UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS - CAMPUS BAURU
DEPARTAMENTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

LUCAS YUKI NISHIMOTO

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA COMO ESTRATÉGIA OPERACIONAL NO MERCADO DE CAPITAIS BRASILEIRO

BAURU
Novembro/2024

LUCAS YUKI NISHIMOTO

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA COMO ESTRATÉGIA OPERACIONAL NO MERCADO DE CAPITAIS BRASILEIRO

Trabalho de Conclusão de Curso do Curso de Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.
Orientador: Prof. Dr. Mateus Roder
Coorientador: Prof. Dr. André Luis Debiaso Rossi

BAURU
Novembro/2024

N724a

Nishimoto, Lucas Yuki

Aplicação de técnicas de aprendizado de máquina como estratégia operacional no mercado de capitais brasileiro / Lucas Yuki Nishimoto. -- Bauru, 2024

70 p. : il., tabs., fotos

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências, Bauru

Orientador: Mateus Roder

Coorientador: André Luis Debiaso Rossi

1. Aprendizado de máquina. 2. Mercado financeiro. 3. Otimização de portfólio. 4. Índice de Sharpe. 5. Inteligência artificial. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Lucas Yuki Nishimoto

Aplicação de técnicas de aprendizado de máquina como estratégia operacional no mercado de capitais Brasileiro

Trabalho de Conclusão de Curso do Curso de Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.

Banca Examinadora

Prof. Dr. Mateus Roder
Orientador
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Ciência da Computação

Prof. Dr. Simone Domingues Prado
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Ciência da Computação

Prof. Dr. Kelton Augusto Pontara da Costa
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Ciência da Computação

Bauru, 12 de novembro de 2024.

Aos meus pais, por serem meus exemplos, sempre os terei como faróis em minha vida. Agradeço por cada palavra de apoio, por acreditarem nos meus sonhos e por tudo o que me ensinaram ao longo da vida. Esta conquista é tão de vocês quanto minha.

Agradecimentos

A realização deste trabalho não seria possível sem o apoio e a presença de pessoas que me acompanharam e incentivaram em cada passo dessa trajetória.

Agradeço primeiramente à minha família, pelo amor, suporte e por sempre depositarem fé em mim. Sua presença constante e incentivo me deram forças para seguir em frente e buscar o melhor meu melhor. Sou imensamente grato pelo cuidado e paciência, que me permitiram chegar até aqui.

À minha namorada, por sua amizade, companheirismo e por ser uma fonte constante de inspiração. Nos momentos difíceis, seu apoio e compreensão foram fundamentais, tornando esse caminho mais leve e significativo.

Aos meus amigos, pela parceria e apoio constante. Agradeço especialmente àqueles com quem compartilhei não apenas a convivência diária durante nossa trajetória acadêmica, mas também aos que, mesmo não morando conosco, fizeram-se presentes como parte indispensável dessa experiência. A todos, sou grato pelos momentos compartilhados, pelo suporte incondicional e pela amizade verdadeira, que foram essenciais para tornar esta conquista possível.

Agradeço também aos meus professores, que ao longo da minha formação acadêmica, compartilharam conhecimentos e me guiaram em minha trajetória. Em especial, expresso minha gratidão aos meus orientadores, que com dedicação, paciência e apoio, foram fundamentais para o desenvolvimento deste trabalho. Suas orientações e incentivo me deram confiança e clareza para superar os desafios e aprimorar meus conhecimentos.

A todos vocês, deixo aqui meu sincero agradecimento por fazerem parte dessa etapa e por todo apoio, amizade e amor.

"É sempre impossível até que seja feito."
Nelson Mandela

Resumo

Este trabalho investiga a aplicação de técnicas de aprendizado de máquina para maximizar o retorno financeiro no mercado de capitais brasileiro, utilizando os algoritmos *Random Forest*, *Support Vector Machine* e *XGBoost*. O estudo visa construir e avaliar portfólios de ativos com base em previsões de movimentações do mercado. Os dados históricos foram obtidos por meio do Yahoo Finance, e indicadores financeiros foram extraídos para alimentar os modelos. A análise compara o desempenho dos modelos em termos de retorno percentual e índice de Sharpe no período de janeiro de 2022 a outubro de 2024. Os resultados revelam que os modelos de aprendizado de máquina ofereceram uma relação risco-retorno muito superior ao Ibovespa, mais estável em casos como o Random Forest, mas com destaque em retorno acumulado para o XGBoost, também com o maior índice de Sharpe. Comparados ao Ibovespa, todos os modelos apresentaram retornos mais elevados e maior consistência, o que aponta para o potencial do aprendizado de máquina em estratégias de investimento seguras e rentáveis, uma vez que todos os modelos levaram a valores de retorno percentual acumulado de mais de 150% em pouco menos de três anos.

Palavras-chave: ML, IA, mercado financeiro, otimização de portfólio, índice de Sharpe, RF, SVM, XGBoost.

Abstract

This work investigates the application of machine learning techniques to maximize financial returns in the Brazilian capital market, using the *Random Forest*, *Support Vector Machine* and *XGBoost* algorithms. The study aims to build and evaluate asset portfolios based on forecasts of market movements. Historical data were obtained through Yahoo Finance, and financial indicators were extracted to feed the models. The analysis compares the performance of the models in terms of percentage return and Sharpe ratio in the period from January 2022 to October 2024. The results reveal that the machine learning models offered a much higher risk-return ratio than the Ibovespa, more stable in cases such as the RF, but with a highlight in cumulative return for XGBoost, also with the highest Sharpe ratio. Compared to Ibovespa, all models showed higher returns and greater consistency, which points to the potential of machine learning in safe and profitable investment strategies, since all models led to cumulative percentage return values of more than 150% in just under three years.

Keywords: ML, IA, financial market, portfolio optimization, Sharpe ratio, RF, SVM, XGBoost.

Lista de figuras

Lista de quadros

Lista de tabelas

Lista de abreviaturas e siglas

IA Inteligência Artificial

ML Machine Learning

RF Floresta Randômica (Random Forest)

SVM Máquina de Vetores de Suporte (Support Vector Machine)

XGBoost Extreme Gradient Boosting

API Interface de Programação de Aplicação (Application Programming Interface)

RRL Aprendizado por Reforço Recorrente (Recurrent Reinforcement Learning)

ETFs Fundos Negociados em Bolsa (Exchange-Traded Funds)

MDD Máxima Redução de Valor (Maximum Drawdown)

MAE Erro Médio Absoluto (Mean Absolute Error)

RMSE Raiz do Erro Médio Quadrático (Root Mean Square Error)

CNN Rede Neural Convolucional (Convolutional Neural Network)

LSTM Memória de Longo Curto Prazo (Long Short-Term Memory)

IFA Algoritmo de Vaga-lumes Melhorado (Improved Firefly Algorithm)

RSI Índice de Força Relativa (Relative Strength Index)

VWAP Preço Médio Ponderado pelo Volume (Volume Weighted Average Price)

CCI Índice de Canal de Commodities (Commodity Channel Index)

B3 Bolsa de Valores de São Paulo (Brasil, Bolsa, Balcão)

Sumário

1 Introdução

Na atualidade, diversas áreas do conhecimento e do cotidiano têm se beneficiado dos avanços tecnológicos promovidos pelos extensos estudos em hardware, software, e sistemas ciber-físicos (GOODFELLOW; BENGIO; COURVILLE, 2016). Grandes exemplos são as aplicações de inteligência artificial (IA) na medicina, computação gráfica, simulação, robótica, e mercado financeiro, para citar alguns (RUSSELL; NORVIG, 2016).

Com o avanço da digitalização, essas áreas experimentaram um crescimento exponencial em eficiência e precisão. Na medicina, por exemplo, algoritmos de aprendizado de máquina permitem diagnósticos mais rápidos e precisos, revolucionando a forma como doenças são detectadas e tratadas (ESTEVA et al., 2017). A robótica, tanto em áreas industriais quanto em tarefas cotidianas, já é uma realidade que aumenta a produtividade e reduz erros operacionais (SICILIANO et al., 2016). Simulações cada vez mais precisas têm sido utilizadas tanto em pesquisas científicas quanto no desenvolvimento de novos produtos, minimizando custos e maximizando o conhecimento (FUJIMOTO, 2015).

No mercado financeiro, a transformação também é notável. A capacidade de processar grandes volumes de dados em tempo real permite análises que antes eram impensáveis. Atualmente, algoritmos avançados analisam históricos de preços, tendências de mercado e até mesmo o comportamento de investidores para gerar previsões (PETERS; MARR, 2016). A automatização de operações financeiras, impulsionada por esses algoritmos, facilita a tomada de decisões estratégicas e a otimização de portfólios, possibilitando a maximização de retornos e minimização de riscos (HIRANSHA et al., 2018).

Esse cenário é particularmente evidente quando analisamos as ações. A complexidade e volatilidade do mercado de capitais têm demandado o uso de ferramentas tecnológicas cada vez mais sofisticadas, permitindo que tanto investidores experientes quanto os novatos tomem decisões baseadas em dados precisos e previsões embasadas em algoritmos de aprendizado de máquina, elevando a competitividade e dinamismo do setor (MARKOWITZ, 1952).

Ações são parcelas representativas do capital social de uma empresa, que conferem ao seu detentor (acionista) uma parte da propriedade e, em alguns casos, o direito de receber dividendos proporcionais aos lucros da organização (MALKIEL, 1999). A prática de emissão de ações surgiu como uma maneira de empresas financiarem suas atividades sem depender exclusivamente de empréstimos bancários (NORTH, 1973). O conceito moderno de ações começou a se consolidar no início do século XVII com a criação da Companhia das Índias Orientais, que permitiu que investidores adquirissem participações na empresa em troca de uma fração dos lucros futuros. Esse modelo revolucionou o financiamento corporativo, abrindo caminho para o mercado de capitais como conhecemos hoje (KINDLEBERGER, 2005).

As bolsas de valores foram estabelecidas para facilitar a negociação das ações emitidas pelas empresas. Essas instituições passaram a atuar como intermediárias entre compradores e vendedores, assegurando que as negociações ocorressem de forma eficiente e segura, garantindo, assim, liquidez aos ativos e confiança aos investidores (LEVINE, 1997).

1.1 Quais são os desafios de operar no mercado

Operar no mercado de ações envolve diversos desafios que podem impactar diretamente os investidores. Um dos maiores obstáculos é a volatilidade, já que os preços das ações podem variar significativamente em curtos períodos, influenciados por fatores como mudanças econômicas, políticas e empresariais (SHILLER, 2015). Essa flutuação de preços torna difícil prever o comportamento futuro dos ativos, exigindo que os investidores tenham uma boa compreensão de análise de mercado e gestão de riscos (DAMODARAN, 2012).

Outro desafio é a assimetria de informações. Muitos investidores individuais não têm acesso aos mesmos dados ou ao nível de análise que grandes instituições financeiras possuem, o que pode dificultar a tomada de decisões informadas (AKERLOF, 1978). Além disso, a falta de conhecimento técnico sobre o funcionamento do mercado e as estratégias de investimento pode levar a erros que resultam em perdas financeiras (BERNHEIM, 1998).

O fator psicológico também desempenha um papel crucial, pois emoções como medo e ganância podem influenciar decisões precipitadas, como a venda de ações em momentos de baixa ou a compra impulsiva durante altas rápidas (KAHNEMAN, 2011). Dessa forma, operar no mercado exige não apenas conhecimento técnico, mas também autocontrole e planejamento de longo prazo (THALER, 2015).

1.2 A relação do brasileiro com o mercado financeiro

Historicamente, o brasileiro pertencente às classes de renda intermediária, com acesso limitado à educação financeira e preferências por investimentos conservadores, tem uma relação distante com o mercado financeiro, especialmente com o mercado de ações (SEVERINO; COIMBRA, 2016). O cenário econômico brasileiro, marcado por décadas de inflação elevada, crises financeiras e incertezas políticas, levou muitos a buscarem segurança em ativos considerados mais estáveis, como a poupança e imóveis, ao invés de arriscar investir em ações. Essa aversão ao risco, aliada à falta de conhecimento sobre o funcionamento do mercado de capitais e à escassa educação financeira, resultou em uma baixa participação da população na Bolsa de Valores (FAVERO; MEIRELES, 2013).

Em contrapartida, a população americana tem uma relação significativamente mais próxima com o mercado financeiro. Nos Estados Unidos, o mercado de ações é amplamente visto como uma das principais formas de investimento para crescimento patrimonial ao longo do

tempo (GOETZMANN; ROUWENHORST, 2005). Parte disso se deve à cultura de investimento amplamente difundida, onde desde cedo as pessoas têm acesso a informações sobre o mercado e são incentivadas a poupar e investir para aposentadoria por meio de instrumentos como o 401(k), fundos mútuos e ações individuais (GREENWOOD; HANSON; STEIN, 2019). Além disso, há uma maior confiança nas instituições financeiras e nas ferramentas disponíveis para o pequeno investidor, o que faz com que mais da metade da população americana invista diretamente ou indiretamente no mercado de ações (INVESTMENTS, 2021).

Nos últimos anos, entretanto, essa realidade começou a mudar no Brasil. A queda nas taxas de juros, especialmente após 2016, com a implementação de uma política monetária de juros mais baixos, fez com que produtos tradicionais, como a poupança, deixassem de ser atrativos para o investidor que busca retornos superiores à inflação (DE, 2020). Ao mesmo tempo, o aumento do acesso à informação, seja por meio de conteúdos digitais ou pela popularização de plataformas de investimento acessíveis ao público, contribuiu para um crescimento no número de brasileiros investindo em ações (SEVERINO; COIMBRA, 2016). Ainda assim, a participação no mercado de ações continua significativamente mais baixa em comparação aos Estados Unidos.

O cidadão de classe média ainda enfrenta desafios como a falta de educação financeira generalizada e a desconfiança em relação às instituições financeiras, um legado de crises anteriores e instabilidade econômica (FAVERO; MEIRELES, 2013). Ao contrário dos americanos, que têm maior familiaridade com o mercado, o investidor brasileiro é muitas vezes movido por ciclos de euforia e pessimismo, influenciados por crises políticas e econômicas (GARCIA, 2011). Esse comportamento muitas vezes impede uma abordagem mais estratégica e de longo prazo na construção de patrimônio através de investimentos no mercado de capitais.

Portanto, enquanto nos Estados Unidos o investimento em ações é visto como parte essencial do planejamento financeiro de longo prazo, no Brasil, ainda há um caminho a ser percorrido para que a população, de modo geral, adote uma visão mais próxima dessa realidade. Para isso, iniciativas de educação financeira, a construção de uma cultura de poupança e investimento e a redução da desconfiança nas instituições financeiras são fundamentais (BERNHEIM, 1998).

1.3 Inteligência Artificial

IA é um campo da ciência da computação que se dedica ao desenvolvimento de sistemas e algoritmos capazes de realizar tarefas que, normalmente, requerem inteligência humana (RUSSELL; NORVIG, 2016). Isso inclui atividades como reconhecimento de padrões, processamento de linguagem natural, tomada de decisões e aprendizado. A IA busca replicar, em algum nível, o processo de raciocínio, percepção e cognição que caracterizam a mente humana.

O conceito de IA remonta ao século XX, com os primeiros trabalhos de Alan Turing, que propôs a possibilidade de máquinas pensarem, e de John McCarthy, que cunhou o termo "inteligência artificial" em 1956 durante a Conferência de Dartmouth (MCCARTHY, 2006). Desde então, a IA tem evoluído significativamente, abrangendo áreas como robótica, visão computacional, jogos, diagnósticos médicos e, mais recentemente, assistentes virtuais como Siri e Alexa.

As abordagens de IA podem ser divididas em dois grandes grupos: IA fraca e IA forte. A IA fraca refere-se a sistemas projetados para realizar uma tarefa específica, como jogar xadrez ou reconhecer imagens, sem ter uma compreensão geral ou consciência. Já a IA forte se refere à inteligência que, teoricamente, seria capaz de compreender, aprender e aplicar conhecimentos em diversas áreas, de maneira similar a um ser humano. No entanto, a IA forte permanece uma aspiração e um campo de pesquisa, ainda distante de ser plenamente alcançada (RUSSELL; NORVIG, 2016).

1.4 Machine Learning

Machine Learning (ML), ou aprendizado de máquina, é um subcampo da IA que se concentra no desenvolvimento de algoritmos capazes de aprender a partir de dados e fazer previsões ou tomar decisões sem serem explicitamente programados para tal (BISHOP, 2006). Diferentemente dos sistemas tradicionais, onde os comportamentos são definidos por regras fixas, os sistemas de ML ajustam seus próprios modelos com base em padrões detectados nos dados.

A ideia central do ML é que um sistema possa identificar padrões e regularidades a partir da experiência com dados, e com isso, ser capaz de realizar inferências ou previsões sobre novos dados. Algoritmos de ML são aplicados em uma ampla gama de áreas, desde recomendações de filmes e diagnósticos médicos (RODER et al., 2023) até otimização de carteiras de investimentos e previsões de mercado (MURPHY, 2012).

Os métodos de ML podem ser divididos em três categorias principais de acordo com o paradigma de aprendizado:

- **Aprendizado supervisionado**: Os algoritmos aprendem a partir de um conjunto de dados rotulados, ou seja, onde as entradas estão associadas às saídas corretas. Um exemplo comum é a classificação de e-mails como "spam" ou "não spam", onde o modelo aprende com exemplos previamente rotulados.
- **Aprendizado não supervisionado**: Aqui, o algoritmo tenta identificar padrões nos dados sem qualquer rótulo ou classificação pré-definida. Um exemplo é a análise de agrupamentos, onde o algoritmo segmenta os dados em grupos com características similares.

- **Aprendizado por reforço**: Nesse tipo de aprendizado, o algoritmo aprende através de tentativa e erro, recebendo recompensas ou penalidades com base em suas ações. Essa abordagem é comumente utilizada em robótica e jogos (SUTTON; BARTO, 2018).

A área de ML tem revolucionado diversos setores e indústrias ao permitir que sistemas analisem grandes volumes de dados, detectem padrões ocultos e automatizem decisões com uma precisão cada vez maior. No mercado financeiro, por exemplo, algoritmos de ML são amplamente utilizados para prever movimentos de preços e auxiliar na alocação de ativos (HULL, 2019).

1.5 Justificativa

A rápida evolução tecnológica e o aumento da disponibilidade de grandes volumes de dados (Big Data) transformaram a maneira como diversas indústrias, incluindo o setor financeiro, tomam decisões estratégicas. Nesse contexto, os algoritmos de ML tornaram-se ferramentas essenciais para a análise de dados complexos e a criação de previsões mais precisas. O uso de ML no mercado financeiro tem se intensificado, especialmente em áreas como a previsão de preços de ativos, otimização de portfólios e gerenciamento de risco. Essas técnicas possibilitam a identificação de padrões ocultos e comportamentos emergentes que seriam difíceis ou impossíveis de serem detectados por métodos tradicionais (CHEN et al., 2021; MA; HAN; WANG, 2021).

Tradicionalmente, a tomada de decisão financeira dependia fortemente da análise de indicadores clássicos, como médias móveis e Índice de Força Relativa, combinados com a experiência e intuição de analistas financeiros. Embora esses métodos ainda desempenhem um papel fundamental, eles são limitados na sua capacidade de lidar com a crescente complexidade e volume de dados do mercado financeiro global. Com o uso de técnicas de ML, como redes neurais e algoritmos de ensemble (por exemplo: Random Forest e Gradient Boosting), é possível processar uma quantidade significativamente maior de variáveis em tempo real e, assim, identificar relações não lineares entre diversos fatores que impactam o comportamento dos ativos financeiros (MURPHY, 2012).

Um dos principais atrativos da aplicação de ML no setor financeiro é sua capacidade de lidar com ambientes de incerteza e volatilidade. Algoritmos como as Redes Neurais Recorrentes (do inglês *Recurrent Neural Networks* - RNN) e Memória de Longo Curto Prazo (do inglês, *Long Short-Term Memory* - LSTM) têm sido amplamente estudados para previsão de séries temporais, incluindo preços de ações, devido à sua habilidade em capturar dependências temporais de longo prazo. Essas técnicas permitem não apenas uma previsão mais robusta de movimentos de mercado, mas também possibilitam a modelagem de cenários de crise, como mudanças abruptas no mercado provocadas por eventos políticos ou desastres naturais (RUSSELL; NORVIG, 2016; BISHOP, 2006).

Além disso, a justificativa para o uso de ML no mercado financeiro brasileiro se torna ainda mais evidente devido à natureza dinâmica e, frequentemente, volátil do mercado local. A economia brasileira, marcada por ciclos de crescimento e recessão, políticas econômicas instáveis e crises periódicas, apresenta desafios adicionais para investidores, tanto institucionais quanto individuais. Nesse cenário, o ML pode auxiliar no desenvolvimento de estratégias de investimento mais robustas, que levem em consideração não apenas o comportamento histórico dos ativos, mas também a complexidade dos fatores macroeconômicos que afetam o mercado brasileiro (BERNHEIM, 1998; FAVERO; MEIRELES, 2013).

Um outro aspecto importante a ser considerado é a evolução das plataformas de negociação algorítmica, que têm permitido que investidores de diferentes perfis, incluindo pequenos investidores, tenham acesso a ferramentas de análise avançada anteriormente restritas a grandes instituições financeiras. Essas plataformas, muitas vezes alimentadas por algoritmos de ML, permitem a execução de estratégias automatizadas em tempo real, ajustando a composição do portfólio de acordo com as condições de mercado (HULL, 2019). Dessa forma, a democratização do acesso a essas ferramentas reforça a relevância do estudo de ML em decisões de investimento, ampliando o impacto dessas técnicas em diversos níveis do mercado financeiro.

Por fim, é importante destacar que o uso de ML para auxiliar em decisões financeiras está alinhado com a tendência de adoção crescente de tecnologias emergentes no setor financeiro, como IA, *blockchain* e *fintechs*. O estudo contínuo sobre a aplicação dessas técnicas no mercado brasileiro não apenas contribui para a inovação do setor, mas também permite que as empresas e investidores do país se mantenham competitivos em um mercado global cada vez mais interconectado (GOETZMANN; ROUWENHORST, 2005; GREENWOOD; HANSON; STEIN, 2019).

1.6 Problema

Embora o ML tenha sido amplamente adotado em diversas áreas do mercado financeiro global, sua aplicação no mercado brasileiro ainda enfrenta desafios significativos. Um dos principais entraves é a volatilidade do mercado, que torna difícil a criação de modelos preditivos estáveis e precisos. A economia brasileira é marcada por incertezas políticas e oscilações econômicas frequentes, o que contribui para variações abruptas nos preços de ativos financeiros, dificultando a capacidade de algoritmos de ML de identificar padrões consistentes (GARCIA, 2011).

Além disso, a liquidez limitada de alguns ativos no mercado brasileiro também representa um obstáculo. Ao contrário de mercados mais desenvolvidos, onde há um volume maior de negociações diárias, o Brasil ainda apresenta setores e ativos com baixa liquidez, o que pode prejudicar a precisão dos modelos ao lidar com lacunas significativas de dados (GOETZMANN;

ROUWENHORST, 2005).

Outro desafio crucial é a disponibilidade e a qualidade dos dados históricos. A aplicação de ML depende de grandes volumes de dados para treinar os modelos e gerar previsões precisas. No entanto, no Brasil, a coleta e a manutenção de dados financeiros de alta qualidade nem sempre estão disponíveis de forma acessível ou completa. Além disso, os dados podem estar dispersos em diferentes plataformas e não padronizados, dificultando a análise eficiente e a criação de bases de dados robustas que alimentam os modelos de ML (BERNHEIM, 1998).

A regulamentação do mercado financeiro no Brasil também apresenta barreiras. Embora as autoridades brasileiras, como a Comissão de Valores Mobiliários (CVM), tenham buscado modernizar as regulamentações, o ritmo de adaptação para acomodar inovações tecnológicas, como o uso de IA e ML, pode ser lento. Isso cria um ambiente de incerteza regulatória que desencoraja a adoção de soluções baseadas em ML em maior escala (SEVERINO; COIMBRA, 2016).

Adicionalmente, há o desafio do entendimento técnico das decisões geradas por algoritmos de ML. A explicabilidade (ou falta dela) dos modelos de IA é uma preocupação crescente, tanto no Brasil quanto globalmente. Muitas vezes, os algoritmos geram decisões com base em padrões de dados que são difíceis de serem interpretados por profissionais do mercado financeiro sem um profundo conhecimento técnico. A complexidade intrínseca dos algoritmos, como CNNs ou métodos de *ensemble*, pode resultar em um "caixa preta"que é difícil de justificar, especialmente em um ambiente regulatório que exige transparência nas decisões financeiras (RUSSELL; NORVIG, 2016; SUTTON; BARTO, 2018).

Por fim, há a falta de especialização técnica por parte de profissionais do setor financeiro. Embora o ML e a IA estejam ganhando terreno no Brasil, a integração dessas tecnologias no cotidiano das empresas financeiras exige uma equipe com conhecimentos avançados em ciência de dados, estatística e TI. A carência de profissionais com essa formação específica pode retardar a adoção e implementação de modelos de ML no mercado financeiro brasileiro (HULL, 2019; MURPHY, 2012).

Portanto, frente aos desafios citados, este trabalho busca modelar e avaliar abordagens baseadas em ML para a realização de operações de compra e venda de ações do mercado brasileiro, visando melhorar as decisões de investimento e maximizar o retorno em cenários de alta volatilidade.

1.7 Objetivos

O objetivo geral deste trabalho é investigar o uso de técnicas de ML na composição de portfólios de ativos financeiros, com foco na maximização de retornos a médio e longo prazo (5 a 10 anos).

Para atingir esse objetivo, são estabelecidos os seguintes objetivos específicos:

- Identificar e selecionar indicadores importantes (e processá-los) para a caracterização dos diferentes ativos;
- Empregar técnicas de ML para predizer o comportamento dos diferentes ativos com uma abordagem de classificação, a partir dos indicadores selecionados;
- Identificar diferentes agrupamentos de ativos para faixas de retornos financeiros a partir das suas características representadas pelos indicadores;
- Selecionar ativos para compor um portfólio de médio e longo prazo, de acordo com as predições realizadas pelas técnicas de ML.

1.8 Estrutura do Trabalho

Este trabalho está organizado em seis capítulos, cada um estruturado para fornecer uma visão detalhada das etapas de desenvolvimento e análise deste estudo.

O primeiro capítulo, **Introdução**, contextualiza o tema abordado, discute a relevância da pesquisa no contexto atual do mercado financeiro e apresenta os objetivos do trabalho. Essa seção inicial estabelece as motivações do estudo e orienta o leitor sobre o percurso seguido na investigação.

O segundo capítulo, **Fundamentação Teórica**, apresenta os principais conceitos do mercado financeiro, ML e indicadores financeiros. Essa base teórica fornece o suporte necessário para a compreensão das técnicas e estratégias utilizadas ao longo do trabalho, incluindo as metodologias para caracterizar e avaliar os ativos financeiros.

O terceiro capítulo, **Revisão da Literatura**, explora os estudos mais relevantes na aplicação de ML ao mercado financeiro. A revisão concentra-se em estratégias de investimento e métodos de otimização de portfólios, destacando os avanços e as abordagens inovadoras empregadas para aumentar a precisão e a eficiência nas decisões de investimento.

No quarto capítulo, **Material e Métodos**, são detalhadas as técnicas de ML aplicadas, a seleção de dados e o procedimento experimental utilizado para análise. Este capítulo descreve a metodologia de forma sistemática, permitindo a reprodução dos experimentos e a validação dos resultados.

O quinto capítulo, **Resultados Experimentais**, apresenta as predições dos modelos e as análises realizadas sobre o comportamento dos ativos financeiros, bem como a composição dos portfólios formados. São discutidas as implicações práticas dos resultados, incluindo a eficácia das estratégias de investimento e a robustez dos modelos empregados.

Por fim, o sexto capítulo, **Conclusão**, sintetiza as principais descobertas e contribuições do trabalho, avaliando as limitações enfrentadas e sugerindo direções para estudos futuros. Ao final, a seção de **Referências** lista as fontes e estudos que fundamentam o desenvolvimento deste trabalho, assegurando a integridade e o rigor acadêmico da pesquisa.

2 Fundamentação Teórica

Este capítulo apresenta os conceitos fundamentais e as técnicas utilizadas para embasar o desenvolvimento deste trabalho. Primeiramente, discutimos o funcionamento e a importância do mercado de ações, destacando os fatores que influenciam seu comportamento e as principais abordagens para análise de ativos. Em seguida, introduzimos os principais indicadores financeiros, que são amplamente utilizados na avaliação de ações e outros ativos financeiros. Por fim, abordamos os algoritmos de aprendizado de máquina mais utilizados em problemas de previsão e classificação no mercado financeiro, explorando suas características, funcionamento e aplicabilidade. Essa base teórica é essencial para o entendimento e implementação das estratégias e modelos propostos ao longo deste trabalho.

2.1 O Mercado de Ações

O mercado de ações é o ambiente no qual ações de empresas públicas são negociadas. Ele desempenha um papel crucial no financiamento das empresas e na criação de oportunidades de investimento para indivíduos e instituições. As ações representam uma fração da propriedade da empresa e, ao adquiri-las, os investidores se tornam proprietários parciais da companhia, tendo direito a dividendos e ao aumento do valor do capital investido, caso a empresa cresça e seja bem-sucedida. Os principais mercados de ações, como a Bolsa de Valores de São Paulo (B3) e a Bolsa de Nova York (NYSE), operam como plataformas para essas transações, reguladas por órgãos como a Comissão de Valores Mobiliários (CVM) no Brasil e a Securities and Exchange Commission (SEC) nos EUA (SEGAL, 2021; (CVM), 2021).

A dinâmica do mercado de ações é influenciada por diversos fatores, como oferta e demanda, resultados financeiros das empresas, condições econômicas gerais, política monetária, entre outros. O desempenho de uma ação é frequentemente monitorado por investidores através de uma série de indicadores financeiros que auxiliam na identificação de tendências e na tomada de decisões sobre compra ou venda de ativos.

Para analisar o mercado de ações, são empregadas três abordagens principais: a análise fundamentalista, a análise técnica e a análise quantitativa. A análise fundamentalista envolve o estudo detalhado dos fundamentos da empresa, como balanços financeiros, receitas, despesas e potencial de crescimento. Esse tipo de análise ajuda a determinar o valor intrínseco de uma ação, permitindo que os investidores avaliem se o ativo está subvalorizado ou supervalorizado.

A análise técnica, por outro lado, se concentra no comportamento histórico dos preços e volumes negociados, utilizando gráficos e indicadores como médias móveis, Índice de Força Relativa e padrões de *candlestick* para prever movimentos futuros. Esta abordagem é

amplamente utilizada por traders que buscam lucrar com as variações de curto prazo no preço das ações.

Já a análise quantitativa utiliza modelos matemáticos e estatísticos para prever movimentos de mercado e identificar oportunidades de investimento. Com o avanço da tecnologia e a popularização do aprendizado de máquina, essa abordagem se tornou cada vez mais sofisticada, permitindo o uso de algoritmos para processar grandes volumes de dados e identificar padrões complexos que não seriam facilmente percebidos por análises tradicionais.

Além dessas abordagens, o mercado de ações desempenha um papel importante na alocação eficiente de recursos na economia, fornecendo capital para as empresas que desejam expandir suas operações e, ao mesmo tempo, oferecendo aos investidores a oportunidade de participar do crescimento econômico. A interação entre investidores e empresas cria um ciclo virtuoso de crescimento, inovação e geração de riqueza. Contudo, o sucesso no mercado de ações requer um entendimento profundo dos fatores que afetam os preços, das características das empresas e dos setores da economia, além da capacidade de gerenciar riscos e de lidar com a volatilidade.

Por fim, é importante notar que, embora o mercado de ações seja uma poderosa ferramenta de crescimento econômico e investimento, ele também é inerentemente arriscado. A volatilidade e a incerteza fazem parte do processo, e por isso os investidores precisam estar preparados para os desafios e oscilações, compreendendo os riscos envolvidos e utilizando estratégias que os ajudem a mitigar possíveis perdas.

2.2 Indicadores Financeiros

Os indicadores financeiros são ferramentas matemáticas utilizadas por investidores para analisar o comportamento histórico dos preços de ativos e prever seus movimentos futuros. Eles são calculados a partir de séries temporais de preços e volumes de negociação e têm como objetivo simplificar a interpretação de dados de mercado complexos. Entre os principais indicadores, destacam-se os de tendência, volatilidade, força relativa e osciladores.

Indicadores podem ser usados em uma ampla gama de estratégias, desde a análise técnica, que envolve a avaliação de padrões de preços e volumes, até modelos quantitativos mais avançados, que utilizam algoritmos de aprendizado de máquina para prever o comportamento dos ativos. A combinação de diferentes indicadores permite ao investidor ter uma visão mais abrangente sobre o ativo, considerando múltiplas dimensões do seu comportamento.

2.2.1 Indicadores empregados

A seguir, detalhamos alguns dos principais indicadores utilizados em análises financeiras e aplicados nesse trabalho:

- RSI (*Relative Strength Index*): O RSI é um oscilador que mede a velocidade e a mudança dos movimentos de preço. Ele varia de 0 a 100 e é geralmente usado para identificar condições de sobrecompra ou sobrevenda de um ativo. Valores acima de 70 indicam que o ativo está sobrecomprado, enquanto valores abaixo de 30 sugerem sobrevenda (MURPHY, 2012).

- Dir: A Direção do preço no dia é um indicador que captura a variação entre o preço de abertura e o preço de fechamento de um ativo em um determinado dia. Se o preço de fechamento for superior ao preço de abertura, considera-se que o ativo teve uma direção positiva (alta) nesse dia, e se o preço de fechamento for inferior ao de abertura, a direção é negativa (baixa). Esse indicador é útil para observar o comportamento diário dos preços e identificar tendências de curto prazo.

- Dir_p: A Direção do preço em um intervalo de períodos (p) é um indicador que mede a variação entre o preço de fechamento inicial e o preço de fechamento final ao longo de um período p. Isso permite observar a tendência de um ativo em uma janela maior de tempo, considerando flutuações diárias e identificando padrões mais amplos de movimento. Se o preço de fechamento ao final do período for maior que o preço de fechamento inicial, considera-se que houve uma direção positiva, e vice-versa para uma direção negativa.

- MACD (*Moving Average Convergence Divergence*): O MACD é um indicador que combina médias móveis de diferentes períodos para mostrar mudanças na força, direção, momento e duração de uma tendência no preço de um ativo. É amplamente utilizado para identificar pontos de reversão de tendência (HULL, 2019).

- Volatilidade: A volatilidade mede a amplitude das flutuações de preço de um ativo ao longo do tempo. Altos níveis de volatilidade indicam maior risco, mas também podem sugerir maiores oportunidades de ganho. A volatilidade é frequentemente usada em modelos de precificação de opções e em estratégias de *hedge* (GOETZMANN; ROUWENHORST, 2005).

- Lag N: Esses indicadores são usados para introduzir defasagens temporais nos dados, permitindo a análise de tendências passadas e a influência dessas tendências no comportamento futuro dos ativos. É inspirado na análise de estacionariedade com testes como os de *Augmented Dickey Fuller* (ADF) (LOPEZ, 1997)

- VWAP (*Volume Weighted Average Price*): O VWAP é o preço médio ponderado pelo volume de um ativo em um determinado período. Ele é amplamente utilizado por traders institucionais para avaliar a eficiência das suas execuções de negociação (MURPHY, 2012).

- Detrend: A técnica de detrend é utilizada para remover a tendência subjacente dos dados, facilitando a análise de oscilações cíclicas ou padrões de curto prazo.

- Bin RSI e Bin SOSC: Indicadores binários derivados de RSI e de osciladores estocásticos, são usados para simplificar a identificação de condições específicas de mercado, como sobrecompra ou sobrevenda, transformando essas informações em variáveis booleanas.
- O CCI (*Commodity Channel Index*) é um oscilador utilizado para identificar desvios do preço em relação à sua média. Ele foi desenvolvido por Donald Lambert e é frequentemente usado em commodities, mas também é eficaz na análise de ações (LAMBERT, 1980).
- Z-Score: O Z-score é uma medida estatística que indica quantos desvios padrão um ponto de dado está em relação à média. É amplamente utilizado para detectar anomalias nos preços.

Esses indicadores desempenham um papel fundamental na construção de estratégias de investimento baseadas em dados financeiros históricos, fornecendo ideias e entendimentos que ajudam a antecipar movimentos do mercado.

2.3 Algoritmos de Aprendizado de Máquina

Nesta seção, são apresentados alguns dos principais algoritmos de aprendizado de máquina utilizados em problemas de previsão e classificação no mercado financeiro. Os algoritmos apresentados a seguir são amplamente discutidos na literatura, mas não necessariamente serão utilizados neste trabalho.

2.3.1 Random Forest

O algoritmo Random Forest (RF) é uma técnica de aprendizado de máquina baseada em métodos de *ensemble*, que combina múltiplos classificadores simples para melhorar a precisão e a robustez do modelo. Especificamente, o Random Forest utiliza múltiplas Árvores de Decisão (*Decision Trees*) e a técnica de *Bagging* (Bootstrap Aggregating) para construir um modelo mais eficiente (BREIMAN, 2001).

A ideia central da Random Forest é mitigar problemas como *overfitting* e alta variância que ocorrem em Árvores de Decisão individuais. Isso é alcançado através da construção de um grande número de Árvores de Decisão, cada uma treinada em um subconjunto aleatório dos dados de treinamento e utilizando um subconjunto aleatório de características para cada divisão (HO, 1995). O resultado final é obtido agregando as predições individuais das árvores, geralmente por votação majoritária (no caso de classificação) ou pela média (no caso de regressão).

O funcionamento do algoritmo Random Forest pode ser descrito em detalhes pelos seguintes passos:

1. **Amostragem por *Bootstrap***: Dado um conjunto de treinamento original D com N exemplos, criam-se M conjuntos de treinamento $D_1, D_2, \ldots, D_M$ por meio de amostragem com reposição (*bootstrap sampling*). Cada D_i tem o mesmo tamanho que D, mas devido à amostragem com reposição, aproximadamente 63,2% dos exemplos originais estão presentes em cada D_i, deixando cerca de 36,8% dos dados como amostras fora da bolsa (*out-of-bag samples*) (BREIMAN, 1996).

2. **Construção das Árvores de Decisão**:

   a) Para cada nó em uma árvore, um subconjunto aleatório de m características é selecionado a partir das K características totais, onde $m \ll K$.

   b) Entre essas m características, é escolhida a melhor divisão com base em uma medida de impureza, como o índice de Gini ou a entropia.

   c) A árvore é crescida até a máxima profundidade sem poda (*pruning*).

3. **Combinação dos Resultados**:

   - **Classificação**: Cada árvore $h_i(x)$ produz uma predição para a entrada x. A predição final $H(x)$ é obtida pela votação majoritária, conforme a Equação 2.1:

$$H(x) = \text{moda}\{h_1(x), h_2(x), \ldots, h_M(x)\} \tag{2.1}$$

   - **Regressão**: A predição final é a média das predições individuais, como mostrado na Equação 2.2:

$$H(x) = \frac{1}{M} \sum_{i=1}^{M} h_i(x) \tag{2.2}$$

2.3.1.1 Medidas de Impureza

Para determinar a qualidade de uma divisão em um nó da árvore, são utilizadas medidas de impureza como o índice de Gini ou a entropia.

2.3.1.1.1 Índice de Gini

O índice de Gini para um nó t é definido pela Equação 2.3:

$$G(t) = 1 - \sum_{i=1}^{C} p_i^2 \tag{2.3}$$

onde:

- C é o número de classes.
- p_i é a proporção de amostras pertencentes à classe i no nó t.

A redução no índice de Gini ao realizar uma divisão é calculada como mostrado na Equação 2.4:

$$\Delta G = G(t) - \left(\frac{N_L}{N_t} G(t_L) + \frac{N_R}{N_t} G(t_R)\right) \tag{2.4}$$

onde:

- $G(t)$ é a impureza do nó pai.
- $G(t_L)$ e $G(t_R)$ são as impurezas dos nós filho esquerdo e direito, respectivamente.
- N_t, N_L e N_R são o número de amostras no nó pai, filho esquerdo e filho direito, respectivamente.

2.3.1.1.2 Entropia e Ganho de Informação

A entropia é definida pela Equação 2.5:

$$H(t) = -\sum_{i=1}^{C} p_i \log_2 p_i \tag{2.5}$$

O ganho de informação ao realizar uma divisão é dado pela Equação 2.6:

$$\Delta H = H(t) - \left(\frac{N_L}{N_t} H(t_L) + \frac{N_R}{N_t} H(t_R)\right) \tag{2.6}$$

2.3.1.2 Seleção Aleatória de Características

A seleção aleatória de um subconjunto de características em cada nó é crucial para introduzir diversidade entre as árvores e reduzir a correlação entre elas. O valor típico para m (número de características selecionadas aleatoriamente em cada nó da árvore) é:

- Para classificação: $m = \sqrt{K}$
- Para regressão: $m = \frac{K}{3}$

onde K é o número total de características.

2.3.1.3 Estimativa de Erro *Out-of-Bag*

Uma vantagem do Random Forest é a capacidade de estimar o erro generalizado sem a necessidade de um conjunto de validação separado. Isso é feito utilizando as amostras *out-of-bag* (OOB) (BREIMAN, 2001).

Para cada amostra x_i que não foi incluída no conjunto de treinamento de uma árvore específica (ou seja, está fora da bolsa para aquela árvore), é possível obter uma predição $h_j(x_i)$. A estimativa do erro OOB é calculada como a taxa média de erro nessas predições.

2.3.1.4 Importância das Variáveis

A importância de uma variável pode ser medida de duas maneiras:

Redução Média da Impureza

É calculada somando a redução de impureza (por exemplo, Gini) proporcionada por cada variável em todas as árvores da floresta.

Permutação das Variáveis

Para cada árvore, os valores de uma variável X_k são permutados nas amostras OOB, e o aumento no erro de predição é medido. Um grande aumento indica alta importância da variável.

2.3.1.5 Vantagens e Desvantagens

As principais vantagens das Florestas Aleatórias incluem:

- **Redução da Variância**: Ao combinar várias árvores, a variância do modelo é reduzida, melhorando a capacidade de generalização.
- **Robustez a Outliers e Ruído**: Devido à natureza coletiva do modelo, outliers têm menos impacto no resultado final.
- **Capacidade de lidar com Dados Altamente Dimensionais**: A seleção aleatória de características em cada nó permite que o algoritmo seja eficiente em situações com muitas variáveis.
- **Estimativa Interna de Erro**: A possibilidade de estimar o erro OOB elimina a necessidade de um conjunto de validação separado.

Desvantagens:

- **Complexidade Computacional**: Construir um grande número de árvores pode ser computacionalmente intensivo, tanto em termos de tempo quanto de memória.

- **Perda de Interpretabilidade**: Diferentemente de uma única Árvore de Decisão, interpretar o modelo completo de uma Floresta Aleatória é mais complexo.

2.3.2 Aplicações e Melhorias

As Florestas Aleatórias têm sido aplicadas com sucesso em diversas áreas, como detecção de fraudes, bioinformática, reconhecimento de padrões, entre outras. Além disso, várias melhorias e extensões foram propostas, incluindo:

- **Extremely Randomized Trees (ExtraTrees)**: Variante que reduz ainda mais a variância ao aleatorizar também os pontos de corte em cada nó.
- **Random Forests para Dados Desbalanceados**: Métodos que ajustam a penalização de erros ou utilizam técnicas de reamostragem para lidar com classes desbalanceadas.
- **Combinação com Outros Métodos**: Integração com algoritmos como *Gradient Boosting* para melhorar o desempenho.

Como ilustrado na Figura 1, o diagrama mostra o funcionamento de uma Floresta Aleatória, destacando como várias árvores são combinadas para melhorar a robustez do modelo e reduzir *overfitting*.

2.3.3 XGBoost

O XGBoost (*eXtreme Gradient Boosting*) é um dos algoritmos de aprendizado de máquina mais populares e eficientes para problemas de classificação e regressão. Ele é baseado no conceito de *boosting*, onde vários modelos simples (chamados de "árvores fracas") são combinados para formar um modelo robusto e altamente preciso. Desenvolvido por Tianqi Chen e Carlos Guestrin em 2016 (CHEN; GUESTRIN, 2016), o XGBoost ganhou notoriedade por seu desempenho em competições de ciência de dados, além de ser amplamente utilizado em aplicações práticas devido à sua eficiência e precisão.

O *boosting* é uma técnica de *ensemble* em que vários modelos são treinados de forma sequencial. Diferentemente do *bagging* (usado no Random Forest), onde as árvores são treinadas de forma independente, o *boosting* constrói cada nova árvore com base nos erros cometidos pelas árvores anteriores. O XGBoost melhora essa abordagem ao utilizar uma variante de *Gradient Boosting*, que se concentra na otimização da função objetivo usando o método do gradiente (FRIEDMAN, 2001).

O XGBoost segue a ideia de construir modelos iterativamente, com cada novo modelo ajustando-se aos resíduos (*residuals*) dos modelos anteriores. O objetivo é minimizar uma função de perda, tal como o erro quadrático médio para regressão ou o log-loss para classificação. Vamos descrever o processo em detalhes:

1. **Inicialização**: O modelo começa com uma predição inicial, como a média das variáveis-alvo no caso de regressão, ou uma predição constante no caso de classificação. Esta predição inicial $\hat{y}^{(0)}$ é atualizada em cada iteração.

2. **Cálculo dos Resíduos**: A cada iteração t, calculam-se os resíduos, ou seja, a diferença entre os valores reais y_i e as predições atuais $\hat{y}_i^{(t)}$. O objetivo é ajustar uma nova árvore de decisão para prever esses resíduos.

3. **Crescimento de uma Nova Árvore**: Uma nova árvore de decisão é ajustada para prever os resíduos calculados na iteração anterior. A árvore tenta minimizar a função de perda ao ajustar os erros das predições anteriores.

4. **Combinação dos Resultados**: A predição final é atualizada somando-se a predição da nova árvore, multiplicada por um fator de aprendizado η (também conhecido como *learning rate*), conforme mostrado na Equação 2.7:

$$\hat{y}_i^{(t+1)} = \hat{y}_i^{(t)} + \eta \cdot f_t(x_i) \tag{2.7}$$

   Onde $f_t(x_i)$ é a predição da nova árvore para o exemplo x_i, e η controla o quanto cada nova árvore contribui para o modelo final.

5. **Iterações**: O processo é repetido por T iterações ou até que a função de perda não melhore significativamente.

2.3.3.1 Função de Objetivo e Regularização

Um dos aspectos chave do XGBoost é sua função de objetivo personalizada, que inclui termos de regularização para evitar overfitting e melhorar a generalização do modelo. A função de objetivo geral do XGBoost é dada pela Equação 2.8:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{T} \Omega(f_k) \tag{2.8}$$

Onde:

- $\ell(y_i, \hat{y}_i^{(t)})$ é a função de perda que mede a diferença entre o valor real y_i e a predição $\hat{y}_i^{(t)}$.
- $\Omega(f_k)$ é um termo de regularização que penaliza a complexidade do modelo, conforme descrito na Equação 2.9:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2.9}$$

  Onde T é o número de folhas da árvore, w_j é o peso associado à j-ésima folha, γ controla a complexidade da árvore, e λ penaliza o tamanho dos pesos.

2.3.3.2 Importância das Variáveis

O XGBoost oferece várias maneiras de medir a importância das variáveis. A mais comum é baseada na **ganho médio** (*average gain*) obtido ao usar uma variável específica para realizar uma divisão nas árvores. Quanto maior o ganho associado a uma variável, mais importante ela é para o modelo final. Outra abordagem é medir a **frequência de uso** (*frequency*), que indica quantas vezes uma variável foi usada para dividir os nós em todas as árvores.

2.3.3.3 Vantagens e Desvantagens

O XGBoost oferece várias vantagens em comparação com outros métodos de aprendizado de máquina:

- **Alto Desempenho Computacional**: O XGBoost foi projetado para ser eficiente, utilizando otimizações como paralelização durante o crescimento das árvores e técnicas de compressão de dados para lidar com grandes conjuntos de dados (CHEN; GUESTRIN, 2016).
- **Regularização Incorporada**: A inclusão de regularização na função de objetivo ajuda a controlar o *overfitting*, especialmente em conjuntos de dados de alta dimensionalidade.
- **Flexibilidade**: O XGBoost pode ser usado para resolver uma ampla gama de problemas, desde regressão até classificação multi-classe, com diferentes funções de perda.
- **Interpretação**: Ele oferece várias métricas para medir a importância das variáveis e permite visualizar o impacto de cada árvore no modelo final.

No entanto, algumas desvantagens incluem:

- **Complexidade Computacional**: Apesar das otimizações, o treinamento do XGBoost pode ser custoso em termos de tempo e memória, especialmente em grandes conjuntos de dados.
- **Ajuste de Hiper-parâmetros**: O XGBoost tem muitos hiper-parâmetros que precisam ser ajustados para obter o melhor desempenho, o que pode tornar o processo de ajuste do modelo mais complicado.

2.3.3.4 Aplicações Práticas

O XGBoost tem sido amplamente utilizado em diversas áreas, como:

- **Finanças**: Modelagem de risco de crédito, detecção de fraudes e previsão de preços de ativos.

- **Bioinformática**: Predição de interações proteicas e classificação de tipos de câncer.
- **Competição de Ciência de Dados**: O XGBoost tem sido utilizado em muitas competições de ciência de dados, como no Kaggle, devido ao seu desempenho superior em muitos problemas.

Como ilustrado na Figura 2, o esquema de funcionamento do XGBoost mostra como o algoritmo constrói e ajusta suas árvores iterativamente, otimizando o desempenho ao longo das iterações.

2.3.4 Support Vector Machines

As Máquinas de Vetores de Suporte (do inglês *Support Vector Machines*) são uma classe de algoritmos amplamente utilizados para tarefas de classificação e regressão em aprendizado de máquina. O conceito principal por trás das SVMs baseia-se em encontrar um hiperplano que separe os dados de forma ideal, maximizando a margem de separação entre diferentes classes. Este algoritmo foi originalmente proposto por Vapnik e colegas no contexto da Teoria do Aprendizado Estatístico (TAE) (VAPNIK, 1995).

A SVM busca um classificador $f(x)$ que tenha a capacidade de generalizar bem os dados e, ao mesmo tempo, mantenha um bom desempenho no conjunto de treinamento. O desafio principal em aprendizado supervisionado é evitar o supertreinamento (*overfitting*), que ocorre quando o modelo se ajusta demais aos dados de treinamento, e o subtreinamento (*underfitting*), que reflete uma baixa capacidade de aprendizado (CORTES; VAPNIK, 1995).

O objetivo da SVM é encontrar um hiperplano que maximize a margem de separação entre as classes. Dados m exemplos de treinamento $X = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$, onde $x_i \in R^n$ e $y_i \in \{-1, 1\}$, a SVM procura encontrar um vetor w e um escalar θ que definem o hiperplano de decisão $w^T x + \theta = 0$, de modo a separar as amostras com a maior margem possível (VAPNIK, 1998). A Figura 3 ilustra um exemplo de hiperplano de separação com margens de suporte.

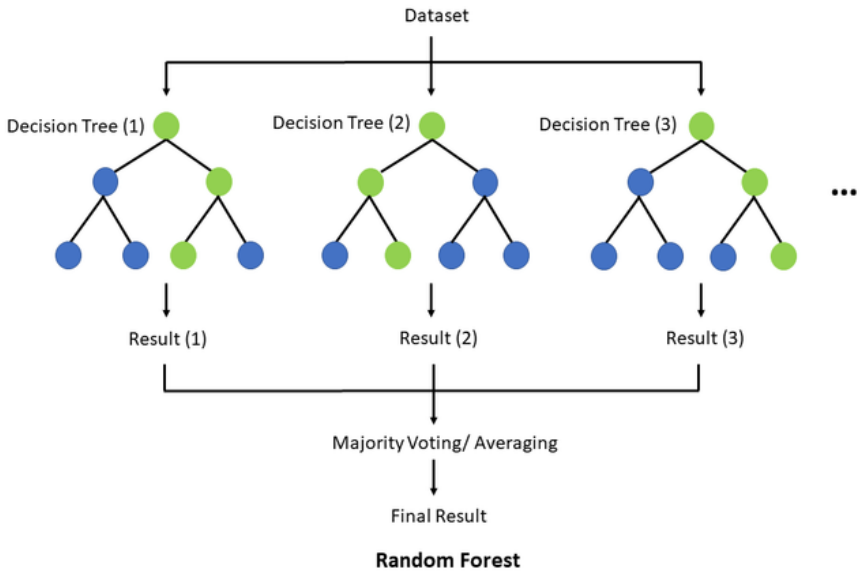
A margem é definida pela distância entre o hiperplano de separação e as amostras mais próximas de cada classe, conhecidas como *vetores de suporte*. Estas amostras são as mais críticas para a determinação do hiperplano (BURGES, 1998).

O problema de otimização subjacente à SVM é descrito pela Equação 2.10:

$$\min_{w,\theta} \frac{1}{2}\|w\|^2 \tag{2.10}$$

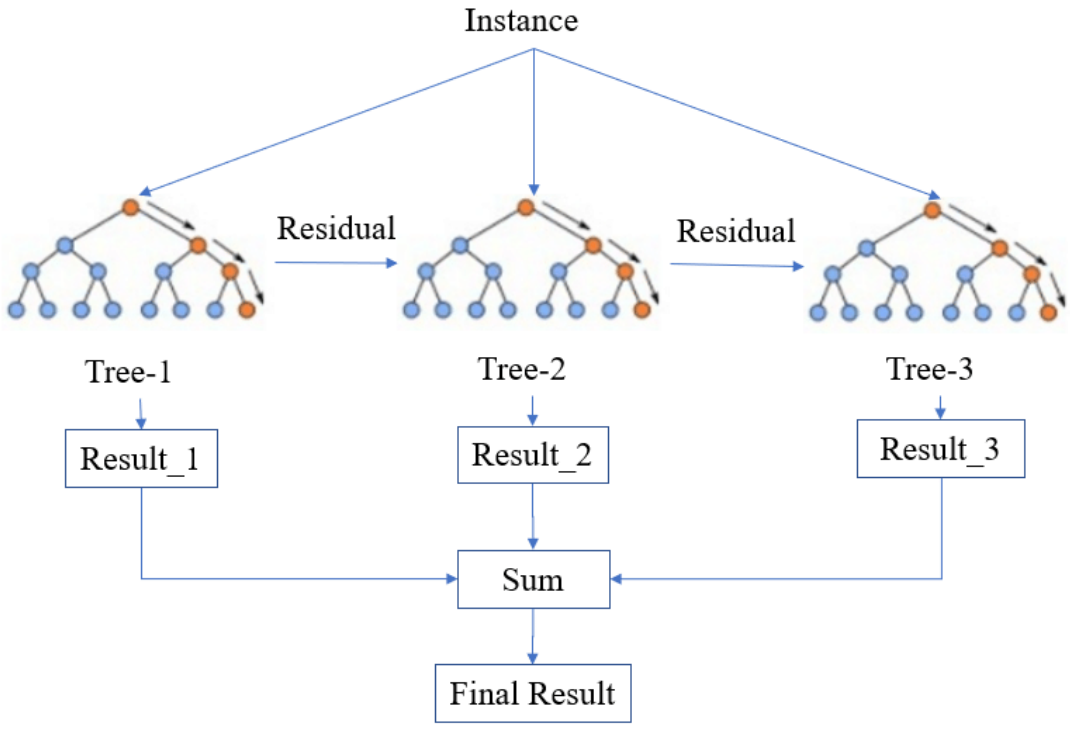
Sujeito às restrições de separabilidade descritas pela Equação 2.11:

Figura 1 – Diagrama ilustrativo do funcionamento de uma Floresta Aleatória.



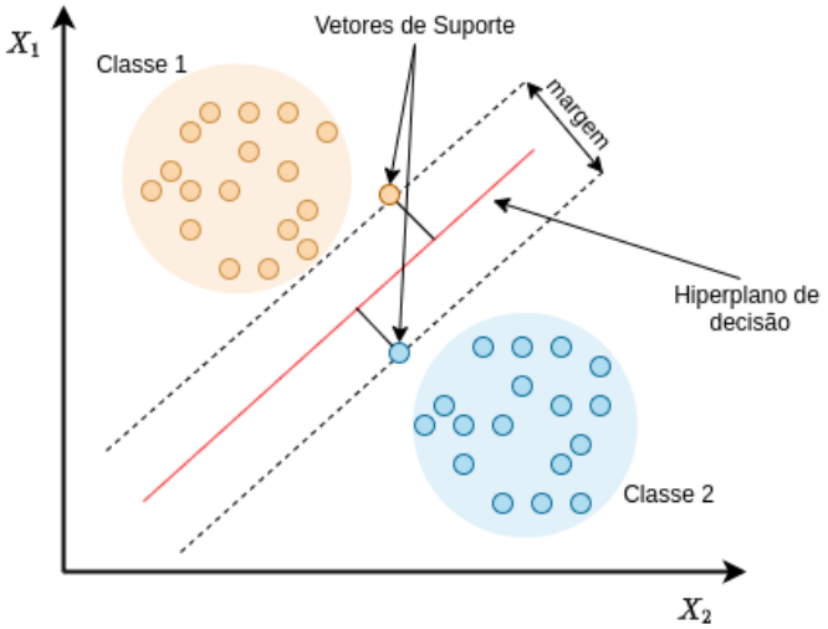
Fonte adaptada de Interactive Chaos.

Figura 2 – Esquema de funcionamento do algoritmo XGBoost.



Fonte: <https://www.researchgate.net/figure/Simplified-structure-of-XGBoost_fig2_348025909>

Figura 3 – Hiperplano de separação em SVM com margens.



Fonte: Barbosa et al. (2021)

$$y_i(w^T x_i + \theta) \geq 1, \quad \forall i = 1, 2, \ldots, m \tag{2.11}$$

Essas restrições garantem que os exemplos de treinamento fiquem do lado correto do hiperplano, com uma margem mínima de 1. A solução para este problema de otimização pode ser obtida utilizando multiplicadores de Lagrange, resultando na forma dual do problema.

A função Lagrangiana é definida pela Equação 2.12:

$$L(w, \theta, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(w^T x_i + \theta) - 1] \tag{2.12}$$

Maximizando esta função em relação a α, podemos reescrever o problema na sua forma dual conforme a Equação 2.13:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{2.13}$$

Sujeito às seguintes restrições:

1. **Restrição de soma zero** (Equação 2.14):

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{2.14}$$

2. **Restrição de não negatividade** (Equação 2.15):

$$\alpha_i \geq 0, \quad \forall i \tag{2.15}$$

A solução para o problema dual fornece os valores ótimos dos multiplicadores de Lagrange α^*, a partir dos quais o vetor w pode ser reconstruído, conforme mostrado na Equação 2.16:

$$w = \sum_{i=1}^{m} \alpha_i^* y_i x_i \tag{2.16}$$

2.3.4.1 SVM com Margens Suaves

Em casos onde os dados não são linearmente separáveis, introduzimos o conceito de margens suaves. A ideia é permitir que algumas amostras violem a margem de separação. Para isso, introduzimos variáveis de folga $\xi_i \geq 0$ que permitem que os exemplos de treinamento possam estar entre ou mesmo do lado errado da margem (SMOLA; SCHÖLKOPF, 1998).

Neste caso, o problema de otimização é reformulado conforme a 2.17:

$$\min_{w,\theta,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \tag{2.17}$$

Sujeito às restrições descritas na Equação 2.18:

$$y_i(w^T x_i + \theta) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i \tag{2.18}$$

Onde C é um parâmetro que controla o trade-off entre a maximização da margem e o erro de classificação no conjunto de treinamento.

2.3.4.2 SVM com Núcleos (Kernel Trick)

Muitas vezes, os dados não são linearmente separáveis no espaço original de entrada. Neste caso, uma abordagem comum é mapear os dados para um espaço de maior dimensão, onde eles possam ser linearmente separáveis. Este mapeamento é realizado implicitamente através de funções *kernel*, sem a necessidade de calcular explicitamente a transformação (SCHÖLKOPF; SMOLA, 2002).

A função de kernel $K(x_i, x_j)$ é definida como o produto interno entre as imagens dos vetores x_i e x_j no novo espaço de características, conforme descrito pela Equação 2.19:

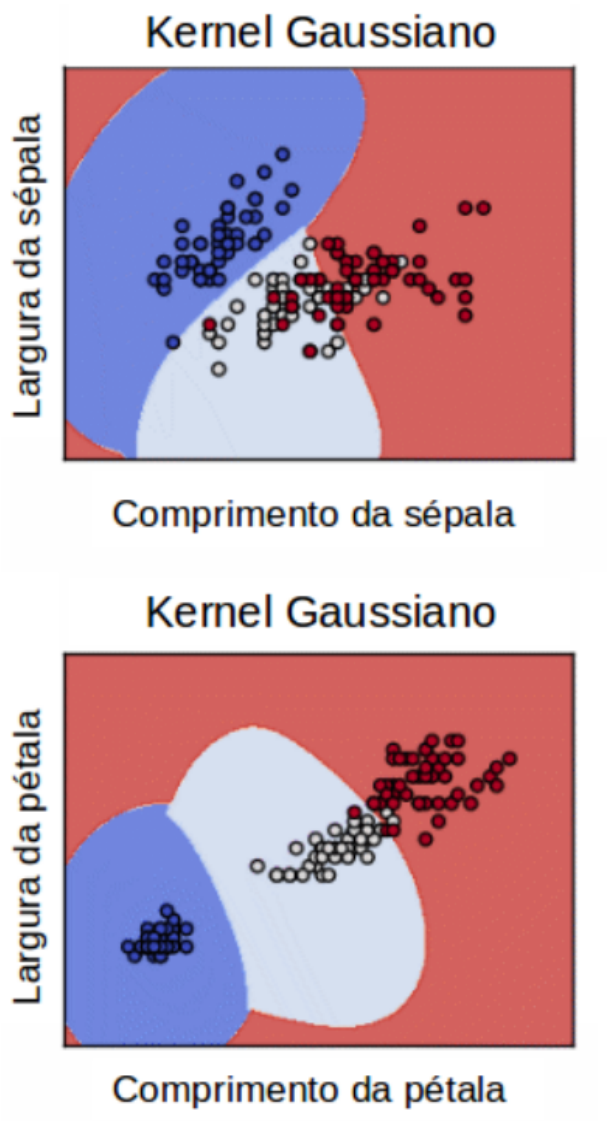
$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \tag{2.19}$$

Alguns exemplos comuns de funções de kernel incluem:

- Kernel linear: $K(x_i, x_j) = x_i^T x_j$
- Kernel polinomial: $K(x_i, x_j) = (x_i^T x_j + 1)^d$
- Kernel Gaussiano (RBF): $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$

A utilização de kernels permite que a SVM seja aplicada em problemas de classificação com fronteiras de decisão complexas, como ilustrado na Figura 4.

Figura 4 – Exemplo de separação não-linear usando um kernel Gaussiano.



Fonte: (FERREIRA, n.d.)

3 Revisão da literatura

Este capítulo apresenta uma análise detalhada de trabalhos acadêmicos e pesquisas que exploram o uso de técnicas de aprendizado de máquina e indicadores financeiros aplicados ao mercado financeiro, especialmente no contexto de otimização de portfólio e previsão de preços de ativos. Inicialmente, na seção de Trabalhos Relacionados, discutimos diversos estudos que implementaram algoritmos de AM, como Random Forest, XGBoost, SVM, LSTM e CNN, para identificar tendências de mercado e construir estratégias de investimento otimizadas.

Além disso, destacamos os principais indicadores financeiros utilizados, como volatilidade, médias móveis e RSI, e as métricas de avaliação que permitem mensurar o desempenho e os riscos associados a essas estratégias. Por fim, na seção de Discussão, sintetizamos as contribuições dos trabalhos analisados e discutimos como esses estudos fundamentam e justificam as escolhas metodológicas e técnicas do presente trabalho.

3.1 Trabalhos relacionados

Nesta seção é realizada uma revisão sobre trabalhos que abordaram esse mesmo âmbito de pesquisa, ou seja, técnicas de AM no mercado financeiro. Os trabalhos citados serão usados como base para o desenvolvimento do presente estudo.

No trabalho desenvolvido por Chen et al. (2021), os autores usaram técnicas de AM com técnicas estatísticas para prever e otimizar a montagem de um portfólio de investimentos. Em geral, o trabalho propõe uma nova abordagem para a construção de carteiras, combinando um modelo de aprendizado de máquina para previsão de ações com o modelo média-variância (MV) para seleção de carteira. O método IFAXGBoost envolve duas etapas principais: previsão de ações e seleção de carteira. Na primeira etapa, o IFAXGBoost combina o XGBoost com um algoritmo Firefly aprimorado (IFA) para prever os preços das ações. Na segunda etapa, as ações com maior potencial de retorno são selecionadas com base nas previsões, e o modelo MV é usado para alocar a proporção de investimento da carteira.

Os indicadores utilizados incluem volatilidade, médias móveis e índice de força relativa. As métricas de avaliação usadas foram o índice Sharpe e o retorno ajustado ao risco. O método foi avaliado com dados da bolsa de valores de Xangai, demonstrando sua eficácia e superioridade em comparação com métodos tradicionais e *benchmarks*.

O trabalho de Ma, Han e Wang (2021) aborda a otimização de carteira com predição de retorno usando aprendizado profundo e AM. Resumidamente, este estudo investiga como a previsão de retorno de modelos de AM clássicos e de aprendizado profundo pode melhorar a formação de carteiras. Usando dados históricos das ações do Índice de Valores Mobiliários

da China 100, descobriu-se que os modelos "Random Forest (RF) + Mean–variance with forecasting (MVF)"e "Support Vector Regression (SVM) + Omega with forecasting (OF)"se destacam, com o RF+MVF sendo recomendado para investimentos diários, apesar do alto turnover.

Os indicadores utilizados neste estudo incluíram médias móveis, RSI e índice Omega. As métricas de avaliação aplicadas foram o índice de turnover, índice Sharpe e a acurácia das previsões. Para a predição dos retornos, os algoritmos empregados foram o Random Forest e o SVM.

Na pesquisa de León et al. (2017), os algoritmos de agrupamento são empregados com técnicas de otimização clássicas para construção de portfólio ajustado ao risco. Este artigo apresenta o desempenho de sete carteiras criadas usando técnicas de análise de agrupamento para classificar ativos em categorias e aplicar otimização clássica dentro de cada agrupamento. Os algoritmos de agrupamento incluem K-means, MBk-medias, spectral clustering, Birch, e agrupamentos hierárquicos como Average Linkage (AL), Complete Linkage (CL) e Ward's Method (WM). A otimização clássica seleciona os melhores ativos em cada categoria para compor a carteira, utilizando a Otimização de Média-Variância de Markowitz.

Os indicadores empregados incluem volatilidade, retorno histórico e correlação entre ativos. As métricas de avaliação utilizadas foram o retorno esperado e a volatilidade do portfólio. Os algoritmos usados foram técnicas de agrupamento como K-means e otimização por Média-Variância.

Os autores Almahdi e Yang (2017) desenvolveram um sistema adaptativo de negociação de portfólio. Neste estudo, o método de aprendizado por reforço recorrente (RRL) é aplicado com um objetivo de desempenho ajustado ao risco, utilizando baixa coerência estatística para gerar sinais de compra/venda e pesos de alocação de ativos ótimos. Em um estudo pioneiro, Moody et al. (1998) introduziram o RRL na construção de um sistema de negociação, concluindo que o índice de Sharpe pode atuar como uma função de utilidade adaptativa. O estudo de Almahdi e Yang também compara o índice Calmar com o índice de Sharpe, demonstrando que o RRL com alocação de ativos de peso variável apresenta um desempenho superior em ETFs altamente líquidos ao longo de cinco anos.

Os indicadores utilizados incluem o índice de Sharpe e o índice Calmar. As métricas de avaliação incluem a máxima rebaixamento esperada ($E(MDD)$) e o desempenho ajustado ao risco. O algoritmo principal utilizado foi o aprendizado por reforço recorrente.

O estudo dirigido por Sbrana e Castro (2023) difere dos anteriores por trabalhar com criptomoedas, entretanto, as técnicas utilizadas podem ser adequadas para o mercado de ações. A pesquisa demonstrou que a combinação da arquitetura N-BEATS com transformações convolucionais, camadas de atenção e a função de ativação Mish é eficaz para prever os preços de criptomoedas em nível de portfólio. O modelo N-BEATS Perceiver, uma versão do N-BEATS

com arquitetura Transformer, se destacou como o melhor em comparação com outras variações.

Os indicadores utilizados foram o preço histórico e o volume de negociação. As métricas de avaliação incluíram o erro médio absoluto (MAE) e o erro médio quadrático (RMSE). Os algoritmos empregados foram N-BEATS com camadas convolucionais e Transformer.

Outro trabalho relevante é o de Liu, Zhang e Wang (2021), onde é explorada a utilização de modelos de aprendizado profundo para previsão de volatilidade e precificação de opções. Neste estudo, os autores combinaram Redes Neurais LSTM com algoritmos de regularização L1 e L2 para prever a volatilidade implícita de opções. A avaliação foi realizada usando métricas como RMSE e Mean Absolute Percentage Error (MAPE). Os resultados demonstraram que o uso de LSTM foi eficaz em relação a modelos tradicionais, especialmente em períodos de alta volatilidade.

No estudo de Zhang, Wang e Zhao (2020), foi utilizado um modelo de Redes Neurais Convolucionais (CNN) para classificação de sentimentos de notícias financeiras e sua correlação com a previsão do movimento dos preços de ações. Os autores utilizaram indicadores como o índice de sentimento e variáveis macroeconômicas, sendo que as métricas de avaliação incluíram a precisão e a f1-score. Os resultados mostraram que o uso de CNN combinado com análise de sentimentos pode melhorar a capacidade de previsão de movimentos de preços.

Tabela 1 – Resumo dos trabalhos de pesquisa revisados.

Autor	**Indicadores Utilizados**	**Métricas de Avaliação**	**Algoritmos Utilizados**
(CHEN et al., 2021)	Volatilidade, Médias Móveis, RSI	Índice Sharpe, Retorno Ajustado ao Risco	XGBoost, IFA, MV
(MA; HAN; WANG, 2021)	Médias Móveis, RSI, Índice Omega	Turnover, Índice Sharpe, Acurácia	Random Forest, SVM
(LEóN et al., 2017)	Volatilidade, Retorno Histórico, Correlação	Retorno Esperado, Volatilidade	K-means, MBk-médias, MV
(ALMAHDI; YANG, 2017)	Índice Sharpe, Índice Calmar	E(MDD), Desempenho Ajustado ao Risco	RRL
(SBRANA; CASTRO, 2023)	Preço Histórico, Volume	MAE, RMSE	N-BEATS, Transformer
(LIU; ZHANG; WANG, 2021)	Volatilidade Implícita	RMSE, MAPE	LSTM, Regularização L1/L2
(ZHANG; WANG; ZHAO, 2020)	Índice de Sentimento, Variáveis Macroeconômicas	Precisão, F1-Score	CNN

Fonte: Elaborado pelo autor

3.2 Discussão sobre os Trabalhos Relacionados

Os trabalhos analisados mostram a relevância de indicadores financeiros como volatilidade, médias móveis e RSI em estratégias de otimização de portfólio e previsão de retornos, corroborando o uso desses indicadores neste trabalho. Os algoritmos de aprendizado de máquina, como Random Forest, XGBoost e SVM, também foram aplicados em previsões financeiras, validando a escolha dessas técnicas para o projeto. Além disso, as métricas de avaliação usadas, como o Índice Sharpe e a Acurácia, ajudam a contextualizar e justificar as métricas adotadas para avaliar o desempenho dos modelos preditivos e das estratégias de otimização deste estudo.

4 Material e Métodos

Este capítulo trata de toda a abordagem metodológica empregada neste trabalho, descrevendo os materiais e métodos aplicados para a experimentação e avaliações dos resultados, bem como a abordagem proposta para a modelagem do problema.

4.1 Sumarização da Metodologia

Aqui, sumarizamos o processo de treinamento, do começo ao fim, bem como a avaliação final, em etapas sequenciais:

- Coleta de dados: Os dados históricos das ações foram obtidos via YFinance para o intervalo de tempo definido.
- Cálculo de Indicadores: Os indicadores foram calculados a partir dos dados históricos de preços, nos diferentes períodos determinados.
- Pré-processamento: Os dados foram limpos para remover valores ausentes ou não numéricos gerados durante o processo de criação dos indicadores.
- Treinamento de Modelos: Foram utilizados os algoritmos para problemas de classificação no processo de busca em grade.
- Avaliação de Desempenho: As previsões foram avaliadas no conjunto de validação usando a acurácia balanceada para definir a política de seleção de ativos. O *backtesting* foi realizado no conjunto de teste a partir da política de seleção de ativos para o portfólio, simulando o desempenho das estratégias ao longo de três anos.

4.2 Linguagem e Ambiente de Desenvolvimento

Todo o desenvolvimento do projeto foi realizado em Python, uma linguagem amplamente adotada na área de ciência de dados e finanças. O ambiente utilizado foi o Jupyter Notebook, uma ferramenta que facilita o desenvolvimento de código interativo e a visualização de resultados em gráficos, tornando o processo de análise de dados mais eficiente.

As seguintes bibliotecas do Python foram empregadas para a análise e implementação dos modelos de aprendizado de máquina:

- Pandas: Para a manipulação de dados, incluindo a leitura de arquivos CSV e a criação de DataFrames, que permitem a organização dos dados de forma tabular (MCKINNEY, 2010).

- Numpy: Para realizar operações matemáticas e cálculos estatísticos necessários para a transformação dos dados e o cálculo de indicadores econômicos (OLIPHANT, 2006).
- Matplotlib: Utilizada para a criação de gráficos e visualizações que facilitam a análise das tendências e padrões nos dados financeiros.
- Scikit-learn (sklearn): Biblioteca usada para a implementação dos modelos de aprendizado de máquina supervisionado. Ela também fornece ferramentas para a validação cruzada e avaliação dos modelos (PEDREGOSA et al., 2011).
- YFinance: Usada para coletar os dados históricos de preços de ações e volumes diretamente da API do Yahoo Finance.

4.3 Base de Dados e Modelagem de Classificação

Para o treinamento dos modelos de aprendizado de máquina, foram utilizados dados financeiros históricos obtidos a partir da API do Yahoo Finance (YFinance). Os dados disponíveis incluem as seguintes características: preços de abertura, fechamento, máxima e mínima, assim como o volume de negociação. Algumas variáveis como a data de dividendos e o preço ajustado foram descartadas, por não apresentarem relevância direta na modelagem aplicada. Os dados foram coletados em uma janela temporal fixa, sem a utilização de uma janela deslizante, para que outras variáveis pudessem ser extraídas sem prejudicar a quantidade de amostras disponíveis para a utilização na modelagem.

A partir das componentes do preço que descrevem o comportamento de cada ação, descritas anteriormente, diversos indicadores financeiros (variáveis/características) foram criados, sendo eles: RSI, MACD, volatilidade, Z-Score, lag de preço, volume médio ponderado pelo preço, índice de canal de commodities, RSI binarizado, SOSC, SOSC binarizado, Dir, Dir_p, RSL e detrend. Além disso, uma vez que os dados foram considerados em janela temporal fixa, o custo de criação dessas variáveis foi minimizado, já que não há diferentes janelas deslizantes de tempo, facilitando o tratamento dos dados em um período de tempo.

No que diz respeito ao horizonte de tempo, foram considerados 15 anos para a criação das bases de dados para as diversas ações selecionadas do IBOVESPA (serão descritas adiante), uma vez que houve certa dificuldade em conseguir dados confiáveis de APIs gratuitas para um intervalo de tempo maior. Considerando cada ativo, os indicadores foram calculados em diferentes intervalos de dias/períodos e armazenados para a etapa de seleção de características automática, visando uma redução da dimensionalidade antes da etapa de treinamento dos modelos. A Tabela 2 descreve todas as características geradas para serem selecionadas e utilizadas.

Tabela 2 – Indicadores empregados na geração de características.

Indicador	Período	Variável do Preço
RSI	[5,10]	[Open, Close]
MACD	[12,26,9]	[Close]
detrend	[10,15,20]	[Open, Close]
volatilidade	[5,10,20]	[Close]
lag_n	[1]	[Close]
VWAP	[5,10]	[Close, Volume]
CCI	[5,6,7,8,9,10]	[Close]
SOSC	[5,6,7,8,9,10]	[Close]
Bin_ SOSC	[5,10]	[SOSC]
zscore	[5,10,15]	[Close]
Dir	[1]	[Open, Close]
Dir_p	[1,2,3,4,5]	[Close]
RSL	[5,10,15]	[Close]

Fonte: Elaborado pelo autor

Os indicadores da Tabela 2 foram utilizados para a seleção de características com o método de Limiar de Variância (*Variance Threshold*[1]), que representa uma abordagem simples e prática para a seleção de características. O método remove todas as características cuja variância não atende a um limiar e, por padrão, remove todos as características de variância zero, ou seja, características que têm o mesmo valor em todas as amostras. O método foi escolhido por ser simples e ter baixo custo computacional, além de não ser o foco do trabalho a seleção automática de características.

Os indicadores financeiros foram adicionados como variáveis independentes para prever o comportamento do ativo de acordo com a modelagem em 4.3.1. A separação desses dados para o treinamento dos modelos utilizou a metodologia amostragem de dados por hold-out sem reamostragem e embaralhamento, com dados em conjuntos de treinamento, validação e teste, a partir de janelas temporais fixas, garantindo que o modelo fosse avaliado em dados não vistos durante o treinamento, aumentando assim a sua capacidade de generalização.

Como dito, a divisão temporal dos dados contemplando as etapas de treino, validação e teste, compreendeu intervalos fixos de tempo, sendo para o período de treino entre 01 de janeiro de 2010 e 31 de dezembro de 2019. Em seguida, o período de validação, de 01 de janeiro de 2020 a 31 de dezembro de 2021, foi utilizado para avaliar os resultados e escolher os hiper-parâmetros e a estratégia de escolha de ativos, evitando o problema de overfitting. Por fim, o período de teste, de 01 de janeiro de 2022 até 10 de outubro de 2024, foi empregado para avaliar o desempenho final dos modelos em um contexto que simula condições reais de mercado, proporcionando uma estimativa justa da capacidade preditiva dos algoritmos.

Para escolher as ações que seriam modeladas, foi escolhida a composição do índice

[1] scikit-learn: <https://encr.pw/ITE7m>

Ibovespa, compreendido por 84 empresas, extraídas do site da B3[2] em setembro/2024. Contudo, nem todas as ações possuem um histórico de dados completo para os 15 anos considerados na análise e, dessa forma, foram aplicados filtros para garantir a consistência e a qualidade dos dados utilizados, resultando em um subconjunto de 72 ações que atendiam aos requisitos estabelecidos. Essa abordagem assegurou que os modelos fossem treinados e avaliados apenas com dados de ações com um histórico robusto e contínuo, evitando problemas de inconsistência e minimizando o risco de vieses indesejados, garantindo, assim, maior robustez nos resultados apresentados.

4.3.1 Abordagem Proposta para Classificação

Entender e modelar as tendências das diversas ações do Ibovespa não é uma tarefa fácil, já que o comportamento dos preços é não-estacionário (média e variância variam ao longo do tempo). Para lidar com essa complexidade, este trabalho empregou os algoritmos Random Forest, SVM e XGBoost, modelos amplamente utilizados em finanças e adequados para problemas de classificação.

Optou-se por transformar um problema de regressão de séries temporais em um problema de classificação, criando uma política de classificação para cenários de movimentação dos ativos com três movimentos: queda, lateralização e alta.

Nessa abordagem, cada ponto da série temporal de preços foi classificado como uma oportunidade de venda (-1), não operação (0) ou compra (1), considerando a comparação entre os dias $d+1$ e d dos preços e o percentual de lateralização definido na otimização de hiper-parâmetros (Seção 4.4), com valores $0,75\%$, 1% e $1,25\%$ da variação entre o dia $d+1$ e o dia d. Quando o preço de um ativo está em queda, temos um retorno percentual negativo da variação do preço, configurando um momento para operações de venda antes do movimento de diminuição do preço. Em contrapartida, quando o preço do ativo está em ascensão, temos um movimento de alta, e uma operação de compra antecipada pode resultar em retornos positivos.

Por fim, entre os dois movimentos há uma tendência de lateralização, ou seja, o preço do ativo não cresce nem decresce para um dado limiar, configurando um opção de não operar naquele dia. O movimento de lateralização é definido por um limiar percentual aceitável para cada ativo, de acordo com suas características, e pode ajudar ou atrapalhar na o balanceamento das classes.

Essa política tem como objetivo suavizar as operações de acordo com as oscilações diárias dos preços, e identificar apenas mudanças significativas, garantindo uma estratégia mais consistente e robusta ao longo do tempo. É importante ressaltar que o valor percentual que define a lateralização é definido como um hiper-parâmetro, ajustado para cada modelo e para cada ativo, afim de encontrar uma boa configuração para estes.

[2] <https://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/ibovespa.htm>

4.4 Algoritmos de ML e ajuste de Hiper-parâmetros

Nessa seção, é discutida a otimização dos hiper-parâmetros dos algoritmos de aprendizado de máquina utilizados, ou seja, quais variáveis foram ajustadas, empregando a técnica de busca em grade (do inglês, *Grid Search*) com validação cruzada (do iglês, *Cross-Validation* - CV), com 5 partições, para encontrar uma boa combinação dos hiper-parâmetros. É uma metodologia amplamente utilizada em diversas áreas, bem como em finanças, como mostrado por Nystrup, Lindström e Madsen (2020) e Zhao, Zhang e Liu (2024).

A busca em grade é uma técnica que realiza uma busca exaustiva por meio de diferentes combinações de hiper-parâmetros em um espaço predefinido, a fim de determinar a melhor configuração para o modelo nas opções definidas. O uso da validação cruzada dentro da busca em grade permite que o conjunto de dados seja dividido em partes (cinco), alternando os subconjuntos de treinamento e validação, proporcionando treinamento e avaliação robustos e menos sensíveis a divisões específicas do conjunto de dados. Esse processo possibilita a escolha da melhor combinação de hiper-parâmetros a partir de uma métrica de avaliação, que pode ser a acurácia, acurácia balanceada, ou qualquer outra relacionada ao problema. As Tabelas 3 ,4 e 5 apresentam os hiper-parâmetros e seus respectivos valores para cada algoritmo, com a adição do percentual de lateralização.

Tabela 3 – Combinação de hiper-parâmetros para a Random Forest.

Random Forest	$n\_estimators$	max_depth	ccp_alpha	percentual lateralização
	$[80, 90, 100, 110, 120, 130]$	$[3, 4, 5, 6, 7, 8]$	$[0.05, 0.01, 0.015, 0.001, 0]$	$[0, 75, 1, 1, 25]$

Fonte: Elaborado pelo autor

Tabela 4 – Combinação de hiper-parâmetros para a SVM.

SVM	C	γ	percentual lateralização
	$[0, 1, 0, 5, 1, 5, 10, 50, 100]$	$[1e-4, 5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 1e-1]$	$[0, 75, 1, 1, 25]$

Fonte: Elaborado pelo autor

Tabela 5 – Combinação de hiper-parâmetros para o XGBoost.

XGBoost	$learning\ rate$	$n\_estimators$	max_depth	percentual lateralização
	$[0.05, 0.04, 0.03, 0.02, 0.01, 0.005, 0.001]$	$[80, 90, 100, 110, 120, 130]$	$[3, 4, 5, 6, 7, 8]$	$[0, 75, 1, 1, 25]$

Fonte: Elaborado pelo autor

É sabido que os modelos de aprendizado de máquina empregados possuem diversos hiper-parâmetros, e aqui consideramos os que têm maior impacto no desempenho destes, como o número de árvores (*n_estimators*), a profundidade máxima (*max_depth*) e o coeficiente de poda (*ccp_alpha*) para a Random Forest; os parâmetros de regularização (C) e de kernel (γ) para a SVM; e a taxa de aprendizado (*learning_rate*), o número de estimadores e a profundidade máxima para o XGBoost.

Para avaliar o desempenho dos modelos empregados, tanto na busca em grade para os hiper-parâmetros, quanto na escolha das ações para a composição do portfólio, foram utilizadas as seguintes medidas:

- Acurácia balanceada: útil em casos de desequilíbrio entre as classes, garantindo uma avaliação justa das classes minoritárias. A acurácia balanceada leva em consideração a taxa de acerto para cada classe e calcula a média aritmética das sensibilidades, permitindo uma avaliação mais justa em cenários desequilibrados.

- Índice de Sharpe: mede o retorno ajustado pelo risco (SHARPE, 1994). Esse indicador é amplamente utilizado em finanças para avaliar o desempenho de um investimento em relação ao risco que ele envolve. O Sharpe é calculado dividindo o retorno excedente do ativo (ou estratégia) pela volatilidade (desvio padrão) dos retornos. Assim, ele reflete o quanto de retorno adicional é obtido por unidade de risco assumido. Quanto maior o Sharpe, melhor o desempenho ajustado ao risco do investimento. Valores baixos ou negativos indicam que o retorno não compensa o risco, o que pode sugerir que a estratégia não é eficaz no longo prazo.

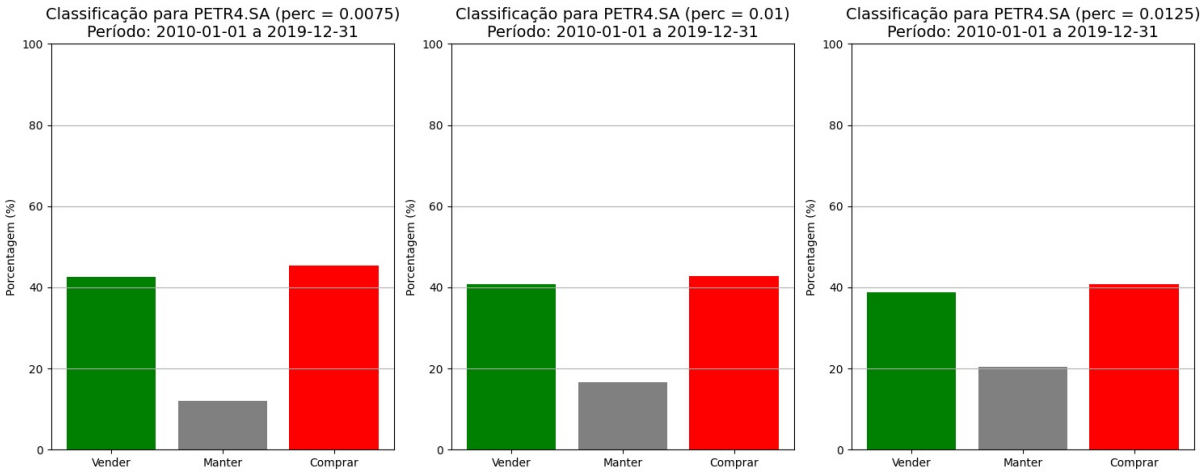
- Rebaixamento Máximo (do inglês, *Maximum DrawDown*): indica a máxima perda acumulada durante o período analisado. O MDD é uma métrica importante para avaliar o risco de um investimento, pois representa a maior queda do valor de um portfólio ou estratégia em relação ao seu pico, antes de se recuperar. Em outras palavras, ele mede a diferença entre o ponto mais alto e o ponto mais baixo que um portfólio atinge durante um período de tempo. Essa métrica é crucial porque ajuda os investidores a entenderem a severidade das perdas possíveis durante fases de mercado desfavoráveis, permitindo avaliar a robustez de uma estratégia de investimento. Estratégias com altos valores de MDD podem ser vistas como arriscadas, mesmo que apresentem retornos elevados em outros períodos.

- Retorno percentual: indica a variação percentual do preço de um ativo entre dois dias consecutivos, ou seja, d e $d-1$. Essa variação é usada para avaliar a rentabilidade percentual de um ativo ao longo do tempo, sendo muito útil e comum em análises de portfólios pois pode ser utilizada de forma cumulativa para acompanhar o crescimento, ou decrescimento, de ativos em base amonetária.

A acurácia balanceada foi utilizada na busca em grade para escolher o melhor conjunto de hiper-parâmetros, uma vez que a distribuição de classes pode ser desbalanceada dependendo do percentual de lateralização. A Figura 5 mostra um exemplo das diferentes distribuições de classes (-1, 0, 1) para os percentuais $0,75\%$, 1% e $1,25\%$ no ativo PETR4.SA.

Por fim, os indicadores Sharpe e MDD foram empregados para avaliar os resultados dos conjuntos de validação, enquanto Sharpe e retorno percentual para o conjunto de teste, a

Figura 5 – Comparação de diferentes percentuais de lateralização para o ativo PETR4.SA.



Fonte: Elaborado pelo autor

fim de analisar e explorar as capacidades dos modelos encontrados pela busca em grade. O conjunto de validação foi utilizado para desenvolver uma abordagem de escolha de portfólio, definida na próxima seção, utilizando os indicadores mencionados.

4.5 Abordagem para a Escolha do Portfólio

Atualmente, a otimização de portfólio no mercado financeiro tem sido um tema de crescente interesse, devido à volatilidade e à complexidade do ambiente econômico. Modelos como o de média-variância, introduzido por Harry Markowitz (MARKOWITZ, 1952), continuam sendo amplamente utilizados, mas avanços significativos têm sido alcançados com abordagens mais robustas e inteligentes. Estudos recentes destacam o uso de técnicas como redes neurais profundas e aprendizado por reforço, que permitem uma maior precisão na seleção de ativos e na gestão do risco, considerando as características específicas como mostrado por Cheng et al. (2024).

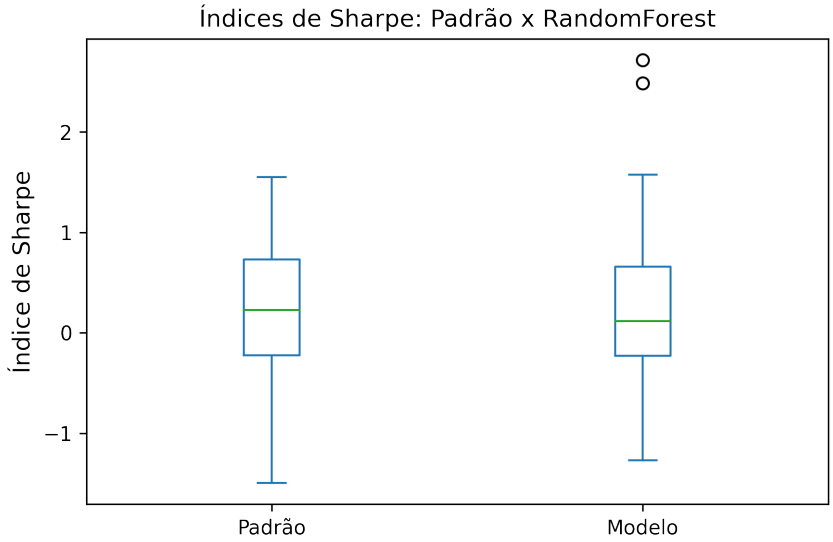
Mesmo com uma vasta gama de técnicas para essa tarefa, esse trabalho buscou investigar se a seleção dos ativos para compor um portfólio pode ser feita de forma simples, utilizando apenas as previsões da movimentação de diferentes ativos do mercado de ações brasileiro feitas pelos algoritmos de aprendizado de máquina empregados. Para isso, foi estabelecida uma regra de escolha baseada em dois indicadores empregados na avaliação dos modelos, o índice de sharpe e o rebaixamento máximo. A estratégia é descrita a seguir:

Dado o conjunto de previsões feitas pelos três modelos, para todos os papeis disponíveis no conjunto de validação ($\forall x \in X_{i=0}^{i=validation\_samples}; f(x_i) = y\_pred_i$), selecionar um limiar (*threshold*) inferior para o índice de Sharpe considerando cada algoritmo. Após a escolha do limiar, filtrar os papeis com índice de Sharpe maior que o limiar e, selecionar os papeis cujo rebaixamento máximo, MDD, dos algoritmos for maior que o rebaixamento dos papeis

selecionados (modelo versus real).

A escolha do valor do terceiro quartil do índice de Sharpe como limiar foi aplicada especificamente aos dados de validação, que continham os resultados das previsões feitas pelos modelos Random Forest, SVM e XGBoost para o comportamento dos ativos financeiros. Com essa estratégia, apenas os ativos que compuseram o conjunto de validação e que apresentaram um índice de Sharpe entre os 25% superiores foram selecionados para o portfólio. O objetivo dessa seleção era avaliar se esses ativos, destacados por seu retorno ajustado ao risco no conjunto de validação, manteriam um desempenho superior também no conjunto de teste, validando a robustez dos modelos e maximizando a potencial consistência da estratégia de portfólio ao longo do tempo. Os gráficos e os respectivos valores dos quartis para cada algoritmo são apresentados nas Figuras 6, 7 e 8.

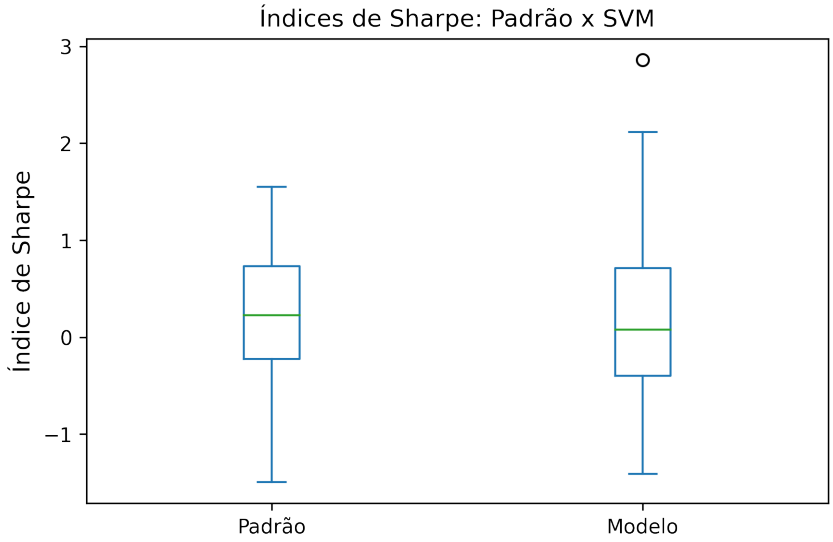
Figura 6 – Box-plot do índice de Sharpe para a Random Forest, com terceiro quartil igual a 0,72.



Fonte: Elaborado pelo autor

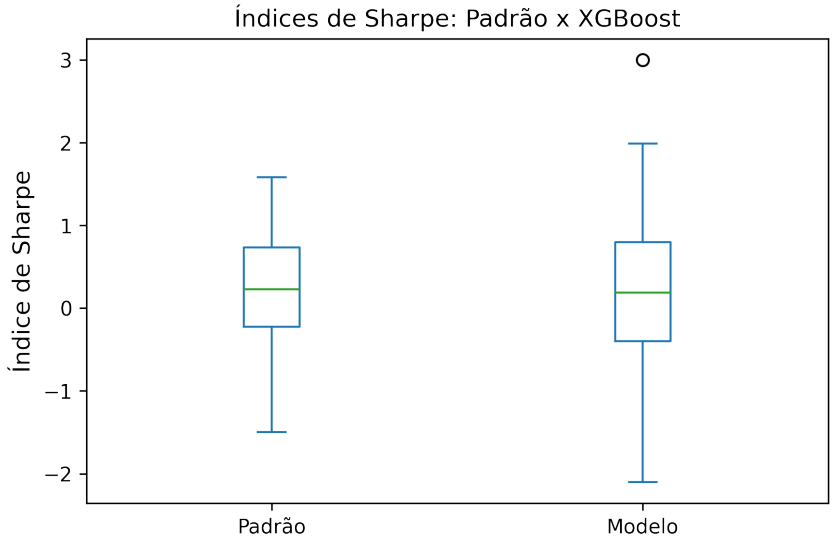
Portanto, os valores de corte dos índices de sharpe para os algoritmos SVM, Random Forest e XGBoost são, respectivamente, 0,72, 0,66, 0,80.

Figura 7 – Box-plot do índice de Sharpe para a SVM, com terceiro quartil igual a 0,66.



Fonte: Elaborado pelo autor

Figura 8 – Box-plot do índice de Sharpe para o XGBoost, com terceiro quartil igual a 0,80.



Fonte: Elaborado pelo autor

5 Resultados experimentais

Neste capítulo, são apresentados os resultados obtidos a partir da aplicação das técnicas de aprendizado de máquina utilizando a estratégia de classificação criada, bem como a estratégia de seleção de ativos para definir os portfólios, como descrito no Capítulo 4.

Primeiramente, são abordados os resultados de desempenho dos modelos Random Forest, SVM e XGBoost em prever o comportamento dos ativos financeiros para o conjunto de validação, considerando o índice de Sharpe e o rebaixamento máximo. Além disso, são destacados os ativos que foram selecionados ao empregar a estratégia proposta.

Por fim, é mostrada e discutida a viabilidade da estratégia de escolha de portfólio proposta, assim como a comparação entre os modelos empregados, ressaltando os pontos fortes e limitações observados. Essa análise experimental permite verificar a robustez e aplicabilidade dos métodos propostos, contribuindo para uma visão prática dos resultados alcançados.

5.1 Análise do Conjunto de Validação

Nessa seção são avaliados os desempenhos dos modelos empregados para a tarefa de classificação considerando o conjunto de validação. Este conjunto desempenha um papel fundamental na validação dos algoritmos e do desenvolvimento de estratégias, uma vez que oferece uma perspectiva de robustez e capacidade de generalização. Os indicadores de desempenho escolhidos foram o índice de Sharpe e o Drawdown Máximo, métricas amplamente utilizadas em finanças para avaliar o retorno ajustado ao risco e a estabilidade de uma carteira de investimentos ao longo do tempo, respectivamente.

5.1.1 Análise do Índice de Sharpe

Dentre os diversos critérios de avaliação, o índice de Sharpe foi escolhido para definir um limiar de retorno ajustado ao risco dos ativos. O índice de Sharpe é utilizado para medir o quanto de retorno adicional um ativo oferece em relação ao risco assumido. Dessa forma, possibilita uma comparação entre ativos de forma clara, eliminando efeitos de flutuações de risco não compensadas por retornos proporcionais.

Para a escolha do limiar inferior do índice de Sharpe, foi realizada uma análise dos box-plots dos três modelos aplicados aos 72 ativos no conjunto de validação. Nessa análise, destacam-se os quartis, que indicam a concentração dos valores do índice de Sharpe para os diferentes ativos selecionados.

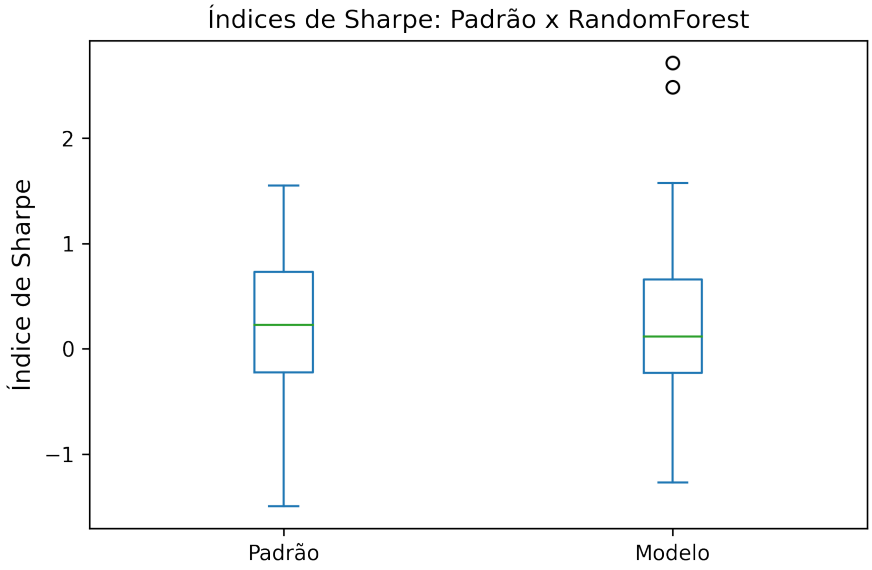
A Figura 9, apresenta o box-plot do índice de Sharpe para o modelo Random Forest.

Nela, observa-se que o modelo teve um menor intervalo de variação dos valores de Sharpe quando comparado aos ativos (padrão), uma vez que os quartis estão mais próximos e o intervalo de mínimo e máximo são menores. Porém, esse modelo apresenta dois valores discrepantes (*outliers*), acima de 2, sendo interessante ao considerarmos os dados acima do terceiro quartil, indicando que podem gerar bons resultados, ou seja, ativos que mantiveram um nível consistente de retorno em relação ao risco para essa seleção.

A Figura 10 mostra o box-plot do índice de Sharpe para o modelo SVM, mostrando uma dispersão diferente dos valores obtidos com o Random Forest. O modelo SVM apresentou maior variação nos quartis que os ativos (padrão), sugerindo uma volatilidade diferente em relação aos retornos ajustados ao risco, o que impacta na estabilidade da seleção, porém, também possui ativos com Sharpe promissor ao se considerar valores acima do terceiro quartil.

Por fim, na Figura 11, o modelo XGBoost é analisado. Pode-se notar uma maior variabilidade nos valores do índice de Sharpe do modelo em relação aos ativos padrão, também com a presença de valores discrepantes, sugerindo que existem ativos melhores com as previsões do XGBoost.

Figura 9 – Box-plot do índice de Sharpe para o modelo Random Forest.

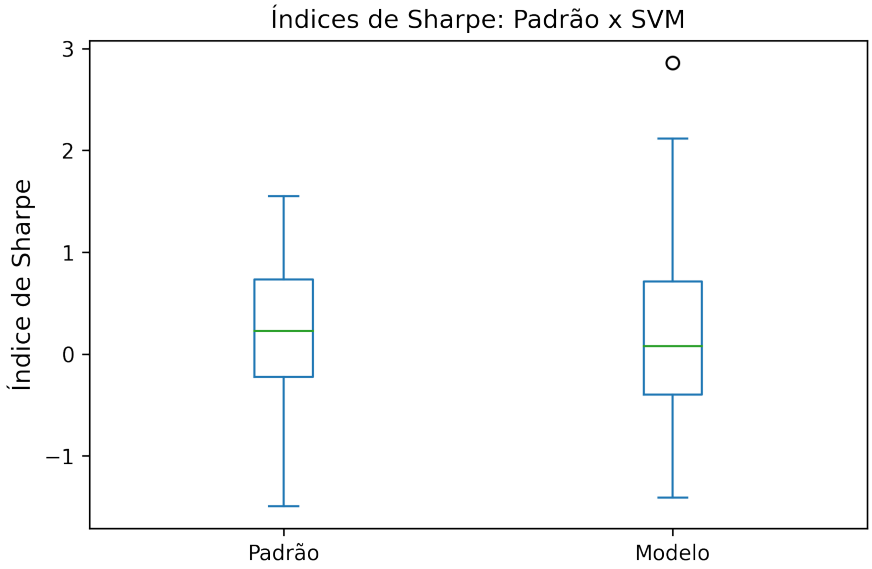


Fonte: Elaborado pelo autor

5.1.2 Análise do rebaixamento máximo

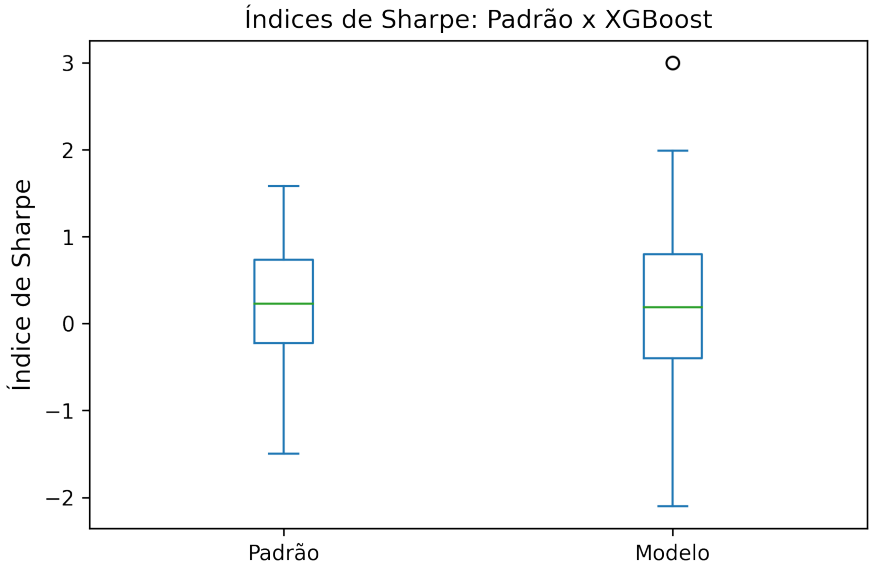
Além do índice de Sharpe, o rebaixamento máximo foi outro critério essencial para a escolha dos ativos que compõem o portfólio. O MDD reflete o maior pico de perda de valor de um ativo antes que ele se recupere, sendo uma métrica importante para avaliar o potencial de risco das operações e a resiliência dos ativos ao longo do tempo.

Figura 10 – Box-plot do índice de Sharpe para o modelo SVM.



Fonte: Elaborado pelo autor

Figura 11 – Box-plot do índice de Sharpe para o modelo XGBoost.

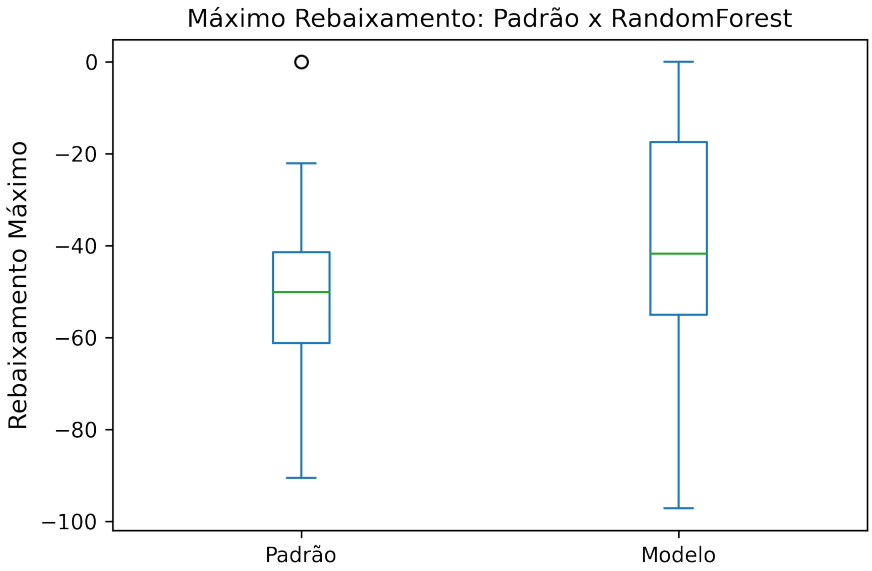


Fonte: Elaborado pelo autor

Aqui, a análise dos box-plots também foi utilizada para destacar a distribuição dos valores e os quartis. Porém, a estratégia foi ligeiramente diferente, ou seja, selecionar ativos cujo rebaixamento máximo dos algoritmos fosse maior que o rebaixamento dos próprios papéis selecionados (modelo versus padrão).

A Figura 12 apresenta o rebaixamento máximo para o modelo Random Forest. Nela, observam-se valores mais próximos de zero, indicando maior estabilidade dos ativos preditos com o modelo, possuindo um perfil de risco menor que os ativos padrão, corroborando ao

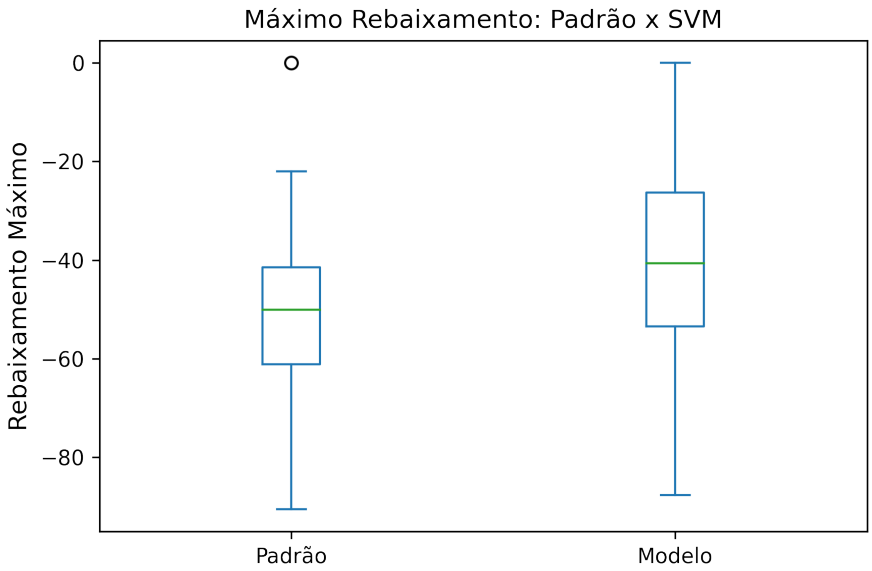
Figura 12 – Box-plot do drawdown máximo para o modelo Random Forest.



Fonte: Elaborado pelo autor

analisar o corpo do box-plot do modelo que está acima do corpo do box-plot padrão.

Figura 13 – Box-plot do drawdown máximo para o modelo SVM.

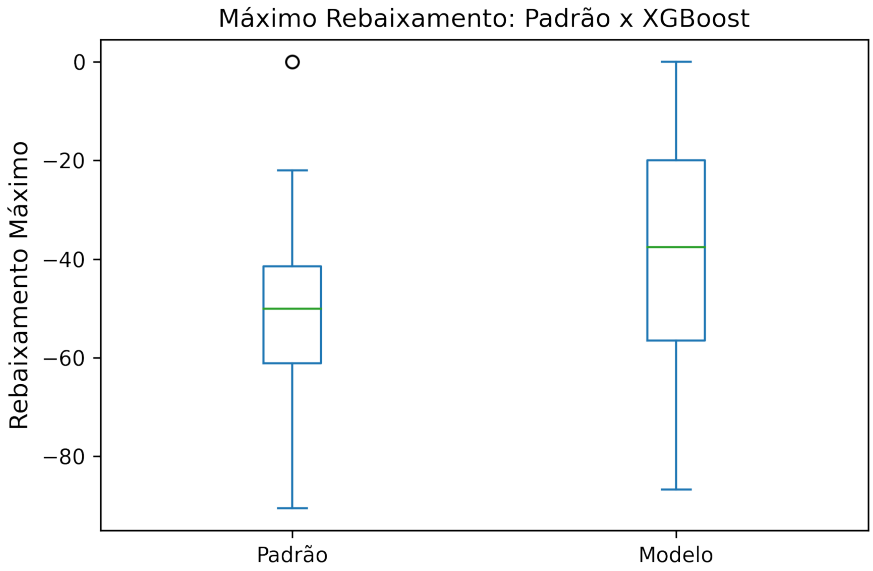


Fonte: Elaborado pelo autor

Na Figura 13, é analisado o rebaixamento para o modelo SVM, onde se nota um comportamento semelhante à Figura 12 (Random Forest), com maiores valores e amplitude destes em relação ao padrão, indicando que o SVM pode prever ativos com maior exposição a perdas, refletindo maior risco.

Por fim, a Figura 14 mostra o rebaixamento máximo para o XGBoost, onde a variabili-

Figura 14 – Box-plot do drawdown máximo para o modelo XGBoost.



Fonte: Elaborado pelo autor

dade também se mostra acentuada, mas ainda com valores mais próximos de zero, sugerindo um perfil de risco que pode ser ajustado com a seleção de ativos por esse modelo.

5.2 Análise da Estratégia

Considerando a abordagem adotada, descrita anteriormente, foi possível selecionar os ativos para a estratégia e gerar seus resultados para o conjunto de teste. Os ativos selecionados para cada algoritmo estão apresentados no Quadro 1. Embora o filtro aplicado para a construção do portfólio tenha sido o mesmo para todos os algoritmos, o número e a composição dos ativos selecionados variaram. Essa diferença se deve às diferenças nos desempenhos dos ativos em termos das métricas durante os períodos de validação e teste, já que cada algoritmo avaliou os papéis de maneira distinta. Essa abordagem ressalta como os modelos podem gerar estratégias de portfólio diversificadas, mesmo sob os mesmos critérios de seleção.

5.2.1 Resultados quantitativos da estratégia

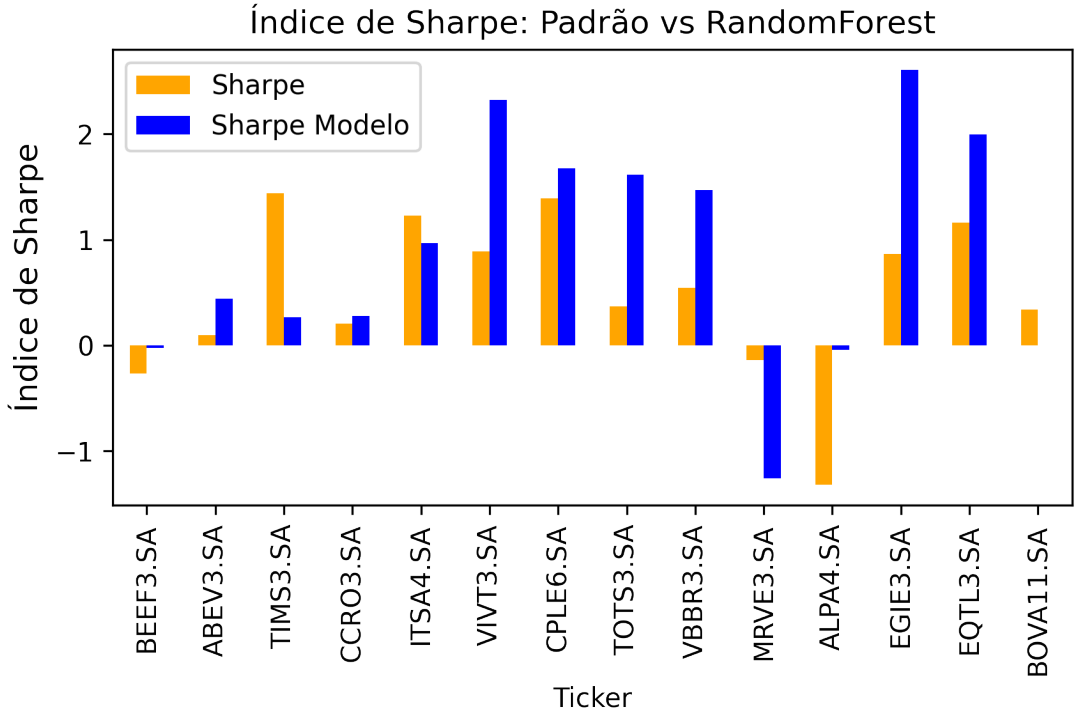
Após a seleção de ativos, a eficácia da estratégia foi validada considerando as análises do índice de Sharpe e do retorno percentual para o conjunto de teste. Inicialmente, são mostrados os resultados do índice de Sharpe. É importante destacar que foi adicionado o ativo BOVA11.SA aos gráficos, uma vez que representa o resultado do Ibovespa e serve de base compartiva para a estratégia. A Figura 15 ilustra os resultados para o modelo Random Forest, com os ativos selecionados com a estratégia proposta. A Figura 16 traz o desempenho do algoritmo SVM com os ativos selecionados, enquanto a Figura 17 mostra os resultados para o XGBoost.

Quadro 1 – Ativos selecionados para cada algoritmo com a estratégia proposta.

Algoritmo	**Ativos Selecionados**
Random Forest (RF)	BEEF3.SA, ABEV3.SA, TIMS3.SA, CCRO3.SA, ITSA4.SA, VIVT3.SA, CPLE6.SA, TOTS3.SA, VBBR3.SA, MRVE3.SA, ALPA4.SA, EGIE3.SA, EQTL3.SA
SVM	TIMS3.SA, EMBR3.SA, CCRO3.SA, ITSA4.SA, RAIL3.SA, CRFB3.SA, ELET3.SA, MULT3.SA, ALPA4.SA, TRPL4.SA, B3SA3.SA, AZZA3.SA, CPLE6.SA, UGPA3.SA, RADL3.SA, VIVT3.SA
XGBoost	BEEF3.SA, CCRO3.SA, BBDC3.SA, ABEV3.SA, BBDC4.SA, VIVT3.SA, ALPA4.SA, RADL3.SA, KLBN11.SA, BRKM5.SA, ELET6.SA, BRFS3.SA, PETR4.SA, PETR3.SA, CPLE6.SA, CMIG4.SA, WEGE3.SA

Fonte: Elaborado pelo autor

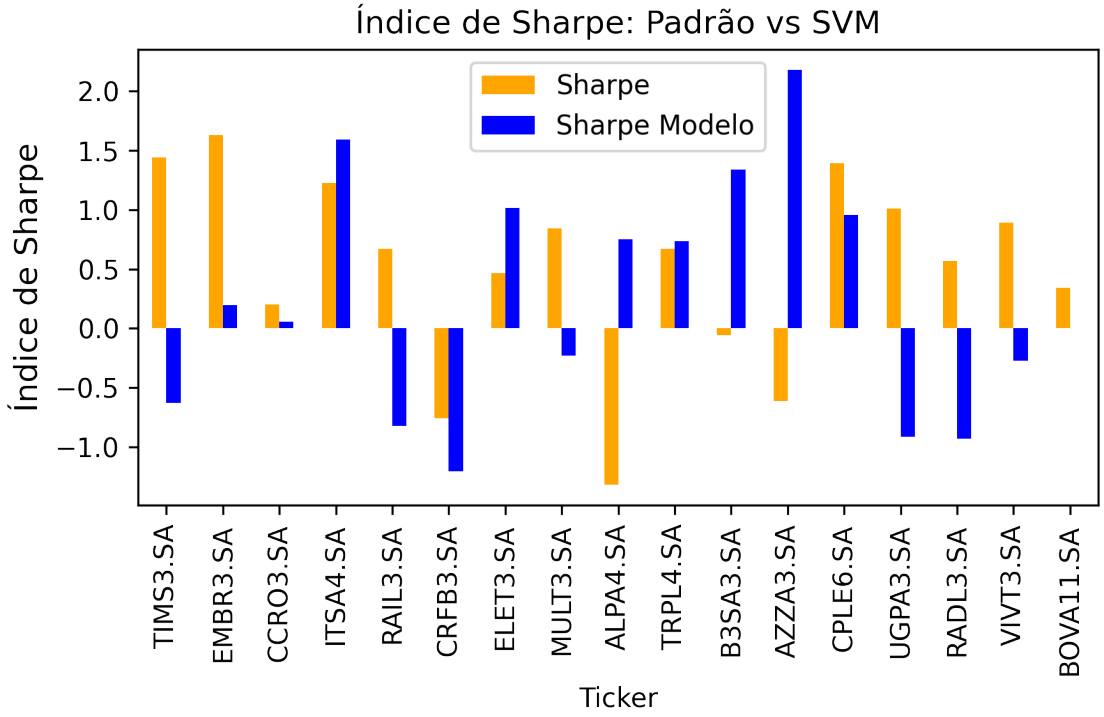
Figura 15 – Gráfico de Sharpe para o modelo Random Forest.



Fonte: Elaborado pelo autor

Da Figura 15, nota-se que o modelo Random Forest obteve, em sua maioria, resultados positivos de Sharpe, com valores acima de 1 e apenas um ativo com valor negativo (MRVE3.SA). Esse cenário pode indicar um bom resultado de retorno percentual acumulado, bem como uma volatilidade controlada, ou seja, uma variância menor que os ativos padrão. Casos como EGIE3.SA e VIVT3.SA se destacam com as previsões do modelo, atingindo Sharpes maiores que 2, por exemplo.

Figura 16 – Gráfico de Sharpe para o modelo SVM.



Fonte: Elaborado pelo autor

Da Figura 16, nota-se que o modelo SVM teve maiores variações nos valores de Sharpe para os diferentes ativos, quando comparado ao Random Forest. Ativos como AZZA3.SA e ITSA4.SA atingiram bons valores de Sharpe, porém, a variabilidade foi alta em sua maioria, indicando um portfólio com oscilações de retorno percentual e maior risco.

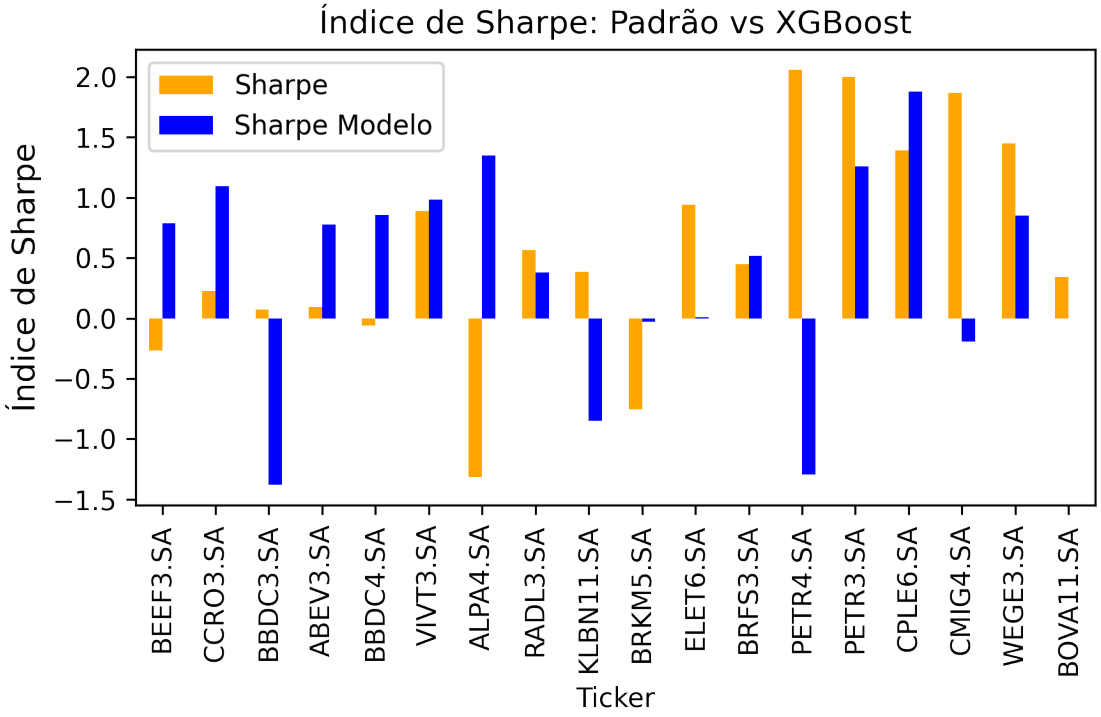
Da Figura 17, observa-se que o modelo XGBoost conseguiu prever a movimentação de ativos que levaram a uma seleção destes com maior possibilidade retorno percentual e variabilidade, uma vez que boa parte dos valores de Sharpe estão próximos de 1. Houve destaque para a CPLE6.SA, onde o valor passou de 1,5.

Adicionalmente, foram gerados os gráficos de retorno percentual acumulado para cada modelo, por ativo, mostrados nas Figuras 18, 19, e 20. Essas figuras comparam os desempenhos percentuais dos ativos ao longo do tempo, permitindo uma análise direta do impacto da seleção de ativos.

Ao se observar a Figura 18, nota-se que o modelo obteve valores de retorno percentual em sua maior parte positivos, com baixa variação entre os ativos, salva exceção da ALPA4.SA, que drasticamente acumulou uma perda de aproximadamente 100%, podendo prejudicar grandemente o portfólio. Entretanto, é um modelo promissor, já que possui estabilidade e consistência nos resultados com a estratégia adotada.

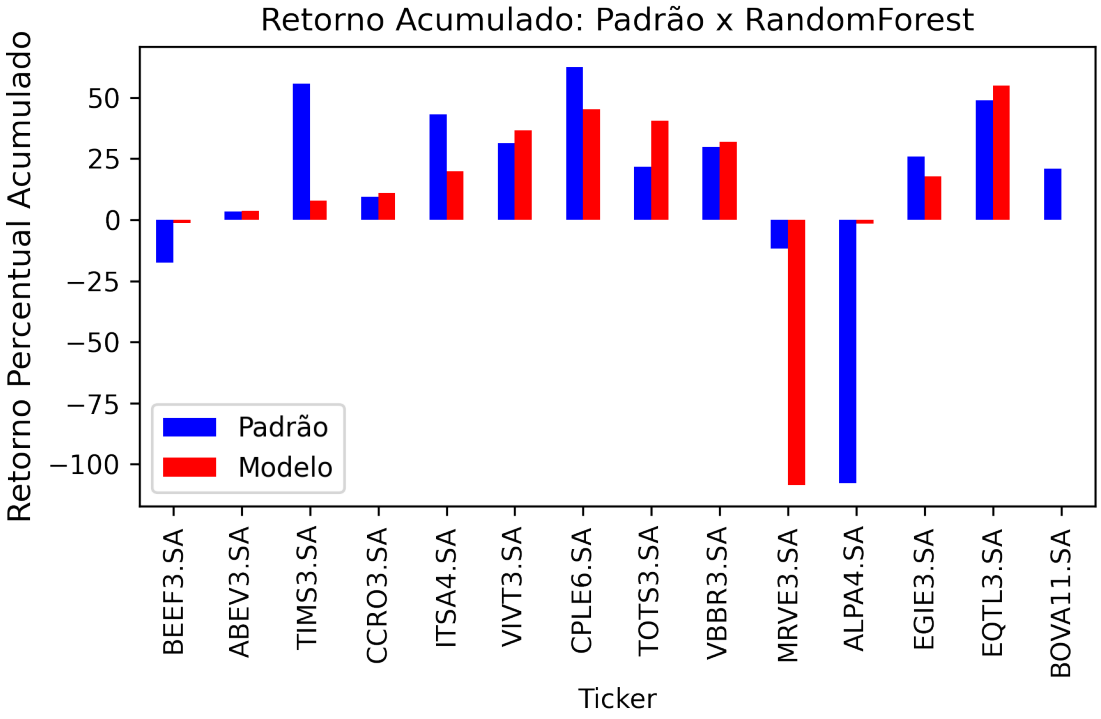
Da Figura 19, observa-se que o modelo SVM apresentou valores de retorno percentual

Figura 17 – Gráfico de Sharpe para o modelo XGBoost.



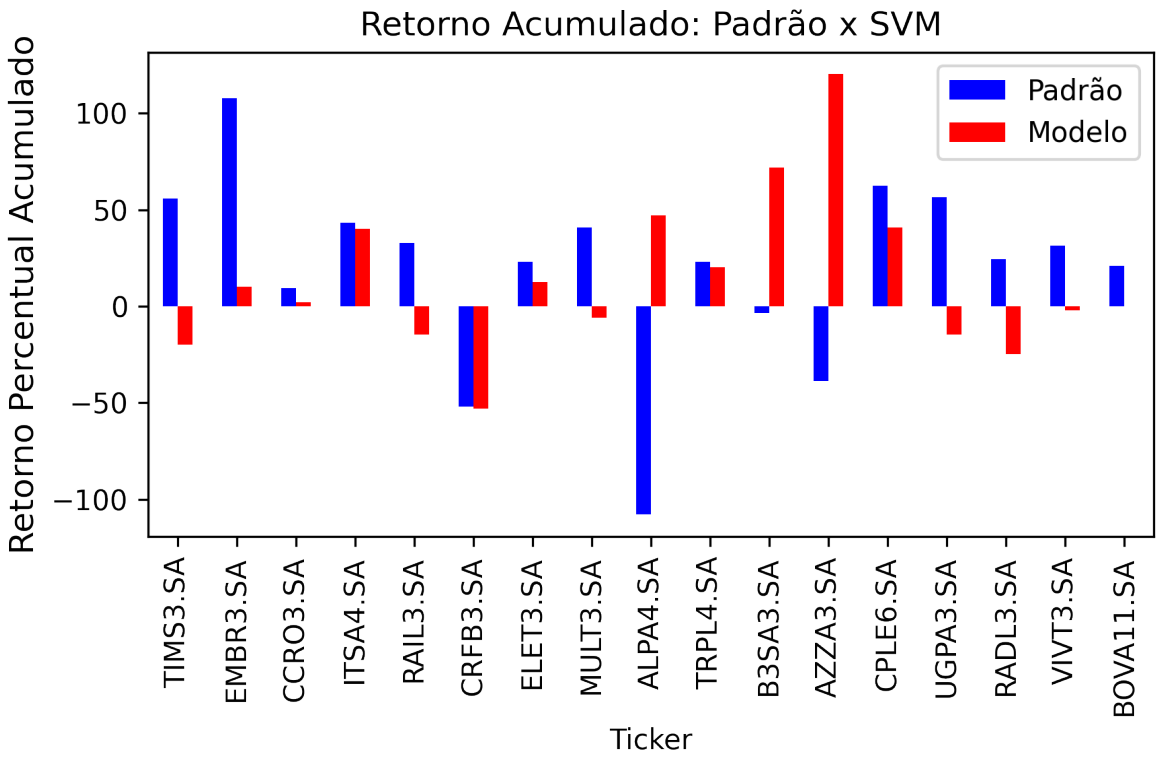
Fonte: Elaborado pelo autor

Figura 18 – Retorno percentual acumulado para o modelo Random Forest.



Fonte: Elaborado pelo autor

Figura 19 – Retorno percentual acumulado para o modelo SVM.



Fonte: Elaborado pelo autor

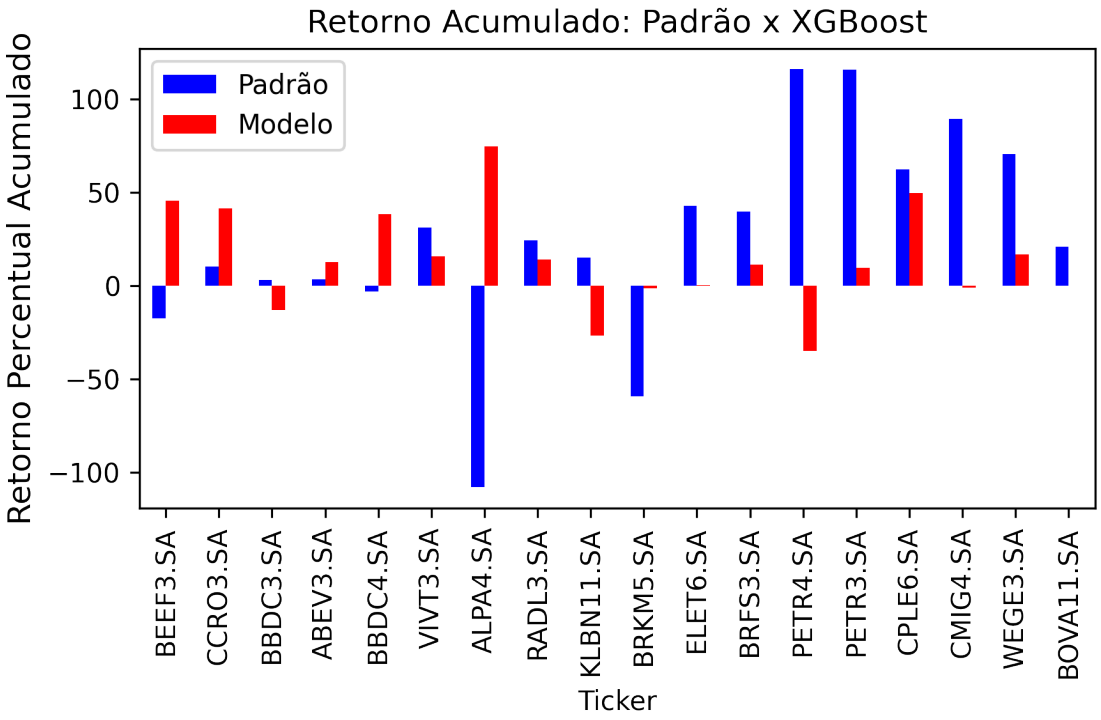
com grande variação, entre -100% e 100%, o que indica um portfólio de alto risco e um desempenho de retorno instável ao longo do tempo, uma vez que operações negativas levam à perda de capital. Curiosamente, o ativo ALPA4.SA também foi escolhido, assim como na Random Forest, e levou a uma perda expressiva de retorno percentual na estratégia.

Por fim, a Figura 20, mostra que o XGBoost conseguiu prever e selecionar ativos com altos retornos percentuais, como no caso da PETR3.SA, PETR4.SA, CMIG4.SA, e WEGE3.SA, parecendo ser o mais rentável dos três modelos. Entretanto, é um portfólio com alta variabilidade, já que possui alguns ativos como ALPA4.SA e MRKM5.SA que acumularam perdas de mais de 50%

Adicionalmente, foram gerados três gráficos dos retornos percentuais acumulados, considerando a soma de todas as operações sugeridas pelos algoritmos, em todos os ativos selecionados na estratégia por cada um. As Figuras 21, 22, e 23 mostram o crescimento da rentabilidade percentual total da carteira para o modelo Random Forest, SVM e XGBoost, respectivamente, ao longo do tempo de teste, permitindo observar tendências e consistências nos retornos.

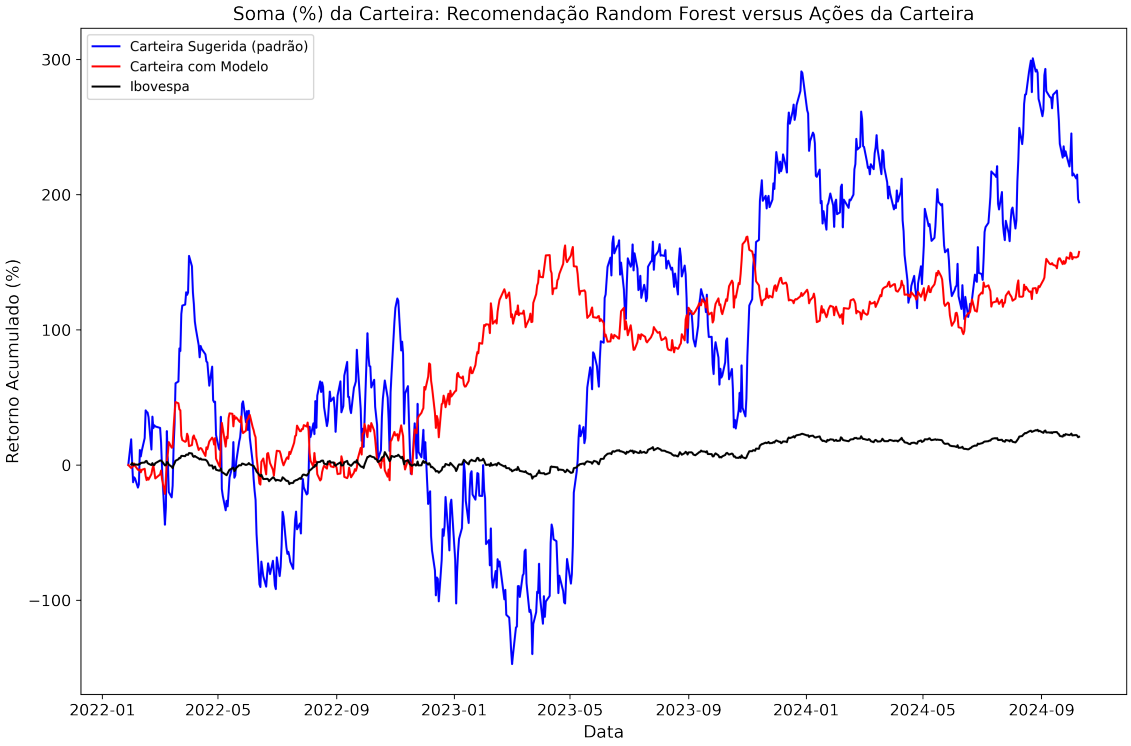
Da Figura 21, nota-se que as previsões feitas pelo modelo Random Forest são, de fato, mais estáveis que os próprios ativos, como discutido anteriormente, e conseguem atingir um bom retorno acumulado para essa carteira de ativos, com pouco menos de 200% em pouco

Figura 20 – Retorno percentual acumulado para o modelo XGBoost.



Fonte: Elaborado pelo autor

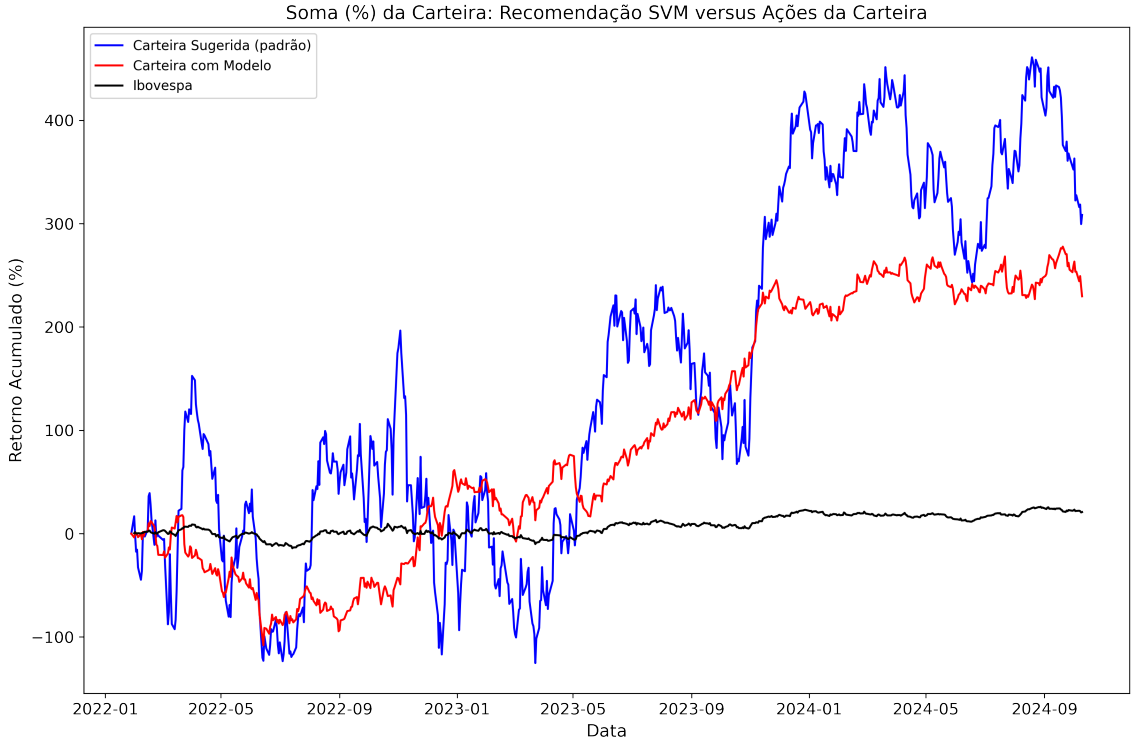
Figura 21 – Retorno percentual acumulado para o modelo Random Forest ao longo do tempo.



Fonte: Elaborado pelo autor

mais de dois anos e meio. Mesmo que os ativos tenham rentabilizado mais nesse período, trata-se de uma boa estratégia já que a oscilação de patrimônio com o Random Forest foi muito menor que os ativos padrão.

Figura 22 – Retorno percentual acumulado para o modelo SVM ao longo do tempo.



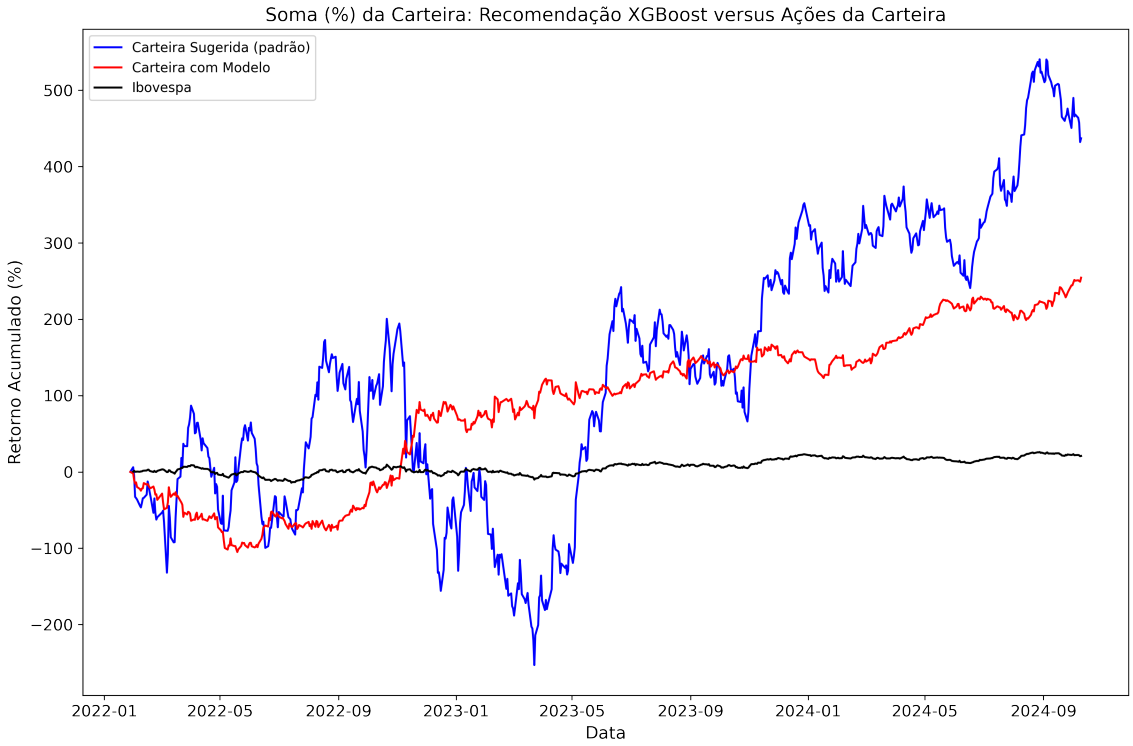
Fonte: Elaborado pelo autor

Da Figura 22, observa-se que as previsões feitas pelo modelo SVM são tão rentáveis quanto os próprios ativos, assim como o Random Forest. Porém, o portfólio tem maior volatilidade e risco, chegando a ter perdas acumuladas de -100%, semelhante aos ativos selecionados por esse modelo. De modo geral, possui um bom retorno mas requer mais capital para conseguir operar com perdas expressivas, o que pode não ser tão eficiente.

Por fim, na Figura 22, observa-se que as previsões feitas pelo modelo XGBoost também levam a uma estratégia rentável e crescente, assim como os próprios ativos selecionados. Esse portfólio tem volatilidade e risco expressivos, com perdas que -100%, mas muito mais estável que as oscilações dos próprios ativos, que acumulam perdas de mais de -200%. É um modelo para uma estratégia mais agressiva, com boa rentabilidade, mas também requer mais capital para conseguir operar com perdas expressivas.

Os resultados apresentados no Quadro 2 destacam um desempenho contrastante entre os portfólios dos modelos e os próprios ativos selecionados no período de 01/2022 até 10/2024. Embora os ativos de cada portfólio tenham apresentado retornos substancialmente superiores aos dos modelos —194,354% para ativos RF, 308,472% para ativos SVM e 436,996% para ativos XGBoost — os modelos mostraram-se mais consistentes ao longo do tempo, com

Figura 23 – Retorno percentual acumulado para o modelo XGBoost ao longo do tempo.



Fonte: Elaborado pelo autor

Quadro 2 – Resumo dos resultados de retorno e índice de Sharpe entre modelos e ativos, de 01/2022 a 10/2024.

Modelo / Ativo	**Retorno (%)**	**Sharpe**
Ibovespa	20,875	0,3404
RF	157,620	1,018
Ativos RF	194,354	0,477
SVM	229,709	1,357
Ativos SVM	308,472	0,560
XGBoost	254,413	**1,731**
Ativos XGBoost	**436,996**	0,892

Fonte: Elaborado pelo autor

uma relação retorno-risco mais favorável, refletida em índices de Sharpe significativamente superiores.

O desempenho de cada modelo em termos de índice de Sharpe revela aspectos importantes na relação entre risco e retorno: enquanto o XGBoost alcançou o maior índice de Sharpe (1,731, em negrito na tabela), indicando uma excelente relação retorno-risco, ele também apresentou um risco relativamente alto em comparação com a Random Forest. Este último, embora com um retorno de 157,620%, menor que o XGBoost, manteve um Sharpe próximo de 1 (1,018), indicando uma combinação favorável de retorno e estabilidade.

Comparando os retornos dos modelos com o Ibovespa, cada modelo superou amplamente

o índice de referência, que obteve 20,875% no período analisado. A Random Forest, por exemplo, superou o Ibovespa em aproximadamente 654%, enquanto o XGBoost teve um desempenho ainda maior, com um retorno 12 vezes superior ao índice. Esse contraste revela que, enquanto os ativos de cada portfólio são ideais para quem busca maximizar o retorno absoluto, os modelos são opções mais adequadas para investidores que valorizam uma estratégia com estabilidade e gestão de risco. Dessa forma, cada abordagem possui seus pontos fortes, sendo o XGBoost o modelo com a melhor relação retorno-risco, e a Random Forest uma alternativa consistente e segura com um Sharpe adequado para investidores focados na segurança.

Sumarizando, foi possível encontrar padrões com as previsões dos três modelos na abordagem proposta e escolher bons ativos para as carteiras. Essas carteiras podem ser ainda melhoradas com estudos posteriores, assim como a ponderação de quais ativos devem ser mais operados, e a escolha de uma delas para operar depende do perfil de cada investidor e do capital disponível.

6 Conclusão

Neste trabalho, investigou-se o uso de técnicas de ML como estratégia para maximizar retornos no mercado de capitais brasileiro. Por meio da construção e avaliação de modelos baseados em RF, SVM e XGBoost, foi possível analisar a eficácia e a estabilidade dos portfólios de ativos selecionados, comparando seus desempenhos em termos de retorno percentual e índice de Sharpe ao longo do período de janeiro de 2022 até outubro de 2024.

Os resultados experimentais, sintetizados na Tabela 2, mostraram que, embora os ativos selecionados para cada portfólio tenham alcançado retornos percentuais mais elevados em comparação com os modelos em si, os modelos ofereceram uma relação risco-retorno mais favorável, como refletido em índices de Sharpe consistentemente mais altos. Especificamente, enquanto os portfólios de ativos obtiveram retornos de 194,354% (ativos RF), 308,472% (ativos SVM) e 436,996% (ativos XGBoost), os modelos apresentaram retornos ligeiramente menores, de 157,620% para RF, 229,709% para SVM e 254,413% para XGBoost.

Esses achados destacam que, apesar dos modelos não maximizarem o retorno, eles geraram estratégias mais estáveis. O índice de Sharpe foi substancialmente maior para todos os modelos em comparação aos portfólios de ativos, com o XGBoost atingindo o valor de 1,731, o maior entre os modelos, seguido pelo SVM com 1,357 e pelo Random Forest com 1,018. Em contrapartida, o índice de Sharpe para os portfólios de ativos foi relativamente mais baixo, indicando uma maior exposição ao risco sem a mesma consistência.

Além disso, a comparação com o Ibovespa, que teve um retorno de 20,875% e um índice de Sharpe de 0,3404, reforça a superioridade das estratégias baseadas em ML em relação ao índice de referência. Cada modelo não apenas superou o Ibovespa em retorno percentual, mas também apresentou uma relação risco-retorno superior, evidenciando o valor dos modelos de ML para estratégias mais seguras e lucrativas em um ambiente volátil como o mercado brasileiro.

Portanto, a análise sugere que, para investidores que buscam maximizar o retorno absoluto, a seleção direta de ativos pode ser atrativa. Entretanto, para aqueles que preferem uma abordagem com menor risco e maior consistência, os modelos de ML são uma escolha mais adequada, oferecendo uma vantagem significativa na gestão de portfólio com foco em estabilidade. Esses insights mostram o potencial das técnicas de ML para aprimorar a tomada de decisões no mercado financeiro e oferecem uma base promissora para futuras pesquisas e aprimoramentos nas estratégias de investimento.

Por fim, em trabalhos futuros, pretende-se explorar outras formas de filtrar ativos potenciais, utilizando outras estratégias ou métricas, como o próprio retorno percentual e a variância de cada ativo, assim como a implementação de abordagens de seleção de portfólios

com algoritmos de ML tomando a decisão, por exemplo. Além disso, pretendemos validar a abordagem aqui proposta em outros mercados, como o mercado americano, para investigar a robustez do método proposto e aplicado no mercado brasileiro.

Referências

AKERLOF, G. A. The market for "lemons": Quality uncertainty and the market mechanism. In: *Uncertainty in economics*. [S.l.]: Elsevier, 1978. p. 235–251.

ALMAHDI, S.; YANG, S. Y. An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. *Expert Systems with Applications*, v. 87, p. 267–279, 2017. ISSN 0957-4174.

BERNHEIM, B. D. Financial illiteracy, education, and retirement saving. *Living with Defined Contribution Pensions*, University of Pennsylvania Press, p. 38–68, 1998.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006.

BREIMAN, L. Bagging predictors. *Machine Learning*, Springer, v. 24, n. 2, p. 123–140, 1996.

BREIMAN, L. *Random forests*. [S.l.]: Springer, 2001. 5–32 p.

BURGES, C. J. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, Springer, v. 2, n. 2, p. 121–167, 1998.

CHEN, T.; GUESTRIN, C. *XGBoost: A scalable tree boosting system*. [S.l.]: ACM, 2016. 785–794 p.

CHEN, W.; ZHANG, H.; MEHLAWAT, M. K.; JIA, L. Mean–variance portfolio optimization using machine learning-based stock price prediction. *Applied Soft Computing*, v. 100, p. 106943, 2021. ISSN 1568-4946.

CHENG, Q.; YANG, L.; ZHENG, J.; TIAN, M.; XIN, D. Optimizing portfolio management and risk assessment in digital assets using deep learning for predictive analysis. *arXiv preprint arXiv:2402.15994*, 2024.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, Springer, v. 20, n. 3, p. 273–297, 1995.

(CVM), C. de V. M. *Relatório Anual da Comissão de Valores Mobiliários 2021*. 2021. Disponível em: <https://www.gov.br/cvm/pt-br>.

DAMODARAN, A. *Investment Valuation: Tools and Techniques for Determining the Value of Any Asset*. [S.l.]: John Wiley & Sons, 2012.

DE, S. The low interest rate environment and its implications for financial markets. *Journal of Financial Perspectives*, v. 7, n. 2, p. 15–29, 2020.

ESTEVA, A.; KUPREL, B.; NOVOA, R. A.; KO, J.; SWETTER, S. M.; BLAU, H. M.; THRUN, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, Nature Publishing Group, v. 542, n. 7639, p. 115–118, 2017.

FAVERO, E.; MEIRELES, F. Cultura e comportamento do investidor brasileiro. *Revista Brasileira de Economia*, 2013.

FERREIRA, E. V. *Support Vector Machines*. n.d. Material de ensino sobre Support Vector Machines. Disponível em: <http://leg.ufpr.br/~walmes/ensino/ML/slides/Support%20Vector%20Machines.pdf>.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, v. 29, n. 5, p. 1189–1232, 2001.

FUJIMOTO, R. M. *Parallel and Distributed Simulation Systems*. [S.l.]: John Wiley & Sons, 2015.

GARCIA, M. G. P. *Crises Econômicas: Causas, Consequências e Desafios*. [S.l.]: Elsevier, 2011.

GOETZMANN, W. N.; ROUWENHORST, K. G. *The Origins of American Financial Markets*. [S.l.]: Oxford University Press, 2005.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016.

GREENWOOD, R.; HANSON, S.; STEIN, J. The world's growing savings glut. *Brookings Papers on Economic Activity*, Brookings Institution Press, p. 377–432, 2019.

HIRANSHA, M.; GOPALAKRISHNAN, E.; MENON, V.; SOMAN, K. Nse stock market prediction using deep-learning models. *Procedia Computer Science*, Elsevier, v. 132, p. 1351–1362, 2018.

HO, T. K. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, IEEE, p. 278–282, 1995.

HULL, J. *Machine Learning in Business and Finance: A Practical Introduction*. [S.l.]: Wiley, 2019.

INVESTMENTS, F. *Fidelity Investments: 2021 Market and Economic Outlook*. 2021. Disponível em: <https://www.fidelity.com>.

KAHNEMAN, D. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.

KINDLEBERGER, C. P. *Manias, Panics, and Crashes: A History of Financial Crises*. [S.l.]: John Wiley & Sons, Inc, 2005.

LAMBERT, D. R. *Commodity Channel Index: Tools for Trading Cyclic Trends*. [S.l.: s.n.], 1980.

LEVINE, R. *Financial Development and Economic Growth: Views and Agenda*. [S.l.]: Journal of Economic Literature, 1997. v. 35. 688-726 p.

LEóN, D.; ARAGóN, A.; SANDOVAL, J.; HERNáNDEZ, G.; ARéVALO, A.; NIñO, J. Clustering algorithms for risk-adjusted portfolio construction. *Procedia Computer Science*, v. 108, p. 1334–1343, 2017. ISSN 1877-0509. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

LIU, W.; ZHANG, J.; WANG, M. Deep learning for volatility prediction and options pricing. *Journal of Financial Markets*, v. 56, p. 1567–1589, 2021.

LOPEZ, J. H. The power of the adf test. *Economics Letters*, Elsevier, v. 57, n. 1, p. 5–10, 1997.

MA, Y.; HAN, R.; WANG, W. Portfolio optimization with return prediction using deep learning and machine learning. *Expert Systems with Applications*, v. 165, p. 113973, 2021. ISSN 0957-4174.

MALKIEL, B. *A random walk down Wall Street: including a life-cycle guide to personal investing*. [S.l.]: WW Norton & Company, 1999.

MARKOWITZ, H. Portfolio selection. *The Journal of Finance*, Wiley Online Library, v. 7, n. 1, p. 77–91, 1952.

MCCARTHY, J. Proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine*, v. 27, n. 4, p. 12–14, 2006.

MCKINNEY, W. *Data structures for statistical computing in Python*. [S.l.: s.n.], 2010. 51–56 p.

MOODY, J.; WU, L.; LIAO, Y.; SAFFELL, M. Performance functions and reinforcement learning for trading systems and portfolios. In: *Proceedings of the Conference on Computational Intelligence for Financial Engineering (CIFEr)*. [S.l.]: IEEE, 1998. p. 3–9.

MURPHY, K. P. *Machine Learning: A Probabilistic Perspective*. [S.l.]: MIT Press, 2012.

NORTH, D. C. *The rise of the western world: A new economic history*. [S.l.]: Cambridge University Press, 1973.

NYSTRUP, P.; LINDSTRöM, E.; MADSEN, H. Hyperparameter optimization for portfolio selection. *The Journal of Financial Data Science*, 2020. Disponível em: <https://www.pm-research.com/content/iijjfds/2/3/40>.

OLIPHANT, T. E. *A guide to NumPy*. [S.l.]: Trelgol Publishing USA, 2006.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O. *Scikit-learn: Machine Learning in Python*. 2011. 2825–2830 p.

PETERS, G. W.; MARR, E. *Financial Signal Processing and Machine Learning*. [S.l.]: John Wiley & Sons, 2016.

RODER, M.; GOMES, N.; YOSHIDA, A.; PAPA, J. P.; COSTEN, F. Multimodal convolutional deep belief networks for stroke classification with fourier transform. In: *2023 36th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. [S.l.: s.n.], 2023. p. 163–168.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. [S.l.]: Pearson Education, 2016.

SBRANA, A.; CASTRO, P. A. Lima de. N-beats perceiver: a novel approach for robust cryptocurrency portfolio forecasting. *Computational Economics*, Springer, p. 1–35, 2023.

SCHÖLKOPF, B.; SMOLA, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. [S.l.]: MIT Press, 2002.

SEGAL, T. Overview of the stock market. *Investopedia*, 2021. Disponível em: <https://www.investopedia.com/articles/basics/06/invest1000.asp>.

SEVERINO, M.; COIMBRA, C. *Investidores Individuais no Brasil: Perfil e Desafios*. Comissão de Valores Mobiliários (CVM), 2016. Disponível em: <https://conteudo.cvm.gov.br/menu/acesso_informacao/serieshistoricas/livros.html>.

SHARPE, W. F. *The Sharpe ratio*. [S.l.: s.n.], 1994. v. 21. 49–58 p.

SHILLER, R. J. Irrational exuberance: Revised and expanded third edition. Princeton university press, 2015.

SICILIANO, B.; SCIAVICCO, L.; VILLANI, L.; ORIOLO, G. *Robotics: Modelling, Planning and Control*. [S.l.]: Springer, 2016.

SMOLA, A. J.; SCHÖLKOPF, B. Learning with kernels. *Springer*, p. 168–204, 1998.

SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning: An Introduction*. [S.l.]: MIT Press, 2018.

THALER, R. H. *Misbehaving: The making of behavioral economics*. [S.l.]: WW Norton & Company, 2015.

VAPNIK, V. N. The nature of statistical learning theory. *Springer*, p. 1–14, 1995.

VAPNIK, V. N. *Statistical Learning Theory*. [S.l.]: Wiley, 1998.

ZHANG, L.; WANG, Y.; ZHAO, M. Sentiment analysis and prediction of stock price movements using cnn. *Journal of Financial Studies*, v. 18, p. 182–195, 2020.

ZHAO, Y.; ZHANG, W.; LIU, X. Grid search with a weighted error function: Hyper-parameter optimization for financial time series forecasting. *Applied Soft Computing*, v. 154, p. 111362, 2024. ISSN 1568-4946. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1568494624001364>.