UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO SOCIOECONÔMICO
DEPARTAMENTO DE ECONOMIA E RELAÇÕES INTERNACIONAIS
CURSO CIÊNCIAS ECONÔMICAS

Gabriel Donadio Costa

Título: Aplicação de *Support Vector Machine* Para Previsão do Comportamento de Ações Negociadas no Mercado Brasileiro.

Florianópolis
2022

Gabriel Donadio Costa

Título: Aplicação de *Support Vector Machine* Para Previsão do Comportamento de Ações Negociadas no Mercado Brasileiro.

Trabalho Conclusão do Curso de Graduação em Ciências Econômicas do Centro Socioeconômico da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Bacharel em Economia.
Orientador: Prof. João Frois Caldeira, Dr.

Florianópolis
2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Costa, Gabriel Donadio
Aplicação de Support Vector Machine Para Previsão do Comportamento de Ações Negociadas no Mercado Brasileiro. / Gabriel Donadio Costa ; orientador, João Frois Caldeira, 2022.
46 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Sócio Econômico, Graduação em Ciências Econômicas, Florianópolis, 2022.

Inclui referências.

1. Ciências Econômicas. 2. Support Vector Machine. 3. Alocação de Capital. 4. Análise Técnica. 5. Análise Fundamentalista. I. Caldeira, João Frois . II. Universidade Federal de Santa Catarina. Graduação em Ciências Econômicas. III. Título.

Gabriel Donadio Costa

Título: **Aplicação de *Support Vector Machine* Para Previsão do Comportamento de Ações Negociadas no Mercado Brasileiro**.

Florianópolis, 04 de Março de 2022.

O presente Trabalho de Conclusão de Curso foi avaliado e aprovado pela banca examinadora composta pelos seguintes membros:

Prof.(a) João Frois Caldeira, Dr.
Universidade Federal de Santa Catarina

Prof.(a) Roberto Meurer, Dr.
Universidade Federal de Santa Catarina

Prof.(a) Pedro Luiz Paolino Chaim, Dr.
Universidade de Federal de Santa Catarina

Certifico que esta é a **versão original e final** do Trabalho de Conclusão de Curso que foi julgado adequado para obtenção do título de Bacharel em Economia por mim e pelos demais membros da banca examinadora.

Documento assinado digitalmente
Joao Frois Caldeira
Data: 09/03/2022 20:09:32-0300
CPF: 034.518.836-51
Verifique as assinaturas em https://v.ufsc.br

Prof. João Frois Caldeira, Dr.
Orientador

Florianópolis, 2022.

Este trabalho é dedicado à minha família, base de tudo, na qual encontro apoio incondicional.

AGRADECIMENTOS

Agradeço meus amigos e familiares, pelo apoio e incentivo em realizar mais uma graduação.

Agradeço os meus colegas de trabalho e a Universidade Aberta do SUS por me proporcionarem cursar as disciplinas.

Agradeço à Universidade Federal de Santa Catarina, pelos ensinamentos e pelo suporte durante todos estes anos. Também agradeço o Prof. João Caldeira, pela paciência e orientação durante a realização deste trabalho.

Por fim, agradeço especialmente o incondicional apoio da minha esposa e da minha mãe, pois sem elas, esta etapa não seria concluída.

RESUMO

Alocação de capital tem sido um grande desafio para os investidores da economia moderna devido à grande quantidade, variedade e volatilidade dos investimentos. Devido à complexidade inerente a este processo, existem ferramentas que podem ser de extrema valia para auxiliar a tomada de decisão dos investidores, entre as quais se destacam os modelos estatísticos e computacionais. O desenvolvimento de técnicas econométricas modernas e os avanços computacionais, proporcionaram o aperfeiçoamento de algoritmos que auxiliam a descoberta de padrões, previsão de rentabilidade e redução de dimensionalidade, como é o caso das máquinas de vetores de suporte. O presente trabalho busca testar ferramentas que possam auxiliar o processo de seleção de portfólio dos investidores. Para isto foram treinados quatro modelos de aprendizado de máquina para as tarefas de classificação (previsão do comportamento do preço da ação) e regressão (previsão do retorno da ação). A amostra compreendeu ações de seis empresas negociadas no mercado brasileiro, durante o período de 27 anos (de 18 de agosto de 1994 à 16 de dezembro de 2021), totalizando 40.562 observações. Os dois modelos de classificação treinados somente com variáveis advindas da análise fundamentalista ou da análise técnica apresentaram acurácia baixa, próximo ao aleatório. Notou-se uma melhora significativa no desempenho do modelo treinado com a combinação das variáveis provenientes das análises fundamentalistas e técnicas ($R^2 = 0,707$). Este resultado vem ao encontro da literatura, a qual aponta que modelos híbridos podem fornecer uma melhor acurácia na previsão das ações. Por fim, o modelo de regressão apresentou resultados satisfatórios somente quando aplicados a amostras de determinadas empresas, como a Vale S.A. ($R^2 = 0,790$) e a Petróleo Brasileiro S.A. ($R^2 = 0,663$). Concluiu-se, portanto, que o algoritmo *support vector machine*, treinado para a tarefa de classificação, com variáveis advindas das análises técnicas e fundamentalistas, pode auxiliar os investidores no processo de decisão de alocação de capital.

Palavras-chave: Support Vector Machine. Alocação de Capital. Análise Técnica. Análise Fundamentalista.

ABSTRACT

Capital allocation has been a major challenge for investors in the modern economy due to the quantity, variety and volatility of investments. Due to the inherent complexity of this process, there are tools that can be extremely valuable to assist investors in decision making, among which statistical and computational models stand out. The development of modern econometric techniques and computational advances have provided the improvement of algorithms that help the discovery of patterns, profitability prediction and dimensionality reduction, as is the case of support vector machines. The present work seeks to test tools that can help the investors' portfolio selection process. For this, four machine learning models were trained for classification tasks (predicting the share price behavior) and regression (predicting the share's return). The sample comprised shares of six companies traded on the Brazilian market, during a period of 27 years (from August 18, 1994 to December 16, 2021), totaling 40,562 observations. The two classification models trained only with variables from fundamental analysis or technical analysis showed low accuracy, close to random. There was a significant improvement in the performance of the trained model with the combination of variables from fundamental and technical analyzes ($R^2 = 0.707$). This result is in line with the literature, which points out that hybrid models can provide better accuracy in predicting stocks. Finally, the regression model showed satisfactory results only when applied to samples of certain companies, such as Vale S.A. ($R^2 = 0.790$) and Petróleo Brasileiro S.A. ($R^2 = 0.663$). It was concluded, therefore, that the support vector machine algorithm, trained for the classification task, with variables arising from technical and fundamental analysis, can help investors in the capital allocation decision process.

Keywords: Support Vector Machine. Capital Allocation. Technical analysis. Fundamental Analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT Associação Brasileira de Normas Técnicas

AM Aprendizado de Maquina

B3 Brasil, Bolsa, Balcão

BM&F Bolsa de Mercadorias & Futuros

BOVA11 iShares Ibovespa Fundo de Índice

BOVESPA Bolsa de Valores de São Paulo

BRAX11 iShares IBrX Indice Brasil

CD Ciência de Dados

CDI Certificado de Depósito Interbancário

CT Capital Turnover

D/E Debt-to-equity ratio

EMA Média Móvel Exponencial

E/P Earning-to-Price

FR Força Relativa

GASVM Algoritmo Genético

GPM Margem de Lucro Bruto

IBGE Instituto Brasileiro de Geografia e Estatística

ISE Istambul Stock Exchange

KOSPI Korea Composite Stock Price Index

MACD Média Móvel Convergente e Divergente

MCAP Market Captalization

ML Machine Learning

OBV On Balance Volume

P/B Price-to-Book

PCA Principal Component Analysis

PIB Produto Interno Bruto

R^2 Coeficiente de determinação

RBF Radial Based Function

RNA Redes Neurais Artificiais

RoA Return on Assets

RoE Return on Equity

RSI Índice de Força Relativa

SELIC Sistema Especial de Liquidação de Custódia

SMA Média Móvel Simples

S/P Sales-to-Price

SVM Support Vector Machine

TIE Cobertura de Juros

TV Trading Volume

SUMÁRIO

1 INTRODUÇÃO

Alocação de capital tem sido um grande desafio para os investidores da economia moderna devido à grande quantidade, variedade e volatilidade dos investimentos. O mercado atual possui desde opções conservadoras, como as poupanças e títulos do governo, até as mais arriscadas, como bolsa de valores e cripto moedas.

Nos últimos anos tem-se notado uma diminuição acentuada na taxa básica de juros. Do período de agosto de 2016 a junho de 2020, a SELIC caiu de 14,25% para 2,25%, registrando o menor patamar da história. Esta política econômica adotada pelo governo, teve como objetivo incentivar o consumo, estimular o investimento das empresas e diminuir a dívida do governo (FEIJÓ et al., 2022).

Devido a necessidade de diversificação dos investimentos e aumento da rentabilidade, observou-se também um aumento da procura pelo mercado de ações. O número de investidores pessoa física na bolsa cresceu 106% em 2019 e 92% em 2020, já o número de instituições financeiras que investem na B3 supera 21 mil (BRASIL, BOLSA, BALCÃO [B3], 2021).

O mercado de ações é uma alternativa de investimento que envolve um alto nível de complexidade, onde há intrínseco um certo grau de incerteza. Muitas vezes, estas incertezas estão associadas a eventos macroeconômicos, como decisões econômicas e políticas, eleições, escândalos de corrupção, já outras vezes podem tratar-se de ruídos, não-estacionaridade e caos determinístico (MARCELINO, 2016).

No âmbito das finanças, diversos estudos consideram que estas incertezas podem ser, em parte, previsíveis (BELTRAMI et al., 2011; HUERTA et al., 2013; ZHANG et al,. 2017). Para Fan e Palaniswami (2001), as séries históricas das cotações do mercado de ações possuem padrões que, se descobertos, possibilitam previsões a respeito do retorno futuro dos ativos.

Devido à complexidade inerente ao processo de alocação de capital e seleção de portfólio no mercado de ações, existem ferramentas que podem ser de extrema valia para auxiliar a tomada de decisão dos investidores, como por exemplo os modelos estatísticos e computacionais (SHATSHAT e AHMED, 2019). O desenvolvimento de técnicas econométricas modernas e os avanços computacionais, proporcionaram o aperfeiçoamento de algoritmos que auxiliam a descoberta de padrões, previsão de rentabilidade e redução de dimensionalidade, como é o caso das redes neurais, máquinas de vetores de suporte (*support vector machines*), métodos de ensemble, entre outros (GÉRON, 2019).

1.1 DELIMITAÇÃO DO TEMA

Nos últimos anos, as tecnologias relacionadas a aprendizado de máquina e ciência de dados tiveram seu uso intensificado na indústria financeira e demonstraram ser ferramentas úteis para geração de valor (SHATSHAT e AHMED, 2019), tanto para investidores "pessoa física" quanto para investidores "pessoa jurídica".

Dentre os investidores pessoa jurídica destacam-se as instituições financeiras (bancos e corretoras), na qual objetivam transferir recursos dos agentes econômicos superavitários para os deficitários (ASSAF NETO, 2001). Estas instituições fazem uso constante e intenso de tecnologias para otimização dos seus processos e negócios, monitorando a relação risco-retorno-alocação de capital, com o objetivo de subsidiar as decisões da equipe de gestão (BRITO, 2003).

O aprendizado de máquina (AM) é uma subárea da inteligência artificial que permite os computadores aprenderem com os dados (GÉRON, 2019). Em uma definição clássica, AM é o campo de estudo que permite que os computadores aprendam sem serem explicitamente programados (SAMUEL, 1959).

No campo da Ciência de Dados existem diversos algoritmos que objetivam o aprendizado de máquina através de exemplos. Os algoritmos têm o papel de reconhecer padrões nos dados, extrair informações e predizer novas observações (MARCELINO, 2016), estando amplamente inseridos nas atividades do cotidiano, como por exemplo, nos sistemas de recomendação de filmes da Netflix ou Youtube, no reconhecimento de *spams*, na direção de carros autônomos e também em sistemas de seleção de alocação de ativos e seleção de portfólios.

A Máquina de Vetores de Suporte é um dos algoritmos mais populares no âmbito do aprendizado de máquina devido a sua versatilidade e eficiência, sendo capaz de realizar tarefas de classificação linear e não linear, regressão e detecção de outliers (GÉRON, 2019). Este algoritmo tem-se mostrado eficiente na previsão do comportamento das ações e na seleção de portfólios, apresentando retornos acima da média de mercado (EMIR et al., 2012; FAN e PALANISWAMI, 2001; HUERTA et al., 2013; KIM, 2003; MARCELINO et al., 2015; ZHANG e ZHAO, 2009).

No âmbito das finanças, alguns dos principais métodos para previsão do comportamento do mercado acionário são a análise técnica e a análise fundamentalista. A análise técnica tenta prever o comportamento do mercado através de indicadores, gráficos e

histórico de cotações. Já a análise fundamentalista utiliza dados das demonstrações financeiras, relatórios gerenciais, características dos gestores de topo e processamento de linguagem natural (NTI et al., 2020a).

Em uma revisão sistemática a respeito da previsão do mercado de ações por meio das análises técnicas e fundamentalistas, NTI et al. (2020a) analisaram 122 artigos de grande impacto na comunidade científica. Os resultados demonstraram que apenas 11% dos artigos utilizaram técnicas combinadas entre as duas análises.

Portanto, este trabalho procura aplicar o algoritmo *support vector machine* na previsão do comportamento de ações negociadas no mercado brasileiro, utilizando a combinação das análises técnicas e fundamentalistas.

1.2 JUSTIFICATIVA

O mercado financeiro apresenta uma diversa gama de ativos, cada um com seu risco e rentabilidade. Via de regra, investimentos com alto risco podem acarretar em ganhos (ou perdas) extraordinários. Por outro lado, os investimentos de baixo risco resultam em uma menor volatilidade, o que garante um retorno menor, porém mais seguro.

O investimento na bolsa de valores apresenta muita volatilidade, podendo ser influenciado por diversas variáveis. Um furo de reportagem noticiado, uma postagem viralizada nas redes sociais ou um evento macroeconômico podem acarretar em uma fuga ou atração de investidores para empresa, elevando ou diminuindo o preço das ações.

Devido à complexidade e diversidade do mercado financeiro, os investidores precisam de informações tempestivas, úteis e precisas que auxiliem na tomada de decisão da forma mais rentável e segura de alocar seu capital. Neves Júnior et al. (2021) ressaltam que, para as instituições financeiras, a alocação de recursos constituem a base para a competitividade na indústria financeira.

Chenhall e Moers (2015) destacam que, com o passar dos anos, o processo de alocação de capital vem se adaptando e evoluindo com as inovações tecnológicas. As novas tecnologias buscam proporcionar redução de custos, aumento de desempenho e implementação de novas práticas, onde muitas vezes processos disruptivos são necessários.

Nos últimos anos, tem-se notado a introdução da inteligência artificial na indústria financeira, visando a detecção de fraudes (corporativas, em *e-commerces* ou em cartões de crédito), seleção de portfólios, entre outros processos. A complexidade destas tarefas faz os

gestores lançarem mão de diversas ferramentas que auxiliam o processo de análise e tomada de decisão, das quais destacam-se os modelos estatísticos e computacionais, como por exemplo as redes neurais e máquinas de vetores de suporte, que objetivam a descoberta de padrões, redução de dimensionalidade e previsão de rentabilidade, auxiliando o processo decisório de gestão, no âmbito das finanças corporativas.

Há cada vez mais indícios na literatura que sugerem que os métodos de aprendizado de máquina podem superar os métodos convencionais em problemas de predição econômica (BUCKMANN et al., 2021), como por exemplo previsão do retorno de títulos do tesouro (BIANCHI et al., 2021), curva de juros (MUN e SOONG, 2021), variáveis macroeconômicas (CHEN et al., 2019), previsão do comportamento e retorno das ações (EMIR et al., 2012; FAN e PALANISWAMI, 2001; HUERTA et al., 2013; KIM 2003; MARCELINO et al., 2015; NTI et al., 2020b; ZHANG e ZHAO, 2009), entre outros. O trabalho de Bianchi et al. (2021) utilizou modelos de *machine learning* como *partial least squares*, regressões lineares penalizadas, *boosted regression trees*, *random forests* e redes neurais para previsão do retorno dos títulos do tesouro de diversas maturidades, através centenas de indicadores macroeconômicos e financeiros. Os resultados demonstraram que as técnicas não lineares de *machine learning*, como as redes neurais, apresentaram os melhores resultados (R^2) na previsão dos retornos dos títulos, se comparados com técnicas lineares de AM e previsões baseadas apenas nas informações da curva de juros.

Já no âmbito da previsão do comportamento das ações, estudos recentes buscam combinar o uso de *machine learning* com análises técnicas e fundamentalistas, como por exemplo o trabalho de Nti et al. (2020b), que testaram a utilização da Máquina de Vetores de Suporte aprimorada com o Algoritmo Genético (GASVM) para seleção de *features*, otimização de parâmetros e previsão do preço das ações da bolsa de valores de Gana. No entanto, poucos estudos combinam as duas análises utilizando mais de uma fonte de dados, como é o caso de Zhang et al. (2017) que procurou testar dados de redes sociais e notícias da web e histórico de cotações das empresas chinesas e Ballings et al. (2015) que utilizou a combinação de dados macroeconômicos e o histórico de cotações na previsão do comportamento das ações.

Portanto, este trabalho pretende testar ferramentas que possam auxiliar o processo de seleção de portfólio dos investidores. Para isto, será aplicado quatro modelos de aprendizado de máquina buscando prever o comportamento de ações negociadas no mercado brasileiro, utilizando as análises técnicas e fundamentalistas.

1.3 OBJETIVOS

Nas seções abaixo estão descritos o objetivo geral e os objetivos específicos deste TCC.

1.3.1 Objetivo Geral

O presente trabalho visa testar ferramentas que possam auxiliar o processo de seleção de portfólio dos investidores, treinando quatro modelos de aprendizado de máquina que buscam prever o comportamento de ações negociadas no mercado brasileiro.

1.3.2 Objetivos Específicos

Coletar dados e variáveis provenientes da base de dados Economática, que auxiliem na previsão das ações negociadas no mercado brasileiro;

Treinar um algoritmo de aprendizado de máquina buscando identificar padrões nos dados;

Identificar qual tarefa (classificação ou regressão) possui o melhor poder preditivo do comportamento das ações;

Identificar qual análise (técnica, fundamentalista ou combinada) possui o melhor poder preditivo do comportamento das ações.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 SELEÇÃO DE PORTFÓLIO

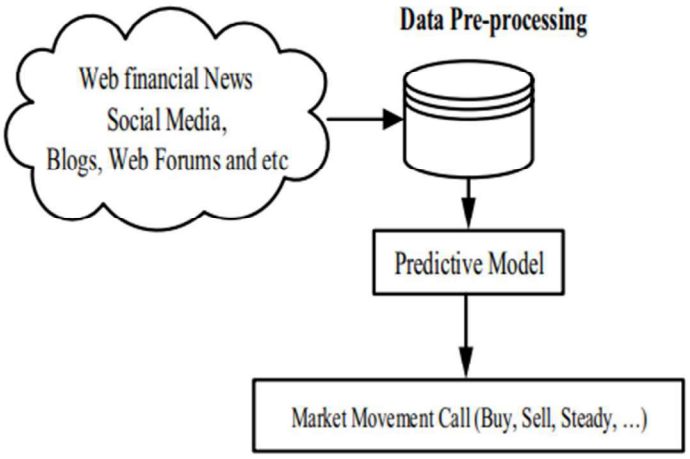
O processo de seleção de portifólio no mercado de ações é uma tarefa complexa devido à vasta variedade de opções de investimentos e ao grande número de fatores externos que podem influenciar no processo de decisão. Devido à alta complexidade, os tomadores de decisão utilizam técnicas para prever o comportamento do mercado de ações. Duas importantes ferramentas para previsão do comportamento da ação são as análises técnicas e fundamentalistas (RENU e CHRISTIE, 2018).

2.1.1 Análise Fundamentalista

A análise fundamentalista presume que o preço da ação, tanto no momento atual, quanto no futuro, depende do valor intrínseco da empresa e do retorno esperado.

Confirme figura abaixo, esta análise utiliza informações públicas disponíveis para analisar as ações com base nos aspectos gerais da economia (inflação, taxas de juros, PIB), do setor em que atua (níveis de competição, ameaça de novos concorrentes, políticas governamentais) e nos fundamentos da própria empresa (características dos gestores de topo, finanças da organização, notícias sobre a empresa, participação no mercado, capacidade de endividamento, entre outros), portanto é indicada para investimentos a médio e longo prazo (OLIVEIRA et al., 2013; RENU e CHRISTIE, 2018).

Figura 1 - Enfoque da análise fundamentalista



Fonte: Nti et al. (2020a)

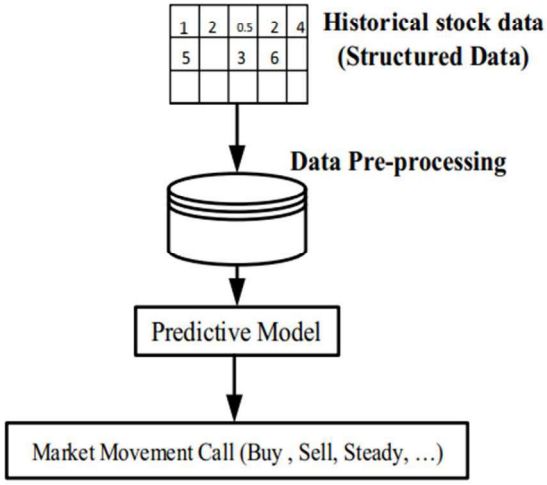
Os principais indicadores utilizados nesta técnica são o valor econômico adicionado – EVA, fluxo de caixa descontado – DCF, Debt-to-Equity – D/E, Price-to-book – P/B, Return-on-Equity – ROE e Return-on-Asset – ROA (CAVALCANTE et al., 2005).

2.1.2 Análise Técnica

Já a análise técnica envolve o emprego de ferramentas para prever o comportamento futuro do preço da ação, com base no padrão histórico. Foca em analisar a evolução dos mercados a partir da representação gráfica do histórico do preço de venda, preço de compra, volume negociado, entre outros.

Portanto, esta análise busca compreender e o comportamento histórico dos preços e volumes, visando determinar a tendência e o preço futuro dos ativos, sendo adequada a investimentos de curto prazo (OLIVEIRA et al., 2013; RENU e CHRISTIE, 2018).

Figura 2 - Enfoque análise técnica



Fonte: Nti et al. (2020a).

Elder (1993) classifica os indicadores da análise técnica em três grupos, são eles: rastreadores de tendência, osciladores e mistos.

Os rastreadores de tendência funcionam quando o mercado apresenta uma direção (tendência) definida. O uso em situações onde o mercado não apresenta uma tendência definida (mercado na horizontal) pode acarretar em resultados pouco fidedignos. São exemplos destes indicadores: média móveis simples - SMA e exponencial - EMA, médias móveis convergente e divergente - MACD, entre outros (ELDER, 1993).

O grupo dos osciladores são recomendados quando o mercado se encontra na horizontal (não há tendência definida). Os indicadores mais utilizados são o índice de força relativa – RSI e o K-estocástico. Por fim, os indicadores mistos são utilizados para um grupo de ações, não sendo possível aplicá-lo a uma única ação. Seu objetivo é sinalizar um consenso dos investidores em relação a um grupo de ações, mercado como um todo ou a um setor específico da economia (ELDER, 1993)

2.2 APRENDIZADO DE MÁQUINA

Aprendizado de Máquina (AM) é uma sub-área da inteligência artificial que permite os computadores aprenderem com os dados (GÉRON, 2019). Em uma definição clássica, o AM é o campo de estudo que permite que os computadores aprendam sem ser explicitamente programado (SAMUEL, 1959)

No campo de Ciência de Dados existem diversos algoritmos que objetivam o aprendizado de máquina através de exemplos. Os algoritmos têm o papel de reconhecer padrões nos dados, extrair informações e predizer novas observações (MARCELINO, 2016).

2.2.1 Tipos de Aprendizado de Máquina

Os algoritmos de aprendizado de máquina são classificados de acordo com o seu funcionamento na fase de treinamento, sendo as mais comuns o aprendizado supervisionado, não supervisionado e semi-supervisionado.

No aprendizado supervisionado os dados fornecidos na fase de treinamento do algoritmo apresentam os rótulos (soluções do problema). Dentre os diversos algoritmos de aprendizado supervisionado, podemos destacar: Regressão Linear, Árvores de Decisão, Redes Neurais, Máquinas de Suporte a Vetor (SVM) e Florestas Aleatórias (GÉRON, 2019).

Já no aprendizado não supervisionado os dados de treinamento que é fornecido ao algoritmo não apresentam os rótulos. Temos como exemplo os algoritmos análise dos componentes principais (PCA) e *k-means* (GÉRON, 2019).

Por fim, no aprendizado semi-supervisionado, o algoritmo aprende com dados parcialmente rotulados. Este é o caso da aplicação redes neurais no caso específico máquinas restritas de Boltzmann, nas quais utilizam parte dos dados não rotulados treinados de forma não

supervisionada e, posteriormente, ajustado utilizando técnicas de aprendizado supervisionado (GÉRON, 2019).

2.2.2 Desafios do Aprendizado de Máquina

Géron, (2019) destaca que os principais desafios do aprendizado de máquina consistem na detecção e eliminação do *overfitting* e *underfitting*, no volume e na qualidade dos dados de treinamento.

Visando o funcionamento de forma satisfatória dos algoritmos de aprendizado de máquina é necessário um grande conjunto de dados, que serão utilizados parte para treinamento do modelo e parte para o teste das previsões.

Os dados também precisam ser representativos do problema a ser estudado, de forma que o algoritmo possa reconhecer padrões, aprender e seja capaz de generalizar os resultados. Portanto, é imprescindível analisar as estatísticas descritivas e, por vezes, retirar do conjunto de treinamento os dados errôneos, ruídos e *outliers*.

O sobreajustamento dos dados (*overfitting*) ocorre quando o modelo funciona bem para os dados de treinamento, no entanto não consegue uma boa predição dos dados de teste. O sobreajuste pode ocorrer quando o modelo é muito complexo, isto é, possui muitas dimensões e poucos dados para o conjunto de treinamento.

As formas mais comuns de corrigir o *overffiting* são com a diminuição do número de dimensões (por exemplo através do *principal component analysis*), seleção de *features* ou o aumento dos dados do conjunto de treinamento.

Já o sub-ajustamento dos dados (*underfitting*) ocorre quando o modelo não possui a complexidade exigida para o conjunto de dados e, portanto, apresente um baixo desempenho. Este problema pode ser corrigido selecionando um modelo mais robusto e complexo e aumentando a dimensionalidade (*feature engineering*).

2.3 SUPPORT VECTOR MACHINES

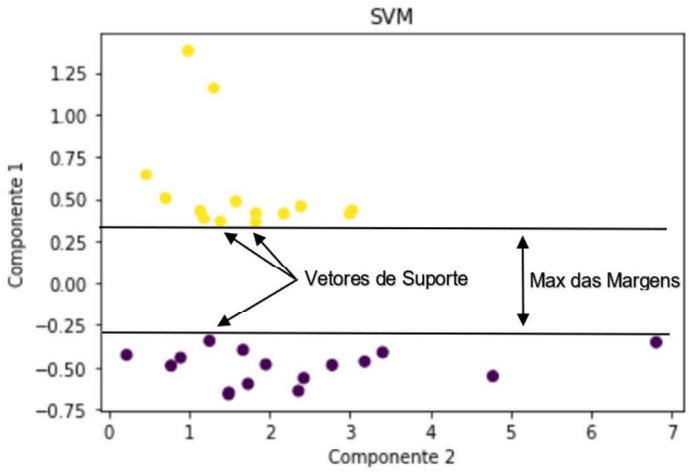
Um dos algoritmos mais populares no âmbito do aprendizado de máquina, a máquina de vetores de suporte foi criado na década de 90, através dos trabalhos de Boser et al. (1992) e Vapnik (1995).

Este algoritmo logo destacou-se pela capacidade de generalização, eficiência, versatilidade, facilidade em lidar com alta dimensionalidade dos dados e evitar *overfitting*, sendo utilizado para realizar tarefas de classificação não linear e linear, regressão, detecção de outliers entre outros (GÉRON, 2019).

Pesquisas na área da teoria de aprendizagem estatística buscavam encontrar um equilíbrio ideal entre o risco empírico e complexidade estrutural. O SVM implementa o princípio da minimização risco estrutural, diferente de outros algoritmos que buscam minimizar o erro empírico, como por exemplo as redes neurais. Portanto, enquanto as redes neurais buscam reduzir o erro de classificação incorreta no conjunto de treinamento, podendo ocasionar *overtraining*, o SVM objetiva minimizar o limite superior do erro de generalização, resultando em uma solução global ótima (FAN e PALANISWAMI, 2001; KIM, 2003).

O principal objetivo do algoritmo SVM é determinar um hiperplano de separação que visa distinguir os dados em duas ou mais classes, visando atingir a separação máxima entre elas, conforme pode ser visto na figura 3 (EMIR et al,. 2012; GUPTA, 2012; MARCELINO, 2016).

Figura 3 - Maximização das margens pelo SVM



Fonte: Elaboração Própria.

Portanto, dado um conjunto de treinamento (x_i, y_i), $onde\ i = 1,2, \dots m$, $x_i \in \mathbb{R}^n\ e\ y_i \in \{-1, +1\}$, para problemas de sepação linear, a classificação correta será dada por:

$$\langle w.x_i \rangle + b \geq +1\ para\ y_i = +1 \quad (1)$$

$$\langle w.x_i \rangle + b \leq -1\ para\ y_i = -1 \quad (2)$$

As equações 1 e 2 podem ser combinadas:

$$y_i(\langle w.x_i \rangle + b) - 1\ \geq 0\ \ \forall\ i \quad (3)$$

Onde w representa o vetor normal do hiperplano e b representa o viés (*bias*). O hiperplano de separação ideal pode ser obtido resolvendo o problema de otimização:

$$Min\frac{1}{2}w^T.w$$

$$sujeito\ a\ \ y_i(\langle w.x_i\rangle + b) - 1\ \geq 0\ \ \forall\, i \tag{4}$$

Utilizando o Lagrangiano (L_P), podemos encontrar o ponto na qual a declividade é nula (ponto de sela):

$$L_P(w,b,\alpha) = \frac{1}{2}w^T.w - \sum_{i=1}^{m}\alpha_i(y_i(\langle w.x_i\rangle + b) - 1) \tag{5}$$

Onde os multiplicadores de Lagrange são $\alpha_i \geq 0, sendo\ i = 1,2 \ldots m$. Utilizando as condições de Karush-Kuhn-Tucker (KKT), pode-se transformar a função de Lagrange L_P em uma função dupla Lagrangiana L_D.

$$Max\ L_D(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j\ \langle x_i.x_j\rangle \tag{6}$$

$$\text{Sujeito a: } \alpha_i \geq 0,\ \ i = 1,2,\ldots m\ \ e\ \sum_{i=1}^{m}\alpha_i y_i = 0$$

Por fim, para encontrar o hiperplano ótimo, a função dupla Lagrangiana deve ser maximizada em relação a α_i não negativo. O resultado determina os parâmetros w^* e b^* do hiperplano ótimo, dado pela equação:

$$f(x,w^*,b^*) = \sum_{i=1}^{m} y_i\alpha_i^*\ \langle w.x_i\rangle + b^* \tag{7}$$

Por diversas vezes os dados podem não ser claramente separáveis e alguns exemplos podem acabar sendo classificados na classe errada. Com o objetivo de diminuir o erro de classificações erradas é incluído as variáveis de folga ($\xi \geq 0$), e o parâmetro C, que representa o peso em fazer uma classificação errada. Este hiperparâmetro pode ser informado ao algoritmo de forma que penalize mais ou menos os erros do classificador (GUPTA, 2012; MARCELINO et al., 2015). Portanto as equações de 1 a 7 podem ser reescritas como:

$$\langle w.x_i\rangle + b \geq +1 - \xi_i\ para\ y_i = +1 \tag{8}$$

$$\langle w.x_i\rangle + b \leq -1 + \xi_i\ para\ y_i = -1 \tag{9}$$

Visando diminuir o erro $\sum_{i=1}^{m}\xi_i$ e inserir o parâmetro C, a função objetivo pode ser reescrita como:

$$Min_{w,b,\xi}\frac{1}{2}w^T.w + C\sum_{i-1}^{m}\xi_i$$

$$Sujeito\ a: y_i(\langle w.x_i\rangle + b) + \xi_i - 1 \geq 0, \quad \xi_i \geq 0, \quad i = 1, \ldots m \tag{10}$$

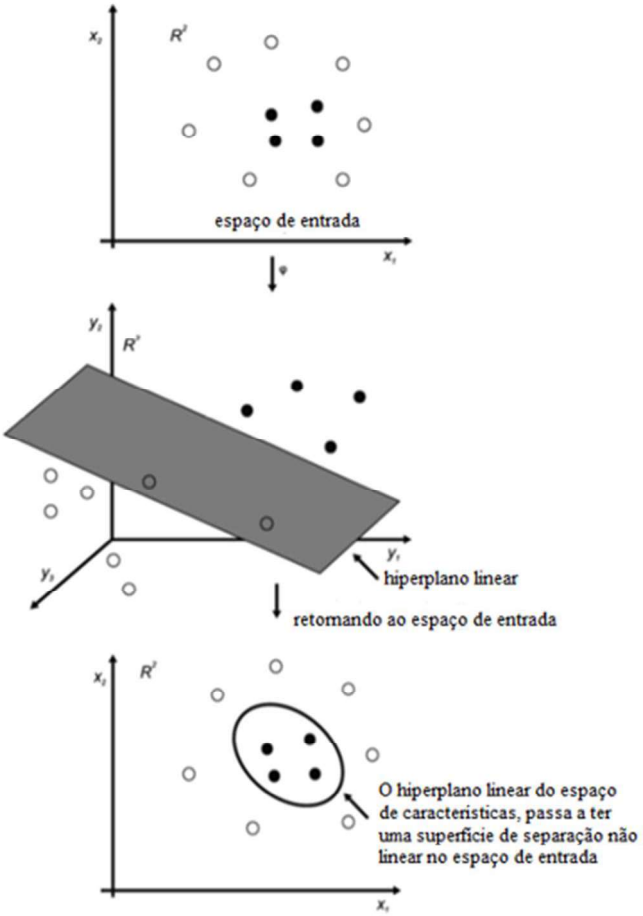
Maximizando a função Lagrangiana L_D, temos:

$$Max\ L_D(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \langle x_i.x_i\rangle$$

$$\text{Sujeito a: } 0 \leq \alpha_i \leq C, \ \ i = 1,2, \ldots m \ \ e\ \sum_{i=1}^{m}\alpha_i y_i = 0 \tag{11}$$

Para problemas em que os dados não possam ser linearmente separados, o algoritmo SVM mapeia o vetor de entrada para um espaço de maior dimensão. Este processo é conhecido como função kernel ou função de mapeamento. O processo de aumento de dimensionalidade pode ser observado na figura 4.

Figura 4 - Função de mapeamento



Fonte: Adaptado de Soman et al. (2009).

Em uma função dupla Lagrangiana L_D, os produtos internos são substituídos pela função kernel, portanto temos:

$$\langle \phi(x_i).\phi(x_j) \rangle = K(x_i, x_j) \quad (12)`$$

A função dupla Lagrangiana L_D, é obtida como:

$$Max\ L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j\ K(x_i, x_j)$$

$$\text{Sujeiro a: } 0 \leq \alpha_i \leq C,\ \ i = 1,2, \dots m\ \ e\ \sum_{i=1}^{m} \alpha_i y_i = 0 \quad (13)$$

Por fim, o hiperplano ideal pode ser obtido utilizando o método de otimização apresentado abaixo:

$$f(x, \alpha^*, b^*) = \sum_{i=1}^{m} y_i \alpha_i^* \langle \phi(x_i).\phi(x) \rangle + b^* = \sum_{i=1}^{m} y_i \alpha_i^* K(x_i, x) + b^* \quad (14)$$

Conforme seção 2.4 deste trabalho, diversos estudos apontam que as máquinas de vetores de suporte conseguem obter bons resultados na previsão de rentabilidade e do comportamento das ações, sendo amplamente utilizado no âmbito das finanças (EMIR et al., 2012; FAN e PALANISWAMI, 2001; HUERTA et al., 2013; KIM 2003; MARCELINO et al., 2015; NTI et al., 2020b; ZHANG e ZHAO, 2009).

2.4 APLICAÇÃO DO SUPPORT VECTOR MACHINE EM FINANÇAS

Um dos autores pioneiros na utilização de maquinas de vetores de suporte para seleção de ações foram Fan e Palaniswami (2001). O trabalho procurou testar a eficácia do SVM na seleção de portfólio de ações da bolsa de valores australiana, durante o período de 1992 ao ano 2000. Para isto, 25% da amostra com o melhor retorno foram classificados com retornos extraordinários, que assumiram o valor +1 e o restante das empresas foram classificados como retornos ordinários (75% da amostra), na qual assumiram o valor de -1. O modelo foi treinado com indicadores financeiros de sete categorias: retorno de capital, lucratividade, alavancagem, crescimento, liquidez de curto prazo, retorno do investimento e risco. Os resultados mostraram a utilidade do SVM para o problema de seleção de portfólio. O portfólio selecionado pelo modelo obteve um retorno de 208%, enquanto o benchmark produziu um retorno acumulado no mesmo período de 71%.

Já Kim (2003) utilizou o SVM para previsão de séries temporais financeiras, comparando os resultados com os modelos de raciocínio baseado em casos e redes neurais *back-propagation.*

O estudo utilizou 12 indicadores provenientes da análise técnica e procurou prever a direção das variações diárias no Korea Composite Stock Price Index (KOSPI). Para isto, quando o índice do próximo dia era menor do que o atual, foi atribuído o valor 0, já quando o índice aumentava de valor, era atribuído o valor 1. O período da amostra foi de Janeiro de 1989 a Dezembro 1998, totalizando 2928 observações (dias). Foram utilizados oitenta porcento dos dados para treinamento a seleção foi feita pelo método *hold-out*.

Os resultados demonstraram que o SVM acertou 57,83% das previsões da direção das variações diárias do índice KOSPI, valor superior a acurácia apresentada pelos outros métodos, tendo a rede neural *back-propragation* uma acurácia de 54,73% e o algoritmo raciocínio baseado em casos uma acurácia de 51,97%. O estudo ainda aponta que o SVM utiliza o princípio da minimização de risco estrutural, ocasionando uma melhor generalização dados temporais do que técnicas tradicionais.

Zhang e Zhao (2009) utilizaram seis indicadores provenientes da análise técnica e o SVM para prever as oscilações no preço do câmbio (euro/dólar) do mercado de ações da China. Para isto foi atribuído a classe +1 quando o preço do câmbio subia, e -1 quando o preço descia. Os dados foram coletados de Julho de 2007 a Julho de 2009, deste foram excluídos as 60 primeiras observações para cálculo da média móvel e os dias onde não houveram alterações nas cotações, totalizando 463 observações. Os resultados mostraram uma alta taxa de precisão na previsibilidade, chegando próximo de 69%. Os autores concluíram que o modelo SVM consegue fazer boas previsões considerando a complexidade do mercado financeiro.

O estudo de Emir et al. (2012) procurou construir um modelo financeiro ideal seleção de portfólios utilizando redes neurais artificiais (RNA) e máquinas de vetor de suporte (SVM).

Para isto, os autores combinaram parâmetros da análise fundamentalista e da análise técnica, durante o período de 2002 a 2010, para as empresas da Bolsa de Valores de Istambul (ISE 30). No final de cada ano, as empresas com melhores taxas de retorno foram classificadas com a classe 1, já as demais com a classe 0.

Visando selecionar as variáveis que realmente são essenciais ao modelo, foi realizado o procedimento de redução de dimensionalidade na fase de pré-processamento dos dados. O procedimento resultou na seleção de 26 indicadores, sendo 14 provenientes da análise fundamentalista e 12 indicadores técnicos.

O modelo RNA utilizou 70% dos dados para treinamento, 10% para validação e 20% para teste. Já para o SVM, a base de dados foi dividida em 75% para treino e 25% para teste. Foram utilizados os kernels linear, polinomial, sigmoidal e *radial based function* (RBF).

Os resultados demonstraram que o modelo RNA apresentou uma acurácia média de 50,35%, já com o SVM a acurácia média foi de 66,19%, durante os anos de 2004 a 2010. A acurácia do modelo SVM também apresentou menor variabilidade (acurácia mínima de 63,33% e máxima de 70%), se comparado com o modelo RNA (acurácia mínima 32,5% e máxima de 67,5%), indicando que a máquina de suporte de vetor parece ser o modelo mais ideal para seleção de portfólios.

Os autores Huerta et al. (2013) procuraram testar se o SVM seria capaz de auxiliar na previsão do preço das ações, utilizando como input as informações contábeis e o histórico de cotações.

Para isto, em vez de utilizarem toda a amostra, foi utilizado uma parcela inicial dos dados. Os autores selecionaram 20% das ações com maior e 20% das ações com pior retorno para treinar o modelo, durante o período de 1981 a 2010. Segundo os autores, esta parcela dos dados é suficiente para o algoritmo estabelecer as relações entre as características das ações e qual classe pertence, ocasionando também a diminuição do custo computacional. A análise dos dados foi realizada considerando também o setor de atuação da empresa. Segundo Huerta et al. (2013), esta estratégia é importe para controlar o impacto econômico sobre o setor de atuação.

Os resultados do estudo evidenciaram que o retorno anual do portfólio nos oitos setores excedeu 15% (desconsiderando os custos de transação), já a volatilidade ficou abaixo dos 8%.

Em um trabalho realizado no Brasil, Marcelino et al. (2015) verificaram se o SVM possuí eficácia para a seleção de portfólios em mercados emergentes. Este estudo utilizou 2 amostras, formadas por ações da BOVA11 (amostra 1) e BRAX11 (amostra 2), durante o período de 20 e 19 trimestres, respectivamente. Foram utilizados 15 indicadores financeiros como variáveis independentes.

Novamente as ações foram classificadas em duas classes, de acordo com sua rentabilidade. A classe 1 foi composta por 25% das ações com maiores retornos (y_i=+1), já as demais foram classificadas na classe 2 (y_i=-1). Para análise dos dados foi utilizado o kernel Gausiano, conforme recomendado pela literatura (HUERTA et al., 2013; EMIR et al., 2012; KIM, 2003).

Os autores usaram o indicador da porcentagem de classificação incorreta das ações da classe 1, para verificar a acurácia do modelo. O algoritmo SVM foi usado para selecionar o

portfólio considerando cada amostra separadamente. Os resultados foram comparados com o *benchmarks* (lucratividade das carteiras BOVA11 e BRAX11, CDI).

O portfólio formado pelo modelo SVM para a amostra 1 retornou uma rentabilidade acumulada nos 20 trimestres de 94,15%. O resultado é consideravelmente superior ao retorno da carteira BOVA11, que é de -14,42% e do retorno do CDI 57,10%. Já o portfólio formado através da amostra 2 teve a rentabilidade de 38,25%. Novamente o resultado do modelo é superior ao obtido pela carteira da BRAX11, na qual computou um retorno de 13,86% para o mesmo período, no entanto, foi inferior ao retorno do CDI (53,75%).

Em uma abordagem mais recente, Nti et al. (2020b) testaram a utilização da Máquina de Vetores de Suporte aprimorada com o Algoritmo Genético (GASVM) para seleção de *features*, otimização de parâmetros e previsão do preço das ações da bolsa de valores de Gana. Posteriormente, os autores compararam a acurácia do modelo GASVM com as técnicas clássicas de aprendizado de máquina.

Os resultados demonstraram que o modelo SVM apresentou a maior acurácia entre os modelos estudados, acertando 93,7% das predições, contra 80,1% das redes neurais, 75,3% das árvores de decisão e 82,3% do *random forest*.

3 METODOLOGIA

3.1 FONTE DE DADOS, AMOSTRA E INDENTIFICAÇÃO DAS VARIÁVEIS

Visando prever o comportamento de ações negociadas no mercado brasileiro, foram utilizadas variáveis técnicas e fundamentalistas. Os dados foram obtidos na base de dados Economática, durante o período de 18 de agosto de 1994 a 16 de dezembro de 2021, com periodicidade diária.

A população deste estudo compreendeu todas as ações da bolsa de valores B3 (antiga BM&F BOVESPA). Foram excluídas as ações que apresentavam dados faltantes ou incompletos durante o período de análise, resultando em 40.562 observações pertencentes a seis empresas, conforme tabela 1.

Tabela 1 - Amostra do estudo

Ação	Tipo	Empresa	Setor
CMIG4	PN	Companhia Energética de Minas Gerais	Energia E Saneamento
LAME4	PN	Lojas Americanas	Varejo
PETR4	PN	Petróleo Brasileiro S.A.	Petróleo E Gás
UNIP6	PN	Unipar Carbocloro	Petroquímico
VALE3	ON	Vale S.A.	Mineração
VIVT3	ON	Telefônica Brasil	Telecomunicações

Fonte: Elaboração Própria

3.1.1 Variáveis Independentes – Análise Fundamentalista

As variáveis independentes provenientes das análises fundamentalistas foram calculadas com as informações das demonstrações financeiras (Balanço Patrimonial e Demonstração do Resultado do Exercício) e do histórico de emissão de ações, todos obtidos na base de dados Economática. Abaixo são apresentadas o indicador/sigla, classes de fatores de acordo com Haugen e Baker (1995) e a fórmula utilizada para o cálculo.

Tabela 2 - Variáveis da análise fundamentalista

Indicador - Sigla	Classes de Fatores (Haugen e Baker, 1995)	Fórmula	
Debt-to-equity ratio - D/E	Risco	$D/E_t = \frac{Passivo\ Circulante_t + Passivo\ não\ circulante_t}{Patrimônio\ Líquido_t}$	(15)
Cobertura de Juros - TIE	Risco	$TIE_t = \frac{EBITDA_t}{Despesas\ Financeiras_t}$	(16)
Market Captalization - MCAP	Liquidez	$MCAP_t = Núm\ ações\ Outstanding_t \times Preço\ Fechamento_t$	(17)
Trading Volume - TV	Liquidez	$TV_t = \frac{Volume_t}{Market\ Captalization_t}$	(18)
Earning-to-Price - E/P	Nível de Preço	$E/P_t = \frac{EBITDA_t}{Peço\ Fechamento_t}$	(19)
Sales-to-Price - S/P	Nível de Preço	$S/P_t = \frac{Receita\ Operacional_t}{Preço\ Fechamento_t}$	(20)
Price-to-Book - P/B	Nível de Preço	$\frac{P}{B_t} = \frac{Market\ Captalization_t}{Patrimônio\ Líquido_t}$	(21)
Margem de Lucro Bruto - GPM	Potencial de Crescimento	$GPM_t = \frac{EBITDA_t}{Receita\ Operacional_t}$	(22)
Capital Turnover - CT	Potencial de Crescimento	$CT_t = \frac{Receita\ Operacional_t}{Ativo\ Total_t}$	(23)
Return on Assets - RoA	Potencial de Crescimento	$RoA_t = \frac{EBITDA_t}{Ativo\ Total_t}$	(24)
Return on Equity - RoE	Potencial de Crescimento	$RoE_t = \frac{EBITDA_t}{Patrimônio\ Líquido_t}$	(25)

Fonte: Elaboração Própria

3.1.2 Variáveis Independentes – Análise Técnica

As variáveis independentes provenientes das análises técnica foram calculadas com as cotações diárias das ações, para isto foi utilizado o preço de fechamento ajustado a splits e dividendos e o volume negociado diariamente. Os dados também foram obtidos na base de dados Economática. Abaixo são apresentadas o indicador/sigla, classes de fatores de acordo com Haugen e Baker (1995) e a fórmula utilizada para o cálculo.

Tabela 3 - Variáveis da análise técnica

Indicador - Sigla	Classes de Fatores (Haugen e Baker, 1995)	Fórmula	
Média Móvel Simples – SMA * Calculado para 3, 7, 15, e 30 dias	Histórico de Preço	$MMS_t = \frac{P_{t-1} + P_{t-2} + \cdots + P_n}{n}$	(26)
Média Móvel Exponencial – EMA *Calculado para 5, 9, 12, 21, 26	Histórico de Preço	$MME_t = P_{t-1}\,K + MME_{t-1} \times (1-K)$ Sendo $K = \frac{2}{n+1}$	(27)
Força Relativa – FR	Histórico de Preço	$FR = \frac{U}{D}$ Sendo: U média das cotações dos últimos 14 dias em que o preço da cotação subiu e D a média das cotações dos últimos 14 dias em que o preço da ação desceu.	(28)
Índice de Força Relativa 14 dias - RSI	Histórico de Preço	$IFR = 100 - \frac{100}{1+FR}$	(29)
Dummies RSI	Histórico de Preço	*Dummies* indicando RSI>70 ou RSI<30	
Média Móvel Convergente e Divergente - MACD	Histórico de Preço	$MACD_t = EMA12_t - EMA26_t$	(30)
Dummy Sinal Tendência Alta MACD	Histórico de Preço	*Dummy* indicando se o $MACD_t > EMA9_t$	
On Balance Volume – OBV	Histórico de Preço	$OBV_t = OBV_{t-1} + \begin{cases} volume_t\,, se\ P_{t-1} > P_{t-2} \\ 0, se\ P_t = P_{t-1} \\ -volume, se\ P_{t-1} < P_{t-2} \end{cases}$	(31)
OBV/Volume	Histórico de Preço	$OBV/Volume_t = \frac{OBV_t}{Volume_t}$	(32)

Fonte: Elaboração Própria

3.1.3 Variável Dependente

Este estudo utilizou o comportamento do preço das ações como variável depende dos modelos de classificação e a retorno das ações para o modelo de regressão.

Para o cálculo do comportamento do preço das ações foi utilizado dados diários do preço de abertura diminuído do preço de fechamento ajustado a splits e dividendos, podendo assumir dois valores (0, caso o preço da ação diminua e 1, no caso do preço da ação não se altere ou suba).

$$Comportamento\ da\ ação_t = \begin{cases} 1, se\ Preço\ de\ fechamento\ ajustado_t \geq Preço\ de\ abertura_t \\ 0, se\ Preço\ de\ fechamento\ ajustado_t < Preço\ de\ abertura_t \end{cases} \quad (33)$$

Já o retorno da ação foi determinado pela equação abaixo:

$$Retorno\ da\ ação_t = Preço\ de\ fechamento\ ajustado_t - Preço\ de\ abertura_t \quad (34)$$

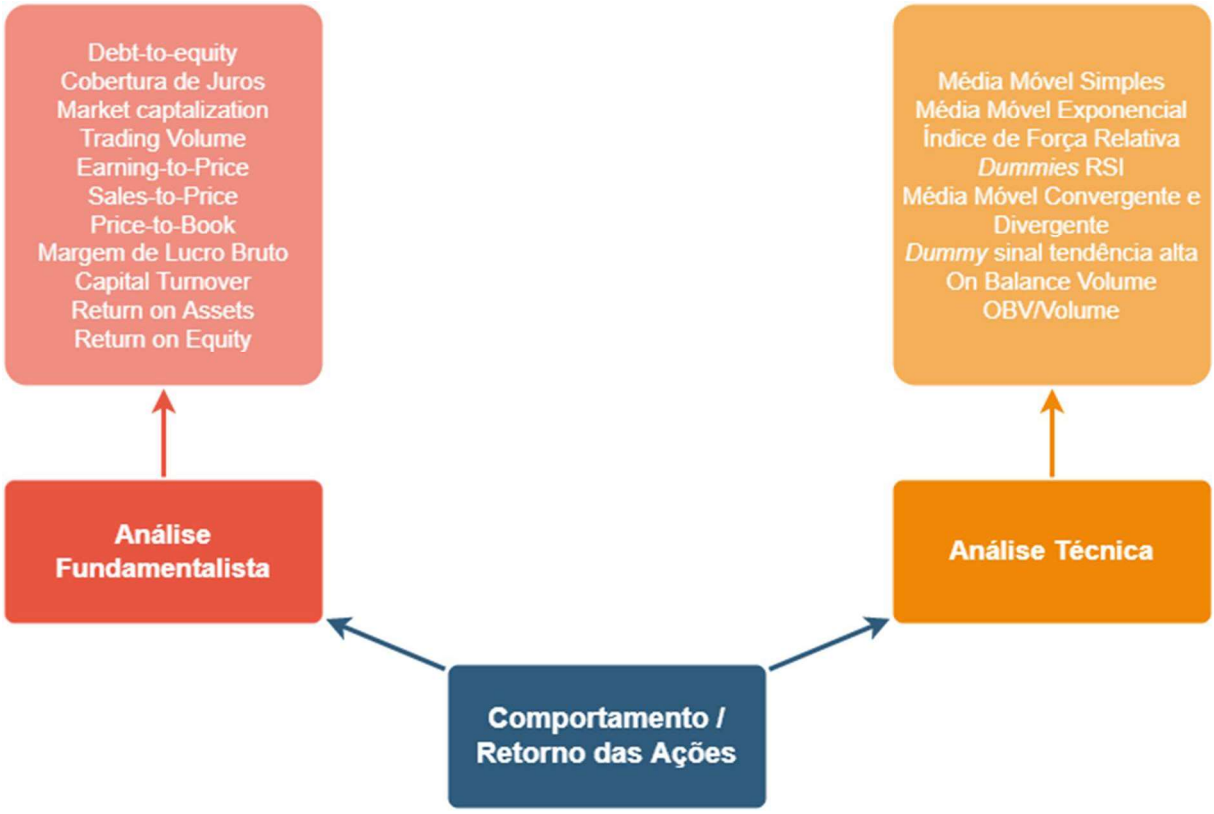
3.2 MODELO EMPÍRICO

O modelo empírico utilizará o algoritmo *support vector machine* na previsão do comportamento e retorno de ações negociadas no mercado brasileiro, utilizando a combinação das análises técnicas e fundamentalistas.

Para definição dos parâmetros ótimos do modelo foi utilizada a ferramenta *gridsearch*, onde foram testados os kernels linear, rbf e sigmoid, parâmetro C com os valores 0.1, 1, 2, 5, 10, 100, 1000, 10000 e 50000 e o parâmetro gamma com os valores 0.001, 0.01, 0.1, 0.5, 1, 1.5, 2, “auto” e “scale”.

Tendo em vista que a variável dependente do modelo de classificação não apresenta desbalanceamento entre as classes, optou-se pela aplicação da validação cruzada com 5 folds. A figura abaixo mostra o esquema do modelo empírico:

Figura 5 - Esquema do modelo empírico



Fonte: Elaboração Própria

3.3 MEDIDAS DE DESEMPENHO DOS MODELOS

Devido ao balanceamento entre as classes da variável dependente do modelo de classificação, o presente estudo utilizará como métrica principal a acurácia e métricas adicionais a precisão (relacionada ao erro tipo I – falsos positivos), revocação (erro tipo II – falsos negativos) e F1 Score, que podem ser calculadas conforme equações abaixo:

$$Acurácia = \frac{Verdadeiro\ Positivo + Verdadeiro\ negativo}{Verdadeiro\ Positivo + Verd \qquad Negativo + Falso\ Positivo + Fals \quad Negativo} \quad (35)$$

$$F1\ score\ = 2 \times \frac{Precisão \times Revocação}{Precisão + Revocação} \quad (36)$$

$$Precisão = \frac{Verdadeiro\ Positivo}{Verdadeiro\ Positivo + Fals \quad Positivo} \quad (37)$$

$$Revocação\ (recall) = \frac{Verdadeiro\ Positivo}{Verdadeiro\ Positivo + Fals \quad Negativo} \quad (38)$$

4 DESENVOLVIMENTO

Nesta seção será apresentada os procedimentos de pré-processamento dos dados e o resultado do modelo empírico.

4.1 PRÉ-PROCESSAMENTO DOS DADOS

Na etapa de pré-processamento foi realizada a limpeza e compilação dos dados extraídos da base de dados. Para isto foram utilizados os softwares Excel e Python, onde realizou-se as seguintes etapas:

1 – Eliminação dos dados faltantes ou incompletos;

2 – Separação da amostra em conjunto de treinamento do algoritmo (80%) e teste (20%);

3 - Os dados de treinamento foram normalizados utilizando o método *StandardScaler*, no qual transforma a distribuição dos dados de forma que a média fique igual a 0 e o desvio padrão igual a 1, conforme equação abaixo:

$$z = \frac{(x-u)}{s} \tag{39}$$

Onde u representa a média da amostra de treinamento e s o desvio padrão da mesma amostra.

4.2 RESULTADO DOS MODELOS EMPÍRICOS

Nesta subseção apresentamos os resultados das quatro aplicações do algoritmo *support vector machine*. Para facilitar o entendimento, um resumo das informações dos modelos foi elaborado na tabela abaixo:

Tabela 4 - Resumo dos Modelos Empíricos

Modelo	Tarefa	Variável Dependente	Variável Independente
A	Classificação	Comportamento das ações	Variáveis Fundamentalistas
B	Classificação	Comportamento das ações	Variáveis Técnicas
C	Classificação	Comportamento das ações	Fundamentalistas e Técnicas
D	Regressão	Retorno das ações	Fundamentalistas e Técnicas

Fonte: Elaboração Própria

4.2.1 Modelo A - Treinamento com Variáveis Fundamentalistas

A primeira aplicação utilizou somente as variáveis fundamentalistas para treinamento do modelo, são elas: debt-to-equity ratio, cobertura de juros, market captalization, trading volume, earning-to-price ratio, sales-to-price, price-to-book, margem de lucro bruto, capital turnover, return on assets e return on equity.

Os parâmetros ótimos definidos pelo *gridsearch* foram, kernel linear, C=1000 e gamma = "auto". Os resultados do modelo foram: acurácia média 56,46%, medida F1 score igual a 68,95%, precisão média de 59,07% e revocação média de 82,78%

4.2.2 Modelo B - Treinamento com Variáveis Técnicas

O segundo modelo utilizou somente as variáveis fundamentalistas para treinamento do algoritmo, são elas: média móvel simples, média móvel exponencial, índice de força relativa 14 dias, *dummies* RSI, média móvel convergente e divergente, *dummy* sinal tendência alta, *on balance volume* e obv/volume.

Os parâmetros ótimos definidos pelo *gridsearch* foram, kernel linear, C=1 e gamma = "auto". Este modelo apresentou os seguintes resultados: acurácia média 55,72%, medida F1 score igual a 67,98%, precisão média de 58,83% e revocação média de 80,50%

4.2.3 Modelo C - Treinado com Variáveis Técnicas e Fundamentalistas

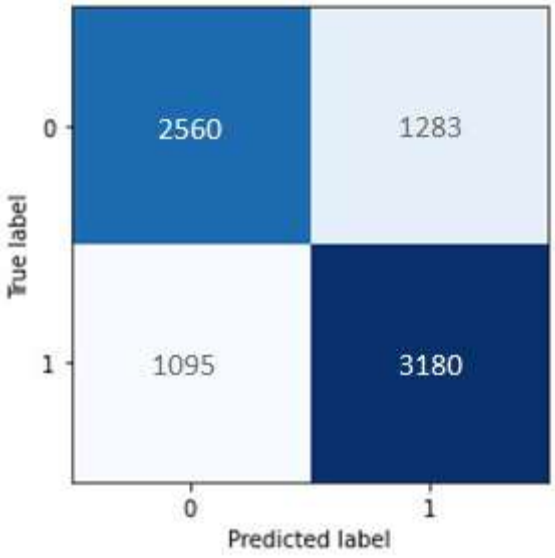
O modelo empírico aplicou o algoritmo *support vector machine* na previsão do comportamento de ações do mercado brasileiro, utilizando a combinação da análise técnica e análise fundamentalista.

Para definição dos parâmetros ótimos utilizou-se a técnica *gridsearch*, onde foram testados os kernels linear, rbf e sigmoid, parâmetro C com os valores 0.1, 1, 2, 5, 10, 100, 1000, 10000 e 50000 e o parâmetro gamma com os valores 0.001, 0.01, 0.1, 0.5, 1, 1.5, 2, "auto" e "scale".

O modelo apresentou como parâmetros ótimos: kernel *radial based function - rbf*, C=10000, gamma=1. Estes parâmetros foram utilizados para treinar um classificador utilizando validação cruzada com 5 folds.

O modelo de classificação apresentou a média de acurácia nos folds de 70,7%, medida F1 score igual a 72,74%, precisão média de 71,25% e revocação média de 74,33%. Abaixo apresentamos a matriz de confusão do modelo de classificação:

Figura 6 - Matriz de confusão do modelo de classificação



Fonte: Elaboração própria

Dentre as 40.562 observações que constituem a amostra, foi utilizado 80% para treinamento do algoritmo e 20% para teste. Podemos observar que dos 8.118 dados de teste, o modelo conseguiu classificar corretamente 5.760, isto é, o algoritmo *support vector machine* treinado com variáveis técnicas e fundamentalistas, conseguiu prever se o preço da ação iria subir, se manter ou descer em 70% das vezes.

4.2.4 Modelo D - Regressão

A última aplicação do algoritmo foi realizada buscando prever o retorno diário da ação em termos quantitativos. Portanto este modelo além de prever a direção (aumento ou diminuição) do preço das ações, busca calcular a variação diária dos preços das ações.

Primeiramente o modelo foi aplicado em toda a base de dados (6 empresas, 40.562 observações), no entanto o algoritmo não conseguiu aprender com os dados e o poder de explicação ficou próximo a zero.

Optou-se então por separar os dados por empresas e testá-las individualmente. Este trabalho apresentará o resultado das duas empresas onde houve capacidade de generalização do algoritmo:

Tabela 5 - Resultado Modelo Regressão

Empresa	R^2	Erro Absoluto Médio	Erro Quadrático Médio	Hiperparâmetros
Petróleo Brasileiro S.A. (PETR4)	0,663	0.1045	0.0334	C=15000, kernel rbf
Vale S.A. (VALE3)	0,790	0.1425	0.0681	C=20000, kernel rbf

Fonte: Elaboração própria

4.3 DISCUSSÃO

Relativamente ao modelo de classificação, foi possível verificar que o treinamento com variáveis provenientes somente da análise fundamentalista (modelo A) teve um resultado ligeiramente superior se comparado com o modelo que utilizou somente variáveis da análise técnica (modelo B). Este resultado foi na contramão do que é indicado na literatura. Oliveira et al. (2013) e Renu e Christie, (2018) destacam que a análise fundamentalista é ideal para investimentos a médio e longo prazo e que a análise técnica conseguiria captar melhor as alterações de curto prazo, como é o caso das variações diárias do preço das ações, objeto deste estudo.

Cabe destacar que o desempenho dos dois primeiros modelos ficou próximo ao aleatório, indicando que o algoritmo conseguiu aprender pouco com os dados e generalizar pouco os resultados.

No entanto, quando combinamos os dados das análises técnica e fundamentalista (modelo C), ganha-se poder de predição e capacidade de generalização dos resultados, atingindo uma acurácia média de 70,7%, medida F1 score igual a 72,74%, precisão média de 71,25% e revocação média de 74,33%. Este resultado vem ao encontro da revisão sistemática de Nti et al. (2020a), na qual aponta que modelos híbridos podem fornecer uma melhor acurácia na previsão das ações.

Por fim, o modelo D procurou prever o retorno diário do preço das ações. Acredita-se que devido a heterogeneidade dos dados, o algoritmo não conseguiu aprender com os dados, apresentando um R^2 próximo a zero.

No entanto, quando dividiu-se a amostra por empresas, notou-se uma melhora no poder preditivo de algumas amostras, apresentando um R^2 próximo a 80% para a empresa Vale s.a. e 66% para a Petrobras. Tendo em vista a grande diferença no poder de predição dos modelos de regressão, recomenda-se cautela na utilização e generalização destes resultados.

Portanto, verificamos que o modelo C, treinado para a tarefa de classificação, com variáveis advindas da combinação das análises técnica e fundamentalista apresenta a melhor acurácia, precisão, revocação e medida F1 score, podendo auxiliar o processo de seleção de portfólio dos investidores.

5 CONCLUSÃO

Este trabalho tem como objetivo testar ferramentas que possam auxiliar o processo de seleção de portfólio dos investidores, treinando quatro modelos de aprendizado de máquina que buscam prever o comportamento de ações negociadas no mercado brasileiro.

Para isto foram treinados quatro modelos de aprendizado de máquina, para as tarefas de classificação (previsão do comportamento do preço da ação) e regressão (previsão do retorno da ação). A amostra compreendeu seis empresas, durante o período de 27 anos (de 18 de agosto de 1994 à 16 de dezembro de 2021), totalizando 40.562 observações.

Os modelos treinados somente com variáveis fundamentalistas (A) e somente com variáveis técnicas (B) apresentaram acurácia baixa, respectivamente 56,46% e 55,72%. Conforme indicado pela literatura esperava-se que o modelo com variáveis provenientes da análise técnica produzisse melhores resultados, tendo em vista que esta análise consegue captar melhor as alterações de curto prazo, como é o caso das variações diárias do preço das ações, objeto deste estudo (OLIVEIRA et al., 2013; RENU e CHRISTIE, 2018).

Foi possível notar uma melhora significativa no desempenho do modelo treinado com a combinação das variáveis provenientes das análises fundamentalistas e técnicas (C). Este resultado vem ao encontro da literatura. Em sua revisão sistemática, Nti et al. (2020) aponta que modelos híbridos podem fornecer uma melhor acurácia na previsão das ações.

O último modelo (D) procurou aplicar a tarefa de regressão visando prever o retorno das ações. Os resultados não foram satisfatórios para a amostra contendo as seis empresas juntas. Acredita-se que a heterogeneidade dos dados pode ter dificultado o aprendizado do algoritmo.

No entanto, quando o algoritmo foi aplicado para cada empresa individualmente notou-se uma melhora no poder preditivo de algumas amostras. Destacou-se a empresa Vale S.A. na qual apresentou um R^2 próximo de 79% e a Petróleo Brasileiro S.A. com um R^2 de 66%.

Concluiu-se que o modelo C apresentou a melhor acurácia, precisão, revocação e medida F1 score, portanto, o algoritmo *support vector machine* treinado para a tarefa de classificação, com variáveis advindas da combinação das análises técnica e fundamentalista pode auxiliar o processo de seleção de portfólio dos investidores.

O presente trabalho apresenta outros pontos que se destacam na literatura de previsão do mercado de ações, como é o caso do grande período de análise dos dados (27 anos), na qual auxilia no treinamento e generalização do modelo e, também, a utilização de técnicas combinadas para a previsão do comportamento do mercado de ações. De acordo com Nti et al. (2020a) apenas 11% dos 144 estudos da sua revisão sistemática utilizam a combinação entre as análises técnicas e fundamentalistas, gerando, via de regra, resultados melhores.

Este trabalho limita-se a utilização da base de dados Economática. Portanto, foram analisadas somente empresas que possuíam os dados completos no banco de dados e variáveis que pudessem serem calculadas com os dados disponíveis na base de dados.

Como recomendações para trabalhos futuros sugere-se a ampliação das variáveis provenientes da análise fundamentalistas, testando além das variáveis advindas das demonstrações contábeis das empresas, outros dados que possam influenciar o preço da ação, como por exemplo a análise de sentimento das notícias vinculadas na mídia e nas redes sociais. Também recomenda-se a ampliação e diversificação da amostra e a utilização de outros algoritmos de aprendizado de máquina, como por exemplo as redes neurais e os métodos ensembles.

REFERÊNCIAS

ASSAF NETO, Alexandre. **Mercado financeiro**. 2001.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 10520**: informação e documentação: citações em documentos: apresentação. Rio de Janeiro, 2002.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 6024**: informação e documentação: numeração progressiva das seções de um documento escrito: apresentação. Rio de Janeiro, 2012.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14724**: informação e documentação: trabalhos acadêmicos: apresentação. Rio de Janeiro, 2011.

BALLINGS, Michel et al. Evaluating multiple classifiers for stock price direction prediction. **Expert systems with Applications**, v. 42, n. 20, p. 7046-7056, 2015.

BELTRAMI, Monica; LOCH, Gustavo Valentim; SILVA, A. Comparação das técnicas de support vector regression e redes neurais na precificação de opções. **XLLII Sbpo**, p. 572-583, 2011.

BIANCHI, Daniele; BÜCHNER, Matthias; TAMONI, Andrea. Bond risk premiums with machine learning. **The Review of Financial Studies**, v. 34, n. 2, p. 1046-1089, 2021.

BOLSA DE VALORES DE SÃO PAULO. Segmentos de listagem: novo mercado. **Bovespa**, São Paulo, 08 de out. 2020. Disponível em <http://www.b3.com.br/pt_br/produtos-e-servicos/solucoes-para-emissores/segmentos-de-listagem/novo-mercado/?tabIndex=0>. Acesso em 08 de out. 2020.

BOSER, Bernhard E.; GUYON, Isabelle M.; VAPNIK, Vladimir N. A training algorithm for optimal margin classifiers. In: **Proceedings of the fifth annual workshop on Computational learning theory**. 1992. p. 144-152.

BRITO, Osias Santana. **Controladoria de risco: retorno em instituições financeiras**. Saraiva, 2003.

BRASIL, BOLSA, BALCÃO. A descoberta da bolsa pelo investidor brasileiro: Quem são e como se comportam as mais de 2 milhões de pessoas que aplicaram parte de seus recursos em bolsa no último ano. **B3**, São Paulo, 19 de ago. 2021. Disponível em <http://www.b3.com.br/pt_br/noticias/investidores.htm>. Acesso em 19 de ago. 2021.

BUCKMANN, Marcus; JOSEPH, Andreas; ROBERTSON, Helena. **An interpretable machine learning workflow with an application to economic forecasting**. mimeo, 2021.

CAVALCANTE, Francisco; MISUMI, Jorge Yoshio; RUDGE, Luiz Fernando**. Mercado de capitais: o que é, como funciona**. Elsevier, 2005.

CARVALHO, Marcelino F. Uma contribuição ao estudo da controladoria em instituições financeiras organizadas sob a forma de múltiplo banco. **São Paulo: FEA/USP**, 1995.

CHEN, Jeffrey C. et al. Off to the races: A comparison of machine learning and alternative data for predicting economic indicators. In: **Big Data for 21st Century Economic Statistics**. University of Chicago Press, 2019.

CHENHALL, Robert H.; MOERS, Frank. The role of innovation in the evolution of management accounting and its integration into management control. **Accounting, organizations and society**, v. 47, p. 1-13, 2015.

DRUCKER, Harris et al. Support vector regression machines. In: **Advances in neural information processing systems**. 1997. p. 155-161.

ELDER, Alexander. **Trading for a living: psychology, trading tactics, money management**. John Wiley & Sons, 1993.

EMIR, Senol; DINÇER, Hasan; TIMOR, Mehpare. A stock selection model based on fundamental and technical analysis variables by using artificial neural networks and support vector machines. **Review of Economics & Finance**, v. 2, n. 3, p. 106-122, 2012.

FAN, Alan; PALANISWAMI, Marimuthu. Stock selection using support vector machines. In: **IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)**. IEEE, 2001. p. 1793-1798.

FEIJÓ, Carmem; ARAÚJO, Eliane Cristina; BRESSER-PEREIRA, Luiz Carlos. Política monetária no Brasil em tempos de pandemia. **Brazilian Journal of Political Economy**, v. 42, p. 150-171, 2022.

GÉRON, Aurélien. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. Alta Books, 2019.

GUPTA, Pankaj; MEHLAWAT, Mukesh Kumar; MITTAL, Garima. Asset portfolio optimization using support vector machines and real-coded genetic algorithm. **Journal of Global Optimization**, v. 53, n. 2, p. 297-315, 2012.

HAIXIANG, Guo et al. Learning from class-imbalanced data: Review of methods and applications. **Expert Systems with Applications**, v. 73, p. 220-239, 2017.

HAUGEN, Robert A.; BAKER, Nardin L. Commonality in the determinants of expected stock returns. **Journal of financial economics**, v. 41, n. 3, p. 401-439, 1996.

HE, Haibo; GARCIA, Edwardo A. Learning from imbalanced data**. IEEE Transactions on knowledge and data engineering**, v. 21, n. 9, p. 1263-1284, 2009.

HUERTA, Ramon; CORBACHO, Fernando; ELKAN, Charles. Nonlinear support vector machines can systematically identify stocks with high and low future returns. **Algorithmic Finance**, v. 2, n. 1, p. 45-58, 2013.

KIM, Kyoung-jae. Financial time series forecasting using support vector machines. **Neurocomputing**, v. 55, n. 1-2, p. 307-319, 2003.

LIMA, Fernando Barros de et al. A controladoria em Instituições Financeiras: Estudo de caso no Banco do Nordeste do Brasil SA. **Contabilidade Vista & Revista**, v. 22, n. 1, p. 43-72, 2011.

LIU, Ling et al. A social-media-based approach to predicting stock comovement. **Expert Systems with Applications**, v. 42, n. 8, p. 3893-3901, 2015.

MARCELINO, Sarah; HENRIQUE, Pedro Alexandre; ALBUQUERQUE, Pedro Henrique Melo. PORTFOLIO SELECTION WITH SUPPORT VECTOR MACHINES IN LOW ECONOMIC PERSPECTIVES IN EMERGING MARKETS. **Economic Computation & Economic Cybernetics Studies & Research**, v. 49, n. 4, 2015.

MARCELINO, Sarah Sabino de Freitas. **Formação de portfólio por meio de máquinas de suporte vetorial e redes de camadas profundas**. Dissertação de Mestrado, Universidade de Brasília, Brasília, DF, Brasil. 2016.

MARKOWITZ, Harry. Portfolio selection. **Investment under Uncertainty**, 1959.

MOSIMANN, C. P.; FISCH, S. **Controladoria: seu papel na administração das empresas**. 2. ed. São Paulo: Atlas. 1999.

NEVES JÚNIOR, Idalberto José; DA COSTA, Letícia; MOURÃO, Silva. Controladoria em tempos de pandemia: reflexões e contribuições para a indústria financeira. **Brazilian Journal of Development**, v. 7, n. 7, p. 71760-71780, 2021.

NTI, Isaac Kofi; ADEKOYA, Adebayo Felix; WEYORI, Benjamin Asubam. A systematic review of fundamental and technical analysis of stock market predictions**. Artificial Intelligence Review**, v. 53, n. 4, p. 3007-3057, 2020a.

NTI, Isaac Kofi; ADEKOYA, Adebayo Felix; WEYORI, Benjamin Asubam. Efficient stock-market prediction using ensemble support vector machine. **Open Computer Science**, v. 10, n. 1, p. 153-163, 2020b.

MUN, Wendy Chong Pooi; SOONG, Varian. Forecasting Yield Curve with Macro-Driven Models: A Comparison Between Machine Learning and Traditional Statistical Approaches. **RPubs**, Disponível em < https://rpubs.com/WendyChongPooiMun/YieldCurve>. Acesso em 25 de fev. 2022.

OLIVEIRA, Fagner A.; NOBRE, Cristiane N.; ZÁRATE, Luis E. Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index–Case study of PETR4, Petrobras, **Brazil. Expert systems with applications**, v. 40, n. 18, p. 7596-7606, 2013.

RENU, Isidore R.; CHRISTIE, P. Fundamental Analysis versus Technical Analysis–a Comparative Review. **International Journal of Recent Scientific Research**, v. 9, n. 1, p. 23009-23013, 2018.

SAMUEL, Arthur L. Some studies in machine learning using the game of checkers. **IBM Journal of research and development**, v. 3, n. 3, p. 210-229, 1959.

SOMAN, K. P.; LOGANATHAN, R.; AJAY, V. **Machine learning with SVM and other kernel methods**. PHI Learning Pvt. Ltd., 2009.

SHATSHAT, M. A. H. I.; AHMED, Sohail. Information Technology Governance Linkage to the Financial Report Quality in Libyan Commercial Banks. 2019.

WIDEMAN, R. Max. **Project and program risk management: a guide to managing project risks and opportunities.** 1992. Tese de Doutorado. Univerza v Mariboru, Ekonomsko-poslovna fakulteta

ZHANG, Zuoquan; ZHAO, Qin. The application of SVMs method on exchange rates fluctuation. **Discrete Dynamics in Nature and Society**, v. 2009, 2009.

ZHANG, Xi et al. Improving stock market prediction via heterogeneous information fusion. **Knowledge-Based Systems**, v. 143, p. 236-247, 2018.