Dissertação de Mestrado

APRENDIZADO DE MÁQUINA NA PRECIFICAÇÃO DE EMPRESAS: UMA ESTRATÉGIA DE INVESTIMENTO

Brendo Henrique de Lima

Orientador:

Prof. Dr. Evandro de Barros Costa

Maceió, Janeiro de 2025

Brendo Henrique de Lima

APRENDIZADO DE MÁQUINA NA PRECIFICAÇÃO DE EMPRESAS: UMA ESTRATÉGIA DE INVESTIMENTO

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Curso de Mestrado em Informática do Instituto de Computação da Universidade Federal de Alagoas.

Orientador:

Prof. Dr. Evandro de Barros Costa

Maceió, Janeiro de 2025

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Taciana Sousa dos Santos – CRB-4 – 2062

L732a Lima, Brendo Henrique.
Aprendizado de máquina na precificação de empresas : uma estratégia de investimento / Brendo Henrique Lima. – 2025.
48 f. : il. color.

Orientador: Evandro de Barros Costa.
Dissertação (Mestrado em Informática) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2025.

Bibliografia: f. 42-44.
Apêndices: f. 45-48.

1. Valuation. 2. Empresas – Precificação. 3. Aprendizado de máquina. 4. Estratégia de investimento. I. Título.

CDU: 004 : 658

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO
Av. Lourival Melo Mota, S/N, Tabuleiro do Martins, Maceió - AL, 57.072-970
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO (PROPEP)
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Folha de Aprovação

BRENDO HENRIQUE DE LIMA

APRENDIZADO DE MÁQUINA NA PRECIFICAÇÃO DE EMPRESAS: UMA ESTRATÉGIA DE INVESTIMENTO

APPLICATION OF ALGORITHMS OF MACHINE LEARNING FOR PREDICTION MARKET VALUE OF COMPANIES

Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas e aprovada em 7 de março de 2025.

Banca Examinadora:

Documento assinado digitalmente
gov.br
EVANDRO DE BARROS COSTA
Data: 23/04/2025 21:55:19-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. EVANDRO DE BARROS COSTA
UFAL – PPGI- Instituto de Computação
Orientador

Documento assinado digitalmente
gov.br
LUCIANA PEIXOTO SANTA RITA
Data: 25/04/2025 10:18:10-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. LUCIANA PEIXOTO SANTA RITA
UFAL-Faculdade de Economia, Administração e Contabilidade-FEAC
Examinador Externo

Documento assinado digitalmente
gov.br
THALES MIRANDA DE ALMEIDA VIEIRA
Data: 25/04/2025 14:00:25-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. THALES MIRANDA DE ALMEIDA VIEIRA
UFAL – PPGI- Instituto de Computação
Examinador Interno

Documento assinado digitalmente
gov.br
JOSE ANTAO BELTRAO MOURA
Data: 25/04/2025 10:35:30-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. JOSÉ ANTÃO BELTRÃO MOURA
UFCG– Universidade Federal de Campina Grande.
Examinador Externo

Resumo

Os modelos tradicionais de *valuation* apresentam desafios significativos. Suas limitações incluem o viés, que compromete a imparcialidade devido a informações prévias e pressões externas; a incerteza, comum em startups e empresas em transição; e a complexidade, intensificada pela globalização e mudanças nas normas contábeis. Além disso, técnicas como o fluxo de caixa descontado podem ser complexos demais e de pouca capacidade de generalização, tornando-se rapidamente obsoletos por dependerem de parâmetros específicos de cada empresa e momento. Este trabalho propõe precificar empresas listadas na bolsa de valores B3 e, a partir desse método de precificação, selecionar ativos para compor uma estratégia de investimento. Utilizando algoritmos de Aprendizado de Máquina, o estudo adota o valor de mercado (Market Cap) como variável dependente e explora diversos modelos para determinar o preço justo de empresas. A precisão dos modelos foi avaliada pela métrica Mean Absolute Percentage Error (MAPE), indicando a necessidade de dados adicionais para mitigar sobreajuste e subajuste. A estratégia de investimento, que seleciona empresas subvalorizadas e rebalanceia o portfólio trimestralmente, mostrou-se eficaz em superar o benchmark em quase todos os grupos de dados e algoritmos testados. Por fim, as evidências estatísticas indicaram que não há relação entre a métrica MAPE e os retornos obtidos pelas estratégias de investimento.

Palavras-chave: Valuation; Aprendizado de Máquina; investimento; bolsa de valores

Abstract

Traditional valuation models present significant challenges. Their limitations include bias, which compromises impartiality due to prior information and external pressures; uncertainty, common in startups and transitioning companies; and complexity, intensified by globalization and changes in accounting standards. Additionally, techniques like discounted cash flow may be overly complex and not easily generalizable, quickly becoming obsolete due to their dependence on company and time-specific parameters. This work proposes pricing companies listed on the B3 stock exchange and, based on this pricing method, selecting assets to compose an investment strategy. Using machine learning algorithms, the study adopts market capitalization (Market Cap) as the dependent variable and explores various models to determine the fair price of companies. The accuracy of the models was evaluated using the Mean Absolute Percentage Error (MAPE) metric, indicating the need for additional data to mitigate overfitting and underfitting. The investment strategy, which selects undervalued companies and rebalances the portfolio quarterly, proved effective in outperforming the benchmark in almost all data groups and algorithms tested. Finally, statistical evidence indicated that there is no relationship between the MAPE metric and the returns achieved by the investment strategies.

Key-words: Valuation; Machine Learning; investments; stock exchange

Lista de Tabelas

Lista de Figuras

Conteúdo

1

Introdução

No campo das finanças corporativas, como destacado por Assaf Neto and Lima (2017), são tratados temas de gestão financeira de instituições e as variáveis que influenciam a tomada de decisão, a avaliação de empresas é essencial. Ela serve como base para decisões de investimento, orienta transações comerciais, facilita a captação de recursos, apoia a definição de estratégias e promove transparência no mercado. A avaliação de empresas, também conhecida como *valuation* ou precificação de ativos, fornece uma base sólida para a tomada de decisões financeiras e ajuda a maximizar o valor e o sucesso de uma empresa [26].

Também é importante discorrer sobre o mercado financeiro, um ambiente dinâmico onde investidores individuais e grandes instituições financeiras buscam oportunidades de lucro. Para isso, uma variedade de estratégias de investimento são empregadas e continuam a evoluir constantemente, especialmente com o surgimento do big data. A evolução tecnológica desempenha um papel crucial na capacitação dos investidores, oferecendo ferramentas e recursos indispensáveis para desenvolver e implementar estratégias de investimento mais sofisticadas e eficazes. O acesso a grandes volumes de dados financeiros, tanto históricos quanto em tempo real, abriu caminho para a aplicação de técnicas estatísticas avançadas e aprendizado de máquina [2]. Essas abordagens permitem aos investidores identificar padrões, tendências e oportunidades de investimento nos mercados financeiros de maneira mais precisa e rápida.

Goodell et al. (2021) abordam o crescente interesse na aplicação do aprendizado de máquina (ML) em finanças, destacando sua superioridade em relação aos modelos econométricos tradicionais. Empresas como bancos, fundos de investimento e fintechs estão intensificando investimentos em ciência de dados, impulsionadas pelo crescimento exponencial da geração de dados e do poder computacional. O ML permite identificar padrões ocultos, prever tendências, inferir relações causais e visualizar conexões complexas entre variáveis financeiras. Essa capacidade oferece novas perspectivas para a análise de dados e aprimora a tomada de decisões estratégicas.

1.1 Motivação

Há diversos métodos de se avaliar uma empresa, Feuser (2021) diz que não há razão para escolher entre os métodos de avaliação, pois nada impede que sejam realizadas avaliações para um mesmo ativo com diversas abordagens, ele fala que a chance de se obter sucesso em um investimento é aumentada caso o ativo resulte estar subavaliado sob as diversas óticas de avaliação.

Diversos autores apontam limitações nos modelos tradicionais de avaliação de empresas. Damodaran (2015) aponta três grandes desafios no valuation: o viés, que dificulta uma avaliação imparcial devido a informações prévias e, em muitos casos, à pressão de atender expectativas do cliente; a incerteza, característica inerente especialmente em startups e empresas em transição, que leva analistas a buscarem atalhos; e a complexidade, ampliada pela globalização e por mudanças nas normas contábeis, que introduzem desafios como risco país e estruturas de propriedade.

Além disso, Vayas-Ortega et al. (2020) também fizeram algumas pontuações: abordagens tradicionais de fluxo de caixa descontado, que é uma técnica de valuation, pode ser complexa ou estar muito relacionada a alguns parâmetros chave desenvolvidos especialmente para aquela empresa específica e momento específico, o que dificilmente é generalizado ou facilmente torna-se obsoleto.

Este trabalho propõe um método baseado exclusivamente em dados para a precificação de empresas, contribuindo para a solução dos desafios apresentados. Os métodos tradicionais de precificação de empresas observam as empresas com perspectivas diferentes, já utilizando algoritmos de aprendizado de máquina é possível incluir diversas variáveis que tratam da empresa, do setor que ela está inserida e do cenário econômico, buscando eliminar o viés que os métodos tradicionais impõem. Assim como no trabalho de Huang et al. (2021), que utiliza aprendizado de máquina para selecionar ações, formar carteiras e obter retornos monetários, o presente estudo também busca retornos financeiros a partir de uma estratégia de investimento em ações brasileiras. A estratégia utiliza o método de precificação proposto para identificar e selecionar as empresas mais baratas, compondo um portfólio de ações.

Em busca de criar um método de avaliação de empresas (*Valuation*), os autores Vayas-Ortega et al. (2020) utilizaram modelos de aprendizado de máquina não paramétricos e supervisionados, a fim de obter uma certa resposta contínua, e assim caracterizar a avaliação da empresa pretendida, eles adotaram o *Enterprise Value* (Valor de Firma) como variável dependente dos modelos, enquanto neste artigo foi utilizado o *Market Cap* (Valor de Mercado) em busca do mesmo objetivo.

É importante compreender a diferença entre preço e valor, ressaltando que este trabalho trata apenas de preço. Feuser (2021) diz que o valor ou preço justo de um ativo é calculado pelos benefícios futuros esperados ao se realizar o investimento, independentemente da abordagem adotada, o autor caracteriza que o preço de um ativo deriva do preço que o comprador e o vendedor negociaram em determinado instante, definindo como uma relação de oferta e

demanda.

1.2 Objetivo

Este trabalho tem como objetivo principal desenvolver, analisar e validar modelos de aprendizado de máquina para prever o preço de empresas a partir de dados contábeis e macroeconômicos. Além disso, estas previsões são utilizadas para identificar ativos subvalorizados, implementar uma estratégia de investimento e avaliar sua performance por meio de backtests.

Já os objetivos específicos são:

- Coletar e tratar os dados necessários para a modelagem preditiva;
- Realizar análise descritiva para explorar os padrões nos dados;
- Desenvolver e testar diferentes técnicas de aprendizado de máquina para prever preços de empresas;
- Avaliar o desempenho dos modelos considerando métricas de erro e capacidade preditiva;
- Selecionar ativos subvalorizados com base nas previsões dos modelos;
- Aplicar uma estratégia de investimento utilizando os ativos selecionados;
- Realizar backtests para validar a estratégia de investimento;
- Comparar os retornos acumulados da estratégia com benchmarks do mercado.

2

Fundamentação Teórica

Este capítulo trata de conceitos da área de finanças corporativas e de técnicas de aprendizagem supervisionada que serão essenciais para entendimento da metodologia proposta no trabalho, bem como trabalhos já realizados nas diversas áreas que venham a contribuir na construção dessa pesquisa.

2.1 Demonstrações Contábeis

Segundo Ribeiro (2017) a análise de balanço é uma ferramenta fundamental utilizada para avaliar a saúde financeira e o desempenho de uma empresa, ela envolve a interpretação e o estudo dos elementos contábeis presentes no Balanço Patrimonial (BP), Demonstração do Resultado do Exercício (DRE), Fluxo de Caixa da empresa (DFC), Demonstração do Valor Adicionado (DVA), Demonstração das Mutações do Patrimônio Líquido (DMPL) e as Notas explicativas. As empresas de capital aberto, aquelas cujas ações são negociadas em bolsas de valores, geralmente são obrigadas a divulgar esses dados financeiros de forma regular, em intervalos trimestrais.

Embora o Balanço Patrimonial (BP), a Demonstração do Resultado do Exercício (DRE) e a Demonstração do Fluxo de Caixa (DFC) sejam os principais demonstrativos financeiros, os relatórios adicionais e as notas explicativas fornecem detalhes e informações importantes para uma compreensão mais aprofundada da empresa, porém, neste capítulo irei abordar apenas os principais demonstrativos sabendo que já suprem o conhecimento teórico deste trabalho.

Diniz (2015) comenta que, por meio da análise de balanço, podemos extrair informações das demonstrações contábeis para a tomada de decisões. As demonstrações fornecem uma série de dados sobre a empresa e a análise visa transformar esses dados em informações, onde o processo de análise será mais eficiente à medida que as informações produzidas possuírem mais qualidade. A autora também fala sobre a nomenclatura desse tipo de análise, que tem

sido chamada pelos estudiosos de Análise das demonstrações contábeis, Análise econômico-financeira, Análise de Balanços, Análise contábil e Análise Financeira.

O Balanço Patrimonial (BP) indica os resultados das atividades de investimento e financiamento em um momento no tempo, sendo composto por três seções principais:

1. Ativos: representam os recursos controlados pela empresa, que têm valor econômico e podem gerar benefícios futuros. O ativo é dividido principalmente em duas subcategorias:
    - Ativos circulantes: incluem itens de curto prazo, como dinheiro em caixa, contas a receber de clientes, estoques e investimentos de curto prazo.
    - Ativos não circulantes: referem-se a ativos de longo prazo, como investimentos de longo prazo, imobilizado (como propriedades, equipamentos, veículos), ativos intangíveis (como patentes, marcas registradas) e outros ativos de longo prazo.
2. Passivos: representam as obrigações financeiras, dívidas e despesas da empresa. O passivo também é dividido em duas subcategorias:
    - Passivos Circulantes: são as obrigações de curto prazo que devem ser liquidadas dentro de um ano ou no ciclo operacional normal da empresa. Exemplos de passivos circulantes incluem contas a pagar, empréstimos de curto prazo e impostos a pagar.
    - Passivos Não Circulantes: são as dívidas e obrigações financeiras que a companhia tem a saldar e que vencerão após um ano. Isso pode incluir empréstimos de longo prazo, títulos de dívida e outras obrigações de longo prazo.
3. Patrimônio Líquido: tem relação direta com os ativos e passivos, ele pode ser obtido pela diferença entre o total do ativo e do passivo, indicando o retorno financeiro que os empreendedores tiveram com a empresa ao término de um período. Ou seja, o patrimônio líquido compreende todo o dinheiro que os acionistas investiram acrescido do que ganharam após e deixaram na empresa para que ela continuasse funcionando. Ele inclui o capital social emitido, lucros ou prejuízos acumulados, reservas de capital e outras variações do patrimônio líquido.

A Demonstração do Resultado do Exercício (DRE) reflete o sucesso da empresa na utilização de ativos para gerar lucros, visando fornecer de maneira esquematizada os resultados (lucro ou prejuízo) auferidos durante um período. Segundo Diniz (2015), a DRE é uma demonstração contábil que possibilita observar os aumentos e as reduções causados no Patrimônio Líquido pelas operações executadas pela companhia. Normalmente, as receitas representam aumento do ativo, e aumentando o ativo, aumenta o patrimônio líquido. As despesas representam redução do patrimônio líquido, por meio de um entre dois caminhos possíveis: redução do Ativo ou aumento do Passivo Exigível. A autora simplifica que a DRE é o resumo do movimento de certas entradas e saídas no Balanço Patrimonial, entre duas datas. Algumas das informações contidas na DRE são:

- Receita total ou receita bruta: é o volume de produtos ou serviços vendidos por uma empresa. Ou seja, é tudo que entrou de dinheiro para a empresa a partir de suas vendas.
- Custos dos bens vendidos: representa os gastos diretamente relacionados à produção ou à entrega dos produtos ou serviços vendidos pela empresa. Incluem custos de materiais, mão de obra, despesas com fabricação, entre outros.
- Lucro bruto: é obtido pela diferença entre a receita total e os custos dos bens vendidos, é a quantia de dinheiro que a empresa ganhou com sua receita total menos os custos de matéria-prima e mão de obra utilizadas para produzir os bens.
- Despesas operacionais: são os gastos necessários para manter as operações diárias da empresa, mas que não estão diretamente relacionados à produção, como despesas administrativas, despesas de vendas e marketing, aluguéis, salários e benefícios dos funcionários, entre outros.
- EBITDA: a sigla vem do inglês e significa earnings before interest, taxes, depreciation and amortization. Em português é conhecido como lucro antes de juros, impostos, depreciação e amortização. É calculado para que seja possível ter uma noção clara de como anda a operação de fato da empresa, buscando refletir a geração de caixa efetivo da empresa.
- Resultado operacional: conhecido como EBIT (lucro antes dos juros e impostos) ou do inglês earnings before interest and taxes, é similar ao EBITDA, porém não é adicionado o valor da depreciação e da amortização. Ele representa o lucro que a empresa obteve com as atividades efetivamente ligadas ao negócio, excluindo ganhos ou despesas não relacionadas a isso.
- Lucro líquido: é o resultado final obtido após deduzir todas as despesas e os impostos da receita da empresa. Representa o lucro ou prejuízo líquido da empresa durante o período.

A Demonstração do Fluxo de Caixa (DFC) apresenta as entradas e as saídas líquidas de caixa das atividades operacionais, de investimentos e de financiamento para um período. Segundo Assaf Neto and Lima (2017), a demonstração do fluxo de caixa tem como principal objetivo a análise da capacidade financeira da empresa em cumprir suas obrigações com terceiros (empréstimos e financiamentos) e acionistas (dividendos), além de avaliar a geração de fluxos de caixa futuros a partir das operações atuais e a situação de liquidez e solvência financeira. As três principais categorias que a DFC é dividida são:

1. Atividades operacionais: apresentam, em essência, todas as entradas e saídas de recursos financeiros da empresa relacionadas às atividades operacionais principais. Alguns exemplos de entradas de caixa incluem o recebimento de dinheiro de vendas de produtos

ou serviços, o recebimento de pagamentos de vendas a prazo, o ganho de receitas financeiras provenientes de investimentos no mercado financeiro, dividendos de investimentos em ações de outras empresas e royalties ou licenças sobre a propriedade intelectual da empresa. Por outro lado, exemplos de saídas de caixa incluem pagamentos a fornecedores por mercadorias ou matérias-primas, o pagamento de salários e benefícios aos funcionários, o pagamento de juros sobre empréstimos ou dívidas da empresa, além dos pagamentos de impostos sobre vendas, imposto de renda e outras obrigações fiscais, entre outros.

2. Atividades de investimento: geralmente se referem às transações relacionadas a investimentos em ativos fixos, mais especificamente os não circulantes, visando a atividade operacional de produção e venda da empresa. São exemplos de entrada de caixa o recebimento de dinheiro decorrente da venda de ativos fixos, como imóveis, equipamentos e veículos. As saídas de caixa são determinadas pelo pagamento para aquisição dos mesmos bens de ativos fixos mencionados acima.

3. Atividades de financiamento: trata das transações relacionadas à captação e pagamento de recursos financeiros utilizados para financiar as operações e investimentos da empresa. Alguns exemplos de entrada de caixa são: emissão de ações ordinárias ou preferenciais e emissão de títulos de dívida, como debêntures ou notas promissórias, resultando no recebimento de dinheiro dos investidores. Já as saídas de caixa, se inserem os pagamentos de dividendos aos acionistas, recompra de ações próprias da empresa no mercado e pagamento de juros e principal de empréstimos contraídos pela empresa.

2.2 Análise das Demonstrações Contábeis

Consiste em uma técnica utilizada para avaliar a saúde financeira e o desempenho de uma empresa, levando em consideração tanto os aspectos econômicos quanto os aspectos financeiros dela. A técnica é guiada pelas demonstrações financeiras já mencionadas anteriormente, a fim de criar indicadores que também são chamados de índices econômicos. Ferrari (2013) define essa técnica como a obtenção, comparação e interpretação de indicadores que se apresentam sob forma de coeficientes, números índices ou quocientes calculados a partir de itens extraídos das demonstrações contábeis [8, p. 65].

Diniz (2015) menciona que a utilização de índices tem como principal objetivo permitir que o analista extraia tendências e consiga comparar os índices com padrões preestabelecidos, é importante que seja feita uma comparação dos indicadores da empresa analisada com os indicadores de outras empresas do mesmo setor em que ela está inserida. Vale mencionar que eles são utilizados para avaliação de empresas com destaque para o método de Valuation por Múltiplos, também conhecido como Avaliação Relativa. Dito isso, fica evidente a importância dos índices

para as diversas pesquisas das áreas que abordam empresas.

2.3 Aprendizado de Máquina

Fontana (2020) define que aprendizado de máquina (machine learning) é um termo geral utilizado para definir uma série de algoritmos que extraem informação a partir de um conjunto de dados. A partir de um conjunto de dados de treinamento, esses algoritmos buscam um padrão relacionando entradas e saídas, permitindo utilizar esse padrão para realizar predições [10, p. 5]. Além de prever valores futuros, o aprendizado de máquina é amplamente utilizado para identificar padrões complexos, detectar anomalias e auxiliar em processos de tomada de decisão. O principal desafio nessa área é garantir que os modelos aprendam de forma eficaz e consigam generalizar para novos dados. As técnicas de aprendizado de máquina são classificadas em três.

O aprendizado de máquina possibilita que sistemas se adaptem e aprendam automaticamente com os dados disponíveis. Seu objetivo é desenvolver modelos capazes de identificar tendências e padrões relevantes, utilizando essas informações para fazer previsões ou auxiliar na tomada de decisões.

1. Aprendizagem Supervisionada: Queiroz Filho (2020) diz que para problemas de aprendizagem supervisionada, os dados precisam ser estabelecidos previamente e devem ser rotulados, tanto os dados de entrada como de saída. O algoritmo determina uma maneira de prever qual o rótulo de saída com base nos dados de entrada.

2. Aprendizagem Não-Supervisionada: para os algoritmos não supervisionados, não são atribuídos rótulos ou classes aos dados. Assim como na aprendizagem supervisionada, o algoritmo busca padrões e similaridades entre os dados, que permitem identificar grupos de itens similares ou similaridade de itens novos com grupos já definidos, ou seja, ele tenta classificar ou predizer por conta própria.

3. Aprendizagem por Reforço: nesse tipo de aprendizagem, o algoritmo aprende qual ação a ser tomada ou resultado a ser obtido, de modo que haja uma recompensa a ser recebida de acordo com o resultado da ação, sendo o objetivo do algoritmo obter a maior recompensa possível.

Entre as abordagens da aprendizagem supervisionada, destacam-se os algoritmos de regressão, cujo objetivo principal é prever valores numéricos, e os algoritmos de classificação, voltados para prever categorias ou classes de dados.

2.3.1 Regressão Linear

Esse algoritmo faz parte dos modelos lineares, que são modelos que fazem predição a partir de uma função linear dos dados de entrada. Modelos lineares para regressão podem ser caracteri-

zados como modelos de regressão para os quais a previsão é uma linha quando houver apenas uma variável independente, um plano ao usar duas variáveis ou um hiperplano em dimensões maiores, ou seja, quando houver mais variáveis. Géron (2019) representou da seguinte maneira:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

- $\hat{y}$ é o valor predito
- n é o número de variáveis independentes
- x_i é o i-ésimo valor das variáveis independentes
- θ_j é o j-ésimo parâmetro

Existem muitos modelos lineares diferentes para regressão, a diferença entre eles é como os parâmetros do modelo são aprendidos a partir dos dados de treinamento e como a complexidade do modelo pode ser controlada. A regressão linear encontra os parâmetros que minimizam o erro quadrático médio entre as previsões e os verdadeiros alvos da regressão no conjunto de treinamento, sabendo que o erro quadrático médio é a soma das diferenças quadradas entre as previsões e os valores verdadeiros.

Uma das vantagens de se usar regressão linear é que se trata de um modelo simples de entender e interpretar, além de ser computacionalmente eficiente, porém como desvantagem há a necessidade de haver uma relação linear entre as variáveis independentes e a dependente, caso contrário, o modelo não fornecerá resultados precisos. A presença de outliers também afeta negativamente o modelo.

2.3.2 Árvores de Decisão e *Random Forest*

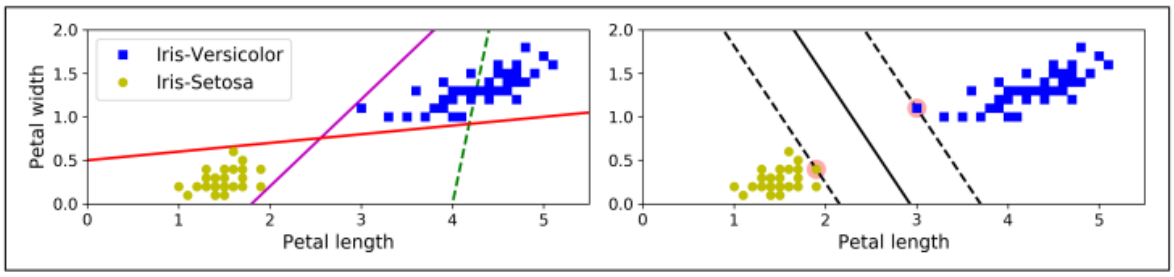
As Árvores de Decisão são modelos de aprendizado de máquina supervisionado amplamente utilizados para tarefas de classificação e regressão. Kelleher et al. (2015) descreveram que esses algoritmos aprendem uma hierarquia de questões "if-else" dentro de uma estrutura em forma de árvore que levam a uma decisão, as decisões são baseadas em uma série de perguntas que levam a diferentes caminhos dentro da árvore até chegar a uma conclusão. Treinar uma árvore de decisão significa aprender uma sequência de perguntas com suas respectivas respostas, onde nem sempre a resposta é uma variável categórica sim/não, podendo ser um valor contínuo onde o algoritmo determina um limite e divide em grupos.

Random Forest são essencialmente uma coleção de Árvores de Decisão, onde cada árvore é ligeiramente diferente das outras. A ideia é que cada árvore pode fazer um trabalho relativamente bom de previsão, mas são suscetíveis a overfitting (ajuste excessivo em parte dos dados). Construindo muitas árvores, todas funcionando bem e ajustadas de maneiras diferentes, são capazes de produzir uma previsão mais precisa e robusta. Essa técnica oferece maior poder preditivo, redução de overfitting e capacidade de medir a importância dos atributos.

2.3.3 *Support Vector Machine* (SVM)

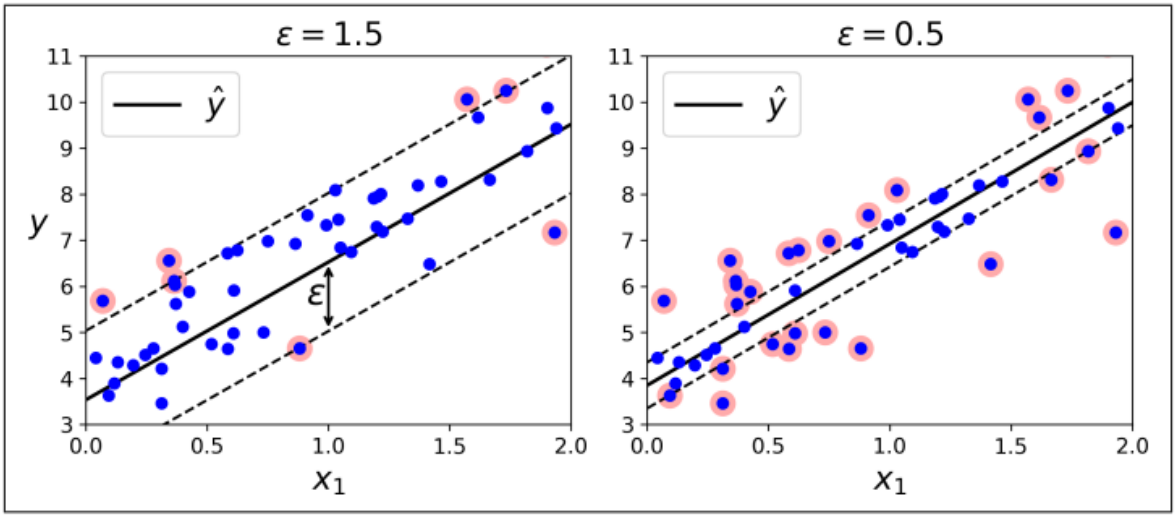
SVM é um modelo de aprendizado de máquina poderoso e versátil, que pode ser utilizado para problemas linear e não linear de classificação e regressão, e até para detectar outlier. Souza et al. (2014) resumem que a ideia do algoritmo para classificação é encontrar um hiperplano que separe os dados de diferentes classes de forma ótima, o melhor hiperplano será o que tiver maior margem para o conjunto de dados, sendo a margem a distância entre o hiperplano e o ponto de dado mais próximo de cada classe dentro dos dados. Géron (2019) apresentou figuras para representar o algoritmo, na Figura 2.1 o gráfico à esquerda mostra os hiperplanos de três possíveis classificadores lineares. O modelo cujo limite de decisão é representado pela linha tracejada é tão ruim que nem mesmo separa adequadamente as classes. Os outros dois modelos funcionam perfeitamente neste conjunto de treinamento, mas seus hiperplanos ficam tão próximos das classes que esses modelos provavelmente não terão um desempenho tão bom para separar novos dados. Em contraste, a linha sólida no gráfico à direita representa o limite de decisão (margem) de um classificador SVM, esta linha não apenas separa as duas classes, mas também permanece o mais longe possível das classes de treinamento mais próximas.

Figura 2.1: SVM - Problema de claffisicação [13, p. 156]



Para regressão, o algoritmo é conhecido como *Support Vector Regression* (SVR), ele busca uma função de regressão que se ajuste aos dados de treinamento dentro de uma margem de tolerância, tentando ajustar o maior número possível de dados dentro da margem enquanto limita as violações da mesma, essa margem é definida por hiperplanos em que os pontos de dados que estão além dela são considerados erros. O algoritmo formula a função como um problema de otimização, buscando minimizar o erro de previsão.

Figura 2.2: SVR - Problema de regressão [13, p. 165]



2.3.4 Métodos *Ensemble*

Kelleher et al. (2015) diz que grande parte do foco da aprendizagem de máquina está em desenvolver o modelo de previsão mais preciso possível para uma determinada tarefa. As técnicas de *ensemble* adotam uma abordagem diferente, em vez de criar um único modelo, elas geram um conjunto de modelos e, em seguida, fazem previsões agregando as saídas desses modelos. A ideia é que um *ensemble* pode ser altamente preciso, mesmo que os modelos individuais que o compõem tenham um desempenho apenas ligeiramente melhor do que uma previsão sem lógica. Os dois tipos mais famosos deses métodos são:

1. *Bagging*: é uma técnica que usa o mesmo algoritmo diversas vezes em diferentes conjuntos de dados aleatórios, depois que todos os modelos individuais são treinados, as previsões de cada modelo são combinadas para formar a previsão final. Géron (2019) exemplifica da seguinte maneira: treine um grupo de classificadores de Árvore de Decisão, cada um em um subconjunto aleatório diferente do conjunto de treinamento. Para fazer previsões, basta obter as previsões de todas as árvores individuais e, em seguida, prever a classe que recebe mais votos, esse conjunto de Árvores de Decisão é chamado de *Random Forest*.

2. *Boosting*: segundo Kelleher et al. (2015), essa técnica consiste na combinação de múltiplos modelos mais simples, visando a criação de um modelo mais robusto. A estratégia envolve o treinamento sequencial de diversos modelos, nos quais cada modelo subsequente é direcionado a corrigir os erros dos modelos anteriores. Em cada iteração, são atribuídos pesos às instâncias classificadas erroneamente, de modo que, na próxima iteração ou modelo, haja uma maior atenção nas instâncias com maior peso, buscando aprimorar seu desempenho. No caso de regressão, os pesos são atribuídos às instâncias cujas

previsões do modelo estão mais distantes dos valores reais, ou seja, aquelas que apresentam os maiores erros residuais. Para realizar predição, são atribuídos pesos a cada modelo do ensemble com base em sua performance, onde modelos mais eficazes recebem pesos maiores. Dessa forma, todos os modelos são empregados, cada um com seu respectivo peso, para realizar a predição final.

2.4 Preparação dos Dados

A preparação dos dados é uma etapa fundamental no desenvolvimento de modelos de aprendizado de máquina e análise de dados. Antes de aplicar qualquer técnica de modelagem, é necessário organizar, transformar e ajustar os dados para garantir que estejam em um formato adequado para os algoritmos. Esta etapa reduz ruídos, elimina inconsistências e garante maior precisão nos resultados dos modelos [13, 15].

Os dois processos mais comuns durante a preparação dos dados são a normalização e a padronização:

1. Normalização: é o processo de escalonar os valores das variáveis para que eles estejam dentro de um intervalo específico, geralmente entre 0 e 1. Este método é útil quando os dados possuem diferentes escalas e as diferenças absolutas podem influenciar excessivamente os algoritmos. A fórmula utilizada para normalização é:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2.1}$$

   Onde:

   - x é o valor original da variável.
   - $x_{\min}$ é o menor valor da variável.
   - $x_{\max}$ é o maior valor da variável.
   - x' é o valor normalizado.

2. Padronização: transforma os dados de forma que eles tenham uma média igual a 0 e desvio padrão igual a 1. Esse método é especialmente relevante quando os dados apresentam diferentes escalas, mas se deseja preservar as distribuições e comparações estatísticas. A fórmula para a padronização é dada por:

$$z = \frac{x - \mu}{\sigma} \tag{2.2}$$

   Onde:

   - x é o valor original da variável.

- μ é a média dos valores da variável.
- σ é o desvio padrão dos valores da variável.
- z é o valor padronizado.

2.5 Métrica de Avaliação

As métricas servem para medir o desempenho dos modelos com base em previsões. Apesar de existir várias métricas para regressão, este trabalho foca apenas em uma, é utilizado o *Mean Absolute Percentage Error* (MAPE), pois ele expressa a precisão do modelo em valores percentuais e é de fácil interpretação. O cálculo consiste na obtenção da média das porcentagens absolutas de erro entre as previsões do modelo e os valores reais, a equação é:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100, \tag{2.3}$$

onde n é o número total de observações; y_i é o valor real da i-ésima observação; $\hat{y}_i$ é o valor previsto da i-ésima observação; e o fator 100 é utilizado para expressar o erro em termos percentuais [16, 21].

Um MAPE baixo nos dados de treino indica que o modelo está capturando bem os padrões e tendências nos dados de treino, enquanto um MAPE alto nos dados de treino indica que o modelo não está capturando os padrões e está muito simples, indicando que há problema de subajuste (*underfitting*). Para os dados de teste, um valor baixo da métrica indica que o modelo generaliza bem e pode fazer previsões precisas em dados novos. Porém um MAPE significativamente mais alto nos dados de teste em comparação com os dados de treino indica sobreajuste (*overfitting*) [15, 20].

2.6 Seleção de Variáveis

A seleção de features é uma etapa crucial no desenvolvimento de modelos preditivos, pois visa identificar as características mais relevantes em um conjunto de dados que influenciam significativamente a previsão da variável alvo. Esse processo consiste em avaliar e selecionar variáveis com maior poder explicativo, ao mesmo tempo em que elimina aquelas consideradas irrelevantes ou redundantes.

Além de melhorar a acurácia do modelo preditivo, a seleção de features reduz a dimensionalidade do conjunto de dados, simplificando a complexidade computacional e facilitando a interpretação do modelo. Estudos mostram que variáveis irrelevantes não apenas aumentam o tempo de processamento, mas também podem prejudicar o desempenho do modelo, gerando vieses ou resultados inconsistentes [15, 21].

Diversas abordagens podem ser utilizadas para a seleção de variáveis, como algoritmos baseados em medidas de importância de variáveis, métodos estatísticos ou técnicas baseadas em aprendizado de máquina, como o uso do Random Forest com a métrica de importância de Gini. Essa métrica é uma medida que quantifica a relevância de cada variável na construção das árvores de decisão dentro do Random Forest. Para problemas de regressão, a importância de Gini é baseada na redução do erro da previsão nas divisões realizadas na árvore. A redução do erro ocorre quando uma divisão nos dados resulta em subgrupos que têm previsões mais precisas, ou seja, uma menor variação nos valores previstos. A importância de Gini de uma variável é calculada pela soma das reduções de erro em todas as árvores onde essa variável foi utilizada. Quanto maior a redução do erro, maior a importância da variável [15, 21].

2.7 Trabalhos Relacionados

Em busca de prever o valor de firma de empresas, que é o valor de mercado mais a dívida líquida, Vayas-Ortega et al. (2020) utilizaram modelos de aprendizado de máquina não paramétricos e supervisionados, a fim de obter uma certa resposta contínua, e assim caracterizar a avaliação da empresa pretendida. A partir do banco de dados do Thomson Reuters, os autores montaram diversos modelos de aprendizado de máquina, no total de 18, com características e técnicas diferentes, destacando: o uso de variáveis endógenas e exógenas individualmente e em conjunto, abrangeram o mercado de diferentes países e setores, utilizaram métodos de normalização, vários algoritmos de aprendizado de máquina e avaliaram os modelos de 4 maneiras diferentes em termos de erro. Após testar os modelos, os autores destacaram quais características se complementam para uma melhora nos algoritmos testados e também aquelas que pioram os mesmos. Eles concluíram que os métodos e variáveis incorporados em diferentes análises mostraram uma capacidade de generalização desigual, assim, propuseram a utilização de processos de aprendizagem especializados para fins específicos.

Utilizar aprendizado de máquina para auxiliar na seleção de ativos como estratégia de investimento é bem comum no universo de gestão de capitais. Queiroz Filho (2020) utilizou regressão logística e rede neural para selecionar ações de empresas listadas na bolsa de valores B3, tendo como objetivo obter um retorno elevado nas operações de mercado. Para isso, ele utilizou dados dos Balanços e Demonstrativos Financeiros trimestrais retirados do software Economática e calculou índices financeiros, que foram usados como variáveis independentes dos modelos. A variável dependente tem valor binário “investe” ou “não investe” calculada de maneira empírica: para cada ativo específico, verificou-se o retorno dos doze meses subsequentes ao período analisado. Dentro da amostra de ações de um trimestre específico, as ações foram divididas em dois grandes grupos: o primeiro possuindo as ações com maiores retornos em doze meses (investe) e o segundo possuindo as ações com piores retornos nos doze meses subsequentes (não investe). A pesquisa contou com dados do período de dezembro de 1999 até

dezembro de 2013, totalizando 4.588 dados de entrada para treino dos modelos, os trimestres de dezembro de 2013 até dezembro de 2017 ficaram fora do treinamento dos modelos pois foram utilizados para estimar o retorno real da estratégia, com 1936 linhas de entrada. A acurácia do modelo de regressão logística foi de 0,5625 e para o modelo de rede neural 0,5510, os modelos foram utilizados para construir carteiras a partir de uma estratégia de investimento que consiste na formação de portfólios trimestrais de ações, que foram balanceados de doze em doze meses. O retorno anual médio das ações que ele obteve no período de dezembro de 2013 a dezembro de 2017, foi de 11,4% no modelo de regressão logística e 7,7% no modelo de rede neural, contra um retorno de 9,3% do Ibovespa no mesmo período.

Ndikum (2020) utilizou aprendizado de máquina para fazer previsão de retorno anuais de ativos financeiros. Na tentativa de superar o Modelo de Precificação de Ativos Financeiros ou *Capital Asset Pricing Model* (CAPM), o autor testa diversos algoritmos de aprendizado de máquina com otimização de hiperparâmetros. O estudo conta com uma base de dados do período de 1983 ao início de 2019, abrangendo todas as ações negociadas publicamente dos Estados Unidos e que estavam disponíveis no Wharton Research Data Services, formando um total de 782 ações. Para o aprendizado de máquina, 30% do dataset foi utilizado como dados de teste, correspondendo aos anos de 2012 a 2018 aproximadamente. Nos modelos haviam aproximadamente 200 variáveis relacionadas ao desempenho da empresa, além de variáveis macroeconômicas dos EUA, onde o autor destaca a importância do cenário macro para a predição anual dos retornos nos dados de teste (2012 a 2018). Os modelos de aprendizado de máquina superaram o CAPM: com o uso do *Mean Squared Error* (MSE) foi possível ver que a taxa de erro do CAPM foi maior dentre os modelos testados, o autor destaca como vantagem na flexibilidade do aprendizado de máquina em incorporar diversas variáveis.

Marçal and Batista (2019) buscaram descrever a capacidade de explicação do preço das ações por meio de índices financeiros derivados de informações contábeis. Para averiguar a capacidade da informação contábil no que diz respeito à predição de valores no mercado de capitais, eles utilizaram a técnica de análise de regressão, com uma abordagem estatística que busca estimar as relações entre as variáveis. A variável dependente para o estudo foi o preço das ações das empresas. Todos os índices financeiros utilizados como variáveis explicativas são compostos exclusivamente de informações presentes em demonstrações contábeis. Tal escolha decorreu da intenção de isolar fatores de especulação de mercado e apresentar um modelo valendo-se apenas de informações contábeis que, consequentemente, seriam utilizadas na análise fundamentalista. As empresas selecionadas foram as que compunham a carteira teórica do IBrX-50 para o quadrimestre de setembro a dezembro de 2017. Os dados coletados para tais empresas foram extraídos da base de dados Economática e são referentes aos exercícios de 2011 a 2016, exceto o preço das ações, que está presente no intervalo de 2012 a 2017, dado que a data limite de publicações dos demonstrativos contábeis ocorre em período posterior ao ano a que se referem. Para o modelo final de regressão foi selecionado apenas um índice por grupo (Atividade; Endividamento; Liquidez; Rentabilidade). Portanto, foram utilizadas como variá-

veis explicativas apenas 4 dentre os 15 índices inicialmente observados. O critério de seleção foi a maior correlação com a variável de resposta (preço das ações) por grupo. Os autores então concluíram que foi possível confirmar a hipótese geral de que a utilização exclusivamente de informações contábeis na análise fundamentalista é capaz de explicar os preços das ações, os resultados apontaram que o modelo é estatisticamente significativo com 99% de confiança, e explicaria ainda, aproximadamente 29% dos preços das ações.

O método proposto por Milosevic (2016) usa uma combinação de algoritmos de aprendizado supervisionado para criar modelos a fim de predizer o preço das ações como um problema de classificação, onde o ativo é classificado como "bom" se tiver uma valorização de 10% no período de 1 ano e "ruim" caso contrário. Para as variáveis independentes dos modelos, foram utilizados dados de indicadores financeiros das empresas, ou seja, dados fundamentalistas. A base de dados foi coletada de várias empresas e índices, somando um total de 1298 empresas. Quanto aos diversos algoritmos utilizados, foram diversos: Árvores de Decisão, SVM, *Random Forest*, Regressão logística, Naive Bayes, entre outros. Nos resultados, a melhor performance foi do algoritmo *Random Forest*, com métricas de avaliação de modelos de classificação como Precision, Recall e F-score atingindo 75,1%, no entanto, o autor fez uma seleção das *features* mais importantes e conseguiu aumentar a performance para 76,5%.

Entre os trabalhos mencionados nesta subseção, Marçal and Batista (2019) ressaltam a importância das informações contábeis, enquanto os demais utilizam técnicas de aprendizado de máquina para realizar predições, há essa similaridade entre os trabalhos mencionados e o presente estudo. Dentre esses, o trabalho de Vayas-Ortega et al. aplica predições ao valor de firma de empresas, que é uma forma de fazer *valuation*, apresentando maior similaridade com o presente estudo. Uma diferença é que os autores não propuseram nem testaram estratégias de investimento baseadas nas previsões geradas pelos modelos, apenas utilizaram métricas de avaliação para medir o desempenho dos mesmos. O presente trabalho avança ao investigar a aplicabilidade das previsões na seleção de ativos e na tomada de decisão, contribuindo para a conexão entre técnicas de aprendizado de máquina e estratégias de investimento.

3

Metodologia

Algoritmos de aprendizado de máquina supervisionado, especificamente os algoritmos de regressão, foram escolhidos para prever o valor de mercado de uma empresa (Y) com base em um conjunto de características ou variáveis (X).

3.1 Algoritmos

Todos os algoritmos abaixo foram utilizados para aumentar a robustez do estudo, proporcionando uma análise mais sólida e confiável. A inclusão de uma ampla variedade de algoritmos busca não apenas validar os resultados sob diferentes abordagens, mas também oferecer uma visão mais abrangente das técnicas disponíveis e de suas aplicações.

- Regressão Linear
- *Random Forest*
- *Support Vectors Regression* (SVR)
- *Adaboost Regressor*
- *Catboost Regressor*

Para selecionar os parâmetros ótimos dos modelos, utilizou-se a metodologia de *Grid Search*, que realiza uma busca exaustiva por todas as combinações possíveis de hiperparâmetros definidos previamente. Os dados foram divididos aleatoriamente em 5 partes de mesmo tamanho, seguindo o método de validação cruzada, garantindo que cada modelo fosse avaliado em diferentes conjuntos de treino e teste. As combinações de hiperparâmetros testadas estão descritas na Tabela 3.1, e a seleção dos melhores valores foi feita com base no desempenho obtido em cada iteração [13].

Tabela 3.1: Valores considerados para parametrização de cada modelo

Algoritmo	Parâmetro	Valores
Regressão Linear	-	-
Random Forest	n_estimators	10; 20; ... ; 90; 100
	max_depth	None; 10; 20; 30; 40
	min_samples_split	2; 5; 10; 12
	min_samples_leaf	1; 2; 4; 6
SVR	kernel	rbf
	C	0.1; 0.5; 1; 2; 5; 8; 15
	gamma	0.001; 0.01; 0.1; 1
Adaboost	n_estimators	10; 20; ... ; 90; 100
	learning_rate	0.01; 0.05; 0.1; 0.5; 1
Catboost	depth	1; 2; 5; 8; 10
	iterations	1; 2; 3; 4; 5; 10; 20
	learning_rate	0.01; 0.05; 0.1; 0.5; 1

3.2 Base de Dados

Foi criado um painel (dados longitudinais) no período trimestral, abrangendo o intervalo do segundo trimestre de 2008 até o quarto trimestre de 2022, exceto para o valor de mercado e preço de fechamento das ações das companhias, que foram adiantados em 1 trimestre a fim de captar a reação do mercado sobre a divulgação dos resultados trimestrais das empresas, que ocorrem em período posterior ao fechamento do trimestre. Exemplificando, para o primeiro trimestre de 2022 (31/03/2022) os dados de valor de mercado e preço de fechamento da ações datam do último pregão do segundo trimestre (30/06/2022). É como se todos os dados do primeiro trimestre (31/03/2022) fossem divulgados somente na data do segundo trimestre (30/06/2022), o valor de mercado e preço das ações correspondem à data do segundo trimestre. Em resumo, os dados são compostos por 59 trimestres, iniciando em 2T2008 e terminando em 4T2022, para o valor de mercado e preço se inicia em 3T2008 e termina em 1T2023, a data do valor de mercado e do preço é a referência para o desenvolvimento deste trabalho.

3.2.1 Dados das Demonstrações Contábeis

Os dados contábeis trimestrais de empresas listadas na B3 foram coletados a partir do Comdinheiro, um sistema integrado de informações financeiras via web. Esses dados refletem os resultados financeiros das empresas ao longo do tempo. [1]

[1] As empresas foram selecionadas a partir da seguinte sequência de consultas no próprio site da B3: Produtos e serviços > Renda variável > Ações > Consultas > Classificação setorial. Foi baixado o arquivo que contém informações setoriais das empresas e seus respectivos códigos de negociação, o download do arquivo foi feito no mês de abril de 2023. Em seguida foram inseridas as iniciais (4 primeiras letras) dos códigos de negociação de todas as empresas do arquivo no site da Comdinheiro, com a condição que se a empresa tiver mais de um tipo de

1. Valor de mercado: variável dependente (Y) dos modelos que corresponde ao valor de mercado da companhia.
2. Preço de fechamento das ações.
3. Setor Classificação Comdinheiro: variável categórica que classifica os setores das empresas segundo o Comdinheiro.
4. Patrimônio Líquido (PL).
5. Lucro Líquido.
6. Receita Líquida.
7. Dívida Líquida.
8. Lucro Bruto.
9. EBIT.
10. EBITDA.
11. Retorno sobre Patrimônio Líquido (ROE):

$$\frac{\text{Lucro Líquido}}{\text{Patrimônio Líquido}} \tag{3.1}$$

12. Margem Bruta:

$$\frac{\text{Lucro Bruto}}{\text{Receita Líquida}} \tag{3.2}$$

13. Margem Líquida:

$$\frac{\text{Lucro Líquido}}{\text{Receita Líquida}} \tag{3.3}$$

14. Margem EBIT:

$$\frac{\text{EBIT}}{\text{Receita Líquida}} \tag{3.4}$$

15. Margem EBITDA:

$$\frac{\text{EBITDA}}{\text{Receita Líquida}} \tag{3.5}$$

16. Dívida Líquida sobre Patrimônio Líquido:

$$\frac{\text{Dívida Líquida}}{\text{Patrimônio Líquido}} \tag{3.6}$$

ação, ele não se repete e fica apenas um tipo, já que as informações contábeis não diferem pelo tipo de código de negociação. Por fim, foram coletados os dados contábeis das empresas

17. Dívida Líquida sobre EBITDA:

$$\frac{\text{Dívida Líquida}}{\text{EBITDA}} \tag{3.7}$$

18. Preço sobre Lucro:

$$\frac{\text{Valor de Mercado}}{\text{Lucro Líquido}} \tag{3.8}$$

19. EV sobre EBITDA:

$$\frac{\text{Valor de Mercado + Dívida Líquida}}{\text{EBITDA}} \tag{3.9}$$

20. Preço sobre Valor Patrimonial:

$$\frac{\text{Valor de Mercado}}{\text{Patrimônio Líquido}} \tag{3.10}$$

3.2.2 Dados Macroeconômicos - Banco Central do Brasil (BCB)

Os dados deste subgrupo foram coletados do Banco Central do Brasil (BCB) via interface de programação de aplicações (API).

21. Taxa de juros - Meta Selic definida pelo Copom: trata da taxa na data do último dia de cada trimestre.

22. Taxa de câmbio livre - Dólar americano venda: taxa correspondente da data do último pregão de cada trimestre.

23. Índice nacional de preços ao consumidor amplo (ipca): com periodicidade mensal, esses dados foram transformados para a periodicidade trimestral a partir da soma acumulada do índice no período de três meses de cada trimestre.

3.2.3 Dados Macroeconômicos - Instituto Brasileiro de Geografia e Estatística (IBGE)

24. Taxa de Desocupação: trata dos dados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) do final de cada trimestre.

As séries históricas da PNAD contínua começam somente em 2012, sendo necessário igualar com as datas dos outros dados. Para complementar os dados mais antigos, é realizada uma interpolação com a série histórica da Pesquisa Mensal de Emprego (PME), que foi descontinuada em 2016 e também é conduzida pelo IBGE. Bacciotti and Marçal (2020) complementam que a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD contínua), resultado da

integração entre a PME e a PNAD anual, passou a ser a única referência do instituto na produção e disseminação de estatísticas sobre o mercado de trabalho.

A interpolação é um método utilizado para estimar valores desconhecidos dentro de um intervalo de dados conhecidos, assim, foram feitas previsões para os valores da PNAD (Y) em função da PME (X). A abordagem utilizada foi linear, em que os valores desconhecidos são estimados como uma média ponderada dos valores conhecidos, assumindo uma relação linear entre eles.

Figura 3.1: Séries Históricas PNAD e PME

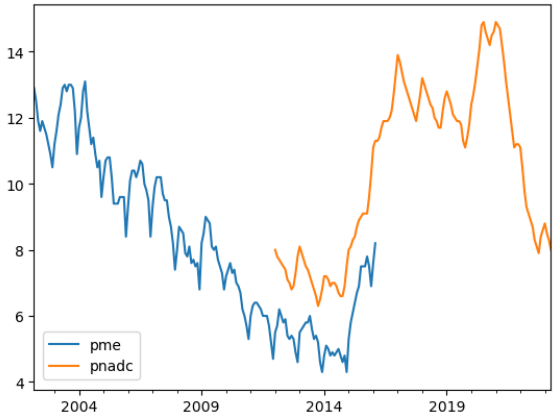
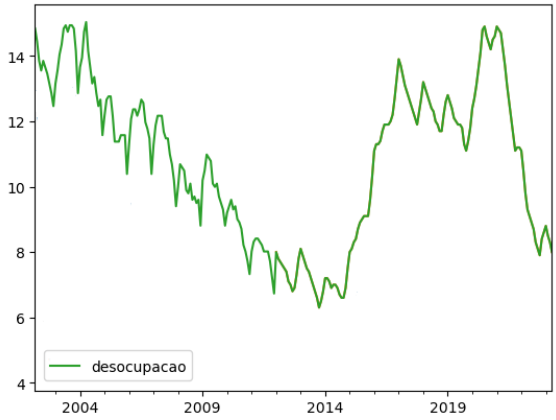


Figura 3.2: Série Histórica Interpolada



3.3 Padronização de Dados

A padronização foi escolhida como método de preparação dos dados devido à diferença significativa nas escalas das variáveis utilizadas no estudo. Quando as variáveis apresentam ordens de magnitude distintas, existe o risco de que aquelas com valores maiores tenham uma influência desproporcional no modelo, o que pode comprometer a qualidade das previsões. A padronização garante que todas as variáveis contribuam de forma equitativa para o aprendizado do modelo, independentemente de suas escalas originais.

3.4 Seleção de Variáveis

A seleção de *features* foi aplicada a alguns modelos para dar mais robustez ao estudo e será detalhada na seção dedicada aos modelos de predição. Neste estudo, a técnica escolhida para a seleção de variáveis foi o algoritmo *Random Forest*, sendo adotado a medida de importância de Gini para seleção de variáveis. Essa abordagem foi empregada sem a aplicação de parâmetros adicionais no algoritmo [18].

3.5 Desenvolvimento dos Modelos de Predição

A metodologia aplicada consiste na criação de vários modelos de aprendizado de máquina, cada um com sua especificidade, a fim de dar mais robustez para a pesquisa, lidando com diferentes cenários e, assim, proporcionar uma compreensão mais abrangente do tema. A seguir, é apresentado o passo a passo da construção dos modelos:

1. Para treinar todos os modelos, foi feito um filtro com objetivo de eliminar empresas muito pequenas e também aquelas com dificuldades financeiras. Foram eliminadas da base de dados todas as instâncias (linhas) onde o valor de mercado das companhias corresponde a um valor menor que o percentil 20 (R$ 195.755.720,00), também foram excluídas as linhas que apresentam patrimônio líquido negativo, o que indica que as dívidas da empresa superam seus ativos, comprometendo sua capacidade de operação. Em seguida, foi aplicado redução de grandeza dos números, feita uma divisão por 10^8 nas respectivas variáveis: patrimônio líquido, lucro líquido, receita líquida, dívida líquida, lucro bruto, ebit, ebitda e valor de mercado.

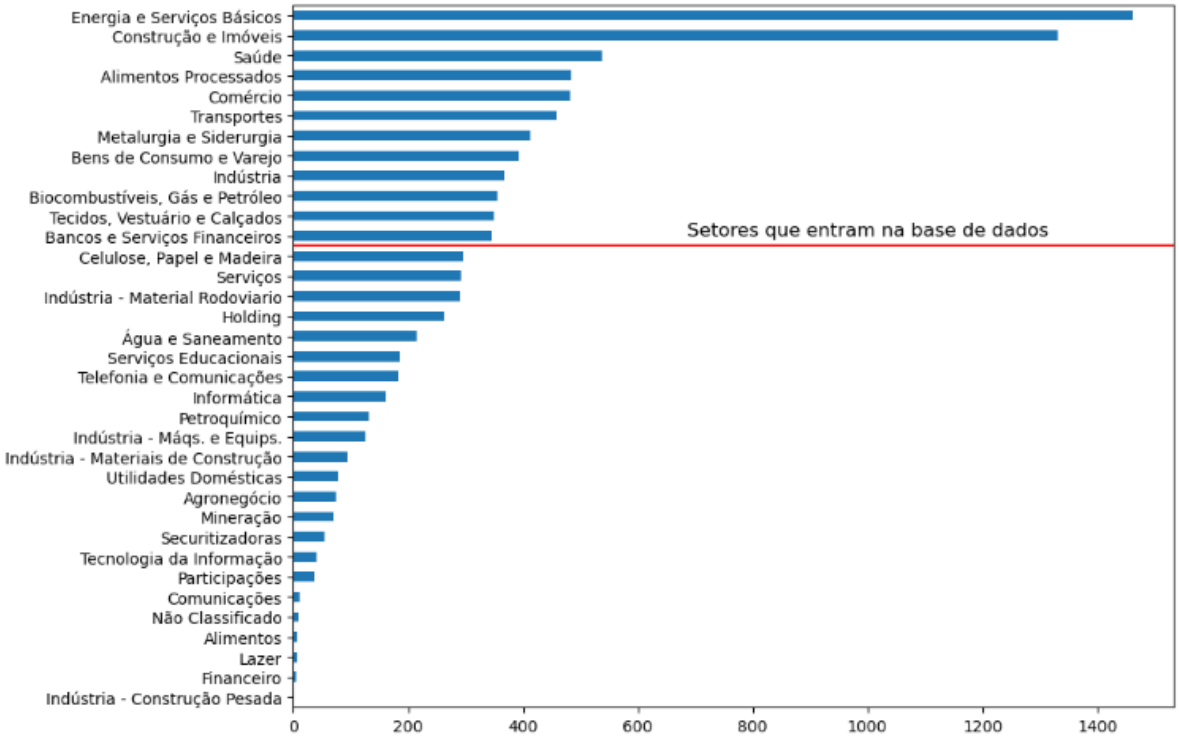
2. Com a premissa de modelar a precificação de empresas similares, todos os modelos foram treinados por setor, utilizando a *feature* "Setor Classificação Comdinheiro", porém, há 35 setores e alguns são escassos de dados, assim, foram selecionados os 12 setores com maior quantidade de instâncias para compor a base de dados, como mostra a Figura 3.4.

3. Com a mesma premissa de modelar a precificação de empresas similares, os dados de cada setor foram divididos no meio (mediana), ou seja, no percentil 50 com base na *feature* valor de mercado, o primeiro grupo com os 50% menores valores de mercado e o segundo grupo os 50% maiores valores de mercado. Reforçando que essa divisão é feita individualmente para cada setor. Modelos sem a divisão de dados também são criados e testados, a fim de comparação.

4. A seleção de variáveis é feita antes de padronizar os dados, já que para o algoritmo *Random Forest* não é necessário, assim, cada modelo após a divisão por setor e mediana, passa pela seleção de variáveis. As 7 melhores features de cada modelo foram selecionadas na composição. Também foram criados modelos com todas as variáveis a fim de comparar resultados.

5. A padronização dos dados é realizada para cada modelo, e em seguida os dados são divididos em treino e teste. Os dados de treino são divididos em 42 trimestres, iniciando no terceiro trimestre de 2008 (30/09/2008) e terminando no quarto trimestre de 2018 (31/12/2018), os dados de teste se iniciam no primeiro trimestre de 2019 (31/03/2019) e terminam no primeiro trimestre de 2023 (31/03/2023), total de 17 trimestres.

Figura 3.3: Quantidade de Instâncias por Setor



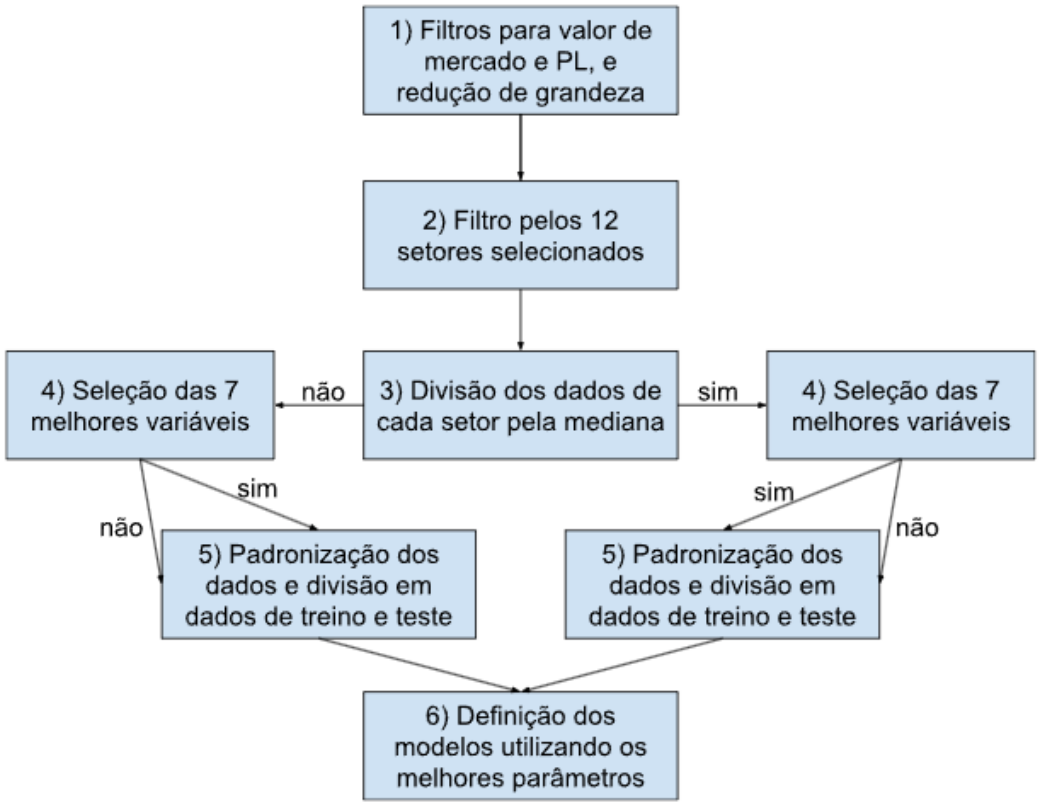
6. Por fim, cada um dos diversos conjuntos de dados de treino é processado pelos cinco algoritmos mencionados anteriormente, cada um utilizando seus melhores parâmetros. A predição do valor de mercado para os dados de teste é realizada com base em cada modelo.

3.5.1 Estratégia de Investimento

A estratégia consiste em selecionar ativos para compor um portfólio de ações, que é rebalanceado trimestralmente, onde todos os ativos têm o mesmo peso dentro do portfólio. É realizado o teste da estratégia no período de 31/03/2019 a 31/03/2023, os dados de teste dos modelos de predição foram utilizados propositalmente na condução da estratégia (*backtest*) a fim de minimizar os problemas de overfitting (sobreajuste), assegurando que os modelos não tivessem conhecimento prévio dos dados.

Para selecionar os ativos da carteira, utilizou-se o critério de comprar empresas subvalorizadas. Foi criada uma nova variável para garantir esse critério, que consiste na razão entre o valor de mercado real e o valor de mercado predito. A ideia por trás dessa variável é que o valor predito representa o valor ideal ou justo da empresa. Quando essa razão é calculada, é possível medir o quão descontada está cada empresa com base no valor de mercado real, que na condução da estratégia é considerado como o valor de mercado atual, correspondente à data

Figura 3.4: Fluxograma dos Modelos de Predição



de seleção de ativos e rebalanceamento. Quanto menor o valor dessa variável, mais barata a empresa é considerada.

A composição da carteira é formada pelas duas empresas mais descontadas de cada setor por trimestre, através de uma classificação *cross-sectional*, totalizando vinte e quatro (24) ativos e garantindo uma boa diversificação [4]. A estratégia assume que o portfólio é rebalanceado durante o leilão de fechamento das ações no último pregão de cada trimestre, já que só há dados do preço de fechamento das ações.

Por fim, foram armazenados os retornos simples nominais brutos da carteira para cada trimestre, sem considerar custos de transação e impostos. O retorno acumulado de todo o período de teste da estratégia foi calculado, levando em consideração o reinvestimento dos retornos em cada período.

4

Resultados

4.1 Análise Descritiva

A base de dados após a aplicação dos filtros mencionados no ponto 1 e a seleção por setores do ponto 2, ficou com 6.982 instâncias. Em relação à quantidade de empresas, a média foi de aproximadamente 118 empresas por trimestre, considerando todos os doze setores. Ao isolar por setor, foi utilizado a mediana como medida central para ter os resultados como números inteiros. Abaixo, a Tabela 4.1 mostra a mediana do número de empresas por trimestre para cada setor, e a Tabela 4.2 exibe a quantidade de instâncias utilizadas para treino e teste de cada modelo.

Tabela 4.1: Mediana de Empresas por Trimestre

Setor	Quantidade
Energia e Serviços Básicos	26
Construção e Imóveis	23
Saúde	7
Alimentos Processados	8
Comércio	8
Transportes	8
Metalurgia e Siderurgia	7
Bens de Consumo e Varejo	5
Indústria	7
Biocombustíveis, Gás e Petróleo	7
Tecidos, Vestuário e Calçados	6
Bancos e Serviços Financeiros	6

Tabela 4.2: Observações por Setor e Divisão

Setor	Observações Treino	Observações Teste	Divisão
Energia e Serviços Básicos	358	99	50% Menores Valores
Energia e Serviços Básicos	595	409	50% Maiores Valores
Energia e Serviços Básicos	953	508	100% dos Valores
Construção e Imóveis	663	307	50% Menores Valores
Construção e Imóveis	234	128	50% Maiores Valores
Construção e Imóveis	897	435	100% dos Valores
Saúde	129	102	50% Menores Valores
Saúde	160	147	50% Maiores Valores
Saúde	289	249	100% dos Valores
Alimentos Processados	89	59	50% Menores Valores
Alimentos Processados	213	124	50% Maiores Valores
Alimentos Processados	302	183	100% dos Valores
Comércio	141	95	50% Menores Valores
Comércio	133	114	50% Maiores Valores
Comércio	274	209	100% dos Valores
Transportes	143	81	50% Menores Valores
Transportes	132	102	50% Maiores Valores
Transportes	275	183	100% dos Valores
Metalurgia e Siderurgia	139	55	50% Menores Valores
Metalurgia e Siderurgia	142	77	50% Maiores Valores
Metalurgia e Siderurgia	281	132	100% dos Valores
Bens de Consumo e Varejo	61	37	50% Menores Valores
Bens de Consumo e Varejo	138	157	50% Maiores Valores
Bens de Consumo e Varejo	199	194	100% dos Valores
Indústria	157	75	50% Menores Valores
Indústria	91	44	50% Maiores Valores
Indústria	248	119	100% dos Valores
Biocombustíveis, Gás e Petróleo	88	28	50% Menores Valores
Biocombustíveis, Gás e Petróleo	128	112	50% Maiores Valores

Continua na próxima página

Setor	Observações Treino	Observações Teste	Divisão
Biocombustíveis, Gás e Petróleo	216	140	100% dos Valores
Tecidos, Vestuário e Calçados	143	104	50% Menores Valores
Tecidos, Vestuário e Calçados	61	42	50% Maiores Valores
Tecidos, Vestuário e Calçados	204	146	100% dos Valores
Bancos e Serviços Financeiros	76	49	50% Menores Valores
Bancos e Serviços Financeiros	143	78	50% Maiores Valores
Bancos e Serviços Financeiros	219	127	100% dos Valores

Para a análise descritiva, vale destacar que diversos setores carecem de informações, a quantidade baixa de empresas por setor e também de observações podem afetar negativamente a generalização dos modelos de predição. Isso sugere que os modelos provavelmente não serão eficazes na precificação de empresas não incluídas nos dados de treino. O sobreajuste (*overfitting*) e o subajuste (*underfitting*) são problemas que também podem ocorrer quando o número de instâncias é pequeno, no caso do sobreajuste, o modelo aprende com muito detalhe e especificidade os dados de treino, capturando até mesmo o ruído, o que faz com que ele tenha um desempenho excelente nos dados de treino, mas péssimo nos dados de teste. Já no caso do subajuste, o modelo é muito simples para captar a complexidade dos dados de treino, falhando em identificar padrões relevantes, resultando em um desempenho fraco tanto nos dados de treino quanto nos dados de teste.

4.2 Modelos de Predição

Na Tabela 4.3 são apresentados os valores da métrica MAPE, a diferença absoluta e a diferença relativa entre os dados de treino e teste para cada modelo. Ao comparar os modelos que realizaram a divisão pela mediana com os que não realizaram, observa-se uma melhora nas métricas de todos os modelos tanto para os dados de teste quanto de treino quando a divisão é aplicada. Em relação à seleção de variáveis, a aplicação também resulta em uma melhora das métricas para os dados de teste em todos os modelos, exceto para o algoritmo *Adaboost* com divisão pela mediana. Para os dados de treino, os algoritmos Regressão Linear e *Adaboost* apresentam uma piora nas métricas tanto para os modelos com divisão pela mediana quanto para os sem divisão. Um valor alto na diferença absoluta e relativa entre as métricas de treino e teste indica um problema de falta de generalização dos modelos. Isso significa que os modelos se ajustaram

Tabela 4.3: *Mean Absolute Percentage Error* (MAPE)

Divisão Mediana	Seleção de Variáveis	Algoritmo	Dados Treino	Dados Teste	Diferença Absoluta	Diferença Relativa
Não	Não	Regressão Linear	158,64%	446,73%	288,09%	181,60%
		Random Forest	8,42%	48,16%	39,74%	471,97%
		SVR	86,28%	219,69%	133,41%	154,62%
		Adaboost	87,54%	104,06%	16,52%	18,87%
		Catboost	39,19%	89,34%	50,15%	127,97%
Não	Sim	Regressão Linear	158,83%	271,35%	112,52%	70,84%
		Random Forest	7,60%	44,71%	37,11%	488,29%
		SVR	61,08%	123,37%	62,29%	101,98%
		Adaboost	104,98%	102,33%	2,65%	2,52%
		Catboost	30,93%	77,08%	46,15%	149,21%
Sim	Não	Regressão Linear	41,11%	178,15%	137,04%	333,35%
		Random Forest	6,74%	44,89%	38,15%	566,02%
		SVR	25,07%	62,51%	37,44%	149,34%
		Adaboost	24,09%	48,04%	23,95%	99,42%
		Catboost	13,92%	51,60%	37,68%	270,69%
Sim	Sim	Regressão Linear	46,92%	122,12%	75,20%	160,27%
		Random Forest	6,13%	42,27%	36,14%	589,56%
		SVR	23,43%	52,53%	29,10%	124,20%
		Adaboost	24,43%	49,92%	25,49%	104,34%
		Catboost	11,97%	48,53%	36,56%	305,43%

bem aos dados de treino, mas não conseguem fazer previsões precisas para novos dados. O algoritmo *Adaboost* apresentou o melhor resultado nesse teste.

4.3 Estratégia de Investimento

Nesta seção, apresentam-se quatro gráficos, cada um mostrando os retornos acumulados de cada grupo de dados, comparados com dois benchmarks. O primeiro benchmark é o índice Bovespa, enquanto o segundo corresponde a uma carteira composta por todas as ações da base de dados, após a aplicação dos filtros de valor de mercado, patrimônio líquido e setor, conforme detalhado nos pontos 1 e 2. Nessa carteira, os ativos são ponderados igualmente, ou seja, todos têm o mesmo peso. Os quatro grupos são descritos abaixo. Reforçando que nos dois primeiros grupos, foram utilizados 12 modelos (um para cada setor). Já no terceiro e quarto grupo, foram aplicados 24 modelos, pois os dados são divididos com base na mediana.

- Grupo Sem Divisão pela Mediana e Sem Seleção de Variáveis
- Grupo Sem Divisão pela Mediana e Com Seleção de Variáveis

- Grupo Com Divisão pela Mediana e Sem Seleção de Variáveis
- Grupo Com Divisão pela Mediana e Com Seleção de Variáveis

Figura 4.1: Retorno Acumulado Bruto (Grupo Sem Divisão pela Mediana e Sem Seleção de Variáveis)

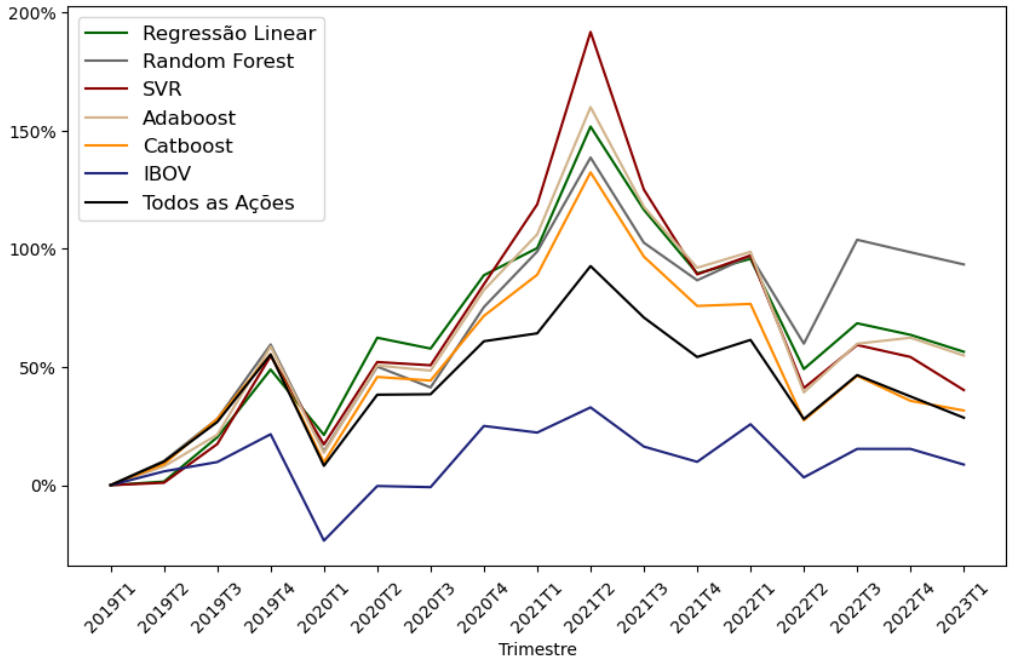


Figura 4.2: Retorno Acumulado Bruto (Grupo Sem Divisão pela Mediana e Com Seleção de Variáveis)

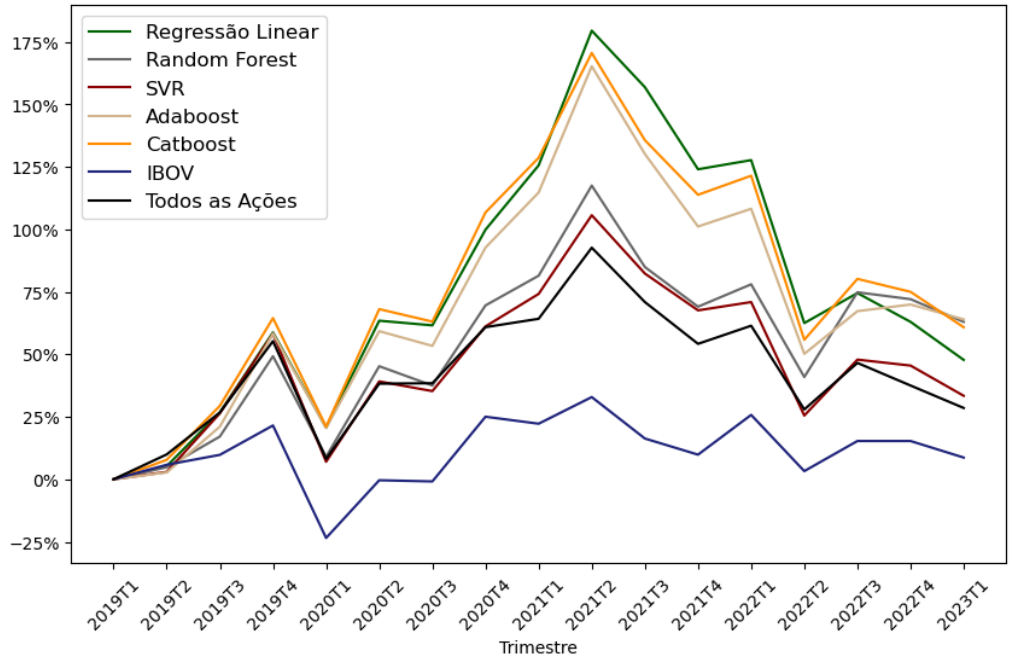


Figura 4.3: Retorno Acumulado Bruto (Grupo Com Divisão pela Mediana e Sem Seleção de Variáveis)

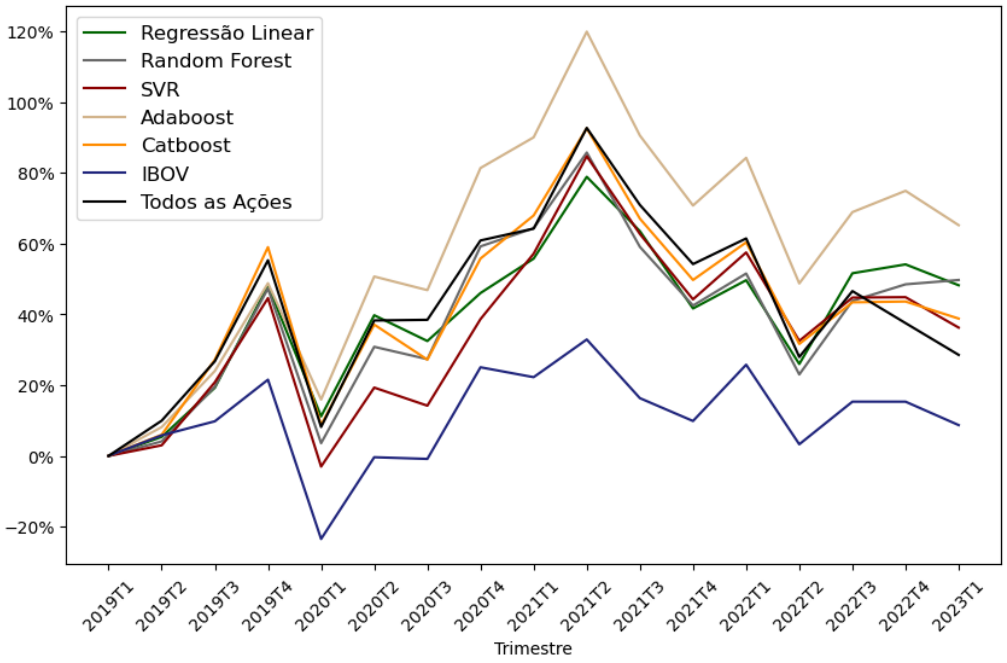
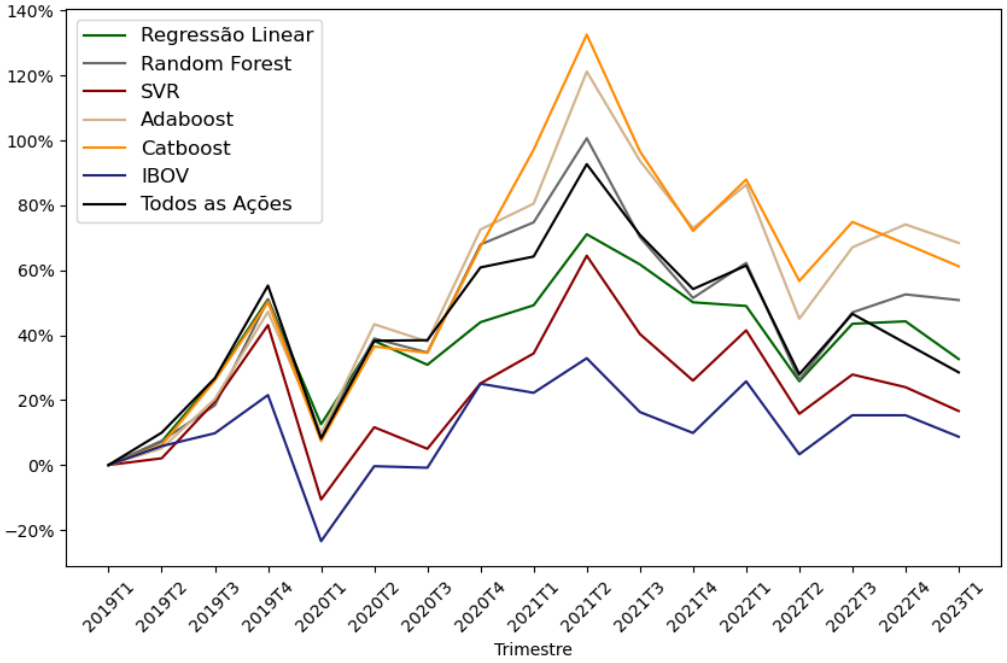


Figura 4.4: Retorno Acumulado Bruto (Grupo Com Divisão pela Mediana e Com Seleção de Variáveis)



O retorno acumulado das estratégias de investimento de todos os algoritmos, em todos os grupos, superou o retorno do índice Bovespa. No entanto, ao comparar com o benchmark da

carteira que contém todas as ações, apenas o algoritmo SVR do grupo com divisão pela mediana e seleção de variáveis não conseguiu obter um desempenho superior.

Para avaliar o desempenho dos grupos e identificar a relação entre um modelo com menor MAPE, que prediz o preço de uma empresa com maior precisão, e o retorno acumulado, foram compilados na Tabela 4.4 os resultados. A tabela apresenta, para cada grupo, a média do retorno acumulado na data final (2023T1) e a média do MAPE, tanto da carteira quanto dos dados de treino e teste. No cálculo da média do MAPE da carteira, foi considerado todo o período de teste, mas a média foi calculada somente a partir dos ativos que constavam na carteira em cada trimestre específico.

Tabela 4.4: Média por Grupo

	Sem Divisão Mediana e Sem Seleção	Sem Divisão Mediana e Com Seleção	Com Divisão Mediana e Sem Seleção	Com Divisão Mediana e Com Seleção
Média Retorno Acumulado	55,28%	53,78%	47,62%	45,93%
Média MAPE Carteira	100,29%	100,69%	98,12%	95,96%
Média MAPE Treino	76,01%	72,68%	22,18%	22,57%
Média MAPE Teste	181,59%	123,76%	77,03%	63,07%

A partir dos dados da Tabela 4.4, é possível observar que parece haver uma relação positiva entre a média de retorno e a média da métrica MAPE, à medida que a média do retorno acumulado diminui com a inclusão da divisão dos dados pela mediana e a seleção de variáveis, a média do MAPE também diminui, exceto para a média do MAPE treino do grupo Com Divisão Mediana e Com Seleção, que aumenta pouco em relação ao grupo Com Divisão Mediana e Sem Seleção, e a média do MAPE carteira do grupo Sem Divisão Mediana e Com Seleção, que aumenta um pouco também em relação ao grupo Sem Divisão Mediana e Sem Seleção.

As correlações de Pearson entre a média do retorno acumulado e o MAPE da carteira para os dados de treino e teste foram, respectivamente, 0,943, 0,983 e 0,947 [5]. Esses valores sugerem uma possível relação entre as variáveis. Para investigar essa relação, foi realizada uma regressão linear simples entre a média do retorno e a média do MAPE da carteira, e o resultado indicou um p-valor de 0,056 para a variável MAPE, sugerindo significância estatística ao nível de 10% [12]. Contudo, ao realizar a regressão com os dados não agrupados, considerando os valores individuais de retorno e MAPE de cada algoritmo em vez da média, o p-valor passa a rejeitar a hipótese de existência de uma relação estatisticamente significativa entre retorno e MAPE.

Foi explorado a variação dentro de cada grupo também através da regressão linear, onde foi possível observar que os p-valores indicaram que não há evidências estatísticas suficientes para

afirmar a existência de uma relação linear significativa entre o retorno e os valores de MAPE da carteira, MAPE dos dados de treino e MAPE dos dados de teste, considerando cada grupo individualmente. A correlação observada nas médias dos grupos sugere que há uma relação linear clara apenas em um nível agregado. No entanto, essa relação não se mantém quando analisamos os dados individualmente, sugerindo que outros fatores podem estar influenciando a relação entre retorno e MAPE dentro de cada grupo. Assim, não é possível afirmar que há uma relação entre as variáveis.

Tabela 4.5: Média por Algoritmo

	Regressão Linear	*Random Forest*	SVR	*Adaboost*	*Catboost*
Média Retorno Acumulado	46,25%	64,21%	31,62%	63,09%	48,01%
Média MAPE Carteira	100,46%	93,54%	95,43%	94,19%	100,21%
Média MAPE Treino	101,37%	7,22%	48,96%	60,26%	24,00%
Média MAPE Teste	254,58%	45,00%	114,52%	76,08%	66,63%

Se tratando dos algoritmos, a Tabela 4.5 mostra a média de retorno e MAPE de todos os algoritmos. O destaque positivo para o retorno fica com o algoritmo *Random Forest* e o negativo com o SVR. Para o MAPE da carteira, o algoritmo que apresentou o menor valor foi o *Random Forest* e o maior valor a Regressão Linear.

5

Conclusão

Buscou-se, neste trabalho, uma forma de precificar empresas se baseando apenas em dados, e também, a partir deste método de precificação, selecionar empresas subvalorizadas e implementar uma estratégia de investimento que é rebalanceada trimestralmente, avaliando a rentabilidade através de backtests.

Uma das principais dificuldades enfrentadas no trabalho foi a obtenção de dados confiáveis de empresas listadas na bolsa do Brasil. A falta de informações de qualidade limitou a análise para a periodicidade trimestral, podendo ter impactado a capacidade preditiva e a generalização dos modelos desenvolvidos. O ideal seria utilizar dados diários para treinar os modelos de predição.

A escassez de dados na pesquisa pode ter ocasionado o sobreajuste (*overfitting*) e subajuste (*underfitting*) dos modelos de aprendizado de máquina. Os resultados mostraram que há boas chances do sobreajuste ter ocorrido em todos os algoritmos, exceto para o Adaboost quando a base de dados não foi dividida pela mediana.

A métrica Mean Absolute Percentage Error (MAPE) apresentou valores relativamente altos, mas que foram diminuindo conforme os dados foram sendo testados em uma base de dados mais homogênea (porém menor), assim, apesar de haver uma melhora da métrica, é necessário a inserção de novos dados para evitar os problemas de sobreajuste e subajuste nos modelos de predição, permitindo avaliar melhor a capacidade de generalização dos modelos.

Quanto à estratégia de investimento, ela se mostrou eficaz em obter retornos acima do índice Bovespa para todos os grupos de dados e algoritmos, o outro benchmark que é uma carteira com todas as ações dos setores analisados, só não foi superado pelo algoritmo SVR do grupo de dados com mediana e com seleção de variáveis. O algoritmo Random Forest foi o destaque no quesito retorno.

Apesar das evidências estatísticas indicarem que não há relação entre a métrica MAPE e o retorno da estratégia de investimento, a metodologia proposta neste estudo reforça o potencial do aprendizado de máquina na área de investimentos. Ao permitir a seleção de empresas de

forma sistemática e baseada em dados, essa abordagem pode auxiliar desde investidores individuais até grandes bancos e gestoras na tomada de decisão, evidenciando o crescimento e a aplicabilidade dessas técnicas no mercado financeiro.

Referências bibliográficas

[1] A. Assaf Neto and F. G. Lima. *Curso de Administração Financeira*. Atlas, 2017.

[2] Susan Athey. The impact of machine learning on economics. *The Economics of Artificial Intelligence: An Agenda*, 2018.
DOI https://doi.org/10.7208/chicago/9780226613475.003.0021.

[3] Rafael Bacciotti and Emerson Fernandes Marçal. Taxa de desemprego no brasil em quatro décadas: retropolação da pnad contínua de 1976 a 2016. *Estudos Econômicos (São Paulo)*, 2020. **DOI** https://doi.org/10.1590/0101-41615035rbe.

[4] Jamil Baz, Nicolas Granger, Campbell R. Harvey, Nicolas Le Roux, and Sandy Rattray. Dissecting investment strategies in the cross section and time series. *SSRN Electronic Journal*, 2015. **DOI** http://dx.doi.org/10.2139/ssrn.2695101.

[5] Jules J. Berman. *Data Simplification: Taming Information with Open Source Tools*. Elsevier, 2016. **DOI** https://doi.org/10.1016/C2015-0-00783-3.

[6] Aswath Damodaran. These 3 issues can sink a business valuation. 2015. URL https://www.forbes.com/sites/aswathdamodaran/2015/08/16/these-3-issues-can-sink-a-business-valuation/?sh=701170363184. Acessado em 8 de Novembro de 2024.

[7] Natália Diniz. *Análise das demonstrações financeiras*. SESES - Estácio, 2015.

[8] Ed Luiz Ferrari. *Análise das Demonstrações Contábeis*. Impetus, 2013.

[9] C. E. P. Feuser. Valuation, 2021. Material didático ou instrucional do Programa de Pós-Graduação da FGV on-line.

[10] Éliton Fontana. Introdução aos algoritmos de aprendizagem supervisionada. 2020. URL https://fontana.paginas.ufsc.br/files/2018/03/apostila_ML.pdf.

[11] John W. Goodell, Satish Kumar, Weng Marc Lim, and Debidutta Pattnaik. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 2021. **DOI** https://doi.org/10.1016/j.jbef.2021.100577.

[12] Damodar N. Gujarati and Dawn C. Porter. *Econometria Básica*. AMGH, 2011.

[13] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly, 2019.

[14] Yuxuan Huang, Luiz Fernando Capretz, and Danny Ho. Machine learning for stock prediction based on fundamental analysis. *IEEE Symposium Series on Computational Intelligence*, 2021. **DOI** https://doi.org/10.1109/SSCI50451.2021.9660134.

[15] John D. Kelleher, Brian Mac Namee, and Aoife D'arcy. *Fundamentals of Machine Learning for Predictive Data Analytics – Algorithms, Worked Examples, and Case Studies*. The MIT Press, 2015.

[16] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, 2013.

[17] Ronan Reis Marçal and Tayná Cruz Batista. Relevância da informação contábil no mercado de capitais brasileiro: uma perspectiva por meio da análise fundamentalista em empresas do índice ibrx-50. *Revista de Contabilidade e Gestão Contemporânea*, 2019. URL https://periodicos.uff.br/rcgc/article/view/38395.

[18] Bjoern H Menze, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 2009. **DOI** https://doi.org/10.1186/1471-2105-10-213.

[19] Nikola Milosevic. Equity forecast: Predicting long term stock price movement using machine learning. *Journal of Economics Library*, 2016. **DOI** https://doi.org/10.48550/arXiv.1603.00751.

[20] Tom M. Mitchell. *Machine Learning*. McGraw-Hill Education, 1997.

[21] Andreas C. Müller and Sarah Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly, 2016.

[22] Philip Ndikum. Machine learning algorithms for financial asset price forecasting. *arXiv*, 2020. **DOI** https://doi.org/10.48550/arXiv.2004.01504.

[23] Ricardo Cezar de Queiroz Filho. Desenvolvimento de estratégia de investimento utilizando factor-based investing e aprendizado de máquina. Master's thesis, Fundação Getulio Vargas, São Paulo, 2020.

[24] Osni Moura Ribeiro. *Estrutura e análise de balanços*. Saraiva Educação SA, 2017.

[25] Davi L. Souza, Matheus H. Granzotto, Gustavo M. de Almeida, and Luís C. Oliveira-Lopes. Fault detection and diagnosis using support vector machines - a svc and svr comparison. *Journal of Safety Engineering*, 2014.
DOI 10.5923/j.safety.20140301.03.

[26] Germania Vayas-Ortega, Cristina Soguero-Ruiz, Margarita Rodríguez-Ibáñez, José-Luis Rojo-Álvarez, and Francisco-Javier Gimeno-Blanes. On the differential analysis of enterprise valuation methods as a guideline for unlisted companies assessment (ii): Applying machine-learning techniques for unbiased enterprise value assessment. *Applied Sciences*, 2020. URL https://www.mdpi.com/2076-3417/10/15/5334.

Apêndice

Tabela 5.1: Retorno Bruto - Sem Mediana e Sem Seleção de Variáveis

Trimestre	**Regressão Linear**	**Random Forest**	**SVR**	**Adaboost**	**Catboost**	**IBOV**	**Todas as Ações**
2019T2	1,50%	10,24%	1,01%	7,93%	8,92%	5,82%	9,95%
2019T3	18,49%	16,10%	16,22%	12,43%	17,58%	3,74%	15,25%
2019T4	23,79%	24,61%	31,71%	30,71%	21,29%	10,71%	22,52%
2020T1	-18,57%	-28,36%	-24,19%	-28,34%	-29,40%	-37,03%	-30,31%
2020T2	33,91%	31,39%	29,73%	32,49%	32,92%	30,18%	27,80%
2020T3	-2,83%	-5,86%	-0,90%	-1,44%	-1,04%	-0,48%	0,13%
2020T4	19,67%	24,06%	22,73%	22,97%	18,96%	26,11%	16,19%
2021T1	6,08%	13,41%	18,28%	12,90%	10,15%	-2,24%	2,08%
2021T2	25,66%	19,99%	33,34%	26,10%	22,91%	8,72%	17,32%
2021T3	-13,98%	-15,10%	-22,86%	-16,23%	-15,34%	-12,48%	-11,29%
2021T4	-12,44%	-7,84%	-15,94%	-11,84%	-10,62%	-5,55%	-9,77%
2022T1	3,26%	5,53%	4,24%	3,52%	0,50%	14,48%	4,70%
2022T2	-23,82%	-18,85%	-28,45%	-29,93%	-27,88%	-17,88%	-20,74%
2022T3	12,98%	27,50%	12,92%	14,79%	14,79%	11,67%	14,54%
2022T4	-2,90%	-2,59%	-3,19%	1,59%	-7,24%	-0,01%	-6,17%
2023T1	-4,38%	-2,60%	-9,08%	-4,64%	-3,02%	-5,74%	-6,53%
Acum.	56,42%	93,36%	40,22%	54,82%	31,56%	8,69%	28,53%

Tabela 5.2: Retorno Bruto - Sem Mediana e Com Seleção de Variáveis

Trimestre	**Regressão Linear**	**Random Forest**	**SVR**	**Adaboost**	**Catboost**	**IBOV**	**Todas as Ações**
2019T2	5,27%	4,83%	2,95%	2,75%	7,80%	5,82%	9,95%
2019T3	20,38%	11,71%	22,77%	17,82%	20,01%	3,74%	15,25%
2019T4	25,28%	27,45%	25,21%	30,75%	27,16%	10,71%	22,52%
2020T1	-23,96%	-26,94%	-32,37%	-23,86%	-26,40%	-37,03%	-30,31%
2020T2	35,40%	33,20%	30,01%	32,15%	38,83%	30,18%	27,80%
2020T3	-1,14%	-5,31%	-2,78%	-3,70%	-2,98%	-0,48%	0,13%
2020T4	23,69%	23,30%	19,13%	25,65%	26,79%	26,11%	16,19%
2021T1	12,87%	7,00%	8,06%	11,43%	10,57%	-2,24%	2,08%
2021T2	23,89%	19,87%	18,08%	23,52%	18,34%	8,72%	17,32%
2021T3	-8,09%	-14,98%	-11,35%	-13,23%	-12,90%	-12,48%	-11,29%
2021T4	-12,79%	-8,58%	-8,08%	-12,61%	-9,27%	-5,55%	-9,77%
2022T1	1,63%	5,30%	1,98%	3,52%	3,56%	14,48%	4,70%
2022T2	-28,63%	-20,84%	-26,56%	-27,87%	-29,64%	-17,88%	-20,74%
2022T3	7,39%	24,07%	17,82%	11,37%	15,64%	11,67%	14,54%
2022T4	-6,60%	-1,58%	-1,61%	1,59%	-2,88%	-0,01%	-6,17%
2023T1	-9,34%	-5,28%	-8,31%	-3,51%	-8,09%	-5,74%	-6,53%
Acum.	47,77%	62,97%	33,40%	63,96%	60,83%	8,69%	28,53%

Tabela 5.3: Retorno Bruto - Com Mediana e Sem Seleção de Variáveis

Trimestre	**Regressão Linear**	**Random Forest**	**SVR**	**Adaboost**	**Catboost**	**IBOV**	**Todas as Ações**
2019T2	5,37%	4,11%	2,99%	8,14%	5,97%	5,82%	9,95%
2019T3	13,19%	14,82%	17,29%	14,65%	20,02%	3,74%	15,25%
2019T4	24,20%	23,23%	19,69%	20,00%	25,01%	10,71%	22,52%
2020T1	-24,94%	-29,67%	-32,92%	-22,09%	-31,62%	-37,03%	-30,31%
2020T2	25,70%	26,28%	23,01%	29,99%	26,12%	30,18%	27,80%
2020T3	-5,22%	-2,67%	-4,28%	-2,55%	-7,23%	-0,48%	0,13%
2020T4	10,24%	25,03%	21,42%	23,50%	22,46%	26,11%	16,19%
2021T1	6,66%	3,21%	13,33%	4,78%	7,81%	-2,24%	2,08%
2021T2	14,85%	13,03%	17,52%	15,72%	14,67%	8,72%	17,32%
2021T3	-8,59%	-14,35%	-11,89%	-13,35%	-13,25%	-12,48%	-11,29%
2021T4	-13,36%	-10,43%	-11,35%	-10,38%	-10,38%	-5,55%	-9,77%
2022T1	5,64%	6,35%	9,19%	7,89%	7,10%	14,48%	4,70%
2022T2	-15,82%	-18,82%	-15,87%	-19,26%	-17,91%	-17,88%	-20,74%
2022T3	20,35%	16,90%	9,25%	13,55%	8,94%	11,67%	14,54%
2022T4	1,65%	3,25%	0,08%	3,56%	0,16%	-0,01%	-6,17%
2023T1	-3,83%	0,83%	-5,94%	-5,55%	-3,36%	-5,74%	-6,53%
Acum.	48,20%	49,70%	36,24%	65,18%	38,79%	8,69%	28,53%

Tabela 5.4: Retorno Bruto - Com Mediana e Com Seleção de Variáveis

Trimestre	**Regressão Linear**	**Random Forest**	**SVR**	**Adaboost**	**Catboost**	**IBOV**	**Todas as Ações**
2019T2	7,24%	7,45%	2,06%	5,12%	6,15%	5,82%	9,95%
2019T3	17,87%	10,17%	17,12%	14,43%	18,72%	3,74%	15,25%
2019T4	19,51%	27,45%	19,72%	22,39%	19,19%	10,71%	22,52%
2020T1	-25,50%	-27,49%	-37,56%	-25,51%	-28,53%	-37,03%	-30,31%
2020T2	22,82%	27,03%	24,93%	30,70%	27,16%	30,18%	27,80%
2020T3	-5,32%	-3,11%	-5,95%	-3,72%	-1,44%	-0,48%	0,13%
2020T4	10,03%	24,72%	19,19%	25,02%	24,34%	26,11%	16,19%
2021T1	3,62%	4,06%	7,37%	4,62%	17,85%	-2,24%	2,08%
2021T2	14,67%	14,82%	22,41%	22,54%	17,98%	8,72%	17,32%
2021T3	-5,43%	-15,17%	-14,66%	-12,42%	-15,45%	-12,48%	-11,29%
2021T4	-7,22%	-11,03%	-10,23%	-10,73%	-12,49%	-5,55%	-9,77%
2022T1	-0,75%	7,14%	12,28%	7,74%	9,22%	14,48%	4,70%
2022T2	-15,57%	-22,16%	-18,18%	-22,15%	-16,63%	-17,88%	-20,74%
2022T3	14,08%	16,41%	10,47%	15,18%	11,60%	11,67%	14,54%
2022T4	0,53%	3,77%	-3,05%	4,22%	-3,86%	-0,01%	-6,17%
2023T1	-8,05%	-1,15%	-5,93%	-3,27%	-4,11%	-5,74%	-6,53%
Acum.	32,63%	50,80%	16,61%	68,41%	61,21%	8,69%	28,53%