiana será comparada com a carteira de máxima razão de sharpe e mínima volatilidade da teoria do portfólio moderno além do índice S&P 500.

Método	**Diferencial de Retorno**
Regressão Bayesiana	81,62
Max Sharpe	84,69
S&P 500	90,21
Min Vol	113,31

Tabela 18: Diferencial de retorno na janela de três anos.

Como demonstrado pela tabela 18, o modelo de regressão proposto teve melhor performance quando comparado aos modelos tradicionais de mercado realizados para a janela de três anos compreendida entre 2020, 2021 e 2022.

Capítulo 5
Conclusão e Trabalhos Futuros

Este trabalho de conclusão de curso teve como objetivo utilizar modelos de aprendizado de máquina para prever preços futuros de ativos e com base nisso montar carteiras que pudessem bater índices e metodologias do mercado financeiro.

No total 5 modelos de AM foram utilizados, dentre eles Regressão Linear Ridge (Regressão Linear com regularização L2), Árvore de Regressão, Regressão de Vetores de Suporte (SVR), Bayesian Ridge e Regressão por Impulso de Gradiente (GBR). Esse métodos foram comparados com as carteiras da teoria do portfólio moderno e índice S&P 500 utilizando duas principais métricas de análise: erro quadrático médio (EQM) e diferença de retorno (*Return Diff*).

Os resultados demonstraram que por mais que não bem sucedidos nos testes de Friedman e t-teste de Student, a estratégia e uso de modelos de AM indicam ao investidor ativos que consistentemente batem as carteiras de Markowitz bem como o índice S&P 500, se provando ser o melhor método em três dos quatro experimentos comparativos realizados.

Dessa forma, os objetivos do trabalho foram concluídos dado que foram propostos modelos simples e eficazes de seleção de ativos para carteiras que batem índices e metodologias do mercado financeiro.

5.1 Limitações Observadas

A principal limitação para o trabalho foi a dificuldade em encontrar dados financeiros de forma fácil para poderem ser utilizados para treino dos modelos. A API de dados utilizada nesta monografia só continha dados a partir de 2020, dessa forma sendo capaz de gerar apenas uma única janela comparativa de 3 anos e três janelas de 1 ano.

A presença de mais dados faz com que os modelos possam ser colocados à prova em situações econômicas mais instáveis do que as já testadas. Inserir períodos de crise ou de grande aquecimento econômico podem testar a solidez dos modelos de forma mais robusta em cenários extremos.

Outra grande vantagem seria a formação de mais janelas tanto em quantidade de tempo quanto em número de janelas, dessa forma poderia ser estudado se existe algum tipo de janela ideal para treino dos modelos.

5.2 Trabalhos Futuros

Para começar, como destacado na seção de limitações, uma possível opção seria o uso de dados históricos de tempos maiores do que três anos, alimentando os modelos com maior quantidade de dados a previsão de carteiras poderia ser testada em cenários de estresse financeiro, bem como para diferentes janelas de tempo.

Em segundo, vale destacar o uso dos múltiplos de mercado para o treinamento dos modelos, no caso deste trabalho de conclusão de curso foram usados P/L, ROE e valor de mercado, contudo outros multiplicadores poderiam ser utilizados em conjunto a esse como margem EBITDA, fatores de dívida, índice de dividendos e P/VP. Com isso, os modelos com certeza teriam outros resultados, podendo superar de forma ainda maior os indicadores de mercado.

Para concluir, além de utilizar os modelos já supracitados, é possível utilizar e treinar redes neurais para fazer a previsão do preço das ações e montagem das carteiras. Esse tipo de modelo de AM possui maior flexibilidade para ser ajustado para o domínio tratado, podendo variar em camadas, funções de ativação, número de neurônios por camadas, funções de perda e taxa de aprendizado.

Capítulo 6

Referências

SHEIKH, H.; PRINS, C.; SCHRIJVERS, E. Artificial Intelligence: Definition and Background. In: Mission AI. Research for Policy. Cham: Springer, 2023.

LIU, C.; SHI, H.; WU, L.; GUO, M. The short-term and long-term trade-off between risk and return: chaos vs rationality. *Journal of Business Economics and Management*, v. 21, n. 1, p. 23-43, 14 Jan. 2020.

TITAN, Alexandra Gabriela. The Efficient Market Hypothesis: Review of Specialized Literature and Empirical Research. *Procedia Economics and Finance*, v. 32, p. 442-449, 2015. ISSN 2212-5671.

AHSAN, Afm. Can Return on Equity be Used to Predict Portfolio Performance? *Economics, Management, and Financial Markets*, v. 7, p. 132-148, 2012.

IRONS, Robert; WEIGAND, Robert. The Market P/E Ratio, Earnings Trends and Stock Return Forecasts. *The Journal of Portfolio Management*, v. 33, p. 87-101, 2007. DOI: 10.3905/jpm.2007.690610.

ZOLOTAREVA, Ekaterina. Aiding Long-Term Investment Decisions with XGBoost Machine Learning Model. Moscow: Financial University under the Government of the Russian Federation, 2021.

Chenyao, Ma. Stock Selection Model Based on Random Forest. Tongji University, Shanghai, China, 2023.

LI, Yuhan. Stock Price Prediction Based on Multiple Regression Models. *Highlights in Science, Engineering and Technology*, v. 39, p. 657-662, 2023. DOI: 10.54097/hset.v39i.6622.

MARKOWITZ, H. Portfolio Selection. *The Journal of Finance*, v. 7, p. 77-91, 1952.

FAMA, E. Efficient Capital Market: A Review of Theory and Empirical Work. *Journal of Finance*, v. 25, p. 382-417, 1970.

ALAM, Azmir. What is Machine Learning? 2023. DOI: 10.5281/zenodo.8231580.

FRANÇA, Reinaldo Padilha; MONTEIRO, Ana Carolina Borges; ARTHUR, Rangel; IANO, Yuzo. Chapter 3 - An overview of deep learning in big data, image, and signal processing in the modern digital age. *Hybrid Computational Intelligence for Pattern Analysis: Trends in Deep Learning Methodologies*. Academic Press, 2021. p. 63-87. ISBN 9780128222263.

EWUZIE, Ugochukwu; BOLADE, Oladotun Paul; EGBEDINA, Abisola Opeyemi. Chapter 9 - Application of deep learning and machine learning methods in water quality modeling and prediction: a review. In: MARQUES, Gonçalo; IGHALO, Joshua O. (ed.). *Intelligent Data-Centric Systems: Current Trends and Advances in Computer-Aided Intelligent Environmental Data Engineering*. Academic Press, 2022. p. 185-218. ISBN 9780323855976.

POTASHEV, K.; SHARONOVA, N.; BREUS, I. The use of cluster analysis for plant grouping by their tolerance to soil contamination with hydrocarbons at the germination stage. *Science of The Total Environment*, v. 485-486, p. 71-82, 1 jul. 2014. DOI: 10.1016/j.scitotenv.2014.03.067. Epub 3 abr. 2014. PMID: 24704958.

HASAN, A. S. M. Mahmudul; SOHEL, Ferdous; DIEPEVEEN, Dean; LAGA, Hamid; JONES, Michael G. K. A survey of deep learning techniques for weed detection from images. *Computers and Electronics in Agriculture*, v. 184, p. 106067, 2021. ISSN 0168-1699

ARNOLD, Christian; BIEDEBACH, Luka; KÜPFER, Andreas; NEUNHOEFFER, Marcel. The role of hyperparameters in machine learning models and how to tune them. *Political Science Research and Methods*, v. 12, p. 1-8, 2024. DOI: 10.1017/psrm.2023.61.

QU, Kecheng. Research on linear regression algorithm. *Shandong Xiehe University*, 250100, JiNan, Shandong, China.

BASAK, Debasish; PAL, Srimanta; PATRANABIS, Dipak. Support Vector Regression. *Neural Information Processing – Letters and Reviews*, v. 11, 2007.

EFENDI, Achmad. A simulation study on Bayesian Ridge regression models for several collinearity levels. *AIP Conference Proceedings*, v. 1913, n. 1, p. 020031, dez. 2017. DOI: 10.1063/1.5016665.

ROKACH, Lior; MAIMON, Oded. Decision Trees. In: *The Data Mining and Knowledge Discovery Handbook*. v. 6, p. 165-192, 2005. DOI: 10.1007/0-387-25465-X_9.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, p. 785-794.

VARGAS JUNIOR, Edson Cilos. Medidas de desempenho para regressão. *Universidade Federal de Santa Catarina*, apresentação de aula de regressão, 2020.

INTERACTIVE BROKERS. Markowitz Model. Disponível em: https://www.interactivebrokers.com/campus/ibkr-quant-news/markowitz-model/.

RASEKHSCHAFFE, Keywan; JONES, Robert. Machine Learning for Stock Selection. *Financial Analysts Journal*, v. 75, n. 3, 2019.

WOLFF, D.; ECHTERLING, F. Stock picking with machine learning. *Journal of Forecasting*, v. 43, n. 1, p. 81–102, 2024.

GAMBIM, Mario; CAMARGO, Heloisa A.; LUCCA, Giancarlo; DIMURO, Graçaliz; ASMUS, Thiago. Uma Estratégia para Alocação de Carteira de Ações usando Algoritmos de Aprendizado de Máquina e Regras Fuzzy. *In*: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 20. , 2023, Belo Horizonte/MG. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023 . p. 1195-1209. ISSN 2763-9061. DOI: https://doi.org/10.5753/eniac.2023.234681.