

**UNIVERSIDADE PRESBITERIANA MACKENZIE
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA ELÉTRICA E COMPUTAÇÃO**

ROBERTO BRUNO LEMES MARRETTI

**SIMULAÇÃO DE NEGOCIAÇÕES EM INSTRUMENTOS DO
ÍNDICE IBOVESPA UTILIZANDO *MACHINE LEARNING*.**

São Paulo
2019

**UNIVERSIDADE PRESBITERIANA MACKENZIE
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA ELÉTRICA E COMPUTAÇÃO**

Roberto Bruno Lemes Marretti

**Simulação de negociações em instrumentos do índice
IBOVESPA utilizando *machine learning*.**

Dissertação de Mestrado apresentada ao
Programa de Pós-Graduação em
Engenharia Elétrica e Computação da
Universidade Presbiteriana Mackenzie como
parte dos requisitos para obtenção do título de
Mestre em Engenharia Elétrica e Computação.

Orientador: Prof. Dr. Nizam Omar

São Paulo
2019

M358s Marretti, Roberto Bruno Lemes
Simulação de negociações em instrumentos do índice IBOVESPA utilizando machine learning / Roberto Bruno Lemes Marretti – São Paulo, 2019.

113 f.: il., 30 cm.

Mestrado (Mestrado em Engenharia Elétrica e Computação) - Universidade Presbiteriana Mackenzie - São Paulo, 2019.

Orientador: Prof. Dr. Nizam Omar

Bibliografia: f. 88-95

1. Bolsa de Valores 2. Mercado de Capitais 3. Inteligência Artificial 4. Aprendizado de Máquina 5. Análise Técnica I. Omar, Nizam, orientador. II.Título.

CDD 005

Bibliotecária Responsável: Maria Gabriela Brandi Teixeira – CRB 8/ 6339

ROBERTO BRUNO LEMES MARRETTI

SIMULAÇÃO DE NEGOCIAÇÕES EM INSTRUMENTOS DO ÍNDICE
IBOVESPA UTILIZANDO *MACHINE LEARNING*

Dissertação de Mestrado apresentada
ao Programa de Pós-Graduação em
Engenharia Elétrica e Computação da
Universidade Presbiteriana Mackenzie,
como requisito parcial para a obtenção
do título de Mestre em Engenharia
Elétrica e Computação.

Orientador: Prof. Dr. Nizam Omar

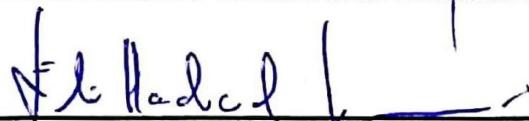
Aprovado em 20 de agosto de 2019.

BANCA EXAMINADORA



Prof. Dr. Nizam Omar

Universidade Presbiteriana Mackenzie



Prof. Dr. Eli Hadad Júnior

Universidade Presbiteriana Mackenzie



Prof. Dr. Paulo André Lima de Castro

Instituto Tecnológico de Aeronáutica - ITA

Dedico este trabalho primeiramente a Deus que me concebeu saúde e força em todos os momentos, e às pessoas responsáveis por contribuir e incentivar na minha formação pessoal e acadêmica.

AGRADECIMENTOS

Ao Programa de Pós-Graduação em Engenharia Elétrica e Computação (PPGEEC) da Universidade Presbiteriana Mackenzie, por todo o suporte para a realização desta pesquisa.

A todos os professores do PPGEEC que compartilharam seus conhecimentos comigo.

Ao meu orientador Prof. Dr. Nizam Omar, por todo o apoio, amizade, auxílio, paciência e incentivo à continuação dos meus estudos.

Aos Professores Eli Hadad e Paulo Castro, cujo os apontamentos foram fundamentais para a melhora deste trabalho como um todo.

Agradeço a todos colegas de turma com quem convivi a experiência de uma produção compartilhada de conhecimento.

Aos meus colegas do Banco Bradesco, agradeço pela confiança, em especial a Thays Andrade, pela constante compreensão e apoio.

Aos colegas de sala de aula, em especial à Cristiano Benites e Bruno Cézar, amizades adquiridas para uma vida toda.

A todos os meus familiares e amigos que me apoiaram.

A minha namorada, por ter caminhado ao meu lado, pela sua paciência e compreensão, especialmente por sempre me apresentar um sorriso, quando sacrificava os dias, noites, fins de semana e feriados em prol da realização deste trabalho.

E, sobretudo, agradeço a Deus, por me permitir e levar a termo este trabalho, e pelas pessoas que Ele colocou em minha vida para que isto fosse possível.

A riqueza de uma nação se mede pela riqueza do povo e não pela riqueza dos príncipes (Adam Smith).

RESUMO

A previsão e compreensão do mercado de capitais é uma tarefa naturalmente desafiadora devido a complexidade e abrangência de variáveis do mercado financeiro. Analistas e investidores utilizam sistemas de *software* com finalidade de ajudá-los na tomada de decisão operacional e estratégica na negociação de instrumentos financeiros com o auxílio da análise técnica, análise fundamentalista e de modelos matemáticos que permitem especificar parâmetros, períodos e regras, possibilitando observar o comportamento de estratégias de negociação, baseando-se em dados de instrumentos financeiros como o histórico de flutuação de preço, volume, volatilidade, *etc.* Estudos recentes sugerem a utilização de técnicas de inteligência artificial aliado aos estudos da análise técnica, fundamentalista e de modelos estatísticos para classificar e identificar instrumentos financeiros com potenciais oportunidades de investimento. Neste aspecto, o presente trabalho tem por objetivo realizar a implementação de métodos de aprendizado de máquina e inferência de resultados em ações do índice IBOVESPA do mercado de ações brasileiro com o uso de dados obtidos da B3. Os resultados obtidos demonstram que a utilização e combinação de diferentes heurísticas computacionais fornecem resultados confiáveis, enfatizando a aplicabilidade de técnicas de inteligência artificial sobre as hipóteses de investimentos tradicionais.

Palavras-chave: *Bolsa de Valores, Mercado de Capitais, Inteligência Artificial, Aprendizado de Máquina, Análise Técnica.*

ABSTRACT

The prediction and understanding of the capital market is a naturally challenging task due to the complexity and breadth of financial market variables. Analysts and investors use software systems to assist them in making strategic and operational decisions in the trading of financial instruments with the help of technical analysis, fundamental analysis and mathematical models that allow the specification of parameters, periods and rules, making it possible to observe the behavior of trading strategies, based on data from the financial instruments such as the historical fluctuation of price, volume, volatility, etc. Recent studies suggest the use of artificial intelligence techniques combined with studies of technical analysis, fundamentalist and statistical models to classify and identify financial instruments with potential investment opportunities. In this aspect, the objective of this work is to implement machine learning methods and inference of results in shares of the IBOVESPA index of the Brazilian stock market with the use of data obtained from B3. The results show that the use and combination of different computational heuristics provides reliable results, emphasizing the applicability of artificial intelligence techniques to traditional investment hypotheses.

Keywords: *Stock Market, Capital Market, Artificial Intelligence, Machine Learning, Technical Analysis.*

LISTA DE ABREVIATURAS E SIGLAS

ADF	<i>Augmented Dickey–Fuller</i>
ARCH	<i>Autoregressive Conditional Heteroskedasticity</i>
ARFIMA	<i>Autoregressive Fractionally Integrated Moving Average</i>
ARIMA	<i>Autoregressive Integrated Moving Average</i>
ARMA	<i>Autoregressive Moving Average Models</i>
AT	Análise Técnica
ATS	<i>Automated Trading System</i>
CPU	<i>Central Processing Unit</i>
GARCH	<i>Generalized AutoRegressive Conditional Heteroskedasticity</i>
GPU	<i>Graphics Processing Unit</i>
HFT	<i>High Frequency Trading</i>
IF	Instrumento Financeiro
IFR	Índice de Força Relativa
IID	<i>Independent and Identically Distributed</i>
KN	<i>k-Nearest Neighbors</i>
KPSS	<i>Kwiat-Phillips-Smith-Shin</i>
MER	Modelo Entidade Relacionamento
ML	<i>Machine Learning</i>
NB	<i>Naive Bayes</i>
PLN	Processamento de Linguagem Natural

PP	<i>Phillips Perron</i>
RF	<i>Random Forests</i>
RL	Rregressão Logística
RLI	Rregressão Linear
RNA	Redes Neurais Artificiais
SARIMA	<i>Seasonal Autoregressive Integrated Moving Average</i>
SARMA	<i>Seasonal Autoregressive Moving Average</i>
SGBD	Sistema de Gerenciamento de Banco de Dados
SW	<i>Shapiro-Wilk</i>
SVM	<i>Support Vector Machine</i>

LISTA DE FIGURAS

Figura 1	Evolução Tecnológica das Bolsas de Valores	4
Figura 2	Gráfico de barra de <i>Candlestick</i>	10
Figura 3	Gráfico demonstrativo de presença de tendência	12
Figura 4	Gráfico de Sinais Bandas de Bollinger	17
Figura 5	Gráfico de Sinais Índice Força Relativa	18
Figura 6	Gráfico de Sinais do Média Móvel Exponencial	19
Figura 7	Gráfico de Sinais Cruzamento Médias Móveis Aritmética	21
Figura 8	Linha do tempo da Inteligência Artificial	30
Figura 9	Frequência de pesquisa do termo <i>machine learning</i> no google.com .	31
Figura 10	Esquema de treinamento e divisão para 5-fold <i>Cross-Validation</i> .	35
Figura 11	Modelo Entidade Relacionamento	40
Figura 12	Exemplo de agrupamento de dados utilizando o método <i>Dollar Bars</i>	43
Figura 13	Frequência média diária de <i>tick</i> , <i>volume</i> e <i>dollar</i> do mini contrato S&P 500	44
Figura 14	Curvas de distribuição normal dos instrumentos financeiros selecionados	47
Figura 15	Exemplo gráfico <i>boxplot</i> de <i>Outliers</i>	50
Figura 16	Exemplo do <i>book</i> de negociações dos intrumentos VALE3 e GGBR4	51
Figura 17	Verificação de <i>outliers</i> em dados de experimento	52
Figura 18	Série temporal com <i>CUSUM Filter</i>	53
Figura 19	<i>Triple-Barrier Method</i>	55
Figura 20	Gráfico do fluxo para concepção do modelo primário	57
Figura 21	Visualização do fluxo para concepção do modelo secundário	58

Figura 22	Ilustração de construção do modelo <i>Random Forests</i>	59
Figura 23	Ilustração da divisão da base de treinamento e teste	62
Figura 24	Resultado de <i>Backtest</i> para VALE3	66
Figura 25	Visualização de retornos para VALE3	67
Figura 26	Importância de atributos do modelo para VALE3	67
Figura 27	Curva ROC para VALE3 .	68
Figura 28	Visualização de desempenho do modelo de <i>Machine Learning</i> (ML) para VALE3 .	69
Figura 29	Resultado de <i>Backtest</i> para PETR4	70
Figura 30	Visualização de retornos para PETR4	71
Figura 31	Importância de atributos do modelo para PETR4	71
Figura 32	Curva ROC para PETR4 .	72
Figura 33	Visualização de desempenho do modelo de ML para PETR4 . . .	73
Figura 34	Resultado de <i>Backtest</i> para ITUB4	74
Figura 35	Visualização de retornos para ITUB4	75
Figura 36	Importância de atributos do modelo para ITUB4	75
Figura 37	Curva ROC para ITUB4 .	76
Figura 38	Visualização de desempenho do modelo de ML para ITUB4 . . .	77
Figura 39	Resultado de <i>Backtest</i> para MULT3	78
Figura 40	Visualização de retornos para MULT3	79
Figura 41	Importância de atributos do modelo para MULT3	79
Figura 42	Curva ROC para MULT3 .	80
Figura 43	Visualização de desempenho do modelo de ML para MULT3 . . .	81
Figura 44	Resultado de <i>Backtest</i> para VVAR3	82
Figura 45	Visualização de retornos para VVAR3	83

Figura 46	Importância de atributos do modelo para VVAR3	83
Figura 47	Curva ROC para VVAR3	84
Figura 48	Visualização de desempenho do modelo de ML para VVAR3	85

LISTA DE TABELAS

Tabela 1	Definição de parâmetros Bandas de Bollinger	16
Tabela 2	Definição de parâmetros do Índice de força relativa	17
Tabela 3	Definição de parâmetros de médias móveis	19
Tabela 4	Definição de parâmetros de cruzamento de médias móveis	20
Tabela 5	Ações selecionadas do índice IBOVESPA para os experimentos realizados	41
Tabela 6	Frequência de dados	45
Tabela 7	Resultado do teste estatístico de <i>Shapiro-Wilk</i>	49
Tabela 8	Parâmetros utilizados no método <i>Triple-Barrier Method</i>	56
Tabela 9	Atributos utilizados como <i>inputs</i> do modelo secundário	60
Tabela 10	Parâmetros utilizados no método <i>Triple-Barrier Method</i>	65
Tabela 11	Parâmetros utilizados no método <i>Triple-Barrier Method</i>	69
Tabela 12	Parâmetros utilizados no método <i>Triple-Barrier Method</i>	73
Tabela 13	Parâmetros utilizados no método <i>Triple-Barrier Method</i>	77
Tabela 14	Parâmetros utilizados no método <i>Triple-Barrier Method</i>	81

SUMÁRIO

1	Introdução	1
1.1	Objetivo do Capítulo	1
1.2	Análise de investimento de ações do mercado financeiro	2
1.3	Justificativa	3
1.4	Motivação	4
1.5	Objetivo Geral	5
1.6	Objetivos Específicos	5
2	Trabalhos Relacionados	7
3	Referencial Teórico	9
3.1	Análise de Investimentos Clássica	9
3.1.1	Análise Técnica	11
3.2	Estratégias Operacionais	14
3.2.1	Bandas de Bollinger	16
3.2.2	Índice de Força Relativa	17
3.2.3	Médias Móveis	18
3.2.4	Cruzamento de Médias Móveis	20
3.2.5	Prazos Operacionais	21
3.3	Métricas de Risco	22
3.3.1	<i>Sharpe Ratio</i>	22
3.3.2	Volatilidade Histórica	23
3.3.3	Máximo <i>Drawdown</i>	23
3.4	Séries Temporais Financeiras	24

3.5	<i>Automated Trading System</i>	26
3.6	Finanças Quantitativas	28
3.7	Inteligência Artificial	29
3.7.1	<i>Machine Learning</i>	31
3.7.2	Tipos de Dados	33
3.7.3	<i>Overfitting</i>	34
4	Metodologia	36
4.1	Objetivo do Capítulo	36
4.2	Implementação Técnica	36
4.3	Preparação dos Dados	37
4.3.1	Base de Dados	37
4.3.2	Frequência dos Dados	41
4.4	Análise Estatística	45
4.4.1	<i>Shapiro-Wilk</i>	47
4.4.2	Detecção de <i>Outliers</i>	49
4.5	<i>CUSUM Filter</i>	53
4.6	<i>Meta-Labeling</i>	54
4.7	<i>Triple-Barrier Method</i>	55
4.8	Algoritmos de Aprendizagem	57
4.8.1	<i>Random Forests</i>	59
4.8.2	Métricas de Desempenho	60
5	Resultados	63
5.1	Ambiente de Simulação e Testes	63
5.2	Avaliação dos Resultados	65

5.2.1	Bandas de Bollinger - Reversão à Média	65
6	Considerações Finais	86
6.1	Trabalhos Futuros	87
	Referências Bibliográficas	88

1 Introdução

1.1 Objetivo do Capítulo

O objetivo deste capítulo é apresentar as justificativas e os motivos que levaram ao desenvolvimento deste trabalho. Também são apresentados os objetivos específicos e a estrutura desta dissertação que possui uma série de tópicos interconectados e os apresenta de uma forma ordenada, pressupondo a leitura do capítulo anterior para compreensão do tema.

- Capítulo 1 - Este capítulo é dedicado a fornecer uma breve introdução sobre o tema, apresentar as justificativas, motivações e objetivos que levaram ao desenvolvimento deste trabalho;
- Capítulo 2 - O segundo capítulo tem como objetivo apresentar os trabalhos relacionados na área;
- Capítulo 3 - O terceiro capítulo apresenta o referencial teórico da pesquisa, e cada capítulo a seguir fornece maiores detalhes sobre cada seção abordada neste capítulo;
- Capítulo 4 - Este capítulo apresenta a metodologia, os detalhes da implementação técnica, discussões e resultados da implementação. Este capítulo também aborda o motivo da escolha do método proposto e a validade da solução encontrada durante a pesquisa;
- Capítulo 5 - O quinto capítulo apresenta os resultados finais das abordagens utilizadas e gráficos comparativos com a estratégia *buy-and-hold*, comparado com o próprio instrumento financeiro;
- Capítulo 6 - Este capítulo final apresenta um breve resumo dos resultados, observações finais e aborda estudos futuros potencialmente relevantes.

1.2 Análise de investimento de ações do mercado financeiro

As teorias clássicas de finanças, em sua maioria, acreditam que o mercado é formado sobre a Hipótese do Mercado Eficiente (EMH, do inglês) elaborado por Fama (1970) defende que os preços dos instrumentos financeiros são definidos conforme notícias e informações são impulsionadas e incorporadas ao mercado financeiro, isto é, o valor de um instrumento reflete toda informação disponível no mercado, logo o preço será sempre justo e equivalente ao seu valor fundamental. Ainda segundo Fama (1970), os preços são imprevisíveis.

A hipótese de Fama (1970) é bastante desafiadora pois o advento e a massificação dos computadores permite a utilização de novas metodologias para a análise de instrumentos do mercado de capitais, com objetivo de obter melhor relação de eficiência do risco envolvido se comparado com estratégias de investimento tradicionais. Para Simões (2010) isto é devido à capacidade de sistemas de *software* em analisar grandes quantidades de dados em um curto período de tempo, ao crescimento do poder computacional e desenvolvimento de ferramentas e sistemas de *software* destinados ao mercado de capitais.

Além disso, Slovic (1972) e Elder (2004) argumentam que fatores psicológicos podem afetar o processo estratégico de tomada de decisão, pois costumam estar pautados em crenças e pela inviabilidade humana de processar tanta informação disponível do mercado financeiro.

Ainda segundo Elder (2004), cada pessoa possui um padrão único de características de personalidade, existindo uma consistência de raciocínio que perdurará numa identificação e organização de traços psicológicos que interagem entre si, o que significa que nem a análise fundamentalista e nem a análise técnica podem ser utilizadas para prever o comportamento de séries temporais, como exemplo disto, podemos tomar os vieses que causam erros na tomada de decisão, colocando em dúvida a hipótese econômica dos mercados eficientes (MILANEZ, 2003).

Conforme o estudo de Alves (2015), tem havido atenção e desenvolvimento da academia e do mercado em geral na área de inteligência artificial, análise sentimental e *performance* de técnicas existentes para predição em instrumentos financeiros da bolsa de valores.

Segundo (WONG; DU; CHONG, 2005), o uso da inteligência artificial tem vindo a ser cada vez mais estudada e utilizada no mercado de capitais devido à atual capacidade de processamento dos computadores, sendo possível realizar predição em modelos de redes neurais com precisão superior a 70%.

(QIAN; RASHEED, 2007), demonstram ser possível a obtenção de precisão superior a 50% com a combinação de modelos gerados por métodos indutivos de aprendizado de máquina.

Larsen (2007) afirma que um *automated trading system* tem a capacidade de substituir 90% os operadores da bolsa de valores, suprimindo a tomada de decisão operacional com base em crenças.

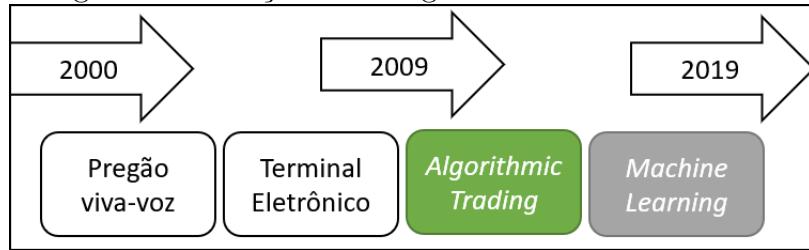
Com base nestas abordagens, pode-se perceber que a utilização de novas técnicas para análise e predição em instrumentos financeiros são relevantes e abrange a interdisciplinaridade das diferentes áreas de conhecimento como Computação, Finanças, Estatística e Psicologia.

O presente trabalho consiste em estudo empírico de estratégias de investimentos com base em modelos indutivos de aprendizagem de máquina em diversas condições do mercado, os resultados podem ser utilizados para auxílio à tomada de decisão operacional e estratégica de compra e venda de instrumentos financeiros derivativos ou do mercado de ações, entretanto o escopo deste trabalho foi delimitado ao mercado de ações.

1.3 Justificativa

A publicação realizada pela revista de tecnologia do *MIT Technology Review*, Byrnes (2017) representa a evolução tecnológica do mercado de capitais e demonstra que indústria vem automatizando cada vez mais os processos operacionais, em busca de eficiência e redução de custos, e consequentemente as funções ocupadas por cargos operacionais estão deixando de existir, sendo cada vez mais requisitados as áreas de computação, matemática e engenharias.

Figura 1: Evolução Tecnológica das Bolsas de Valores



Fonte: Elaborado pelo autor.

Conforme o estudo da Coherent (2019), uma fração (acima de 80%) de todos os negócios realizados na NYSE¹ são efetuados por processos operacionais altamente informatizadas (geralmente comandados por grandes *hedge funds*) e que empregam análises extremamente sofisticadas, processadas em seus sistemas e computadores de última geração, agindo no mercado financeiro em milissegundos ou mesmo em microssegundos com base na análise de algoritmos.

Ainda segundo a pesquisa da Coherent (2019), estima-se que 60% à 73% todas as operações realizadas na NYSE² são realizadas por sistemas automatizados e por sistemas inteligentes.

Portanto, a evolução do mercado financeiro, justifica utilização de técnicas de *machine learning* neste trabalho.

1.4 Motivação

Larsen (2007) afirma que o mercado de ações é operado principalmente por ganância e medo, quando um deles se torna predominante, a racionalidade humana é deixada de lado, como por exemplo bolhas que historicamente foram formadas na bolsa de valores de diversos países.

A utilização de técnicas quantitativas utilizando ML tornam-se importantes aliados na tomada de decisão operacional e estratégica, uma vez que não há fatores emocionais e pessoais envolvidos, e consequentemente, por utilizar algoritmos computacionais, possui

¹NYSE: New York Stock Exchange - Bolsa de Valores de New York - <https://www.nyse.com>

²NYSE: New York Stock Exchange - Bolsa de Valores de New York - <https://www.nyse.com>

alto desempenho de processamento de informações se comparado à capacidade humana.

Com o surgimento de robôs investidores, de sistemas de tomada de decisão de forma autônoma, e com a crescente técnicas operacionais de *trades* de curto período de tempo, como por exemplo a técnica de *High Frequency Trading* (HFT), surge a demanda de agilidade e robustez no processo de tomada de decisões e também no processamento de informações do mercado financeiro de forma rápida e eficaz.

Algumas estratégias de negociação estenderam suas pesquisas para além dos dados de mercado, para incluir dados do setor de consumo, imagens de satélite e até mesmo postagens de mídia social. Futuros sistemas operados por essas organizações certamente serão mais sofisticados do que os que estão sendo utilizados hoje (PRADO, 2018).

A maior parte dos estudos disponíveis na literatura aborda os mercados financeiros de países com a economia desenvolvida (ARTHUR, 2018; BLANCHARD et al., 2018), onde o comportamento dos instrumentos financeiros é diferente do observado em mercados emergentes, onde existe maior volatilidade (RAZA et al., 2016).

Existe *déficit* de pesquisa relativo aos mercados emergentes, muito do que já foi abordado está consolidado apenas para os mercados desenvolvidos, ou seja, ainda há muito o que se pesquisar para mercados emergentes.

1.5 Objetivo Geral

O objetivo geral deste trabalho foi apresentar os principais componentes utilizados para a análise de instrumentos financeiros da Bolsa de valores de São Paulo utilizando métodos de *machine learning* por meio de uma pesquisa tecnológica.

1.6 Objetivos Específicos

Prever o comportamento de instrumentos financeiros da bolsa de valores de São Paulo utilizando métodos de *machine learning* com objetivo de obter estratégias operacionais com retornos acima do *benchmark buy-and-hold* em comparação com o próprio instrumento financeiro.

- Implementação e desenvolvimento de modelos de *machine learning* utilizando técnicas de Análise Técnica (AT);
- Avaliação de métricas de desempenho;
- Automatização do processo de análise de investimentos para auxílio na tomada de decisão operacional e estratégica;
- Comparação de retornos dos modelos ao *benchmark buy-and-hold* em relação ao próprio instrumento financeiro.

2 Trabalhos Relacionados

Alves (2015) realizou um estudo analisando o sentimento das pessoas em redes sociais e de *microblogs* por meio de ferramentas que adotam o uso de técnicas de ML, Processamento de Linguagem Natural (PLN) e uso de indicadores de AT para tomada de decisão de compra ou venda em instrumentos da bolsa de valores de São Paulo. Trata-se de abordagens de coleta de dados, processamento de dados, segmentação, classificação sentimental e análise estatística de uma determinada empresa por meio do uso de dados públicos obtidos da Internet, por meio do uso de simuladores e de algoritmos determinísticos, obteve-se retornos significativos acumulados de 277,86% para o instrumento PETR4 e de 224,28% para o instrumento VALE3 pelo período compreendido de agosto de 2013 à abril de 2015³. A autora concluiu que análise sentimental da multidão utilizado dados de redes sociais como o Twitter⁴ foi possível obter retornos acima de técnicas convencionais de análise técnica, evidenciando a realação estatística das Finanças Comportamentais (YOSHINAGA et al., 2008), com possibilidades de aprofundamento de estudos futuros para o mercado brasileiro.

Pimenta (2017) propôs a utilização de estratégias operacionais envolvendo o uso de programação genética multiobjeto com a combinação de regras de indicadores de análise técnica utilizando abordagens determinísticas para seleção de regras na concepção de um *Automated Trading System* (ATS). Obteve-se retornos significativos acumulados de 77.61% para o instrumento BBAS3 e de 48,50% para o instrumento GGBR4, em comparativo com o *benchmark buy-and-hold* em relação ao próprio instrumento financeiro, obteve-se 0.60% e -44.37% respectivamente, superando o *benchmark buy-and-hold*⁵. O autor também realizou comparativo em outros quatro instrumentos financeiros (BOVA11, CMIG4, EMBR3, VALE5) e em todos os instrumentos financeiros selecionados a abordagem adodata utilizando programação genética multiobjeto superaram o *benchmarch buy-and-hold* durante o período compreendido de fevereiro de 2013 à julho de 2016 da bolsa de valores de São Paulo. O autor verificou que as técnicas abordadas geraram lucros durante o período da análise, também deixou evidente que deve-se levar em con-

³Retorno bruto, não está sendo considerado taxas, emolumentos e outros encargos operacionais.

⁴<https://www.twitter.com>

⁵Retorno bruto, não está sendo considerado taxas, emolumentos e outros encargos operacionais.

sideração os resultados negativos obtidos durante os testes realizados em um período de situação de crise do mercado financeiro.

O trabalho de Giacomel (2016) propõe a utilização de Redes Neurais Artificiais (RNA) em conjunção com técnicas de *ensemble* no mercado de ações brasileiro e norte-americano conforme o perfil do investidor: moderado ou agressivo. O autor também utilizou indicador de análise técnica na composição das redes neurais para o perfil de investimento agressivo, uma observação importante realizada neste trabalho é de que a utilização do indicador de análise técnica não deve ser utilizado separadamente das técnicas propostas, pois não obteve desempenho acima do *benchmark* para nenhuma das estratégias abordadas quando utilizado isoladamente, entretanto a utilização em conjunto com as técnicas propostas foi possível obter resultados satisfatórios para cada estratégia operacional utilizada.

A pesquisa de Szyszka (2017) realizou simulações para o instrumento financeiro VALE3 do mercado brasileiro de ações utilizando o método *Random Forests* (RF) para diferentes intervalos de tempo em séries temporais do *intraday*, os resultados demonstraram que o método pode ser utilizado como ferramenta preditiva dos retornos com grau de significância estatística para intervalos de 15 minutos durante o período estudado.

O estudo de Páscoa (2018) utilizou técnicas de ML para estratégias direcionais no mercado de criptomoedas, especificamente para Bitcoin⁶ e no índice Europeu PSI-20. Obteve resultados com retorno de 69% em um intervalo de 5 anos para o índice PSI-20, utilizando o método de RF, em contrapartida, o índice acumulou prejuízo de -7% no período, ou seja, o método demonstrou retornos significativos acima do *benchmark buy-and-hold*, a simulação também foi realizada para o mercado de criptomoedas para o mesmo intervalo de tempo e o método RF acumulou rentabilidade de 33% no período, enquanto o Bitcoin acumulou rentabilidade de 421% no mesmo período, ou seja, abaixo do *benchmark buy-and-hold*. O autor concluiu que o método se mostrou ineficiente para o mercado de Bitcoin. Vale ressaltar que a dinâmica dos dois instrumentos são completamente diferentes, o índice possui menor volatilidade em comparação ao Bitcoin, conforme a observação do autor, uma análise realizada pela Bloomberg⁷, considerou o Bitcoin como sendo o instrumento mais volátil do mundo.

⁶<https://bitcoin.org>

⁷<https://www.bloomberg.com/graphics/2017-bitcoin-volume/>

3 Referencial Teórico

O referencial teórico traz conceitos de finanças sobre o tema abordado e fundamenta a utilização de técnicas de ML para a análise e simulação séries temporais em instrumentos financeiros do índice IBOVESPA.

3.1 Análise de Investimentos Clássica

Segundo (LEMOS; CARDOSO, 2010) a análise de investimentos é uma atividade realizada por empresas, investidores profissionais e pessoas em geral, seja por busca de lucro do capital disponível ou então apenas por correção monetária, tem por objetivo compreender o comportamento de um instrumento, sendo categorizada em três escolas: Teoria de Dow, Análise Fundamentalista e Análise Técnica (ELDER, 2004).

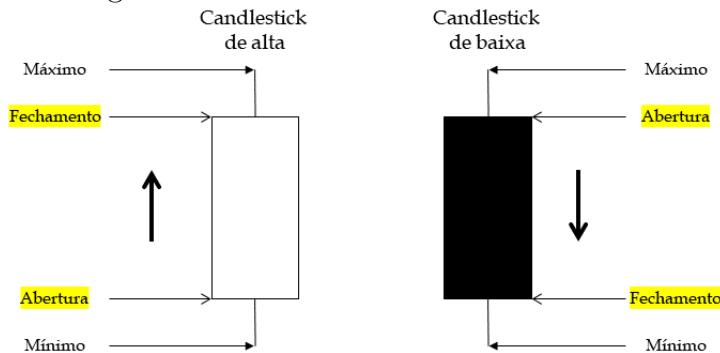
A análise fundamentalista consiste em prever a tendência de uma ação baseado em dados financeiros da empresa, cenários do mercado microeconômico e macroeconômico, setor de atuação com objetivo determinar um *valor justo* do instrumento (LEMOS; CARDOSO, 2010).

Conforme definição de (PINHEIRO, 2016), pode-se determinar a análise fundamentalista como toda informação disponível no mercado com objetivo de antecipar o comportamento e valor futuro do ativo.

A análise técnica teve origem no Japão (NORONHA, 2003), com o gráfico conhecido como candelabro japonês ou *candlestick*⁸, no entanto foi por meio da Teoria de Dow que este campo teórico se difundiu, a análise técnica busca compreender o cenário atual do instrumento com o uso de recursos gráficos com objetivo de antecipar movimentos futuros.

⁸Segundo Lemos e Cardoso (2010) estima-se que os *candlesticks* são utilizados pelos japoneses desde o século XVII.

Figura 2: Gráfico de barra de *Candlestick*



Fonte: Elaborado pelo autor.

As barras de *candlestick* são obtidas por amostragem de dados em intervalos de tempo fixos, por exemplo, uma vez a cada minuto. Os dados obtidos formam uma série de tempo discreto (Seção 3.4) e geralmente incluem:

- Data e horário
- Preço de abertura
- Preço de fechamento
- Preço máximo
- Preço mínimo
- Volume de negócios
- Volume financeiro, *etc.*

De acordo com Chaves (2004), a análise técnica sugere que a tendência de preços pode ser determinada com base no comportamento histórico, isto é, por meio da identificação de padrões recorrentes, seja pelo comportamento dos *candlesticks*, modelos gráficos ou matemáticos, é possível dizer que a ação terá uma tendência de alta ou baixa de preço, pressupondo o melhor momento para comprar ou vender um Instrumento Financeiro (IF).

3.1.1 Análise Técnica

Saffi (2003) descreve que os indicadores de análise técnica utilizam referência o preço histórico, volume, quantidade de negócios, valor mínimo e máximo de uma série, usualmente os indicadores são exibidos na forma de linhas, histograma, *etc*, sobre os gráficos de *candlestick*, tornando os dados graficamente observáveis para um investidor ou analista.

No entanto, não há garantia que um indicador pode gerar bons resultados, além disso, cada pessoa pode ter uma interpretação diferente dependendo por exemplo de sua experiência, interpretação de risco, personalidade, crenças ou estado emocional.

Noronha (2003) afirma que os indicadores técnicos auxiliam no processo de tomada de decisão, ajudando a identificar tendências e cenários de reversão de mercado. (BARBOSA, 2007), afirma que indicadores técnicos são usualmente classificados em duas categorias, sendo: indicadores de tendência e indicadores de oscilação.

Para Matsura (2013) os indicadores são uma confirmação de direcionamento do mercado que podem auxiliar na decisão operacional e estratégica.

Saffi (2003) afirma que os indicadores de tendências mais utilizados no mercado por analistas e investidores são as médias móveis, o índice de força relativa (Relative Strength Index, RSI), indicador William's %R, Estocástico e o indicador de Média Móvel Convergência-Divergência (*Moving Average Convergence-Divergence*).

Baptista e Pereira (2009) concluíram que por meio da análise técnica é possível obter resultados superiores da estratégia *buy-and-hold* no índice IBOVESPA, se for considerado os custos operacionais, visto que estes podem reduzir substancialmente os lucros obtidos.

O mercado é formado por ciclos econômicos assimétricos, analistas buscam por padrões recorrentes e tentam alcançar lucros pela repetição destes padrões com o auxílio de gráficos e indicadores técnicos, assume-se que uma tendência prossegue até se ter indicação do contrário, demonstrando o direcionamento do estado atual do mercado.

Indicadores de Tendência

Os indicadores de tendência são utilizados quando o mercado demonstra algum direcionamento claro, isto é, apenas analisando o gráfico é possível identificar predominância

de alta ou baixa do instrumento financeiro, estes indicadores têm por objetivo sinalizar se uma determinada tendência pode prevalecer durante um período de tempo.

Figura 3: Gráfico demonstrativo de presença de tendência



Fonte: Elaborado pelo autor.

- **Média Móvel:** Lemos e Cardoso (2010) afirmam que a média móvel aritmética é um dos indicadores de tendência mais difundido pela simplicidade do cálculo, tem como principal característica a suavização de ruídos de flutuações de preço, foi concebido pelos operadores de mercado antes da era computadorizada:

$$MMA = \frac{Pf1 + Pf2 + Pf3 + \dots + Pf(n)}{n} \quad (1)$$

Na qual:

Pf é o preço de fechamento no período de 1 a n ;

n é o número de períodos da média móvel aritmética (MMA).

Para Lemos e Cardoso (2010) a média móvel aritmética é utilizada para mostrar o melhor momento para negociar um IF com base no comportamento histórico, ainda segundo Lemos e Cardoso (2010) a principal diferença entre outros indicadores é de que a média móvel não antecipa uma tendência, apenas acompanha o movimento do mercado.

Wong, Du e Chong (2005) aplicaram a análise técnica utilizando o cálculo de médias móveis no mercado de capitais da China, Hong Kong e Taiwan entre o período de

1992 e 2004 e obtiveram retornos acima da estratégia *buy-and-hold* e concluíram sendo um bom indicador para determinar o *timing* certo de compra e venda de um ativo. Para Luiz (2009) as médias móveis não são bons indicadores em cenários de alta volatilidade, pois em séries temporais de adversidade, o indicador pode gerar tendências errôneas, por exemplo, em cenário de crise econômica.

- ***Moving Average Convergence-Divergence:*** O MACD foi concebido em 1979, por Gerald Appel, este indicador é formado por duas linhas de médias móveis exponenciais, conforme afirma Matsura (2013) as linhas das duas médias móveis exponenciais tende a se cruzar em algum momento, quando isso ocorre é formado um de tendência que pode significar de alta ou baixa do preço do instrumento.

$$MACD = MME(C, n) - MME(C, n) \quad (2)$$

Na qual:

C = Preço de fechamento;

MME = Média móvel exponencial de n períodos.

n é o número de períodos

Indicadores de Oscilação

Os indicadores de oscilação são utilizados quando o mercado não demonstra algum direcionamento claro, ou seja, quando não é possível identificar uma tendência, este grupo de indicadores pode por exemplo indicar se o mercado está sobrecomprado ou sobrevenido.

- **Índice de Força Relativa:** Segundo Lemos e Cardoso (2010) o IFR (Índice de Força Relativa) é um indicador que compara a magnitude de qualquer movimento do ativo, em outras palavras é utilizado para demonstrar se um ativo está acumulando ganhos ou perdas, assim podendo desenvolver uma tendência.

$$IFR = 100 - \left(\frac{100}{1 + FR} \right) \quad (3)$$

Na qual:

FR = média aritmética do preço máximo de n períodos dividido pela média aritmética do preço de mínimo de n períodos.

- **Willian's %R:** Lemos e Cardoso (2010) afirmam que o Willian's %R é um indicador de oscilação que também pode confirmar uma tendência secundária de um indicador de tendência, isto é, o investidor pode procurar por pontos de sobrecompra para uma determinada decisão estratégica de compra ou venda do ativo.

$$\%R = 100 \left(\frac{H_p - P_t}{H_p - L_p} \right) \quad (4)$$

Na qual:

H_p = Preço máximo de n períodos.

P_t = Preço de fechamento de n períodos.

L_p = Preço mínimo n períodos.

- **Stochastic Oscillator:** Segundo (LEMOS; CARDOSO, 2010) foi concebido por George C. Lane nos anos 50, é um indicador de oscilação que tem por objetivo medir a amplitude do preço de fechamento de um ativo em relação aos valores máximos e mínimos do período.

$$\%K = C - \left(\frac{L_p(n)}{H_t - L_p(n)} \right) \quad (5)$$

Na qual:

C = Preço de fechamento;

$L_p(n)$ = Preço mínimo de n períodos;

H_t = Preço Máximo de n períodos.

3.2 Estratégias Operacionais

As estratégias operacionais podem ser definidas como regras que são utilizadas para definir entrada/saída de um instrumento financeiro.

- **Arbitragem estatística:** Conforme a definição de (CALDEIRA, 2013) a arbitragem estatística pode ser entendida como uma oportunidade de negócio livre de risco (ou de baixo risco) isto é, quando o mercado apresenta alguma "falha" na especificação do instrumento, geralmente as arbodagens utilizadas tem como base o uso de modelos estatísticos.
- ***Buy-and-hold:*** Fama (1970) afirma ser uma estratégia que visa um objetivo de longo prazo, a cerne deste tipo de operação é comprar o instrumento de uma ou de várias empresas (carteira de investimentos) compartilhando os valores da companhia, acreditando de que esta terá crescimento substancial, aumentando assim o valor dos ativos na bolsa de valores e consequentemente os ativos dos investidores, a análise deste tipo de estratégia também pode utilizar as teorias da análise técnica e fundamentalista.
- **Indicadores de Tendência:** Para Lemos e Cardoso (2010) quando uma tendência é identificada, seja por meio de cálculo matemático ou de padrão gráfico, o investidor define uma estratégia, bem como os riscos envolvidos e o objetivo a ser alcançado, em seguida é realizado uma ordem de compra visando o preço alvo ou então até que seja identificado uma reversão de tendência. Para Lemos e Cardoso (2010) o preço de um instrumento é formado por movimentos cíclicos, e por meio de indicadores de tendência é possível identificar o movimento do mercado baseado nas flutuações do preço histórico e predizer se um determinado comportamento é relevante o suficiente para que alguma estratégia seja definida.
- **Reversão à média:** Parte do pressuposto que os preços tendem a retornar à média histórica, pode ser utilizado como parâmetro de entrada ou saída de um instrumento financeiro.
- **Cruzamento de médias:** Por meio da utilização de duas médias móveis de períodos diferentes, quando há cruzamento, entende-se que pode ser um ponto de entrada/saída.

Abaixo, são apresentados as estratégias, indicadores e parâmetros escolhidos para a concepção do modelo de ML primário.

3.2.1 Bandas de Bollinger

Segundo Lemos e Cardoso (2010) as bandas de bollinger pode ser utilizadas como um indicador de entrada ou saída de tendência, é formado por duas linhas de desvio padrão da média móvel conforme demonstrado na figura 4.

Tabela 1: Definição de parâmetros Bandas de Bollinger

Estratégia	Período	Desvio Padrão
Reversão à média	50	2

Fonte: Elaborado pelo autor.

A regra para geração de sinais de compra e venda está demonstrado no algoritmo 1, quando o menor valor cruzar a linha inferior da banda, é gerado sinal de compra, quando o maior valor cruzar com a linha superior, é gerado sinal de venda conforme demonstrado na figura 4.

Algoritmo 1: Sinal de compra ou venda Banda de Bollinger

Entrada: $high$ maior valor do *candle*, low menor valor do *candle*, $bupper$ cálculo

banda superior, $bbottom$ cálculo banda inferior

Saída: $signal$ Indicadores de sinal: *BUY*, *SELL* ou *NULL* = Neutro

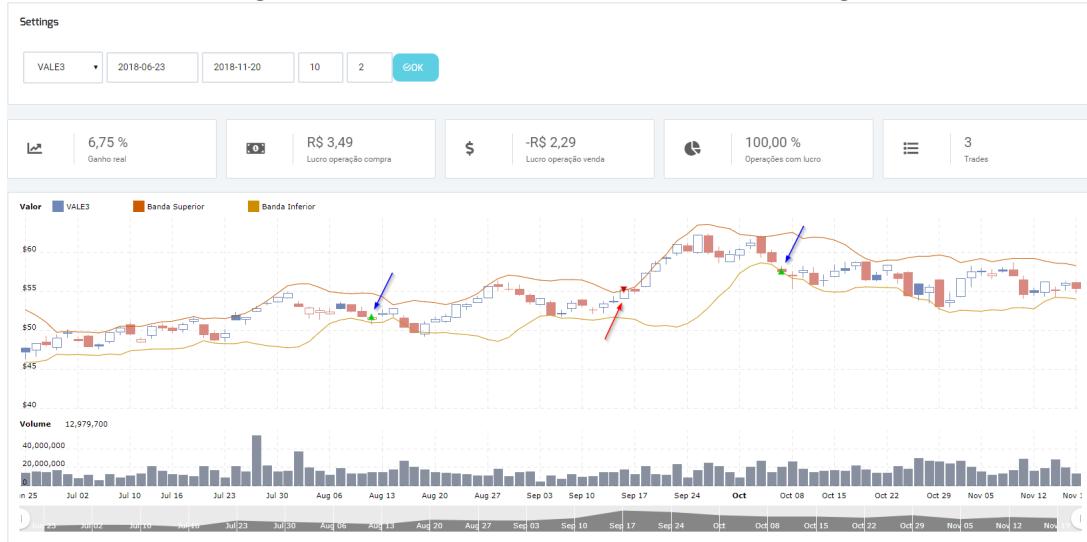
início

```

 $signal \leftarrow NULL;$ 
if  $bbottom \geq low$  then
    |  $signal \leftarrow BUY$ 
    if  $bupper \leq high$  then
        |  $signal \leftarrow SELL$ 
fim
retorna  $signal$ 

```

Figura 4: Gráfico de Sinais Bandas de Bollinger



Fonte: Elaborado pelo autor.

3.2.2 Índice de Força Relativa

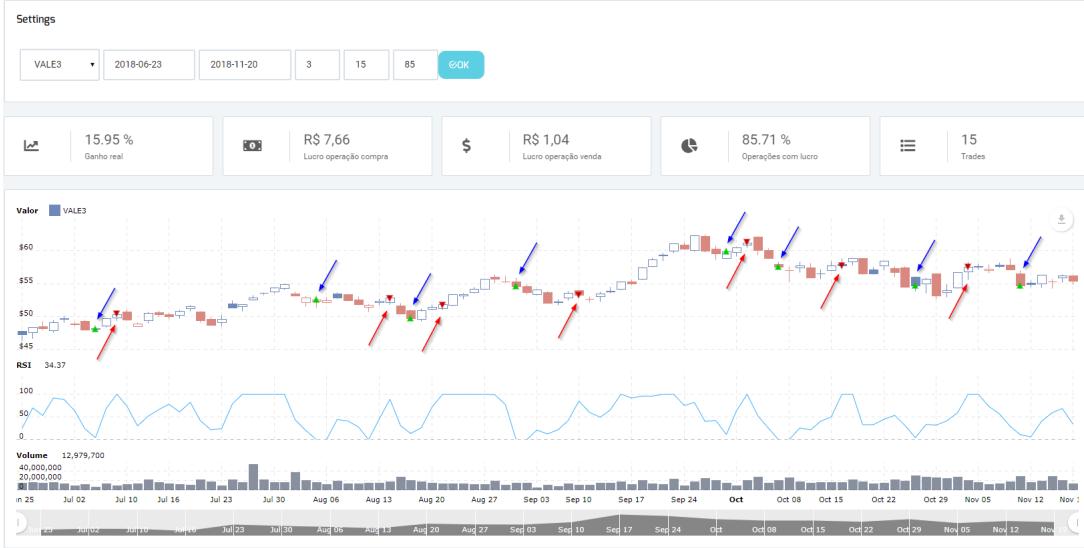
De acordo com Lemos e Cardoso (2010) o Índice de Força Relativa (IFR) é um indicador de oscilação utilizado para medição de força de flutuações de preço entre dois extremos (IFR de 0 e IFR de 100), foi concebido principalmente para identificar uma possível reversão de tendência ou então indicar quando um IF está sobrecomprando ou sobrevendido.

Tabela 2: Definição de parâmetros do Índice de força relativa

Estratégia	IFR superior	IFR inferior	Período
Reversão do IFR	10	90	20

Fonte: Elaborado pelo autor.

Figura 5: Gráfico de Sinais Índice Força Relativa



Fonte: Elaborado pelo autor.

A regra para geração de sinais de negociação utilizando o IFR é demonstrada no algoritmo 2.

Algoritmo 2: Sinal de compra ou venda Índice de Força Relativa

Entrada: $ifrSuperior$ parâmetro IFR superior, $ifrInferior$ parâmetro IFR inferior, ifr cálculo índice força relativa

Saída: $signal$ Indicadores de sinal: BUY , $SELL$ ou $NULL$ = Neutro
início

```

 $signal \leftarrow NULL;$ 
if  $ifr \geq ifrSuperior$  then
    |  $signal \leftarrow BUY$ 
if  $ifr \leq ifrInferior$  then
    |  $signal \leftarrow SELL$ 
fim
retorna  $signal$ 

```

3.2.3 Médias Móveis

Para Elder (2004) as médias móveis acompanham o mercado e mostram a direção que as flutuações do preço tendem a seguir, é utilizada como indicador de tendência, quando

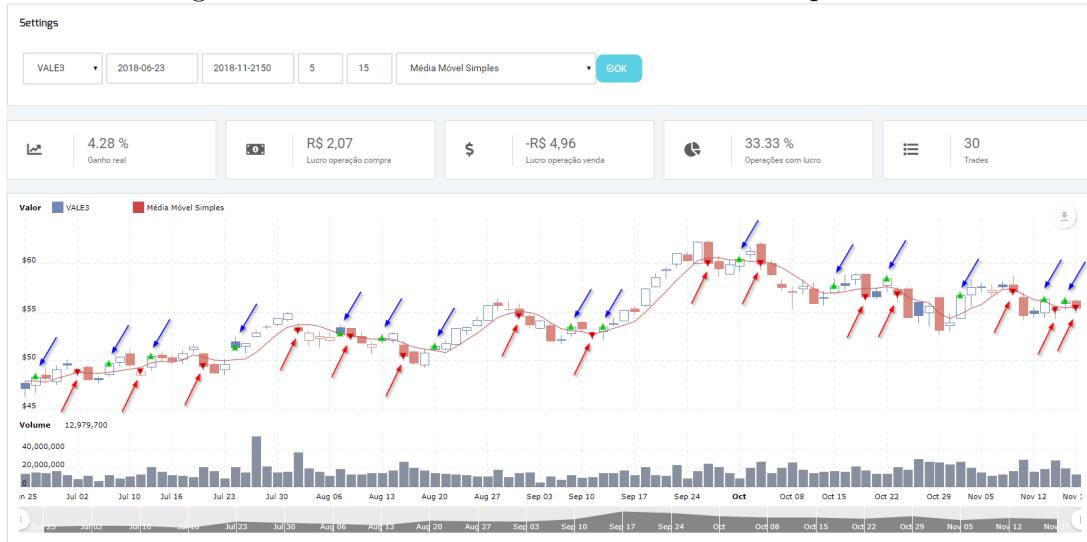
os preços estão abaixo da média, é um indicativo de mudança do mercado e demonstra uma oportunidade de compra.

Tabela 3: Definição de parâmetros de médias móveis

Estratégia	Indicador	Período
Tendência	Média Móvel Exponencial	17
Tendência	Média Móvel Exponencial	34

Fonte: Elaborado pelo autor.

Figura 6: Gráfico de Sinais do Média Móvel Exponencial



Fonte: Elaborado pelo autor.

A regra para geração de sinais de negociação utilizando o média móvel é demonstrada no algoritmo 3.

Algoritmo 3: Sinal de compra ou venda baseado em médias móveis

Entrada: *close* valor de fechamento do *candle*, *media* cálculo de média móvel

Saída: *signal* Indicadores de sinal: *BUY*, *SELL* ou *NULL* = Neutro

início

```
    signal ← NULL;  
    if media ≥ close then  
        | signal ← BUY  
    if media ≤ close then  
        | signal ← SELL  
fim  
retorna signal
```

3.2.4 Cruzamento de Médias Móveis

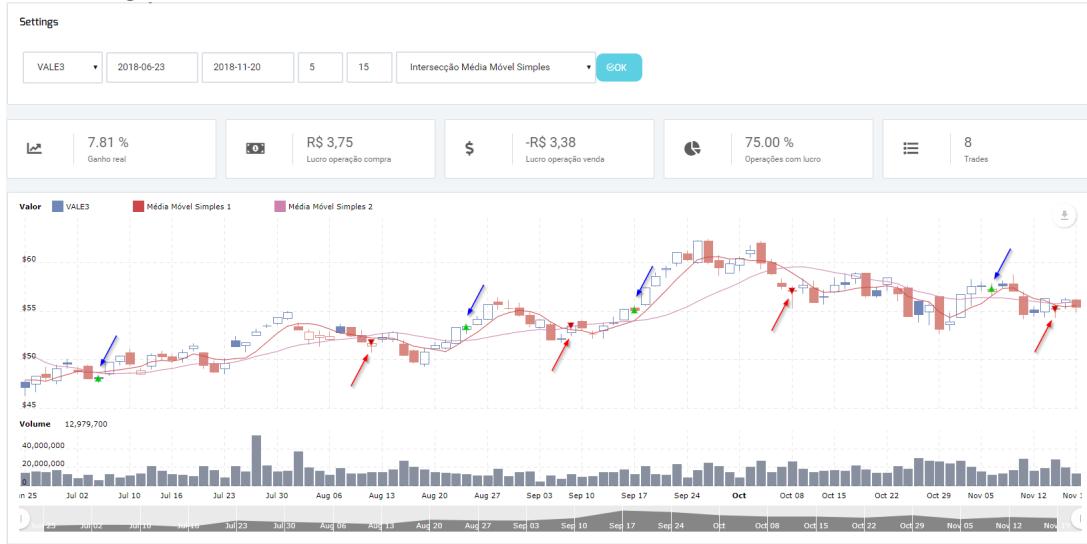
O cruzamento de médias móveis é outra maneira de analisar a tendência de um instrumento, para a simulação de cruzamento de média móveis foi escolhido o cálculo de média móvel aritmética, sendo a primeira média mais rápida e a segunda a mais lenta. Se houver um cruzamento de baixo para cima da média mais rápida em relação a mais lenta, então, há a indicação de compra, caso contrário, venda.

Tabela 4: Definição de parâmetros de cruzamento de médias móveis

Estratégia	Indicador	Período média móvel lenta	Período média móvel rápida
Cruzamento médias	Média Móvel Simples	9	17

Fonte: Elaborado pelo autor.

Figura 7: Gráfico de Sinais Cruzamento Médias Móveis Aritmética



Fonte: Elaborado pelo autor.

A regra para geração de sinais de negociação utilizando o cruzamento de médias móveis é demonstrada no algoritmo 4.

Algoritmo 4: Sinal de compra ou venda baseado no cruzamento de médias móveis

Entrada: *media1* cálculo da primeira média móvel, *media2* cálculo da segunda

média móvel

Saída: *signal* Indicadores de sinal: *BUY*, *SELL* ou *NULL* = Neutro
início

```

signal ← NULL;
if media1 ≥ media2 then
    | signal ← BUY
if media2 ≤ media1 then
    | signal ← SELL
fim
retorna signal
```

3.2.5 Prazos Operacionais

- **Day trade:** Segundo Luiz (2009) são operações realizadas ao longo do dia na bolsa de valores, visando obter retornos momentâneos de uma tendência, esta modalidade é recomendada para investidores mais experientes já que é preciso ter um conhe-

cimento profundo do mercado para operar com este tipo de modalidade, também pode ser definido um objetivo, bem como o *take-profit* e *stop-loss*, dependendo da estratégia utilizada.

- ***Swing Trade***: São operações realizadas na bolsa de valores com duração maior de um dia, o tempo operacional depende da estratatégia utilizada, é recomendada para operações cujo tenha um objetivo a ser alcançado, bem como o *take-profit* e *stop-loss*, dependendo da estratégia utilizada.
- ***High Frequency Trading HFT***: Conforme definição de Aldridge (2009) são operações de baixa latência totalmente automatizadas por algorítmos de negociação e operacionalizado sem intervenção humana, as ordens geralmente são realizadas em grande quantidade de negócios em vários momentos do dia, dependendo da estratégia utilizada, tem por objetivo obter ganhos momentâneos no mercado.

3.3 Métricas de Risco

Os indicadores de risco são utilizados para medir a relação de risco/retorno de uma operação no mercado financeiro, entretanto segundo Mandelbrot e Hudson (2010) as teorias clássicas de finanças nem sequer começaram a capturar ou medir toda a gama de riscos do mercado de capitais.

Abaixo são apresetandos os indicadores de risco utilizados no mercado financeiro e uma breve descrição sobre cada um.

3.3.1 *Sharpe Ratio*

Segundo Sharpe (1966), o *Sharpe Ratio* apresenta uma relação simples de desempenho em relação à um instrumento livre de risco, comumente utilizado para calcular a *performace* de portfólio de instrumentos ou então o índice de resultado em relação a uma determinada estratégia ou carteria de investimentos, quando comparado em relação à outro instrumento de mercado, (relação alto risco x alto risco), aquele que possuir o maior índice de *sharpe ratio* proporcionará um melhor retorno financeiro para o mesmo risco envolvido.

$$SR = \frac{Rp - rf}{\sigma(p)} \quad (6)$$

Na qual:

SR = Sharpe Ratio

Rp = Rentabilidade do portifólio

rf = Rentabilidade livre de risco

$\sigma(p)$ = Desvio padrão do portifólio

3.3.2 Volatilidade Histórica

Para Elder (2004) pode ser entendida como uma medida de risco do movimento de preços de um instrumento, quanto menor a dispersão de preço, menor o risco envolvido.

Lemos e Cardoso (2010) complementam como sendo o desvio padrão dos preços diárias.

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (R_i - \bar{r})^2}{n - 1}} \quad (7)$$

Na qual:

σ = Volatilidade histórica

n = Número de observações, geralmente utilizado o valor 252 (aproximação de dias úteis para um ano).

Ri = Retorno do período

r = Retorno médio

3.3.3 Máximo *Drawdown*

Conforme definição de Caldeira (2013) o máximo *drawdown* é uma medida entre o máximo global e o mínimo global da rentabilidade acumulada, pode-se entender como a relação entre o maior topo e fundo do comportamento histórico do IF.

$$MDD(N) = \frac{P - L}{P} \quad (8)$$

Na qual:

MDD = Máximo Drawdown

N = Número de observações

P = Valor máximo global em função de (N)

L = Valor mínimo global em função de (N)

3.4 Séries Temporais Financeiras

Uma série temporal é caracterizada como uma sequência de observações sobre uma variável ordenada no tempo (WOOLDRIDGE, 2006), no campo de finanças, podemos exemplificar as séries temporais como flutuações de preços, séries de volatilidade, séries de retorno percentual, *etc.*

As séries temporais financeiras apresenta algumas características particulares e vários desafios, como por exemplo prever a volatilidade de retornos futuros, modelos como os modelos *Autoregressive Conditional Heteroskedasticity* (ARCH) e *Generalized AutoRegressive Conditional Heteroskedasticity* (GARCH) foram desenvolvidos para lidar com estes desafios (SHUMWAY; STOFFER, 2017), também comuns a outros tipos de séries temporais:

- **Presença de tendência:** as variações de preços das séries temporais financeiras seguem um comportamento aleatório, o que caracteriza um padrão não estacionário e imprevisível, naturalmente como é o mercado financeiro, podendo conter tendências de alta e de baixa, com durações variáveis dependendo dos parâmetros do mercado, conforme demonstra figura 3;
- **Sazonalidade:** as motivações para a ocorrência de sazonalidade no mercado são diversas e este comportamento pode ter sua causa determinada por eventos periódicos, como por exemplo uma eleição presidencial, tem como efeito primário o aumento da quantidade de negociações e do volume financeiro na bolsa de valores;

- **Outliers:** são variáveis que se encontra a uma distância anormal das demais e provavelmente irá causar anomalias nos resultados obtidos por meio de algoritmos e sistemas de análise, em mercados financeiros podemos destacar como exemplo a presença de *gap*⁹ na cotação de preço;
- **Heteroscedasticidade:** a variância dos valores de entrada e de saída da série temporal não é constante com o passar do tempo, tornando o comportamento da série temporal aleatório;
- **Não-linearidade:** devido à sua complexidade de percorrer múltiplos caminhos e sentido e ao seu comportamento estocástico, não é possível modelar este tipo de série temporal utilizando uma equação diferencial linear;
- **Quebra estrutural:** entende-se que há uma ou mais mudanças para cima ou para baixo, segundo (STOCK; WATSON, 2004) pode surgir de um evento específico ou por meio de evolução vertiginosa, podendo interferir nas inferências estatísticas e no desempenho de testes de raiz unitária caso esta informação seja desconsiderada.

Conforme a definição de (BARROS, 2018), existem vários tipos testes estatísticos de raiz unitária, dentre eles:

- *Augmented Dickey-Fuller* (ADF) (1976)
- *Phillips Perron* (PP) (1988)
- *Kwiat-Phillips-Smith-Shin* (KPSS) (1992)

Séries temporais não estacionárias são difíceis de trabalhar quando queremos fazer análises inferenciais, como média e variância de retornos, ou então determinar a probabilidade de perda. Uma das razões pelas quais os métodos de ML de falham é porque baseiam-se no pressuposto de ter uma série *Independent and Identically Distributed* (IID), sendo esse pressuposto irreal no caso de séries temporais financeiras (PRADO, 2018), um dos desafios de utilizar ML em finanças é de que as séries temporais de preços têm tendências ou médias não constantes, e isso torna a série temporal não estacionária (SHUMWAY; STOFFER, 2017).

⁹pelo gráfico do instrumento, é possível visualizar um grande deslocamento na cotação.

Para contornar esta questão, utiliza-se métodos de padronização, como o método linear em diferença a afim de obter uma série IID, sabe-se que por forças de arbitragem, as séries temporais financeiras demonstram baixa relação de ruído (EASLEY; PRADO; O'HARA, 2012a), cada movimento do mercado depende de uma longa história de níveis anteriores, este fenômeno é denominado de memória da série histórica, entretanto, o método de primeira diferença reduz ainda mais o sinal de ruído, no sentido de que a memória histórica é desconsiderada inteiramente depois de uma janela de amostra (PRADO, 2018).

Entretanto, Russell e Norvig (2009), Bonaccorso (2017) citam a existência de métodos ML que possuem abordagens para trabalhar em amostras com distribuição normal denominados de algoritmos paramétricos de aprendizagem de máquina, se o tipo de distribuição é uma curva não normal, adota-se a utilização de métodos não paramétricos, a seção 4 aborda em detalhes o tipo de implementação utilizado, conforme o resultado do tipo de amostra obtido para os instrumentos financeiros selecionados.

3.5 *Automated Trading System*

Um ATS, também conhecido como *Mechanical Trading Systems*, *Algorithmic Trading*, *Automated Trading* ou *System Trading* é um sistema de *software* que implementa algorítmos determinísticos de estratégias operacionais para negociação na bolsa de valores, executando ordens no mercado sem que haja a intervenção humana, auxiliando no processo operacional de negociação, utilizando modelos e regras que foram previamente definidas (HENDERSHOTT; JONES; MENKVELD, 2011; TILLY; MONTESANO; SMITH, 2016).

Prado (2018) afirma que o desenvolvimento de um *automated trading system* é um processo complexo que deve ser feito por etapas, cada etapa deve ser concebida de maneira cautelar e cuidadosa.

Os passos descritos por Prado (2018) como parte do ciclo de desenvolvimento de um projeto de ATS devem ser divididos por funções específicas e existir o domínio de especialidades multidisciplinares para diferentes áreas de pesquisa e conhecimento, sendo:

- **Processamento e Armazenamentos dos Dados:** Consiste pelos procedimentos

de coleta de dados, correções, indexação, armazenamento e consumo, cada instrumento financeiro possui as suas particularidades como por exemplo a realização de divisão, agrupamento, ajustes de bonificação e dividendos, *etc* para o caso de ações e vencimento, rolagem, direito de exercício, dever de recompra, *etc* para o caso de opções, para mitigar deturpação no processo de desenvolvimento e análise do modelo de ML, estes eventos específicos devem ser tratados nesta etapa.

- **Análise de Atributos:** Esta etapa consiste na transformação de dados brutos em sinais informativos por meio de procedimentos de rotulação de dados, atribuição de pesos para determinados sinais ou eventos e análise de relevância estatística de atributos, um erro comum é acreditar que a análise de atributos pode ser utilizada como ponto de partida para o desenvolvimento de estratégias de negociação, ao invés disso, esta etapa consiste na categorização de observações que pode ser útil em várias situações do mercado. (AKANSU; KULKARNI; MALIOUTOV, 2016; PRADO, 2018).
- **Estratégias:** Neste momento, as informações obtidas por meio da categorização dos atributos são transformadas em algoritmos de investimento, o objetivo desta etapa é analisar o comportamento dos atributos em diferentes instrumentos e diferentes momentos do mercado e formular uma teoria geral que as explique, portanto, a estratégia é um experimento projetado para validar uma teoria relacionada às observações dos atributos e responder questões que façam sentido, em específico, identificar mecanismos econômicos como assimetria de informação, viés comportamental do instrumento, mudanças regulatórias setoriais, *etc*. Um item importante a salientar é de que os atributos são descobertos por meio de algoritmos de *ML* denominado como *black-box*¹⁰, entretanto, é recomendável e prudente desenvolver a estratégia de negociação com base em uma tese que deve ser pautada sobre o aspecto de uma *white-box*. Para isso, surge a abordagem de *meta-labeling*, abordado no capítulo 4.
- **Backtesters:** Utilizado para verificar como a estratégia se comportaria se os movimentos futuros repetissem os mesmos padrões observados no passado, entretanto,

¹⁰Em computação, é classificado como um sistema ou dispositivo qual a relação entre o estímulo de entrada e a resposta de saída é desconhecida ou não é levada em consideração (AZIZ; DOWLING, 2019).

o comportamento histórico é apenas um dos possíveis resultados de um processo estocástico, e não necessariamente provável, desta forma, cenários alternativos devem ser avaliados utilizando técnicas empíricas e experimentais, em particular, deve-se considerar que a estratégia pode estar em *overfitting*, tema abordado neste trabalho.

- **Equipe de Implantação** Esta etapa consiste na implementação e integração da estratégia em ambiente produtivo logicamente idêntico ao protótipo constituído durante as fases anteriores, detalhes adicionais devem ser avaliados e considerados para a viabilidade da estratégia como latência de resposta, uso de computação distribuída, utilização de *Graphics Processing Unit* (GPU) para ganho de *performance* no processamento de cálculos estatísticos, redundância e demais itens necessário em ambientes críticos e de alta disponibilidade.
- **Supervisão de Portfólio:** Após o período de integração, o modelo segue outras etapas até ser considerada efetivamente implantada, como por exemplo o monitoramento inicial após o período de *backtest*, testes de negociação, maturação, alocação gradativa de recursos financeiros, e descontinuação uma vez que a teoria inicial deixa de ser suportada por evidência empírica e estatística, maiores detalhes sobre cada um dos processos acima podem ser obtidos no trabalho de Prado (2018).

3.6 Finanças Quantitativas

As finanças quantitativas têm como base o uso da estatística e da econometria, Harry Markowitz foi um dos pioneiros nesta linha de pesquisa (VARIAN, 1993) ao propor a Teoria de Portfólio (MARKOWITZ, 1952) que busca uma alocação ótima de carteira baseado num tratamento matemático que procura maximizar a utilidade esperada como critério de formação de portfólio (um problema de maximização quadrática).

Tradicionalmente, a área de finanças quantitativas para investimentos possui duas subáreas:

- Comportamento e precificação de instrumentos;
- Gerenciamento de risco e portfólio.

Os métodos convencionais para análise quantitativa de instrumentos financeiros são modelos econométricos que trabalham com a previsão de séries temporais, tais como: o *Autoregressive Moving Average Models* (ARMA), o *Autoregressive Integrated Moving Average* (ARIMA), o *Autoregressive Fractionally Integrated Moving Average* (ARFIMA), o *Seasonal Autoregressive Moving Average* (SARMA), o *Seasonal Autoregressive Integrated Moving Average* (SARIMA), o ARCH e o GARCH.

O uso de algoritmos de ML na área de finanças quantitativas não tem a pretensão de substituir os métodos clássicos de econometria, mas sim complementa-los, um algoritmo de ML tem a capacidade de aprender padrões em um espaço de alta dimensionalidade, sem ser especificamente direcionado, como por exemplo o algoritmo *Support Vector Machine* (SVM), além disso, os algoritmos de ML permitem acompanhar alterações repentinas nas oscilações do mercado, aprender com o comportamento dos instrumentos, adaptar-se ao ambiente, o que dificilmente pode ser realizado por um modelo matemático linear e pré-definido (RESENDE, 2016).

3.7 Inteligência Artificial

A inteligência artificial é a nova fronteira de muitas aplicações úteis da vida real, a utilização no mercado de capitais é uma delas, tornando-se importante aliado na tomada de decisão operacional e estratégica uma vez que não há fatores emocionais envolvidos, em mercados financeiros, o sucesso de um investidor depende da qualidade da informação utilizada para auxiliá-lo, e em quão rápido ele consegue chegar a tal decisão (CAVALCANTE et al., 2016).

A figura 8 ilustra a linha do tempo das principais descobertas e evoluções que contribuiram significativamente para a área da Inteligência Artificial.

Figura 8: Linha do tempo da Inteligência Artificial



Fonte: Elaborado pelo autor.

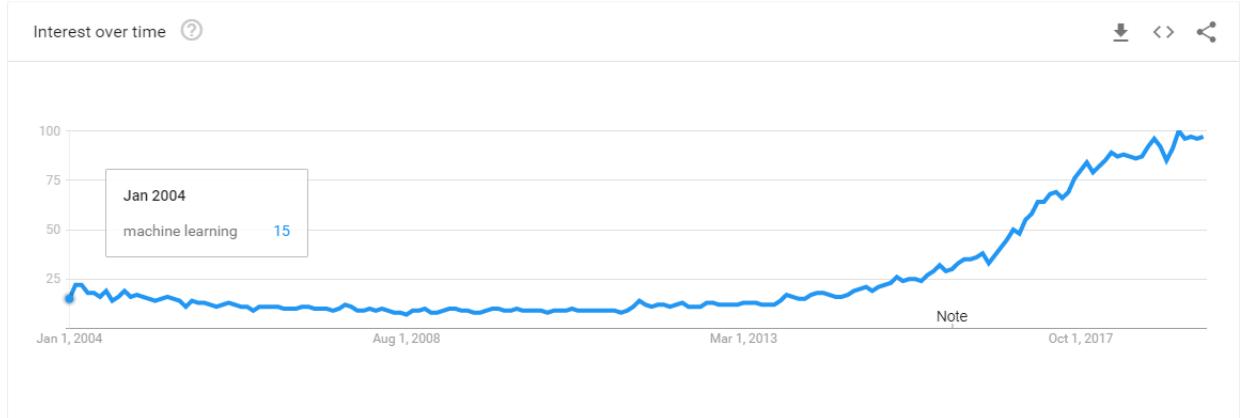
Diversas pesquisas utilizando técnicas de inteligência artificial foram introduzida para tentar prever e compreender o comportamento de instrumentos financeiros, como por exemplo, algoritmos de otimização, *clustering*, *decision tree*, SVM, *K-nearest neighbors*,

online machine learning, deep learning, algoritmos de PLN, etc. (CASTRO P. A. L.; ANNONI JUNIOR, 2018; THENMOZHI; CHAND, 2016; GRIGORYAN, 2017; AKITA et al., 2016; FISCHER; KRAUSS, 2018), a facilidade de adaptação de novos domínios, a capacidade de predição e de lidar com dados em um espaço de alta dimensionalidade foram os principais fatores que motivaram a escolha do método de ML neste trabalho.

3.7.1 *Machine Learning*

Nos últimos anos, o termo *machine learning* tornou-se um dos assuntos mais pesquisados e importantes na área da inteligência artificial. Não é de surpreender que suas aplicações estejam se tornando cada vez mais difundidas em todos os setores de negócios, sempre com novas e mais poderosas ferramentas *open source* e *frameworks* prontos para utilização, junto com centenas de artigos publicados todos os meses (BONACCORSO, 2017).

Figura 9: Frequênciа de pesquisa do termo *machine learning* no google.com



Fonte: <https://trends.google.com>

O objetivo na utilização de técnicas de ML no mercado de capitais busca desenvolver um modelo que explique o comportamento de instrumentos da bolsa de valores de São Paulo, especificamente do mercado à vista, com isso, será verificado a possibilidade de produzir previsões acima do *benchmark buy and hold* em séries temporais fora da amostra de treinamento utilizando um modelo de treinamento supervisionado.

Segundo (RUSSELL; NORVIG, 2009; BONACCORSO, 2017) os tipos de treinamen-

tos disponíveis para ML, são:

- **Aprendizado Supervisionado:** Um cenário supervisionado é caracterizado pelo conceito da existência de um supervisor cujo a principal tarefa consiste em fornecer ao agente uma medida precisa do seu erro diretamente comparável com os valores de entrada e saída esperada, a partir dessas informações, o algoritmo pode corrigir seus parâmetros de modo a reduzir a magnitude de uma função de perda global.
- **Aprendizado Semi-Supervisionado:** Esta técnica envolve o uso de dados rotulados e não rotulados, geralmente é adotada quando é necessário categorizar uma grande quantidade de dados com alguns exemplos completos (rótulos) ou quando há a necessidade de impor algumas restrições a um algoritmo de *clustering*.
- **Aprendizado Não Supervisionado:** Esta abordagem não há a presença de qualquer supervisor e, portanto, baseia-se no erro médio absoluto, é útil quando é necessário aprender como um conjunto de elementos que pode ser agrupado de acordo com a sua similaridade e considerando a medida de distância dos elementos e a posição mútua.
- **Aprendizado Por Reforço:** Mesmo sem a existência de supervisores reais, o aprendizado por reforço também é baseado no *feedback* fornecido pelo ambiente, entretanto neste caso, a informação é mais qualitativa e não ajuda o agente a determinar uma medida precisa do seu erro, este *feedback* é geralmente chamado de recompensa, é útil para entender se uma determinada ação executada em um estado é positiva ou não.

Implementar um modelo de ML para obter resultados fora da amostra exige alguns cuidados, segundo (PRADO, 2018), abaixo estão descritos alguns motivos pelos quais o uso de ML por fundos de investimentos quantitativos tendem a falhar:

1. Trabalhar de forma independente, em *silos*, para assegurar a diversificação;
2. Utilizar o *backtest* como ferramenta de pesquisa;
3. Utilizar série temporal histórica na forma cronológica (amostra ineficiente);

4. Normalizar a série temporal histórica utilizando primeira diferença;
5. Utilizar um valor fixo para o alvo do modelo de predição (\hat{y});
6. Treinar a estratégia e o alvo simultaneamente no mesmo modelo;
7. Atribuir pesos em modelos não IID;
8. Não utilizar *Cross-Validation* adequadamente;
9. Validar a estratégia apenas pelos métodos de *Walk-Forward* e *Backtesting*;
10. *Backtest Overfitting*

Segundo Gendreau e Potvin (2010), a utilização de heurísticas construtivas por meio de regras baseadas em dados do problema pode gerar uma solução factível ou não, portanto, a utilização de métodos clássicos de análise de investimento como a análise fundamentalista, análise técnica e econometria podem ser utilizados para geração de variáveis categórias do modelo predição de *machine learning*, verificando-se o grau de relevância de cada atributo, o capítulo 4 demonstra em detalhes a implementação utilizada.

3.7.2 Tipos de Dados

Embora a grande parte de trabalhos acadêmicos e algorítmos de *trading* realizem principalmente a análise de dados e a derivação para outros tipos de séries temporais utilizando como base a série histórica de preços, há outros tipos de dados que podem ser utilizados para o desenvolvimento de modelos de ML, tais como:

- **Dados não estruturados:** Os dados não estruturados consistem em documentos, como artigos de notícias, publicações em redes sociais, documentos, relatórios, *etc*. A análise e processamento desse tipo dado depende de técnicas de PLN, o uso de dados não estruturados no mercado financeiro é uma tentativa de determinar o sentimento do contexto, como por exemplo classificando os textos como atributos de compra, venda ou neutro, podendo ser utilizado em modelos de ML para a condução de uma estratégia de negociação. O termo denominado para o processo de classificação é análise de sentimento (BOLLEN; MAO; ZENG, 2011).

- **Dados de alta frequência HFT:** Os dados de alta frequência é caracterizado como todos os dados gerados durante o pregão eletrônico de forma bruta, ou seja, os dados são concebidos conforme a movimentação do instrumento, informações como volume, quantidade de ordens executadas, quantidade de ordens canceladas, profundidade do *book* de ofertas, preço, *etc*, podem ser obtidas na ordem de microsegundos, considerando que o instrumento objeto tenha movimentação suficiente nesta ordem de grandeza (CARTEA; JAIMUNGAL; PENALVA, 2015).

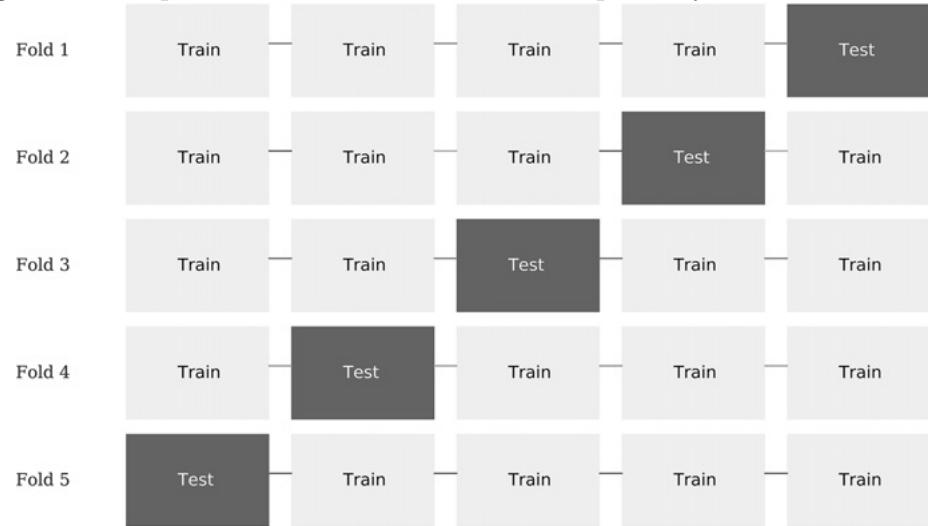
Para a realização deste trabalho, foi considerado a utilização da série temporal histórica de preços na ordem de microsegundos, detalhes sobre a implementação e fracionamento dos dados podem ser encontrados no capítulo 4.

3.7.3 *Overfitting*

Overfitting é um fenômeno que pode ocorrer com todos os tipos de modelos de aprendizagem, mesmo quando a função de alvo não é totalmente aleatória (RUSSELL; NORVIG, 2009), é comum observar a presença de *overfitting* quando se busca otimizar a *performance* de um modelo de ML, e pode ser compreendido quando o desempenho de um modelo é relativamente significante dentro de uma amostra de treinamento no entanto seu desempenho é insignificante fora da amostra de treinamento, O desafio do *overfitting* pode ser visto como apenas um dos exemplos de uma crescente conscientização no mundo para utilização de métodos adequados de pesquisa científica sobre a necessidade de rigor e reproduzibilidade (BAILEY et al., 2014).

Para contornar o desafio de *overfitting* em séries temporais financeiras, será utilizado o modelo de treinamento *Random Forests* que foram projetados para alcançar a função alvo em um conjunto de dados desconhecidos, obtendo-se resultados com baixa variância devido ao seu método de aprendizagem, também são abordados os métodos de *Purged K-Fold Cross-Validation* propostos por Prado (2018) para realização de testes de *overfitting*.

Figura 10: Esquema de treinamento e divisão para 5-fold *Cross-Validation*



Fonte: Prado (2018).

A figura 10 demonstra a validação de um modelo de ML utilizando o método *K-Fold Cross-Validation*.

4 Metodologia

4.1 Objetivo do Capítulo

O objetivo deste capítulo é conceituar e apresentar a metodologia e subsídios utilizados para o desenvolvimento da presente pesquisa, em linhas gerais, segundo (FILHO, 2012), pode-se classificar o presente estudo como pesquisa aplicada, uma vez que visa atingir resultados imediatos.

4.2 Implementação Técnica

A metodologia aplicada para a elaboração deste trabalho compreende as seguintes etapas:

1. Pesquisa bibliográfica;
 2. Proposição, implementação e aprimoramento de técnicas de ML;
 3. Teste exploratório da proposição com a utilização de dados reais da bolsa de valores de São Paulo¹¹.
- A etapa de pesquisa bibliográfica compreende o estudo dos principais autores que desenvolvem estudos relacionados para o mercado de ações, condensar os principais métodos, técnicas e comparar os resultados obtidos dos diferentes métodos de análise.
 - A implementação e proposição de novas técnicas compreende a possível evolução de técnicas existentes, sendo abordagens que diz respeito à aprendizado de máquina e finanças:
 - Obtenção e preparação dos dados observados;
 - Implementação de modelos de *machine learning*;
 - Busca e otimização de atributos do modelo;

¹¹<http://www.b3.com.br>

- Implementação de estratégias de negociação;
- Utilização do método *K-Fold Cross-Validation* para mitigação de *overfitting*;
- Inferência dos resultados obtidos.

4.3 Preparação dos Dados

A reputação de um fornecedor de dados geralmente baseia-se na qualidade (percebida) dos dados. Em termos simples, dados ruins ou ausentes levam a sinais de predição erradas e, portanto, em análises incorretas. Apesar disso, muitos pesquisadores ainda sofrem com uma qualidade de dados ruim ou inconsistente. Assim, há sempre um processo de mineração necessário a ser realizado (SILVA, 2016).

Para este trabalho, os dados foram obtidos dentro do contexto de pregão da Bolsa de Valores de São Paulo¹², ou seja, já está partindo de uma base de dados identificada e que representa o resultado real.

4.3.1 Base de Dados

Após o processo de obtenção dos dados, é realizado a etapa de processamento e armazenamento dos dados, a figura 11 demonstra o Modelo Entidade Relacionamento (MER) da base de dados local por meio do Sistema de Gerenciamento de Banco de Dados (SGBD) Timescale¹³, onde são armazenados e gerenciados os dados de negócio sincronizados da B3 (2018), os modelos de ML e os resultados das métricas de desempenho.

A disposição de tabelas foi realizada de forma a atender as especificidades deste trabalho, entretanto o modelo pode ser reaproveitado para pesquisas semelhantes, a organização das tabelas estão representadas da seguinte forma:

- **Tabela *tintraday*:** Armazena os dados de negociações realizadas em instrumentos financeiros para o período *intraday*:
 - Coluna *tradenumbers*: armazena o número da negociação.

¹²Obtido do FTP oficial da Bolsa de Valores de São Paulo - <ftp://ftp.bmf.com.br/MarketData/Bovespa-Vista/>

¹³<https://www.timescale.com/>

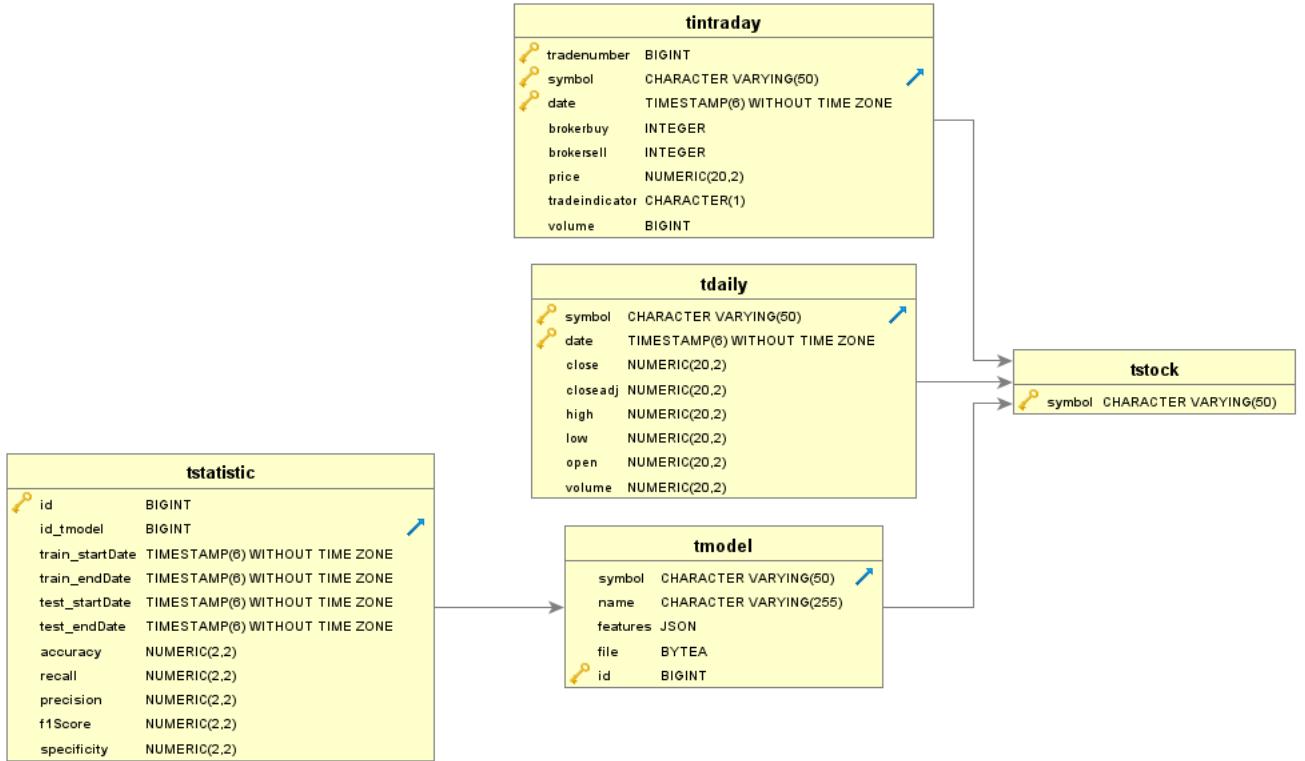
- Coluna *symbol*: armazena o símbolo do instrumento financeiro.
 - Coluna *date*: armazena a data e o horário da negociação.
 - Coluna *brokerbuy*: armazena a identificação da corretora compradora.
 - Coluna *brokersell*: armazena a identificação da corretora vendedora.
 - Coluna *price*: armazena o preço da negociação.
 - Coluna *tradeindicator*: armazena a situação da negociação (1:ativo, 2:canceled).
 - Coluna *volume*: armazena a quantidade da negociação.
- **Tabela *tdaily***: Armazena os dados de negócios realizados diariamente agrupados por dia, contendo o resultado diário do mercado por instrumento:
- Coluna *symbol*: armazena o símbolo do instrumento financeiro.
 - Coluna *date*: armazena a data da negociação.
 - Coluna *close*: armazena o preço de fechamento.
 - Coluna *closeAdj*: armazena o preço de fechamento ajustado.
 - Coluna *high*: armazena o maior preço do dia.
 - Coluna *low*: armazena o menor preço do dia.
 - Coluna *open*: armazena o preço de abertura.
 - Coluna *volume*: armazena o volume de negócios.
- **Tabela *tstock***: Armazena os dados gerais de instrumentos financeiros:
- Coluna *symbol*: armazena o símbolo do instrumento financeiro.
- **Tabela *tmodel***: Armazena os modelos de ML:
- Coluna *symbol*: armazena o símbolo do instrumento financeiro.
 - Coluna *name*: armazena o nome do modelo.
 - Coluna *features*: armazena os atributos do modelo em formato JSON¹⁴.
 - Coluna *file*: armazena o modelo em formato binário.

¹⁴<https://www.json.org/>

- Coluna *id*: armazena o identificador da tabela *tmodel*.
- **Tabela *tstatistic***: Armazena os dados estatísticos dos resultados do treinamento e do teste de validação dos mododelos de ML:
 - Coluna *id*: armazena o identificador da tabela *tstatistic*.
 - Coluna *id_tmodel*: armazena o identificador do modelo (chave da tabela *tmodel*).
 - Coluna *train_startDate*: armazena a data e o hora do início do período de treinamento do modelo.
 - Coluna *train_endDate*: armazena a data e o hora do fim do período de treinamento do modelo.
 - Coluna *test_startDate*: armazena a data e o hora do início do período de teste do modelo.
 - Coluna *test_endDate*: armazena a data e o hora do fim do período de teste do modelo.
 - Coluna *accuracy*: armazena a acurácia do modelo.
 - Coluna *recall*: armazena o *recall* do modelo.
 - Coluna *precision*: armazena a precisão do modelo.
 - Coluna *f1score*: armazena o *F1-Score* do modelo.
 - Coluna *specificity*: armazena a especificidade do modelo.

A figura 11 representa o modelo entidade relacionamento.

Figura 11: Modelo Entidade Relacionamento



Fonte: Elaborado pelo autor.

Para o armazenamento dos dados de negociações, foi escolhido o SGBD Timescale¹⁵ para manipulação de séries temporais de maneira estruturada e eficiente, com a premissa de armazenamento crescente e massivo, o Timescale foi concebido especificamente para uso em séries temporais, para o processo de obtenção, armazenamento e indexação de dados, foi desenvolvido o módulo sincronizador, trata-se de uma aplicação na linguagem de programação Java¹⁶ que tem como objetivo realizar o sincronismo diário dos dados da Bolsa de Valores de São Paulo de forma automatizada, detalhes sobre a implementação técnica e utilização podem ser obtidos acessando o repositório *open-source*: <https://github.com/marretti/stock-market-b3>.

A tabela 5 representa os instrumentos financeiros selecionados entre os mais negociados na bolsa de valores de São Paulo e de diferentes setores de atuação.

A escolha dos instrumentos foi realizada de maneira arbitrária, foram selecionados

¹⁵<https://www.timescale.com>

¹⁶<https://www.java.com>

os instrumentos dentre os que possuem maior volume de negociação para cada setor de atuação.

Tabela 5: Ações selecionadas do índice IBOVESPA para os experimentos realizados

#	Código do Instrumento	Instituição	Período
1	PETR4	Petróleo Brasileiro S.A.	01/04/2018 à 15/06/2019
2	VALE3	Vale S.A.	01/04/2018 à 15/06/2019
3	ITUB4	Itaú Unibanco S.A.	01/04/2018 à 15/06/2019
4	VVAR3	Via Varejo S.A.	01/04/2018 à 15/06/2019
5	MULT3	Multiplan E. I. S.A.	01/04/2018 à 15/06/2019

Fonte: Elaborado pelo autor.

Após o processo de obtenção de dados não estruturados, processamento e armazenamento, podemos agrupar os dados para utilização em algoritmos de ML abordadas na literatura.

A escola de análise de investimentos clássica (ELDER, 2004; TAYLOR, 2011; BROOKS, 2011) sugere a utilização de séries temporais de preços agrupados com informações de *candlesticks* de 1, 5, 15, 30 minutos, diário, semanal, mensal, *etc.*

O tópico a seguir irá abordar uma alternativa para o tratamento adequado em séries temporais financeiras, que acredita-se ser mais eficiente para o mercado financeiro atual.

4.3.2 Frequência dos Dados

Atualmente, os mercados são negociados por algoritmos de *trading* que operam com a supervisão humana, onde os ciclos de processamento da *Central Processing Unit* (CPU) são mais relevantes do que os intervalos cronológicos, embora as séries temporais formadas por tempo cronológico sejam talvez as mais populares entre os praticantes e acadêmicos, eles devem ser evitados (EASLEY; PRADO; O'HARA, 2011), pelos motivos destacados abaixo:

- Os mercados não recebem informações em tempo cronológico, como seres biológicos, faz sentido para os seres humanos organizar o seu dia de acordo com o ciclo da

luz solar, mas os mercados de hoje são operados por algoritmos com a supervisão humana, para a qual os ciclos de processamento de CPU são muito mais relevantes do que intervalos cronológicos (EASLEY; PRADO; O'HARA, 2011). Isso significa que as séries temporais de tempo cronológico superestima as informações durante os períodos de baixa atividade em relação aos períodos de alta atividade.

- Devido ao excesso e sub-amostragem, as séries temporais de tempo cronológico, exibem propriedades estatísticas fracas, como correlação serial, heterocedasticidade e não-normalidade de retornos (EASLEY; PRADO; O'HARA, 2012b). Modelos GARCH foram desenvolvidos, em parte, para lidar com a heterocedasticidade associada à amostragem incorreta.

Com objetivo de obter séries contínuas, homogêneas e estruturadas, foi realizado estudo comparativo dos métodos propostos abaixo:

- ***Tick Bars:*** Desde as pesquisas de (MANDELBROT; TAYLOR, 1967), diversos estudos foram realizados afim de obter dados de séries temporais mais próximo de uma distribuição normal IID (ANÉ; GEMAN, 2000), este método consiste na separação das séries temporais por quantidade de negócios, por exemplo, a cada 1000 negócios realizados, é obtido uma amostra da série temporal correspondente à aquele "evento" do mercado, entretanto, cuidados com a presença de *outliers* devem ser levados em consideração.
- ***Volume Bars:*** Consiste na separação por volume de negociações, qual deve ser definido um valor específico, como por exemplo, a cada 50.000 volumes é gerado uma amostra dependente da série temporal, o agrupamento por volume foi concebido inicialmente por Clark (1973) e obteve-se amostragem com melhores propriedades estatísticas (por exemplo, a aproximação de uma série temporal com distribuição normal IID).
- ***Dollar Bars:*** É estabelecido um *threshold* fixo, onde o valor do *dollar bar* é obtido pela multiplicação do (*preço * volume*) até atingir o *threshold* definido, gerando uma amostra dependente da série temporal, segundo Prado (2018), este tipo de abordagem tende a ser mais robusto ao longo do tempo, devido a particularidades

das ações do mercado financeiro como a distribuição de dividendos, agrupamentos, desdobramentos, *etc*, são ajustados dinamicamente através de uma função de ajuste.

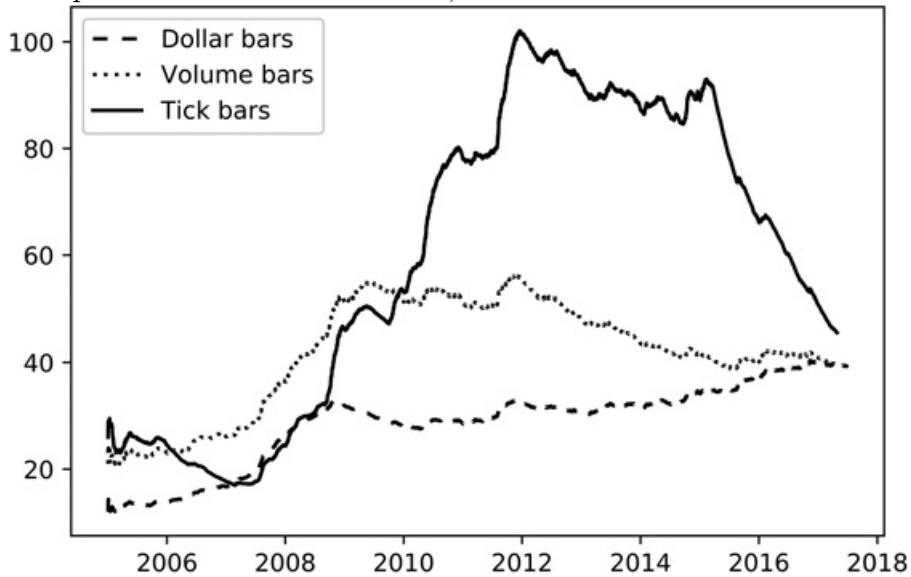
Figura 12: Exemplo de agrupamento de dados utilizando o método *Dollar Bars*

date_time	open	high	low	close	cum_vol	cum_dollar	cum_ticks
2018-04-02 10:03:00.000	21.36	21.36	21.36	21.36	48700	1040232.0	35
2018-04-02 10:03:06.027	21.36	21.37	21.35	21.35	49800	1063549.0	34
2018-04-02 10:03:09.461	21.35	21.35	21.32	21.35	47400	1011113.0	23
2018-04-02 10:03:27.009	21.35	21.37	21.33	21.34	46900	1001199.0	25
2018-04-02 10:04:08.222	21.34	21.37	21.32	21.37	47400	1011903.0	113
2018-04-02 10:04:32.341	21.37	21.38	21.35	21.37	47700	1019066.0	84
2018-04-02 10:04:46.968	21.37	21.39	21.37	21.39	48100	1028346.0	21
2018-04-02 10:04:52.401	21.39	21.39	21.37	21.38	47600	1017614.0	25
2018-04-02 10:05:17.281	21.38	21.40	21.37	21.39	47200	1009380.0	44
2018-04-02 10:05:23.056	21.39	21.40	21.38	21.40	81200	1737570.0	21
2018-04-02 10:05:24.167	21.40	21.42	21.39	21.40	47000	1005902.0	30

Fonte: Elaborado pelo autor.

Podemos observar na figura 12 o exemplo de agrupamento da série temporal utilizando o método *Dollar Bars*, devido a cobertura do período explorado (menos de 2 anos), foi utilizado o valor de *threshold* de 1.000.000, este é o valor de referência a ser seguido, dependendo da movimentação do mercado e do volume das ordens, se por exemplo houver uma única ordem que ultrapasse o resultado da equação (*preço * valor*), será extraído a amostra resultante deste evento, conforme o valor (1.737.570) destacado na figura 12.

Figura 13: Frequência média diária de *tick*, *volume* e *dollar* do mini contrato S&P 500



Fonte: Prado (2018).

A figura 13 demonstra a estabilidade de um período de +10 anos no contrato S&p 500 da NYSE¹⁷, segundo Prado (2018) o *dollar bars* é considerado o mais robusto em termos estatísticos para eventos como *split* de ações, agrupamento que impactam no número de negócios e consequentemente no volume, como a equação é denominada pelo *threshold* ajustado do (*preço * volume*), tais eventos já se encontram incorporados ao longo do tempo.

A tabela 6 ilustra a quantidade de dados das séries para cada instrumento financeiro, compreendido entre o período de estudo (01/04/2018 à 15/06/2019), sendo representado por:

- Código: Representa o código do instrumento financeiro;
- Observações: A quantidade de observações compreendido entre o período de estudo deste trabalho;
- *Tick Bars*: A quantidade de *Tick Bars* processado com *threshold* de 1.000;
- *Volume Bars*: A quantidade de *Volume Bars* processado com *threshold* de 50.000;

¹⁷New York Stock Exchange - <https://www.nyse.com>

- *Dollar Bars*: A quantidade de *Volume Bars* processado com *threshold* de 1.000.000.

Tabela 6: Frequência de dados

#	Código	Observações	<i>Tick Bars</i>	<i>Volume Bars</i>	<i>Dollar Bars</i>
1	PETR4	18.773.990	18.074	331.129	378.574
2	VALE3	10.741.829	10.742	98.660	236.781
3	ITUB4	9.761.047	9.762	77.162	147.651
4	VVAR3	2.667.580	2.668	36.889	9.734
5	MULT3	3.174.503	3.175	11.353	15.366

Fonte: Elaborado pelo autor.

4.4 Análise Estatística

Um ponto importante de decisão ao utilizar amostra de dados é inferir se os dados seguem uma distribuição normal Gaussiana, devido ao fato que uma fração do campo da estatística pressupõem que os dados foram obtidos com base na hipótese de uma distribuição normal Gaussiana (THODE, 2002).

Segundo Bonaccorso (2017), com os algoritmos de ML isto não é diferente, se forem escolhidos métodos que assumem uma distribuição Gaussiana e seus dados forem extraídos de uma distribuição não normal, as descobertas podem ser enganosas ou claramente erradas, conforme a definição de Russell e Norvig (2009), os métodos de aprendizagem paramétricos (Regressão Linear (RL), Regressão Logística (RL), *Naive Bayes* (NB), *etc* - também conhecidos como algoritmos lineares de aprendizado de máquina) dependem que os dados da amostra sigam uma distribuição normal, diferente dos algoritmos de aprendizagem não paramétricos (SVM, RF, *k-Nearest Neighbors* (KN)), *etc*) que utilizam outras formas de abordagem para o domínio do problema.

Conforme ilustra a figura 14, os dados de retornos discretos obtidos dos instrumentos financeiros não seguem uma distribuição normal, apesar de graficamente terem apresentado uma melhora na curva de distribuição após o tratamento utilizado conforme descrito na seção 4.3.2, não pode-se afirmar que os dados seguem uma distribuição ou então simplesmente adotar a utilização de métodos paramétricos de ML, para dirimir qualquer dúvida

sobre o tipo de amostragem dos dados, foi adotado o teste de *Shapiro-Wilk* (SW) conforme detalhado na seção 4.4.1.

Uma variável aleatória x tem distribuição normal com parâmetros μ e σ desconhecidos se a função densidade de probabilidade é dada por:

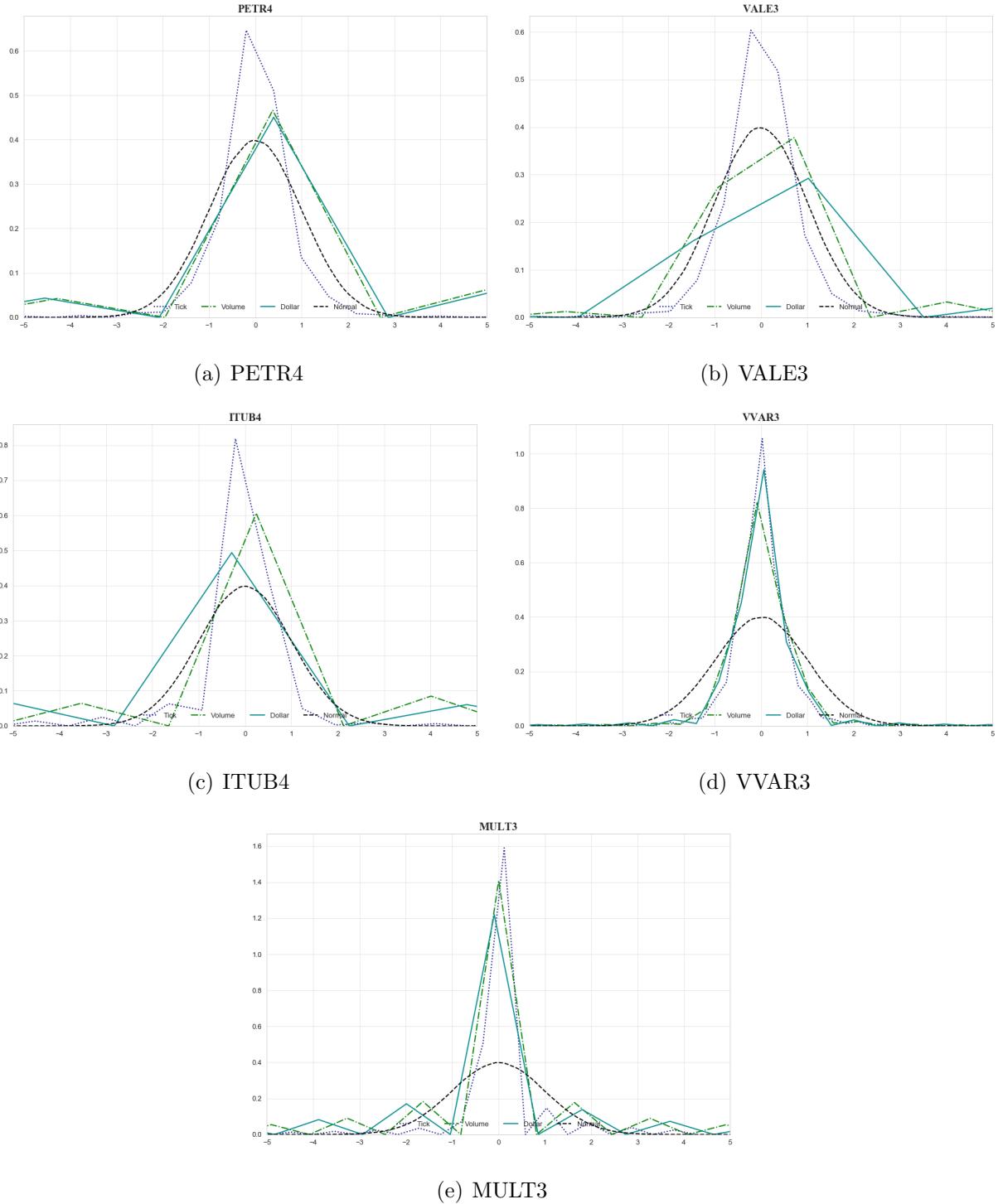
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}, -\infty < x < \infty. \quad (9)$$

Na qual:

μ : é a média da amostra.

σ : é o desvio padrão da amostra.

Figura 14: Curvas de distribuição normal dos instrumentos financeiros selecionados



4.4.1 Shapiro-Wilk

Com a finalidade de determinar se as amostras de dados seguem uma distribuição normal Gaussiana, o teste de SW, tornou-se o preferido devido às suas propriedades estatísticas relevantes (MENDES; PALA, 2003), dada uma amostra aleatória ordenada,

$y_1 < y_2 < \dots < y_{n'}$ o teste estatístico original de (SHAPIRO; WILK, 1965) é definido como:

$$SW = \frac{(\sum_{i=1}^n a_i y(i))^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (10)$$

Na qual:

$y(i)$: é a i^{th} ordem estatística.

y : é a amostra de tamanho n .

a_i : São as constantes geradas por meio do cálculo da variância, covariância e média da amostra de tamanho n .

\bar{y} : é a média da amostra.

Se o *p-value* dos resultados do teste de SW, for maior que (0.05), então não há evidências estatísticas para rejeitar a hipótese nula, portanto os dados são distribuídos normalmente, logo:

H0: os dados seguem uma distribuição normal.

H1: os dados não seguem uma distribuição normal.

Tabela 7: Resultado do teste estatístico de *Shapiro-Wilk*

#	Código	<i>Shapiro-Wilk</i>	<i>p-value</i>	Formato
1	PETR4	0.6532	0.0	<i>Tick Bars</i>
2	VALE3	0.6678	0.0	<i>Tick Bars</i>
3	ITUB4	0.3717	0.0	<i>Tick Bars</i>
4	VVAR4	0.5614	0.0	<i>Tick Bars</i>
5	MULT3	0.1168	0.0	<i>Tick Bars</i>
6	PETR4	0.5643	0.0	<i>Volume Bars</i>
7	VALE3	0.5978	0.0	<i>Volume Bars</i>
8	ITUB4	0.3505	0.0	<i>Volume Bars</i>
9	VVAR4	0.5107	0.0	<i>Volume Bars</i>
10	MULT3	0.1058	0.0	<i>Volume Bars</i>
11	PETR4	0.5563	0.0	<i>Dollar Bars</i>
12	VALE3	0.5763	0.0	<i>Dollar Bars</i>
13	ITUB4	0.3502	0.0	<i>Dollar Bars</i>
14	VVAR4	0.4723	0.0	<i>Dollar Bars</i>
15	MULT3	0.1074	0.0	<i>Dollar Bars</i>

Fonte: Elaborado pelo autor.

Conforme pode-se observar a tabela 7, todos os instrumentos financeiros falharam sobre a hipótese de SW, o resultado do *p-value* é menor que 0.05 para todos os tipos de formato de dados, portanto, não há evidências estatísticas para aceitar a hipótese nula, desta forma, o escopo deste trabalho será delimitado para o uso apenas de métodos de ML não paramétricos.

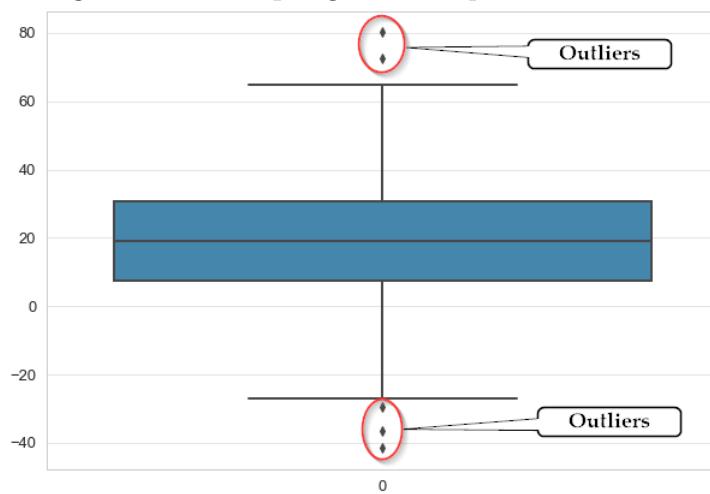
4.4.2 Detecção de *Outliers*

Ao utilizar séries temporais financeiras, é essencial identificar a presença de *outliers*, estas variáveis apresentam afastamento da média amostral, podendo causar viés comportamental em modelos de ML quando o domínio de aplicabilidade não tem como finalidade a identificação de anomalias (ZHAO; WANG, 2015).

Na estatística descritiva, o método do Quartil proposto por Tukey (1977) é amplamente conhecido e citado no meio acadêmico e utiliza o conceito de intervalo de confiança, os valores pertencentes a um determinado intervalo de confiança são considerados aceitáveis, enquanto que valores fora deste (abaixo ou acima) são tratados como *Outliers*. Não é objeto deste trabalho uma ampla exploração de métodos de detecção de *Outliers*, para detalhes desta implementação, é recomendado consultar as referências originais.

O *boxplot* é um método para representar graficamente grupos de dados numéricos através do método do Quartil, onde os *outliers* são exibidos como pontos individuais, conforme demonstra a figura 15, a definição abaixo sugere que, se houver a presença de *outliers*, serão exibidos como pontos separados do demais valores números e o restante dos exemplos numéricos serão agrupados.

Figura 15: Exemplo gráfico *boxplot* de *Outliers*



Fonte: Elaborado pelo autor.

Uma observação realizada por (PRADO, 2018) é a existência de leilão de abertura e de fechamento em cada instrumento financeiro, e ao término do leilão, apenas uma negociação é publicada no *book* de ofertas com um valor descomunal. Esta negociação pode ser o equivalente a milhares de ordens, embora seja reportado apenas como uma negociação, isto significa que, por um período de tempo, a bolsa de valores recebe pedidos de ofertas e acumula sem executá-los, ao abrir ou fechar o pregão eletrônico, registra todas as ofertas em apenas uma negociação.

Este tipo de comportamento pode gerar a presença de *outliers*, foi verificado a presença

de agrupamento de ordens no leilão da bolsa de valores de São Paulo (objeto da pesquisa) e não foi encontrado tal comportamento no *book* de negociações conforme ilustra a figura 16, as ordens foram executadas separadamente na abertura do mercado pela bolsa de valores de São Paulo, o exemplo foi apresentado com a profundidade de 25 negociações, entretanto foi possível constatar a existência de mais de 100 negociações no mesmo milisegundo para esta data.

Também foi verificado para o leilão de fechamento nos mesmos instrumentos (VALE3, GGBR4) e não foi constatado agrupamento de ordens no leilão.

Figura 16: Exemplo do *book* de negociações dos intrumentos VALE3 e GGBR4

	date	price	volume	date	price	volume	
	date	price	volume	date	price	volume	
2018-04-02 10:12:05	539	42.32	200	2018-04-02 10:03:00	004	15.64	300
2018-04-02 10:12:05	539	42.32	200	2018-04-02 10:03:00	004	15.64	200
2018-04-02 10:12:05	539	42.32	1000	2018-04-02 10:03:00	004	15.64	100
2018-04-02 10:12:05	539	42.32	1900	2018-04-02 10:03:00	004	15.64	500
2018-04-02 10:12:05	539	42.32	200	2018-04-02 10:03:00	004	15.64	100
2018-04-02 10:12:05	539	42.32	700	2018-04-02 10:03:00	004	15.64	3400
2018-04-02 10:12:05	539	42.32	100	2018-04-02 10:03:00	004	15.64	1600
2018-04-02 10:12:05	539	42.32	1000	2018-04-02 10:03:00	004	15.64	200
2018-04-02 10:12:05	539	42.32	200	2018-04-02 10:03:00	004	15.64	100
2018-04-02 10:12:05	539	42.32	700	2018-04-02 10:03:00	004	15.64	200
2018-04-02 10:12:05	539	42.32	1000	2018-04-02 10:03:00	004	15.64	1200
2018-04-02 10:12:05	539	42.32	800	2018-04-02 10:03:00	004	15.64	1000
2018-04-02 10:12:05	539	42.32	400	2018-04-02 10:03:00	004	15.64	600
2018-04-02 10:12:05	539	42.32	200	2018-04-02 10:03:00	004	15.64	600
2018-04-02 10:12:05	539	42.32	100	2018-04-02 10:03:00	004	15.64	300
2018-04-02 10:12:05	539	42.32	1000	2018-04-02 10:03:00	004	15.64	100
2018-04-02 10:12:05	539	42.32	800	2018-04-02 10:03:00	004	15.64	100
2018-04-02 10:12:05	539	42.32	200	2018-04-02 10:03:00	004	15.64	100
2018-04-02 10:12:05	539	42.32	700	2018-04-02 10:03:00	004	15.64	100
2018-04-02 10:12:05	539	42.32	200	2018-04-02 10:03:00	004	15.64	400
2018-04-02 10:12:05	539	42.32	100	2018-04-02 10:03:00	004	15.64	1400
2018-04-02 10:12:05	539	42.32	800	2018-04-02 10:03:00	004	15.64	500
2018-04-02 10:12:05	539	42.32	100	2018-04-02 10:03:00	004	15.64	100
2018-04-02 10:12:05	539	42.32	1000	2018-04-02 10:03:00	004	15.64	1000
2018-04-02 10:12:05	539	42.32	100	2018-04-02 10:03:00	004	15.64	400

(a) VALE3

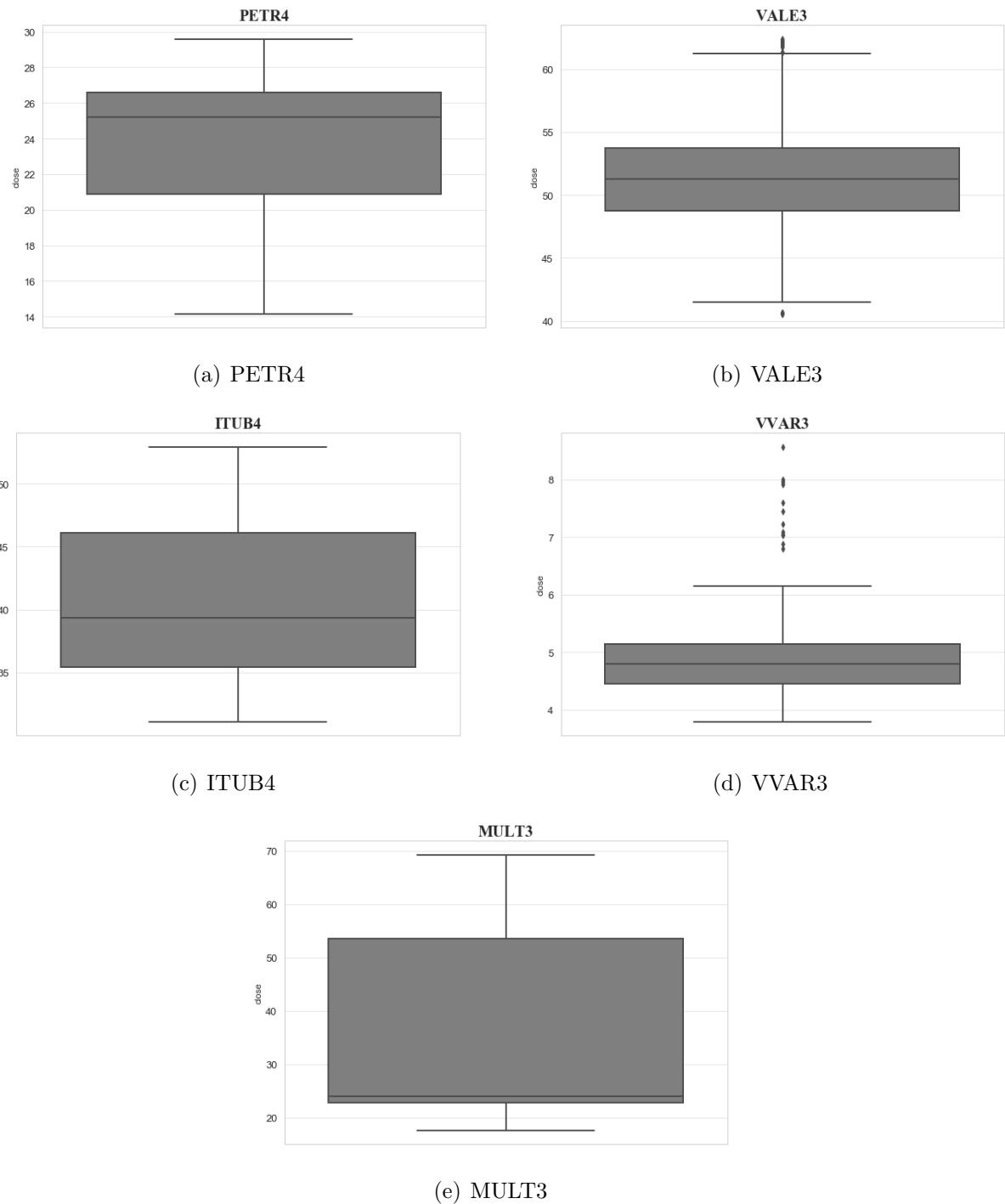
(b) GGBR4

O procedimento de leilão pode ocorrer a qualquer momento do dia, conforme os critérios definidos de cada bolsa de valores, como por exemplo, após a inibição de negociações, devido a de divulgação de fato relevante durante o pregão eletrônico (MOTA, 2013), o leilão decorrente durante dia é um dispositivo de segurança com finalidade de proteger o investidor de oscilações bruscas e distorções para cima ou para baixo, causado

principalmente pelo efeito manada (HIRSHLEIFER; TEOH, 2003).

A figura 17 demonstra a presença de *outliers* nos instrumentos VALE3 e VVAR3 detectados utilizando o método do Quartil (TUKEY, 1977), o capítulo 4.6 descreve o tratamento adequado de *outliers* para o contexto deste trabalho.

Figura 17: Verificação de *outliers* em dados de experimento



4.5 *CUSUM Filter*

O *CUSUM Filter* é um método de controle que tem como finalidade detectar mudança no valor médio dada uma série estacionária e um *threshold* alvo (LAM; YAM, 1997), o método pode monitorar continuamente o erro dos retornos discretos e um aumento significativo do erro é interpretado como uma mudança na distribuição geradas pela amostra ao longo do tempo. Quando uma mudança abrupta é detectada, isto pode ser interpretado como uma possível mudança estrutural da série conforme o *threshold* alvo, quanto menor o valor alvo, maior serão os pontos de mudança identificados na série, conforme ilustra o exemplo da figura 18, os pontos em vermelho são as mudanças detectadas pelo método *CUSUM Filter* utilizando um *threshold* de 0.10.

Em mercados financeiros, Lam e Yam (1997) propõem uma estratégia de negociação utilizando o *CUSUM Filter* como fator de decisão dos sinais de operações de compra e venda, esses autores demonstram que tal estratégia é equivalente ao chamado “Estratégia de negociação por filtro” estudado por Fama e Blume (1966).

Figura 18: Série temporal com *CUSUM Filter*
CUSUM Filter



Fonte: Elaborado pelo autor.

No contexto deste trabalho, o *CUSUM Filter* será utilizado como ponto de entrada

para uma operação, em conjunto com os métodos propostos nas seções 4.6 e 4.7.

4.6 *Meta-Labeling*

O método de rotulagem consiste no mapeamento de exemplos em um conjunto de observações utilizando métodos de ML dentro de um contexto de aprendizado supervisionado, em finanças, é comum encontrar estudos que realizam o mapeamento do alvo baseado um retorno discreto fixo (por exemplo, o algoritmo de ML deve ser capaz classificar um retorno discreto fixo em algum horizonte de tempo x), conforme os estudos realizados por Karthik, Nishanth e Manikandan (2016), Gyamerah, Ngare e Ikpe (2019).

Segundo Prado (2018), há uma consonância de pensamento de que os métodos de ML são interpretados e criticados como verdadeiras caixas-pretas e de difícil entendimento, por isso, este trabalho foi baseado na metodologia de meta-rotulagem (*Meta-Labeling*) proposta por Prado (2018), onde os algoritmos de ML são utilizados para prever o tamanho apropriado a ser negociado no instrumento financeiro, isto possibilita construir um sistema baseado em modelo exógeno fundado na teoria econômica (caixa branca) e o aprendizado de máquina será responsável apenas por predizer o tamanho apropriado da negociação.

Com a utilização do método de *Meta-Labeling*, os efeitos de *overfitting* são limitados (PRADO, 2018) pois o algoritmo de ML não decidirá o lado do mercado, apenas o tamanho, alcançar bons retornos em negociações de lotes pequenos ou obter prejuízos em negociações de lotes grandes pode ser desastroso, por isso a importância de desenvolver um algoritmo de ML exclusivamente focado em obter o dimensionamento adequado, tão importante quanto identificar boas oportunidades.

Ao desacoplar a predição do mercado da predição de dimensionalidade da negociação, pode-se desenvolver um modelo de ML com estratégias exclusivas para posições longas, com base nas recomendações de um modelo primário, e outro modelo com estratégias para posições curtas, baseada nas recomendações de um modelo primário totalmente diferente.

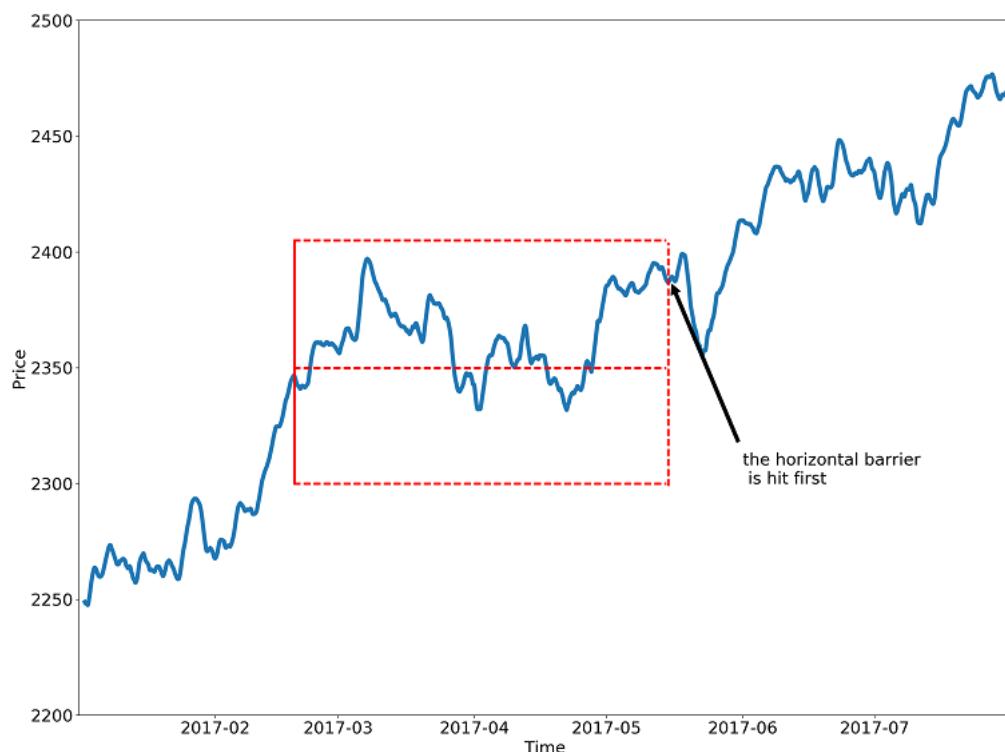
A seção 4.7 irá exemplificar o uso da meta-rotulagem.

4.7 *Triple-Barrier Method*

Quando uma negociação é realizada, os operadores de mercado podem optar por definir preventivamente qual o objetivo daquela operação (preço-alvo ou *take-profit*) e também o quanto está disposto a perder (*stop-loss*) caso a sua análise ou teoria não esteja correta, executando as suas ordens nos preços de *take-profit* e *stop-loss* previamente definidos.

Prado (2018) propõe uma estratégia de negociação denominada de Método de Tripla-Barreira (*Triple-Barrier Method*), que combina a meta-rotulagem com o comportamento real do mercado, isto é, os rótulos são discretizados por barreiras de limites de preço, conforme ilustra a figura 19.

Figura 19: *Triple-Barrier Method*



Fonte: (PRADO, 2018)

Na prática, o método busca definir os limites ganhos e de perdas em uma negociação que são delimitados por meio das barreiras horizontais e são calculadas com base em uma função dinâmica da volatilidade estimada e dos riscos envolvidos em uma negociação,

a barreira vertical pontilhada à direita é definida com base em um tempo de expiração desde que a negociação foi aberta, isto é, quando o alvo do retorno discreto dinâmico de predição do modelo não é atingido antes do tempo de expiração, a operação é encerrada com lucro ou com prejuízo.

O algoritmo de aprendizagem supervisionado realiza a meta-rotulagem da seguinte forma:

- Se a barreira horizontal superior é tocada primeiro desde o início da negociação, o algoritmo de aprendizagem classifica a observação com o rótulo 1.
- Se a barreira horizontal inferior é tocada primeiro desde o início da negociação, o algoritmo de aprendizagem classifica a observação com o rótulo -1.
- Se a barreira vertical da direira é tocada primeiro desde o início da negociação, o algoritmo de aprendizagem classifica a observação com o rótulo do retorno discreto obtido (lucro ou prejuízo).

Os parâmetros utilizados para o método de barreira tripla seguem a sugestão de valores dos exercícios definidos no livro do autor, conforme demonstra a tabela 8.

Tabela 8: Parâmetros utilizados no método *Triple-Barrier Method*

<i>Threshold</i> retorno	<i>CUSUM Filter</i>	Tempo de Exp.	Config. (ptSl)	Período de Volatilidade
0.005	0.03 à 0.05	1 à 5 dias	1,1,1	100

Fonte: Elaborado pelo autor.

A configuração 1,1,1 é padrão utilizada para definir as três barreiras de configuração (duas horizontais e uma vertical à direita, que é utilizada como tempo de expiração da operação).

A utilização do método *Triple-Barrier Method* permite dirimir o risco da operação, já que a operação possui um tempo máximo de duração e alvos de lucro e prejuízo previamente definidos.

4.8 Algoritmos de Aprendizagem

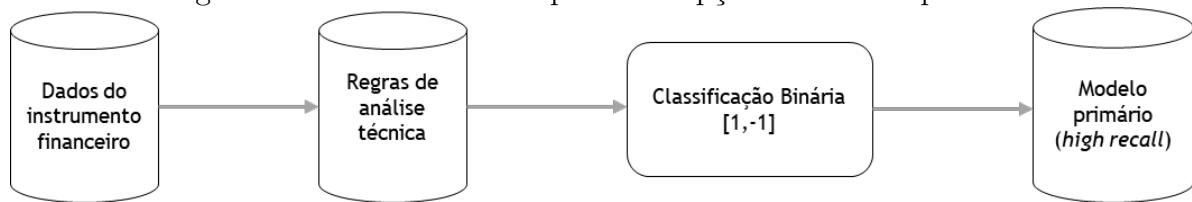
Conforme a definição de Russell e Norvig (2009), os algoritmos de aprendizagem não supervisionados tem a capacidade de aprender padrões complexos em um espaço de alta dimensionalidade, sem necessariamente serem direcionados, já os algoritmos de aprendizagem supervisionados exigem que determinadas observações \mathbf{x} (variável independente) sejam associadas a uma matriz de rótulos ou valores $\hat{\mathbf{y}}$ (variável dependente), de tal modo que estas observações \mathbf{x} possam ser preditas em amostras ainda não observadas.

Neste primeira etapa, o objetivo é obter um modelo primário que atinja um valor satisfatório de *recall* (veja seção 4.8.2), as demais métricas de avaliação não devem ser consideradas como fatores de avaliação de desempenho.

Conforme a discussão abordada por Prado (2018), as tarefas de predição de movimento do instrumento financeiro (ciclo de alta, ciclo de baixa ou ciclo de lateralização), e de predição de retornos discretos devem ser realizadas de forma independente, afim de obter-se especificidade de domínio do problema para cada modelo.

A figura 20 ilustra o fluxo para concepção do modelo primário utilizando regras de análise técnica, conforme descrito na seção 3.2, este modelo é utilizado para definir o lado do instrumento financeiro, e é concebido como modelo exógeno que contém as variáveis de direcionamento do IF utilizadas pelo modelo secundário, sendo: 1 para compra e -1 para venda, neste trabalho foi adotado as teorias da AT, entretanto, pode ser utilizado outras abordagens como a análise fundamentalista, modelos econométricos ou então a adoção de um conjunto de regras de diferentes modelos.

Figura 20: Gráfico do fluxo para concepção do modelo primário



Fonte: Elaborado pelo autor.

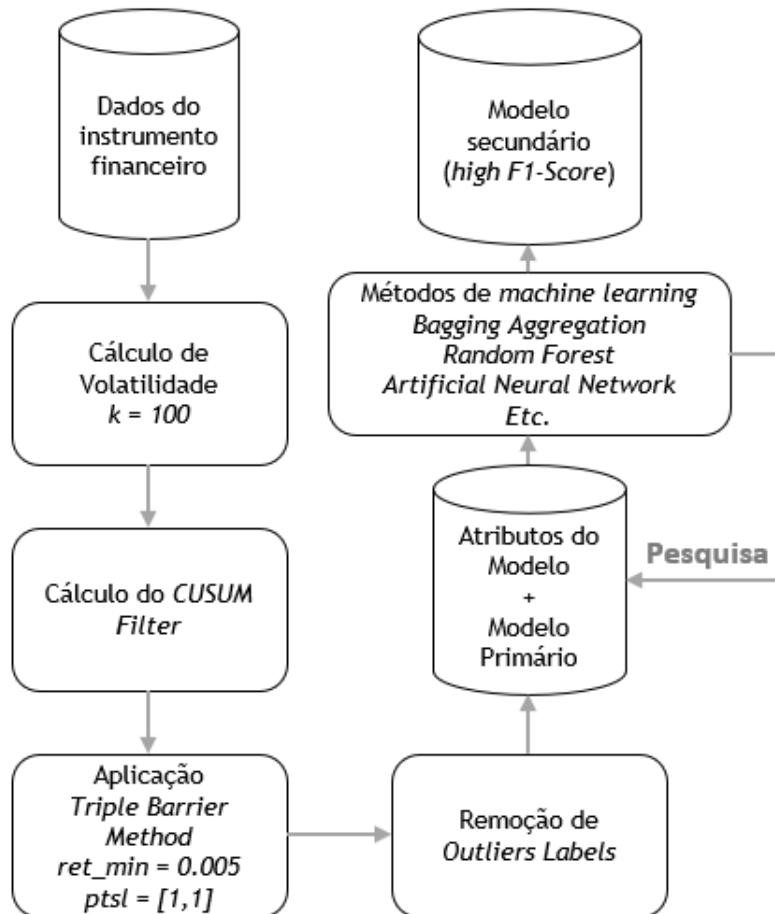
No contexto financeiro, uma abordagem simples para um problema de aprendizado

supervisionado é tentar prever o preço de um instrumento financeiro em algum horizonte fixo no futuro. Observe que essa é uma tarefa de regressão, ou seja, tentamos prever uma variável aleatória contínua.

Este é um problema complexo pois de acordo com Verma e Verma (2007), os preços são notoriamente ruidosos, serialmente correlacionados e o conjunto de todos os valores de preços possíveis é tecnicamente infinito. Por outro lado, podemos abordar isso como um problema de classificação, em vez de prever o preço exato, podemos prever retornos discretos.

A figura 21 ilustra a concepção final do modelo secundário utilizando o método *Triple Barrier Method* descrito na seção 4.7 e de atributos categóricos, neste momento também são removidos as observações raras (*outliers*) presentes no resultado de processamento do *Triple-Barrier Method*.

Figura 21: Visualização do fluxo para concepção do modelo secundário

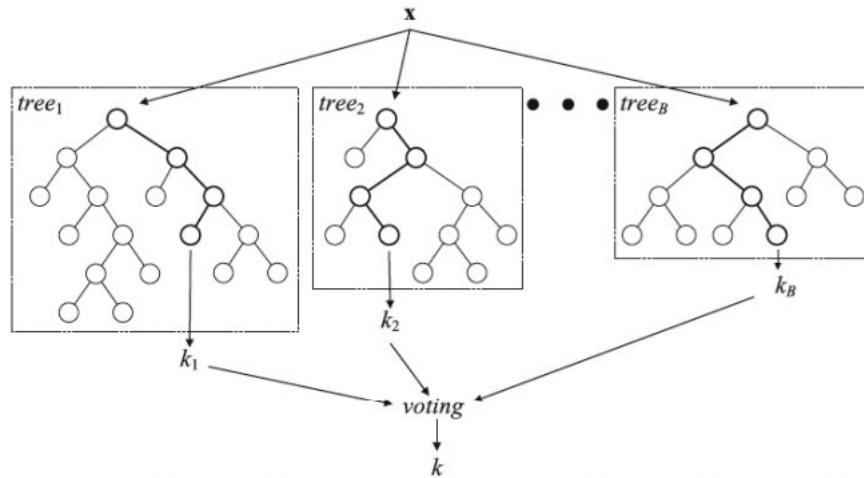


Fonte: Elaborado pelo autor.

4.8.1 Random Forests

Introduzido por Breiman (2001), as *Random Forests* utilizam o método de aprendizado de máquina supervisionado e implementa algoritmos de classificação compostos por *ensembles* de múltiplas árvores de decisão independentes e que dividem os dados em subconjuntos (ramificações) com base nas condições aprendidas para classificar as observações de uma nova instância, as escolhas ótimas locais são calculadas com base no menor erro quadrático local, e as classes finalistas são escolhidas com base em um sistema de votação. Deste modo, as florestas aleatórias reduzem o risco de *overfitting* e determinam soluções ótimas localmente, a figura 22 ilustra o exemplo de construção de um modelo baseado em RF.

Figura 22: Ilustração de construção do modelo *Random Forests*



Fonte: Verikas, Gelzinis e Bacauskiene (2011)

Os atributos categóricos utilizados no modelo secundário foram escolhidos arbitrariamente com referência nas pesquisas de: (CHOUDHRY; GARG, 2008; LIANG et al., 2017; PRADO, 2018) e de acordo com os resultados obtidos durante os testes realizados.

Tabela 9: Atributos utilizados como *inputs* do modelo secundário

#	Nome	Fórmula	Descrição
1	logret	$\ln(\text{close price}_i/\text{close price}_{i-1})$	-
2	mom1	$((\text{close price}_{i-1})/\text{close price}_{i-n})) * 100$	$n=1$
3	mom2	$((\text{close price}_{i-1})/\text{close price}_{i-n})) * 100$	$n=2$
4	mom3	$((\text{close price}_{i-1})/\text{close price}_{i-n})) * 100$	$n=3$
5	mom4	$((\text{close price}_{i-1})/\text{close price}_{i-n})) * 100$	$n=4$
6	mom5	$((\text{close price}_{i-1})/\text{close price}_{i-n})) * 100$	$n=5$
7	corre1	$\ln\left(\frac{\sum_{i=1}^{n-k}(Y_i-\bar{Y})(Y_{i+k}-\bar{Y})}{\sum_{i=1}^n(Y_i-\bar{Y})^2}\right)$	$n=1$
8	corre2	$\ln\left(\frac{\sum_{i=1}^{n-k}(Y_i-\bar{Y})(Y_{i+k}-\bar{Y})}{\sum_{i=1}^n(Y_i-\bar{Y})^2}\right)$	$n=2$
9	corre3	$\ln\left(\frac{\sum_{i=1}^{n-k}(Y_i-\bar{Y})(Y_{i+k}-\bar{Y})}{\sum_{i=1}^n(Y_i-\bar{Y})^2}\right)$	$n=3$
10	corre4	$\ln\left(\frac{\sum_{i=1}^{n-k}(Y_i-\bar{Y})(Y_{i+k}-\bar{Y})}{\sum_{i=1}^n(Y_i-\bar{Y})^2}\right)$	$n=4$
11	corre5	$\ln\left(\frac{\sum_{i=1}^{n-k}(Y_i-\bar{Y})(Y_{i+k}-\bar{Y})}{\sum_{i=1}^n(Y_i-\bar{Y})^2}\right)$	$n=5$
12	logt1	$\ln(\text{close price}_{i-1})$	-
13	logt2	$\ln(\text{close price}_{i-2})$	-
14	logt3	$\ln(\text{close price}_{i-3})$	-
15	logt4	$\ln(\text{close price}_{i-4})$	-
16	logt5	$\ln(\text{close price}_{i-5})$	-
17	vol15	$\ln\left(\sqrt{\frac{\sum(\text{close price}_i - \bar{\text{close price}})^2}{15}}\right)$	-
18	vol30	$\ln\left(\sqrt{\frac{\sum(\text{close price}_i - \bar{\text{close price}})^2}{30}}\right)$	-
19	vol50	$\ln\left(\sqrt{\frac{\sum(\text{close price}_i - \bar{\text{close price}})^2}{50}}\right)$	-
20	side	{1,-1}	Modelo primário exógeno

Fonte: Elaborado pelo autor.

4.8.2 Métricas de Desempenho

As métricas de desempenho dos modelos de previsão são constituídas de fórmulas matemáticas (CHAN, 2009) conforme detalhados a seguir:

- **Matriz de Confusão:** A matriz de confusão é utilizada para representar o es-

tado da classificação dos objetos de uma determinada amostra de dados, utiliza indicadores que contabilizam a quantidade de objetos classificados corretamente ou incorretamente pelo modelo, sendo representada por:

Verdadeiro Positivo (TP): é a quantidade de objetos que foram classificados corretamente com a classe positiva da amostra de dados.

Falso Positivo (FP): é a quantidade de objetos que foram classificados incorretamente com a classe positiva da amostra de dados, entretanto estes objetos pertencem à classe negativa.

Verdadeiro Negativo (TN): é a quantidade de objetos que foram classificados corretamente com a classe negativa da amostra de dados.

Falso Negativo (FN): é a quantidade de objetos que foram classificados incorretamente com a classe negativa da amostra de dados, entretanto estes objetos pertencem à classe positiva.

Por meio do resultado da matriz de confusão, é possível calcular a precisão do modelo, neste caso deve ser considerado que as classes da amostra de dados estão igualmente distribuídas.

$$Matriz = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

- **Acurácia**: é a quantidade de amostras positivas (AP) e negativas (AN) classificadas corretamente dividido pelo total de amostras (TA) da amostra avaliada, conforme a equação 12, quanto maior é a acurácia, menor é o número de falsos positivos do modelo.

$$Acuracia = \frac{AP + AN}{TA} \quad (12)$$

- **Recall**: é a quantidade de amostras positivas (AP) classificadas corretamente sobre o total de amostras classificadas como falsas negativas (FN) mais AP em percentual.

$$Recall = \frac{AP}{FN + AP} \quad (13)$$

- **Precisão:** é o número de previsões corretas (AP) dividido pelo número total de pontos falsos (FP) mais AP em percentual, quanto maior a precisão do modelo, menor é o número de falsos positivos cometidos.

$$Precisao = \frac{AP + AN}{TA} \quad (14)$$

- **F1-score:** calcula a eficiência da média harmônica entre a Precisão e o *Recall*.

$$F1\text{-score} = \frac{2 * Precisao * Recall}{Precisao + Recall} \quad (15)$$

- **Especificidade:** é a quantidade de amostras negativas identificadas corretamente (AN) sobre o total de amostras negativas (TAN).

$$Especificidade = \frac{AN}{TAN} \quad (16)$$

Para a realização das simulações, a base de dados foi dividida em 80% das observações para o período de treinamento do algoritmo e 20% das observações para o período de teste, este valor de referência foi obtido com base nas observações de Jansen (2018), desta forma, os dados foram divididos conforme ilustra a figura 23.

Figura 23: Ilustração da divisão da base de treinamento e teste



Fonte: Elaborado pelo autor.

5 Resultados

Este capítulo demonstra os resultados obtidos com experimentos desenvolvidos durante esta pesquisa.

5.1 Ambiente de Simulação e Testes

A implementação técnica utilizou principalmente os seguintes componentes e sistemas de *software*:

- Linguagem Java 1.9¹⁸;
- Linguagem Python 3.6¹⁹;
- Componente Scikit-learning 0.21.2²⁰;
- Componente Matplotlib 3.1.0 Python²¹;
- Componente Pyfolio 0.9.2 Python²²;
- Componente Numpy 1.15.4 Python²³;
- Componente Pandas Dataframe 0.24.2²⁴;
- Componente Hudson and Thames Quantitative Research²⁵;
- Componente TA-Lib: Technical Analysis Library²⁶;
- Componente HighCharts²⁷;
- Componente amCharts²⁸;

¹⁸<https://www.java.com>

¹⁹<https://www.python.org>

²⁰<https://scikit-learn.org>

²¹<https://matplotlib.org/>

²²<https://github.com/quantopian/pyfolio>

²³<https://numpy.org/>

²⁴<https://pandas.pydata.org>

²⁵<https://github.com/hudson-and-thames/mlfinlab>

²⁶<http://ta-lib.org>

²⁷<https://www.highcharts.com>

²⁸<https://www.amcharts.com>

- SGBD TimescaleDB²⁹;
- Módulo de sincronização de dados para BMF&F Bovespa³⁰.

²⁹<https://www.timescale.com>

³⁰<https://github.com/marretti/stock-market-b3>

5.2 Avaliação dos Resultados

Os resultados demonstrados abaixo estão separados por instrumento financeiro e pelo tipo de estratégia adotada conforme descrito na seção 3.2, que contém o detalhamento das estratégias operacionais.

5.2.1 Bandas de Bollinger - Reversão à Média

- Avaliação de desempenho do instrumento financeiro **VALE3**:

A tabela 10 representa as configurações utilizadas para os testes realizados neste instrumento financeiro.

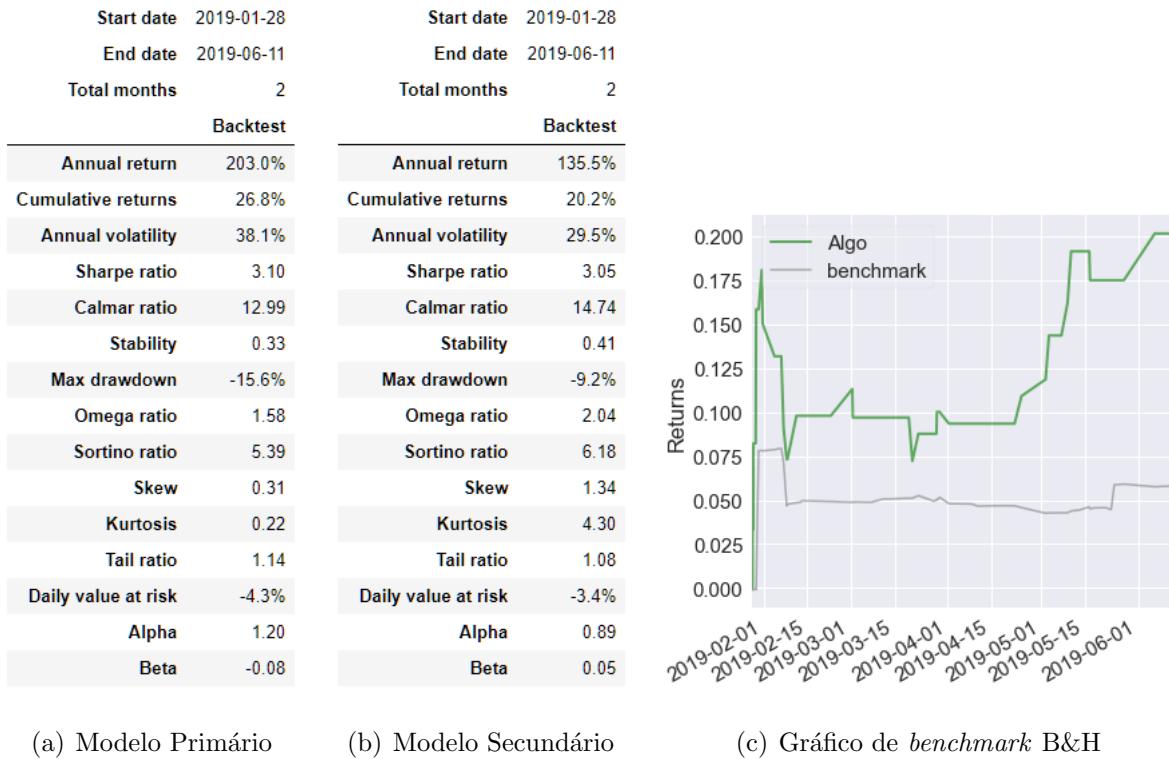
Tabela 10: Parâmetros utilizados no método *Triple-Barrier Method*

<i>Threshold</i> retorno	<i>CUSUM Filter</i>	Tempo de Exp.	Config. (ptSl)	Período de Volatilidade
0.005	0.03191	5 dias	1,1,1	100

Fonte: Elaborado pelo autor.

A figura 24 item (a) ilustra os resultados obtidos pelo modelo primário de AT utilizando o indicador de *Bandas de Bollinger* com a estratégia de reversão à média, o item (b) ilustra os resultados obtidos aplicando o método de *Triple-Barrier Method* e utilizando o modelo primário como entrada exógena, o item (c) ilustra o resultado comparativo entre o modelo secundário e o *benchmark Buy-and-Hold*.

Figura 24: Resultado de *Backtest* para VALE3



Pode-se afirmar que a utilização do *Triple-Barrier Method* demonstrou melhora generalizada nos resultados obtidos para cada indicador conforme ilustra a figura 24 (item (b) Modelo Secundário), apesar de apresentar um retorno acumulado menor, pode-se considerar que houve eficiência na diminuição de prejuízos e na volatilidade histórica.

A figura 25 ilustra os resultados de retornos obtidos durante o período analisado utilizando o modelo secundário, a primeira figura demonstra o *drawdown* no período, as demais visualizações apresentam o retorno mensal e os retornos médios obtidos.

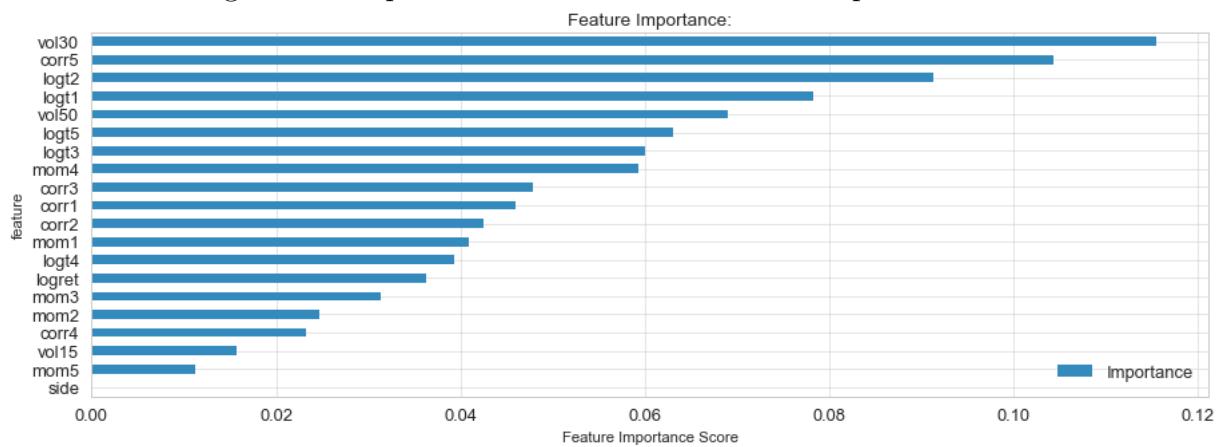
Figura 25: Visualização de retornos para VALE3



Fonte: Elaborado pelo autor.

A figura 26 ilustra o resultado do grau de importância durante o treinamento do modelo ML para cada atribututo utilizado como *input* do modelo secundário.

Figura 26: Importância de atributos do modelo para VALE3

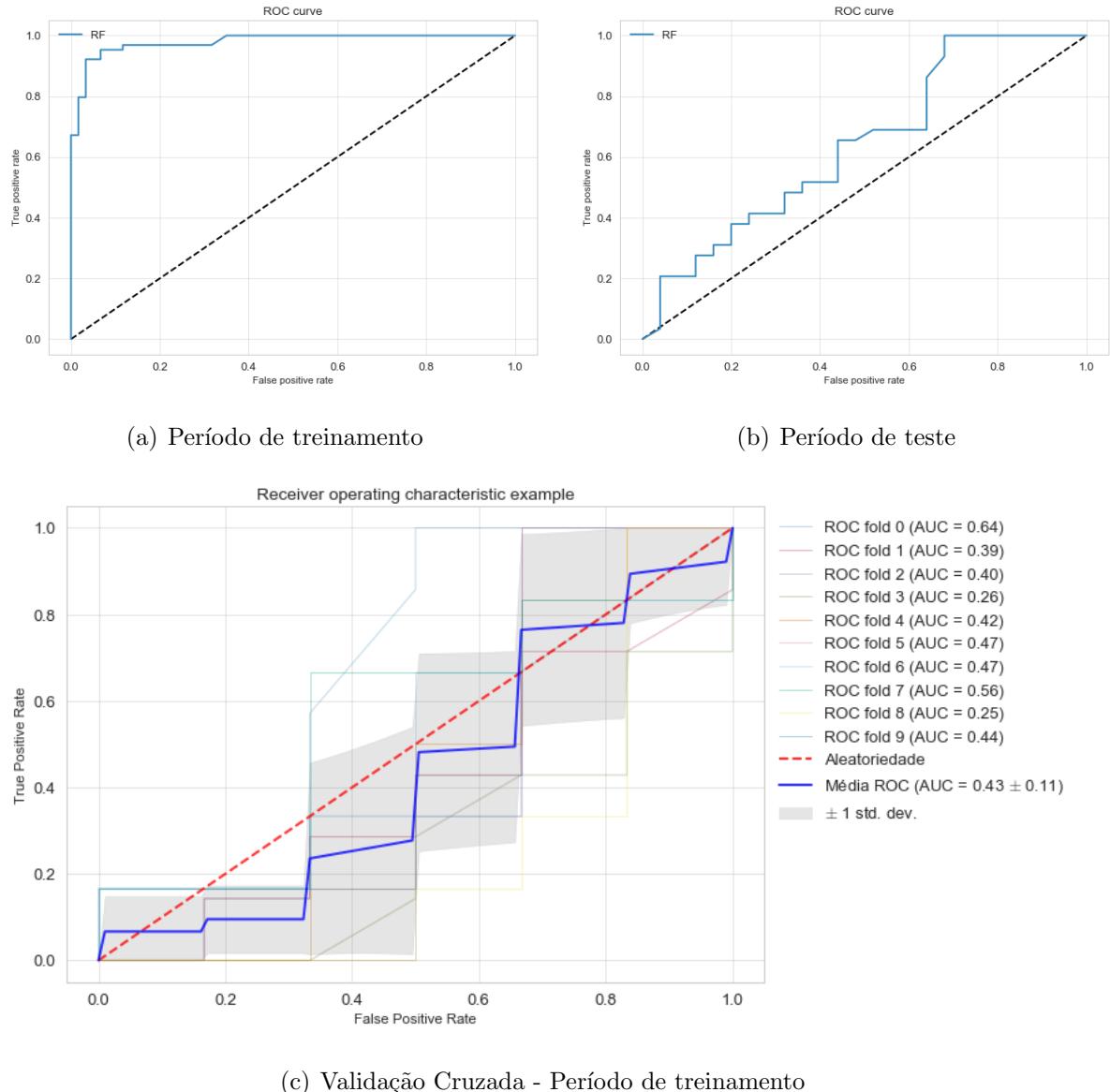


Fonte: Elaborado pelo autor.

A figura 27 ilustra a curva ROC para o período de treinamento - item (a) e para o

período de teste - item (b), e o resultado do teste de validação cruzada - item (c) (para detalhes, veja seção 3.7.3).

Figura 27: Curva ROC para VALE3



Os resultados representados pela figura 27 - item (c) ilustra o teste de validação cruzada realizado na base de treinamento para 10 diferentes períodos, pode-se observar que a média da curva ROC se estabeleceu na maior parte do tempo abaixo da aleatoriedade.

Figura 28: Visualização de desempenho do modelo de ML para VALE3

```

precision    recall   f1-score
0           0.53     0.64     0.58
1           0.62     0.52     0.57
micro avg    0.57     0.57     0.57
macro avg    0.58     0.58     0.57
weighted avg 0.58     0.57     0.57

Confusion Matrix
[[16  9]
 [14 15]]

Accuracy
0.5740740740740741

```

Fonte: Elaborado pelo autor.

A figura 28 ilustra a análise de desempenho do modelo durante o período de treinamento.

- Avaliação de desempenho do instrumento financeiro **PETR4**:

A tabela 11 representa as configurações utilizadas para os testes realizados neste instrumento financeiro.

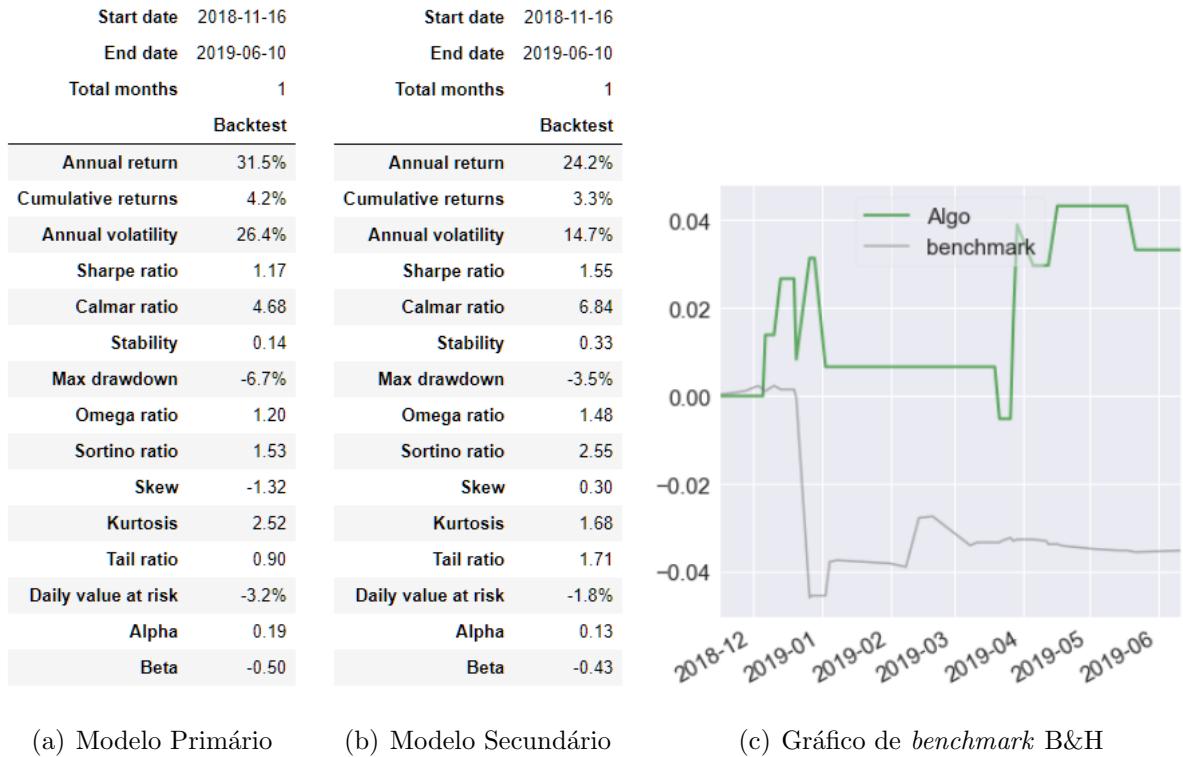
Tabela 11: Parâmetros utilizados no método *Triple-Barrier Method*

<i>Threshold</i> retorno	<i>CUSUM Filter</i>	Tempo de Exp.	Config. (ptSl)	Período de Volatilidade
0.005	0.005	1 dia	1,1,1	100

Fonte: Elaborado pelo autor.

A figura 29 item (a) ilustra os resultados obtidos pelo modelo primário de AT utilizando o indicador de *Bandas de Bollinger* com a estratégia de reversão à média, o item (b) ilustra os resultados obtidos aplicando o método de *Triple-Barrier Method* e utilizando o modelo primário como entrada exógena, o item (c) ilustra o resultado comparativo entre o modelo secundário e o *benchmark Buy-and-Hold*.

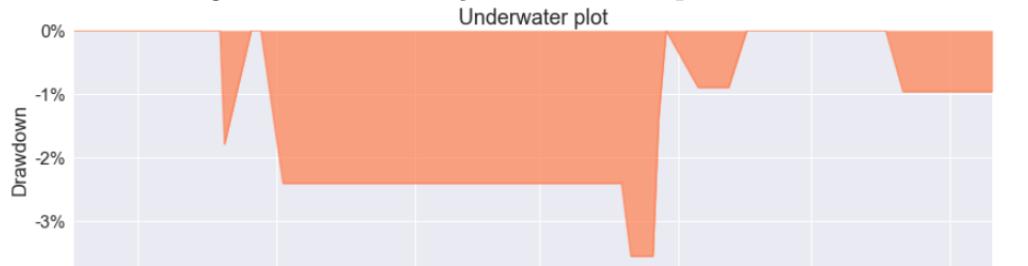
Figura 29: Resultado de *Backtest* para PETR4



Pode-se afirmar que a utilização do *Triple-Barrier Method* demonstrou melhora generalizada nos resultados obtidos para cada indicador conforme ilustra a figura 29 (item (b) Modelo Secundário), apesar de apresentar um retorno acumulado menor, pode-se considerar que houve eficiência na diminuição de prejuízos e na volatilidade histórica.

A figura 30 ilustra os resultados de retornos obtidos durante o período analisado utilizando o modelo secundário, a primeira figura demonstra o *drawdown* no período, as demais visualizações apresentam o retorno mensal e os retornos médios obtidos.

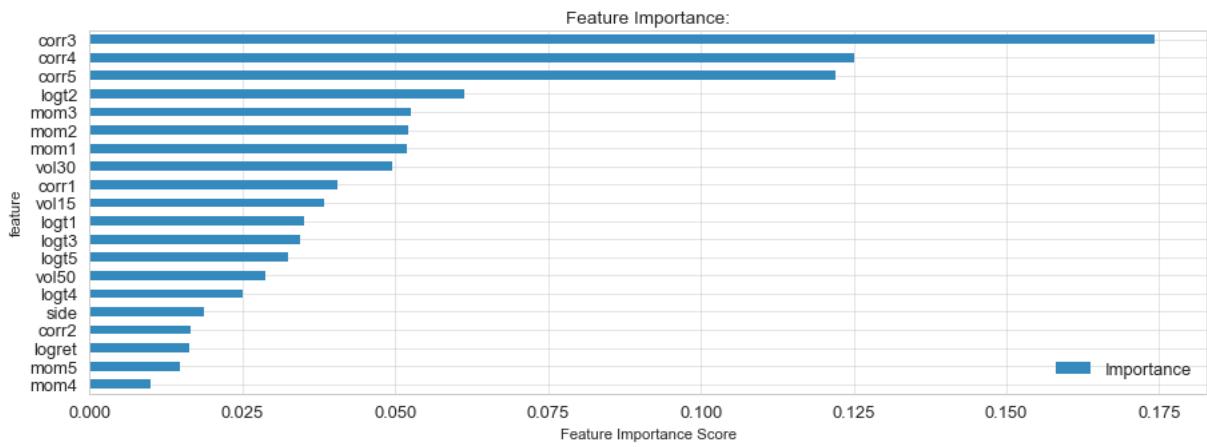
Figura 30: Visualização de retornos para PETR4



Fonte: Elaborado pelo autor.

A figura 31 ilustra o resultado do grau de importância durante o treinamento do modelo ML para cada atribututo utilizado como *input* do modelo secundário.

Figura 31: Importância de atributos do modelo para PETR4

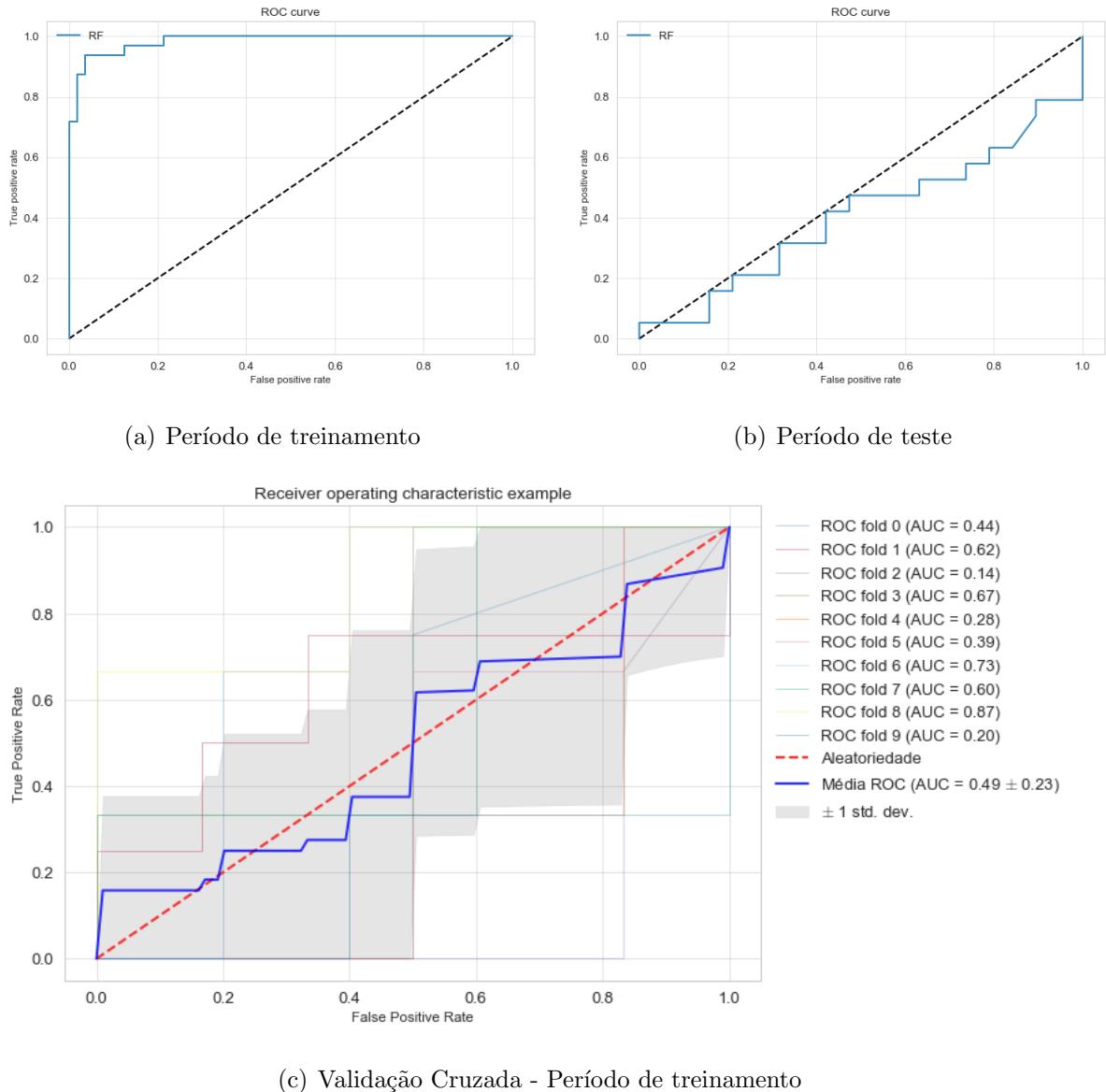


Fonte: Elaborado pelo autor.

A figura 32 ilustra a curva ROC para o período de treinamento - item (a) e para o período de teste - item (b), e o resultado do teste de validação cruzada - item (c)

(para detalhes, veja seção 3.7.3).

Figura 32: Curva ROC para PETR4



Os resultados representados pela figura 32 - item (c) ilustra o teste de validação cruzada realizado na base de treinamento para 10 diferentes períodos, pode-se observar que a média da curva ROC se estabeleceu na maior parte do tempo abaixo da aleatoriedade.

A figura 33 ilustra a análise de desempenho do modelo durante o período de treinamento.

Figura 33: Visualização de desempenho do modelo de ML para PETR4

```

precision    recall  f1-score

0            0.50    0.68    0.58
1            0.50    0.32    0.39

   micro avg     0.50    0.50    0.50
   macro avg     0.50    0.50    0.48
weighted avg     0.50    0.50    0.48

Confusion Matrix
[[13  6]
 [13  6]]

Accuracy
0.5

```

Fonte: Elaborado pelo autor.

- Avaliação de desempenho do instrumento financeiro **ITUB4**:

A tabela 12 representa as configurações utilizadas para os testes realizados neste instrumento financeiro.

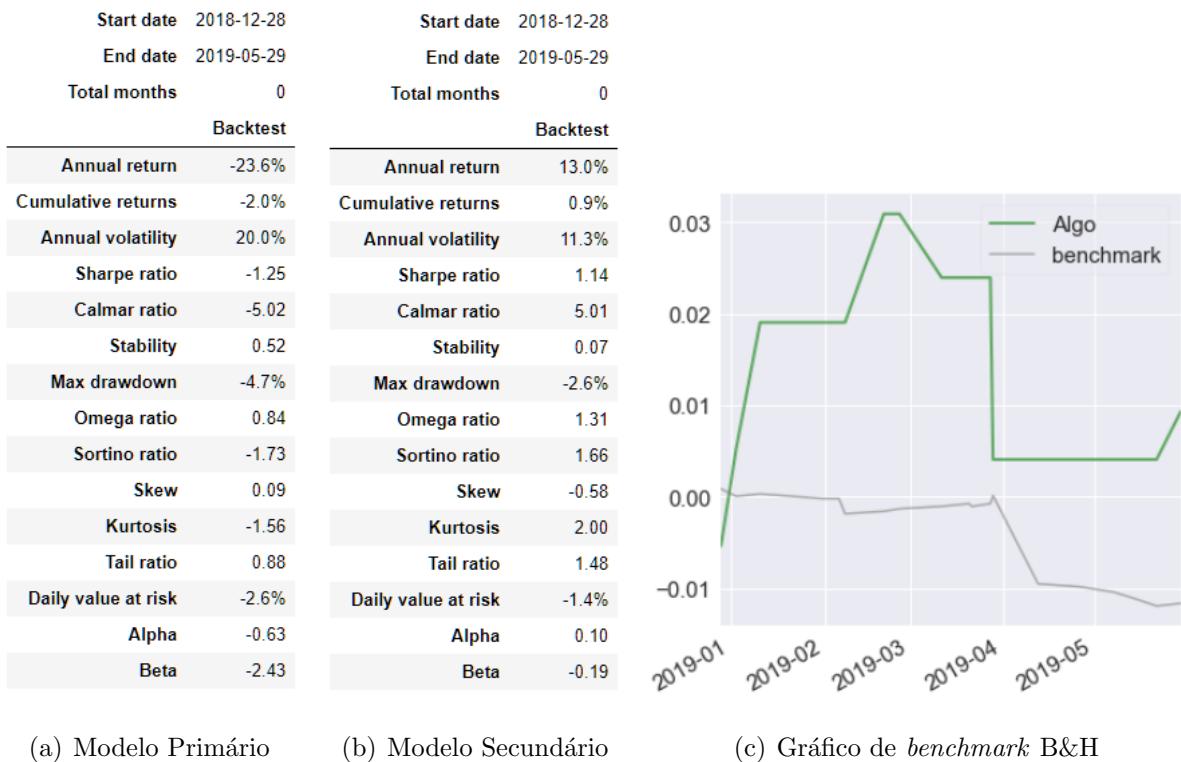
Tabela 12: Parâmetros utilizados no método *Triple-Barrier Method*

<i>Threshold</i> retorno	<i>CUSUM Filter</i>	Tempo de Exp.	Config. (ptSl)	Período de Volatilidade
0.005	0.05	1 dia	1,1,1	100

Fonte: Elaborado pelo autor.

A figura 34 item (a) ilustra os resultados obtidos pelo modelo primário de AT utilizando o indicador de *Bandas de Bollinger* com a estratégia de reversão à média, o item (b) ilustra os resultados obtidos aplicando o método de *Triple-Barrier Method* e utilizando o modelo primário como entrada exógena, o item (c) ilustra o resultado comparativo entre o modelo secundário e o *benchmark Buy-and-Hold*.

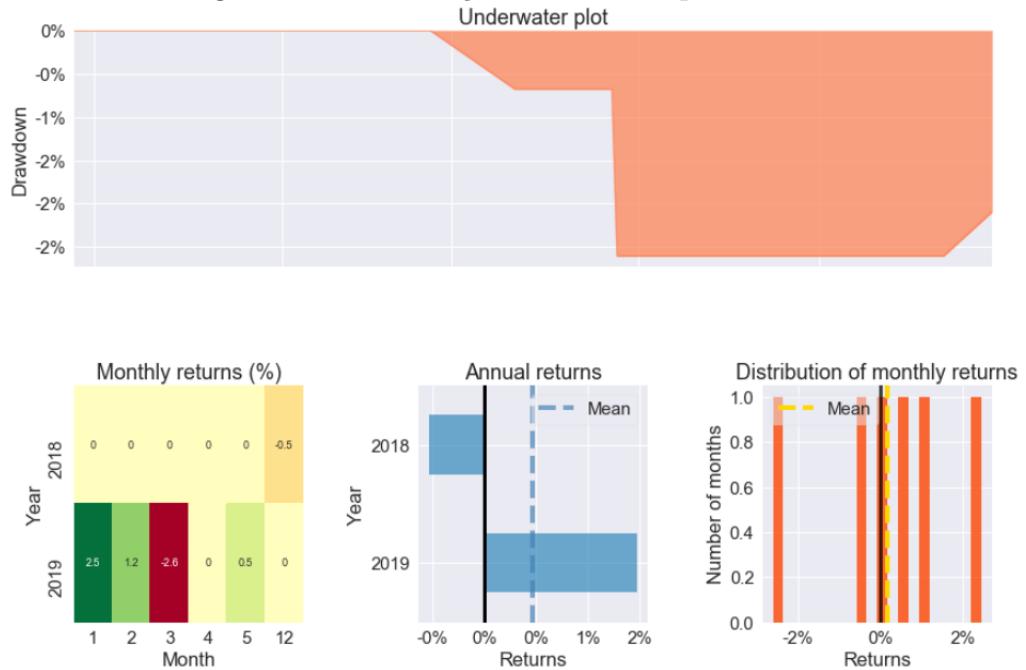
Figura 34: Resultado de *Backtest* para ITUB4



Pode-se afirmar que a utilização do *Triple-Barrier Method* demonstrou melhora generalizada nos resultados obtidos para cada indicador conforme ilustra a figura 34 (item (b) Modelo Secundário), apresentado inclusive um retorno acumulado maior ao *buy-and-hold*, pode-se considerar que houve eficiência na diminuição de prejuízos, na volatilidade histórica e aumento no retorno acumulado.

A figura 35 ilustra os resultados de retornos obtidos durante o período analisado utilizando o modelo secundário, a primeira figura demonstra o *drawdown* no período, as demais visualizações apresentam o retorno mensal e os retornos médios obtidos.

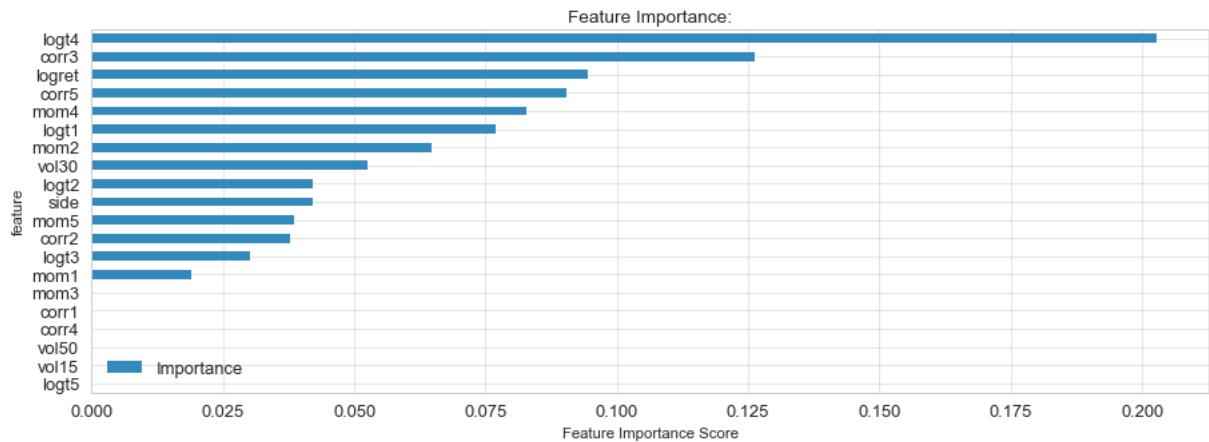
Figura 35: Visualização de retornos para ITUB4



Fonte: Elaborado pelo autor.

A figura 36 ilustra o resultado do grau de importância durante o treinamento do modelo ML para cada atribututo utilizado como *input* do modelo secundário.

Figura 36: Importância de atributos do modelo para ITUB4

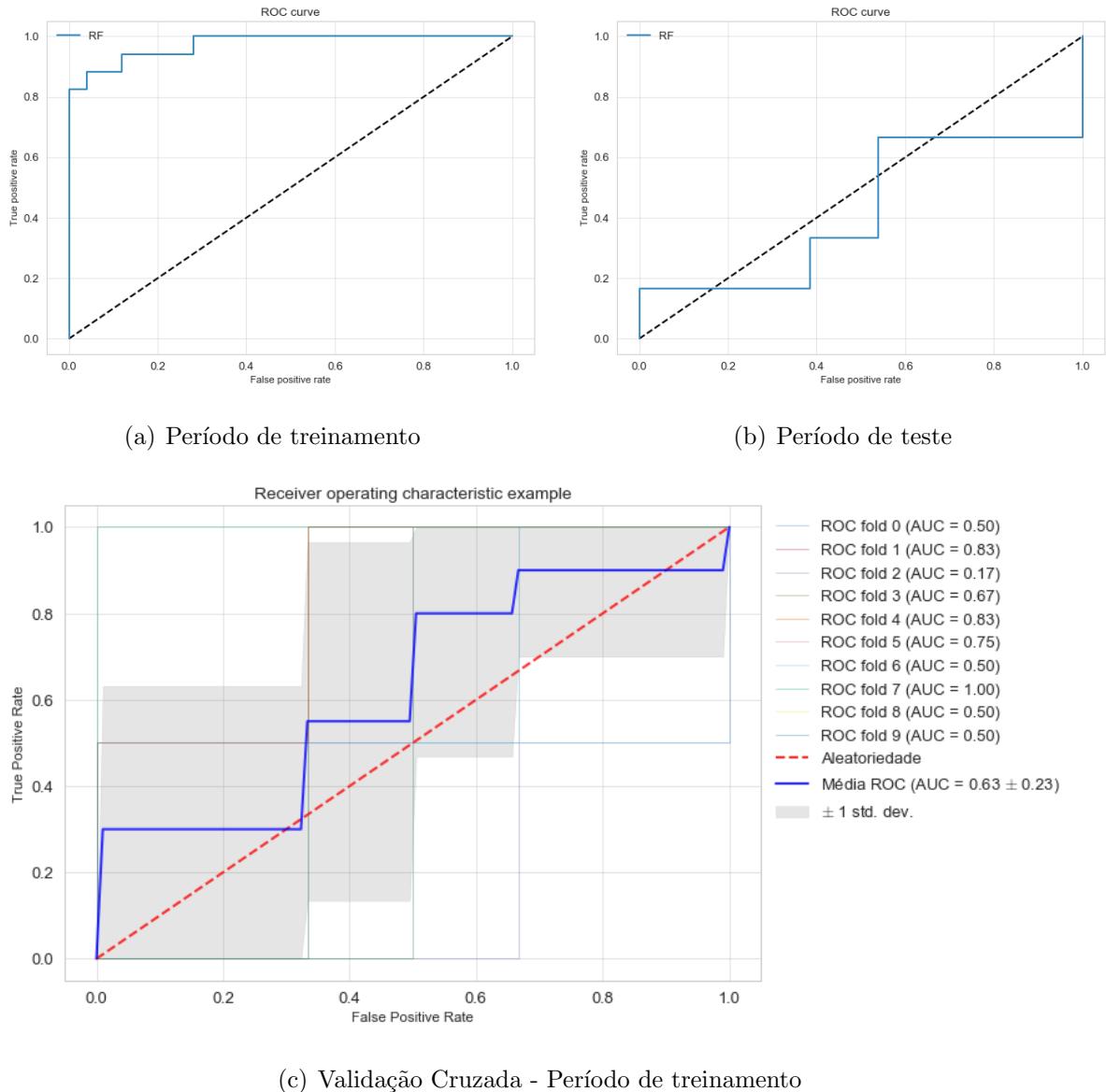


Fonte: Elaborado pelo autor.

A figura 37 ilustra a curva ROC para o período de treinamento - item (a) e para o período de teste - item (b), e o resultado do teste de validação cruzada - item (c)

(para detalhes, veja seção 3.7.3).

Figura 37: Curva ROC para ITUB4



Os resultados representados pela figura 37 - item (c) ilustra o teste de validação cruzada realizado na base de treinamento para 10 diferentes períodos, pode-se observar que a média da curva ROC se estabeleceu na maior parte do tempo acima da aleatoriedade.

Figura 38: Visualização de desempenho do modelo de ML para ITUB4

```

precision    recall   f1-score
0            0.67     0.62     0.64
1            0.29     0.33     0.31

micro avg    0.53     0.53     0.53
macro avg    0.48     0.47     0.47
weighted avg  0.55     0.53     0.54

Confusion Matrix
[[8 5]
 [4 2]]

Accuracy
0.5263157894736842

```

Fonte: Elaborado pelo autor.

A figura 38 ilustra a análise de desempenho do modelo durante o período de treinamento.

- Avaliação de desempenho do instrumento financeiro **MULT3**:

A tabela 13 representa as configurações utilizadas para os testes realizados neste instrumento financeiro.

Tabela 13: Parâmetros utilizados no método *Triple-Barrier Method*

<i>Threshold</i> retorno	<i>CUSUM Filter</i>	Tempo de Exp.	Config. (ptSl)	Período de Volatilidade
0.005	0.03	3 dias	1,1,1	100

Fonte: Elaborado pelo autor.

A figura 39 item (a) ilustra os resultados obtidos pelo modelo primário de AT utilizando o indicador de *Bandas de Bollinger* com a estratégia de reversão à média, o item (b) ilustra os resultados obtidos aplicando o método de *Triple-Barrier Method* e utilizando o modelo primário como entrada exógena, o item (c) ilustra o resultado comparativo entre o modelo secundário e o *benchmark Buy-and-Hold*.

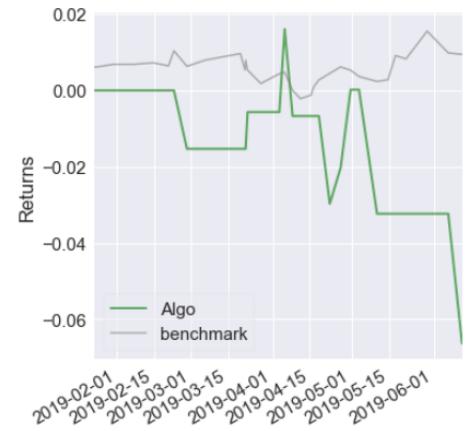
Figura 39: Resultado de *Backtest* para MULT3

	Start date	2019-01-23		Start date	2019-01-23
	End date	2019-06-11		End date	2019-06-11
	Total months	1		Total months	1
	Backtest				
Annual return	-68.1%		Annual return	-42.8%	
Cumulative returns	-13.1%		Cumulative returns	-6.6%	
Annual volatility	36.5%		Annual volatility	19.5%	
Sharpe ratio	-2.95		Sharpe ratio	-2.76	
Calmar ratio	-3.13		Calmar ratio	-5.27	
Stability	0.00		Stability	0.45	
Max drawdown	-21.7%		Max drawdown	-8.1%	
Omega ratio	0.65		Omega ratio	0.48	
Sortino ratio	-3.51		Sortino ratio	-3.16	
Skew	-0.27		Skew	-0.99	
Kurtosis	-1.09		Kurtosis	1.66	
Tail ratio	0.56		Tail ratio	0.55	
Daily value at risk	-5.0%		Daily value at risk	-2.7%	
Alpha	-1.07		Alpha	-0.59	
Beta	-0.09		Beta	0.63	

(a) Modelo Primário

	Start date	2019-01-23		Start date	2019-01-23
	End date	2019-06-11		End date	2019-06-11
	Total months	1		Total months	1
	Backtest				
Annual return	-42.8%		Annual return	-42.8%	
Cumulative returns	-6.6%		Cumulative returns	-6.6%	
Annual volatility	19.5%		Annual volatility	19.5%	
Sharpe ratio	-2.76		Sharpe ratio	-2.76	
Calmar ratio	-5.27		Calmar ratio	-5.27	
Stability	0.45		Stability	0.45	
Max drawdown	-8.1%		Max drawdown	-8.1%	
Omega ratio	0.48		Omega ratio	0.48	
Sortino ratio	-3.16		Sortino ratio	-3.16	
Skew	-0.99		Skew	-0.99	
Kurtosis	1.66		Kurtosis	1.66	
Tail ratio	0.55		Tail ratio	0.55	
Daily value at risk	-2.7%		Daily value at risk	-2.7%	
Alpha	-0.59		Alpha	-0.59	
Beta	0.63		Beta	0.63	

(b) Modelo Secundário



(c) Gráfico de *benchmark* B&H

Pode-se afirmar que a utilização do *Triple-Barrier Method* demonstrou melhora generalizada nos resultados obtidos para cada indicador conforme ilustra a figura 39 (item (b) Modelo Secundário), apesar de apresentar um retorno acumulado menor, pode-se considerar que houve eficiência na diminuição de prejuízos e na volatilidade histórica.

A figura 40 ilustra os resultados de retornos obtidos durante o período analisado utilizando o modelo secundário, a primeira figura demonstra o *drawdown* no período, as demais visualizações apresentam o retorno mensal e os retornos médios obtidos.

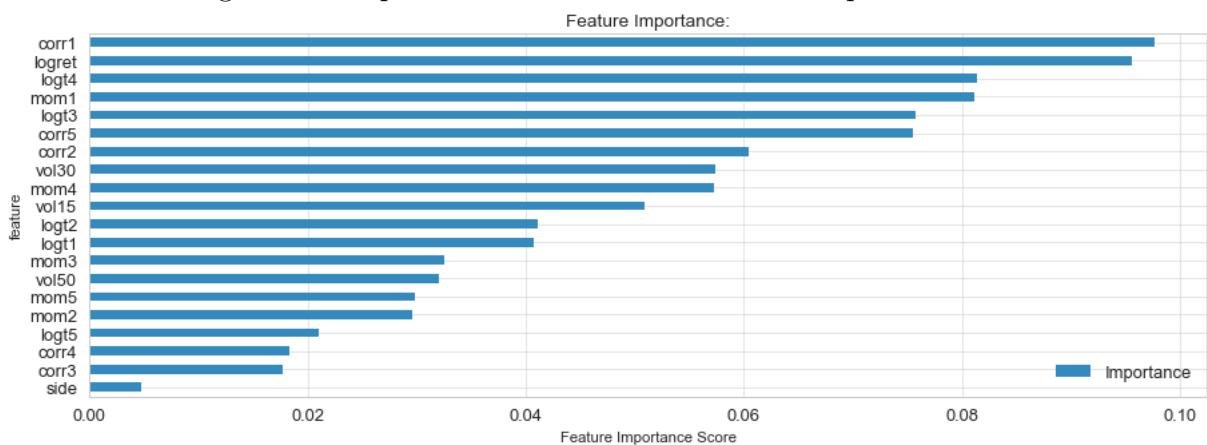
Figura 40: Visualização de retornos para MULT3



Fonte: Elaborado pelo autor.

A figura 41 ilustra o resultado do grau de importância durante o treinamento do modelo ML para cada atributo utilizado como *input* do modelo secundário.

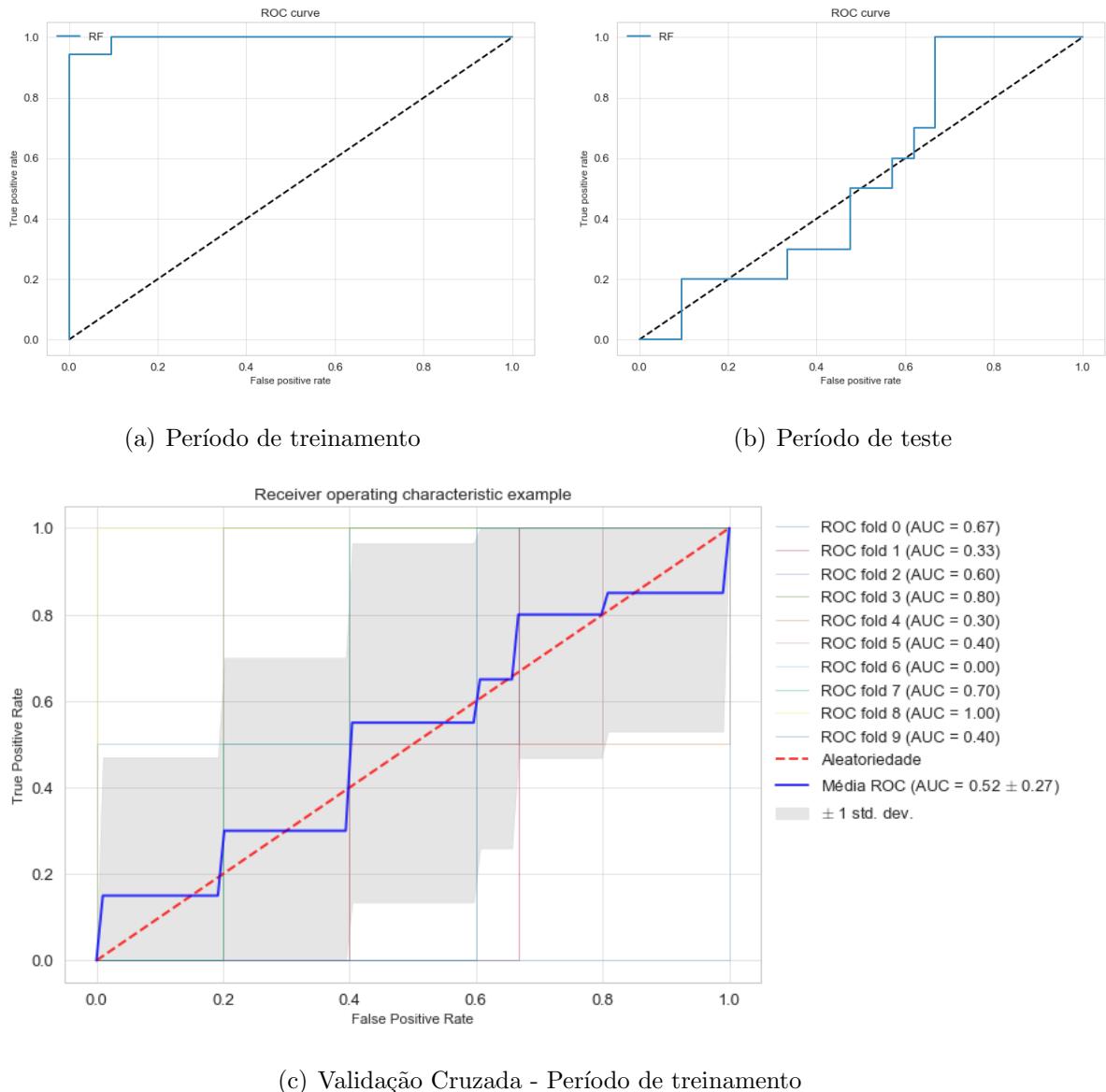
Figura 41: Importância de atributos do modelo para MULT3



Fonte: Elaborado pelo autor.

A figura 42 ilustra a curva ROC para o período de treinamento - item (a) e para o período de teste - item (b), e o resultado do teste de validação cruzada - item (c) (para detalhes, veja seção 3.7.3).

Figura 42: Curva ROC para MULT3



Os resultados representados pela figura 42 - item (c) ilustra o teste de validação cruzada realizado na base de treinamento para 10 diferentes períodos, pode-se observar que a média da curva ROC se estabeleceu na maior parte do tempo acima da aleatoriedade.

Figura 43: Visualização de desempenho do modelo de ML para MULT3

```

precision    recall   f1-score
0           0.64     0.67     0.65
1           0.22     0.20     0.21

micro avg    0.52     0.52     0.52
macro avg    0.43     0.43     0.43
weighted avg  0.50     0.52     0.51

Confusion Matrix
[[14  7]
 [ 8  2]]

Accuracy
0.5161290322580645

```

Fonte: Elaborado pelo autor.

A figura 43 ilustra a análise de desempenho do modelo durante o período de treinamento.

- Avaliação de desempenho do instrumento financeiro **VVAR3**:

A tabela 14 representa as configurações utilizadas para os testes realizados neste instrumento financeiro.

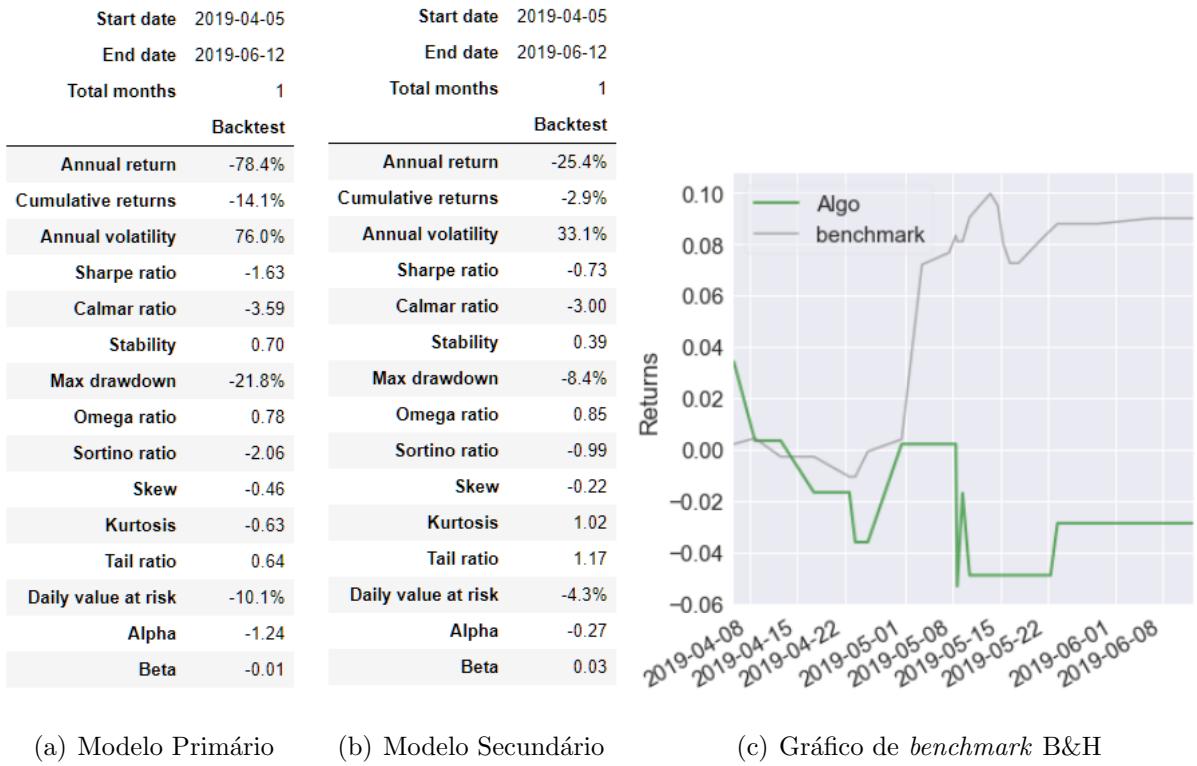
Tabela 14: Parâmetros utilizados no método *Triple-Barrier Method*

<i>Threshold</i> retorno	<i>CUSUM Filter</i>	Tempo de Exp.	Config. (ptSl)	Período de Volatilidade
0.005	0.03939	3 dias	1,1,1	100

Fonte: Elaborado pelo autor.

A figura 44 item (a) ilustra os resultados obtidos pelo modelo primário de AT utilizando o indicador de *Bandas de Bollinger* com a estratégia de reversão à média, o item (b) ilustra os resultados obtidos aplicando o método de *Triple-Barrier Method* e utilizando o modelo primário como entrada exógena, o item (c) ilustra o resultado comparativo entre o modelo secundário e o *benchmark Buy-and-Hold*.

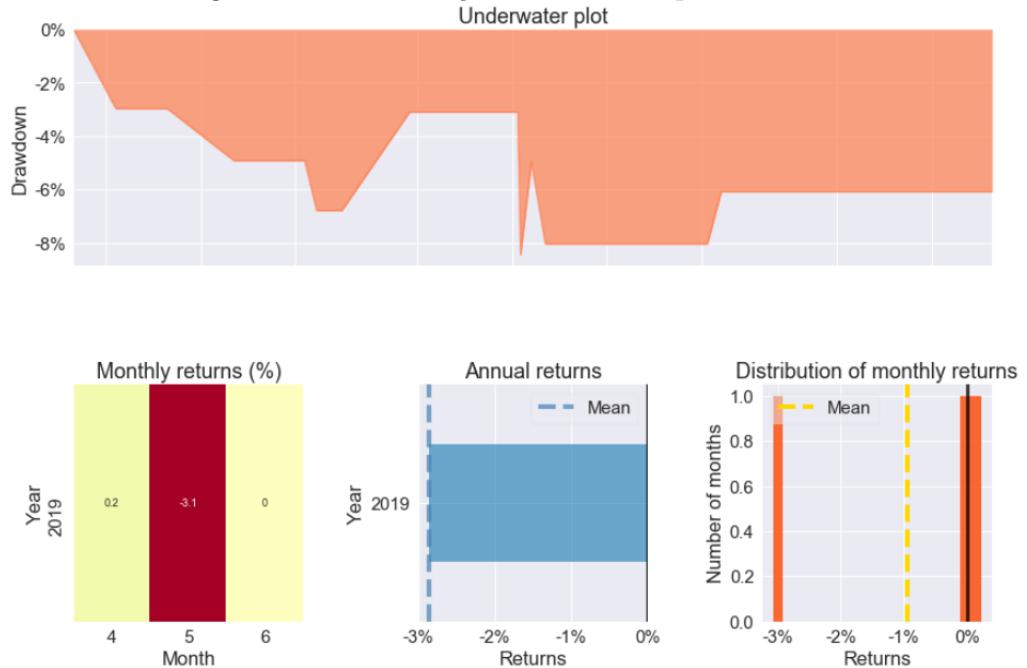
Figura 44: Resultado de *Backtest* para VVAR3



Pode-se afirmar que a utilização do *Triple-Barrier Method* demonstrou melhora generalizada nos resultados obtidos para cada indicador conforme ilustra a figura 44 (item (b) Modelo Secundário), apesar de apresentar um retorno acumulado menor, pode-se considerar que houve eficiência na diminuição de prejuízos e na volatilidade histórica.

A figura 45 ilustra os resultados de retornos obtidos durante o período analisado utilizando o modelo secundário, a primeira figura demonstra o *drawdown* no período, as demais visualizações apresentam o retorno mensal e os retornos médios obtidos.

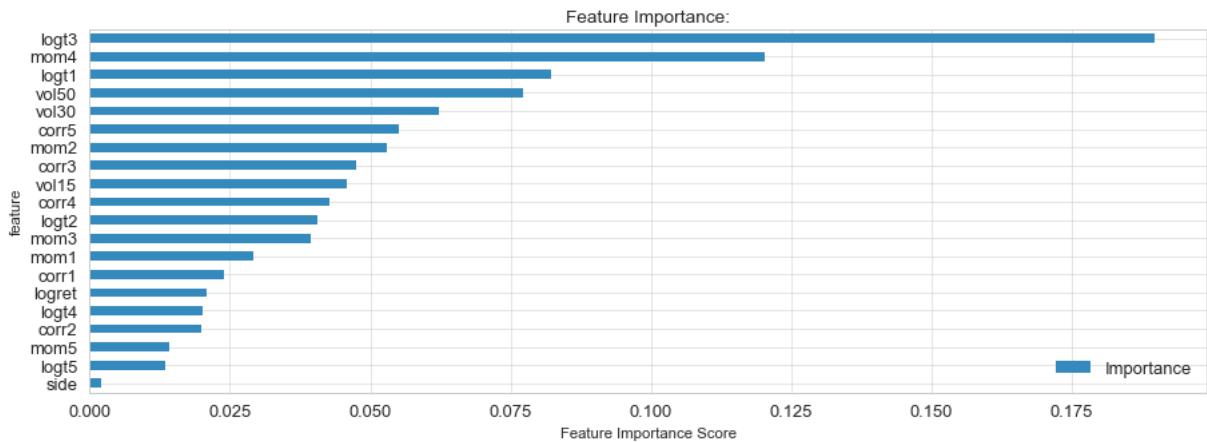
Figura 45: Visualização de retornos para VVAR3



Fonte: Elaborado pelo autor.

A figura 46 ilustra o resultado do grau de importância durante o treinamento do modelo ML para cada atributo utilizado como *input* do modelo secundário.

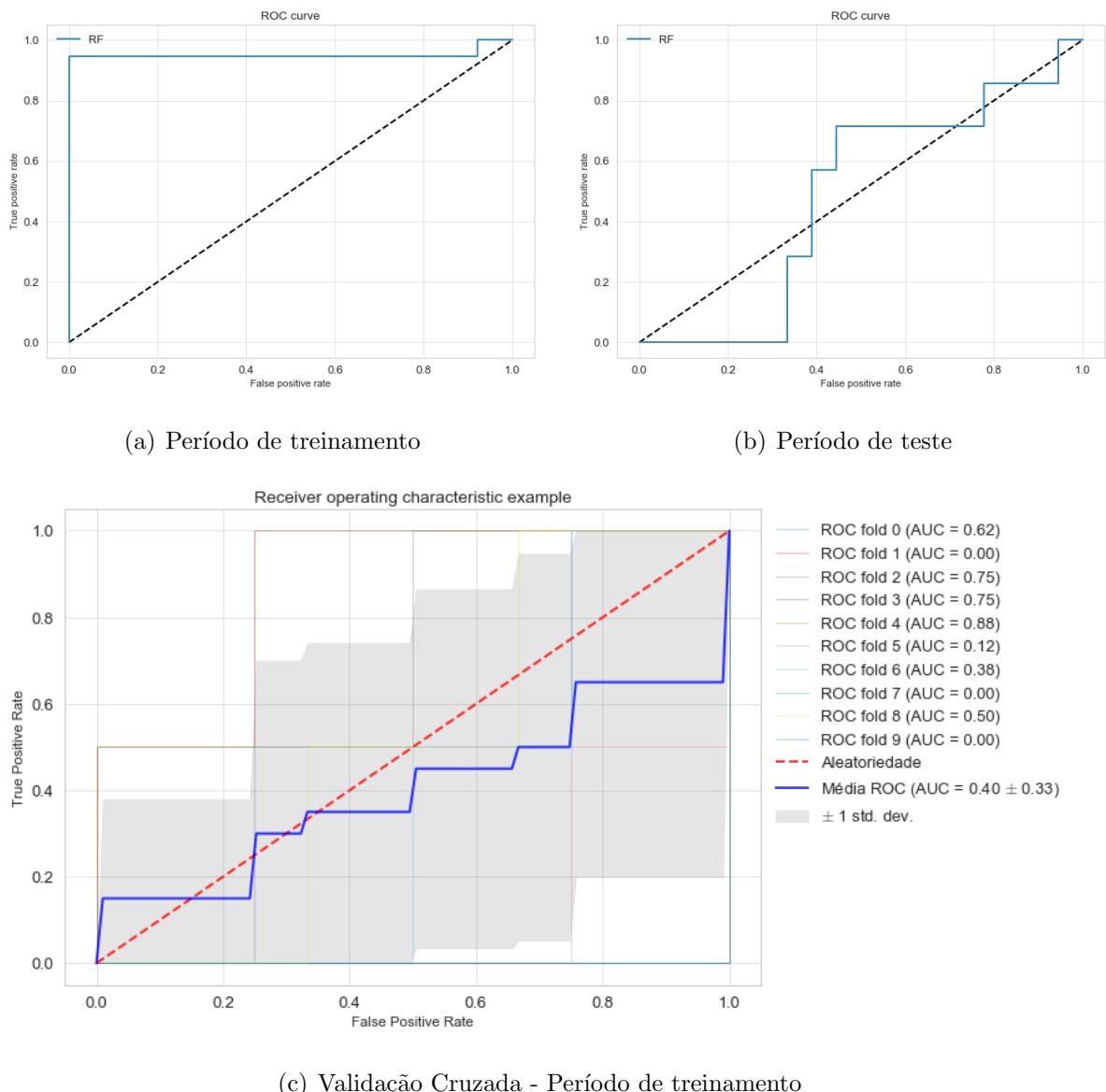
Figura 46: Importância de atributos do modelo para VVAR3



Fonte: Elaborado pelo autor.

A figura 47 ilustra a curva ROC para o período de treinamento - item (a) e para o período de teste - item (b), e o resultado do teste de validação cruzada - item (c) (para detalhes, veja seção 3.7.3).

Figura 47: Curva ROC para VVAR3



Os resultados representados pela figura 47 - item (c) ilustra o teste de validação cruzada realizado na base de treinamento para 10 diferentes períodos, pode-se observar que a média da curva ROC se estabeleceu na maior parte do tempo abaixo da aleatoriedade.

Figura 48: Visualização de desempenho do modelo de ML para VVARD3

```
precision    recall  f1-score

      0       0.69      0.61      0.65
      1       0.22      0.29      0.25

  micro avg       0.52      0.52      0.52
  macro avg       0.45      0.45      0.45
weighted avg     0.56      0.52      0.54

Confusion Matrix
[[11  7]
 [ 5  2]]

Accuracy
0.52
```

Fonte: Elaborado pelo autor.

A figura 48 ilustra a análise de desempenho do modelo durante o período de treinamento.

6 Considerações Finais

Com as recentes técnicas de ML elaboradas por Prado (2018), buscou-se refletir resultados objetivos com a utilização e combinação de modelos desenvolvidos durante a pesquisa no mercado brasileiro de ações.

Os instrumentos financeiros quais o método não foi capaz de superar o *benchmark buy-and-hold* foram MULT3 e VVAR3, acredita-se que pela baixo volume e liquidez destes instrumentos, se comparado os demais, foram os fatores levaram à obtenção de retornos abaixo do esperado, ocasiando o treinamento e predição de observações insignificantes, os demais instrumentos apresentaram *performance* acima do *benchmark buy-and-hold* durante o período estudado.

Um ponto observado durante a pesquisa é de que houve redução unânime da volatilidade e de indicadores cujo demonstram a relação de risco em todos os testes realizados, utilizando o método *Triple-Barrier Method* (PRADO, 2018).

A diminuição dos valores para os indicadores de risco é reflexo devido ao ajuste do tamanho do lote para cada operação, para os mesmos sinais de entrada e saída, concebido pelo modelo de ML que busca a discretização de alvo dinâmico e a limitação do risco por barreiras de limites.

Isto representa um avanço na utilização de métodos de inteligência artificial em finanças, apesar do objetivo deste trabalho foi realizar comparativo de retornos, a redução dos indicadores de risco é um ponto que pode ser observado sob o prisma de *compliance* de uma instituição financeira.

Esta pesquisa teve uma questão referente a limitação de dados, motivo pelo qual o período de análise iniciou-se em 2018, plataformas como Bloomberg³¹ ou Reuters³² não fornecem dados do *intraday*, por este motivo o escopo deste trabalho foi delimitado aos arquivos obtidos do FTP³³ da B3, sendo que os arquivos mais antigos são removidos de tempos em tempos.

A utilização de uma base de dados com um período maior de tempo permite estender

³¹<https://www.bloomberg.com/>

³²<https://www.reuters.com/>

³³<ftp://ftp.bmf.com.br/MarketData/>

este trabalho para a utilização de técnicas de *Embargo* e também fomentar o uso de técnicas de *deep learning* com RNA que exigem uma amostragem de dados significativa para a obtenção de resultados robustos.

É importante lembrar que qualquer processo aqui apresentado é apenas uma orientação sobre o comportamento dos instrumentos financeiros, não devendo ser tratado como algo isolado para fins operacionais, segundo Gil (2008), nem sempre é possível a realização de pesquisas rigidamente explicativas nas ciências sociais, sabe-se o quanto o mercado é variável em razão de aspectos políticos, sociais e econômicos e de difícil controle e previsão de impacto que pode causar nos instrumentos financeiros.

6.1 Trabalhos Futuros

Os resultados apresentados demonstram uma estratégia baseado no modelo exágono de regras de AT utilizando o indicador de Bandas de Bollinger, esta pesquisa pode ser estendida utilizando outros indicadores e regras de AT, conforme detalhado na seção 3.2.

A metodologia proposta é genérica e pode ser aplicada para diferentes métodos de ML, com a flexibilidade de escolher diferentes indicadores para compor o conjunto de atributos de entrada para o modelo, além de ser possível avaliar diferentes técnicas de aprendizado de máquina.

Um ponto importante a ser observado e que pode ser tratado como trabalho futuro é tentar entender se os retornos dos instrumentos MULT3 e VVAR3 estão relacionados à amostragem insignificante de dados.

A base de dados utilizada foi por um período menor de 2 anos, talvez a utilização de um período maior pode demonstrar resultados mais significantes, pois a base de treinamento do modelo poderá ter uma amostragem maior de dados.

Uma sugestão para extensão desta pesquisa é a utilização de métodos de ML utilizando RNA em conjunto com técnicas de *ensemble* para a realização de estudo comparativo utilizando a abordagem do *Triple-Barrier Method* apresentada por Prado (2018).

Este método também poderia ser utilizado para realizar gerenciamento de portfólio de ações.

Referências Bibliográficas

- AKANSU, A. N.; KULKARNI, S. R.; MALIOUTOV, D. M. *Financial Signal Processing and Machine Learning*. [S.l.]: John Wiley & Sons, 2016.
- AKITA, R. et al. Deep learning for stock prediction using numerical and textual information. In: IEEE. *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*. [S.l.], 2016. p. 1–6.
- ALDRIDGE, I. *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems*. [S.l.]: Wiley, 2009. (Wiley Trading). ISBN 9780470579770.
- ALVES, D. S. *Uso de técnicas de Computação Social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores*. 132 p. Doutorado em Engenharia Elétrica — Universidade de Brasília, Brasília, 2015.
- ANÉ, T.; GEMAN, H. Order flow, transaction clock, and normality of asset returns. *The Journal of Finance*, Wiley Online Library, v. 55, n. 5, p. 2259–2284, 2000.
- ARTHUR, W. B. Asset pricing under endogenous expectations in an artificial stock market. In: *The economy as an evolving complex system II*. [S.l.]: CRC Press, 2018. p. 31–60.
- AZIZ, S.; DOWLING, M. Machine learning and ai for risk management. In: *Disrupting Finance*. [S.l.]: Springer, 2019. p. 33–50.
- B3. *Participação dos Investidores*. 2018. Online. Disponível em: <http://www.bmfbovespa.com.br/pt_br/servicos/market-data/consultas/historico-pessoas-fisicas/>. Acesso em: 30 May. 2018.
- BAILEY, D. H. et al. Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, v. 61, n. 5, p. 458–471, 2014.
- BAPTISTA, R. F. de F.; PEREIRA, P. L. V. Análise do desempenho de regras da análise técnica aplicado ao mercado intradiário do contrato futuro do índice ibovespa. *Revista Brasileira de Finanças*, v. 6, n. 2, January 2009.

BARBOSA, M. J. *Análise Gráfica Produz Boas Rentabilidades? Uma Avaliação da Eficácia da Análise Técnica Computadorizada na Geração de Retornos*. 96 p. Mestrado — Universidade Federal de Pernambuco, Pernambuco, 2007.

BARROS, A. C. *Análise de Séries Temporais*. [S.l.]: Editora Elsevier, 2018.

BLANCHARD, O. et al. Why has the stock market risen so much since the us presidential election? 2018.

BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. *Journal of computational science*, Elsevier, v. 2, n. 1, p. 1–8, 2011.

BONACCORSO, G. *Machine learning algorithms*. [S.l.]: Packt Publishing Ltd, 2017.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

BROOKS, A. *Trading price action trends: technical analysis of price charts bar by bar for the serious trader*. [S.l.]: John Wiley & Sons, 2011.

BYRNES, M. T. R. N. *As Goldman Embraces Automation, Even the Masters of the Universe Are Threatened*. 2017. Online. Disponível em: <<https://www.technologyreview.com/s/603431/as-goldman-embraces-automation-even-the-masters-of-the-universe-are-threatened/>>. Acesso em: 16 Jun. 2019.

CALDEIRA, J. F. Arbitragem estatística, estratégia long-short pairs trading, abordagem com cointegração aplicada ao mercado de ações brasileiro. *Economia*, v. 14, n. 1b, 2013. Disponível em: <https://EconPapers.repec.org/RePEc:anp:econom:v:14:y:2013:i:1b:521_546>.

CARTEA, Á.; JAUMUNGAL, S.; PENALVA, J. *Algorithmic and high-frequency trading*. [S.l.]: Cambridge University Press, 2015.

CASTRO P. A. L.; ANNONI JUNIOR, R. S. J. S. Análise autônoma de investimento: Uma abordagem probabilística discreta. *Revista de Informática Teórica e Aplicada (RITA)*, v. 25, p. 23–38, 2018.

CAVALCANTE, R. C. et al. Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications*, Elsevier, v. 55, p. 194–211, 2016.

CHAN, E. *Quantitative trading: how to build your own algorithmic trading business.* [S.l.]: John Wiley & Sons, 2009.

CHAVES, D. A. T. *Análise técnica e fundamentalista: divergências, similaridades e complementaridades.* 119 p. Graduação em Administração de Empresas — Universidade de São Paulo, São Paulo, 2004.

CHOUDHRY, R.; GARG, K. A hybrid machine learning system for stock market forecasting. *World Academy of Science, Engineering and Technology*, v. 39, n. 3, p. 315–318, 2008.

CLARK, P. K. A subordinated stochastic process model with finite variance for speculative prices. *Econometrica: journal of the Econometric Society*, JSTOR, p. 135–155, 1973.

COHERENT, M. I. *Algorithmic Trading Market - Size, Share, Outlook, and Opportunity Analysis.* 2019. Online. Disponível em: <<https://www.coherentmarketinsights.com-market-insight/algorithmic-trading-market-2476>>. Acesso em: 22 Mar. 2019.

EASLEY, D.; PRADO, M. L. D.; O'HARA, M. The microstructure of the flash crash: Flow toxicity, liquidity crashes and the probability of informed trading. *Journal of Portfolio Management*, Euromoney Institutional Investor PLC, v. 37, n. 2, p. 118–128, 2011.

EASLEY, D.; PRADO, M. Lopez de; O'HARA, M. The volume clock: Insights into the high frequency paradigm. *The Journal of Portfolio Management, (Fall, 2012) Forthcoming*, 2012.

EASLEY, D.; PRADO, M. Lopez de; O'HARA, M. The volume clock: Insights into the high frequency paradigm. *The Journal of Portfolio Management, (Fall, 2012) Forthcoming*, 2012.

ELDER, A. *Como se transformar em um operador e investidor de sucesso: Entenda a psicologia do mercado financeiro, técnicas poderosas de negociação, gestão lucrativa de investimentos.* [S.l.]: Elsevier Brasil, 2004. ISBN 9788535249347.

FAMA, E. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, v. 25, n. 2, p. 383–417, 1970. Disponível em: <<https://EconPapers.repec.org/RePEc:bla:jfinan:v:25:y:1970:i:2:p:383-417>>.

FAMA, E. F.; BLUME, M. E. Filter rules and stock-market trading. *The Journal of Business*, JSTOR, v. 39, n. 1, p. 226–241, 1966.

FILHO, J. A. S. D. P. *Metodologia Científica*. [S.l.: s.n.], 2012.

FISCHER, T.; KRAUSS, C. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, Elsevier, v. 270, n. 2, p. 654–669, 2018.

GENDREAU, M.; POTVIN, J.-Y. *Handbook of Metaheuristics*. 2nd. ed. [S.l.]: Springer Publishing Company, Incorporated, 2010. ISBN 1441916636, 9781441916631.

GIACOMEL, F. d. S. Um método algorítmico para operações na bolsa de valores baseado em ensembles de redes neurais para modelar e prever os movimentos dos mercados de ações. 2016.

GIL, A. C. *Métodos e técnicas de pesquisa social*. [S.l.]: 6. ed. Ediitora Atlas SA, 2008.

GRIGORYAN, H. Stock market trend prediction using support vector machines and variable selection methods. In: ATLANTIS PRESS. *2017 International Conference on Applied Mathematics, Modelling and Statistics Application (AMMSA 2017)*. [S.l.], 2017.

GYAMERAH, S. A.; NGARE, P.; IKPE, D. On stock market movement prediction via stacking ensemble learning method. In: IEEE. *2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*. [S.l.], 2019. p. 1–8.

HENDERSHOTT, T.; JONES, C. M.; MENKVELD, A. J. Does algorithmic trading improve liquidity? *The Journal of Finance*, Wiley Online Library, v. 66, n. 1, p. 1–33, 2011.

HIRSHLEIFER, D.; TEOH, S. H. Herd behaviour and cascading in capital markets: A review and synthesis. *European Financial Management*, Wiley Online Library, v. 9, n. 1, p. 25–66, 2003.

JANSEN, S. *Hands-On Machine Learning for Algorithmic Trading: Design and implement investment strategies based on smart algorithms that learn from data using Python*. Packt Publishing, 2018. ISBN 9781789342710. Disponível em: <<https://books.google.com.br/books?id=tx2CDwAAQBAJ>>.

KARTHIK, H.; NISHANTH, V.; MANIKANDAN, J. Stock market prediction using optimum threshold based relevance vector machines. In: IEEE. *2016 22nd Annual International Conference on Advanced Computing and Communication (ADCOM)*. [S.l.], 2016. p. 21–26.

LAM, K.; YAM, H. Cusum techniques for technical trading in financial markets. *Financial Engineering and the Japanese Markets*, Springer, v. 4, n. 3, p. 257–274, 1997.

LARSEN, F. *Automatic stock market trading based on Technical Analysis*. 92 p. Mestrado — Norwegian University of Science and Technology, Trondheim, Norway, 2007.

LEMOS, F.; CARDOSO, C. *Análise Técnica Clássica - Com as Mais Recentes Estratégias da Expo Trader Brasil*. 1. ed. São Paulo: Saraiva, 2010.

LIANG, Q. et al. Restricted boltzmann machine based stock market trend prediction. In: IEEE. *2017 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2017. p. 1380–1387.

LUIZ, A. A. *Teste de Previsibilidade de Mercado Usando Conceitos de Análise Técnica*. 2009. Online. Disponível em: <http://www.creasp.org.br/biblioteca/wp-content/uploads/2012/09/Teste_de_Previabilidade_de_Mercado.pdf>. Acesso em: 27 May. 2018.

MANDELBROT, B.; HUDSON, R. *The (Mis)Behaviour of Markets: A Fractal View of Risk, Ruin and Reward*. [S.l.]: Profile, 2010. ISBN 9781847651556.

MANDELBROT, B.; TAYLOR, H. M. On the distribution of stock price differences. *Operations research*, Informs, v. 15, n. 6, p. 1057–1062, 1967.

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

MATSURA, E. *Comprar ou Vender? - Como Investir na Bolsa Utilizando Análise Gráfica*. 7. ed. São Paulo: Saraiva, 2013.

MENDES, M.; PALA, A. Type i error rate and power of three normality tests. *Pakistan Journal of Information and Technology*, v. 2, n. 2, p. 135–139, 2003.

MOTA, F. d. A. *O dever de divulgar fato relevante na companhia aberta*. Mestrado, 2013.

NORONHA, M. *Análise Técnica: Teorias Ferramentas Estratégias*. 5. ed. São Paulo: Editec, 2003.

PÁSCOA, M. I. F. *Os desafios da Machine Learning: Aplicação ao Mercado Financeiro*. Dissertação (Mestrado), 2018.

PIMENTA, A. *Métodos automatizados para investimento no mercado de ações via inteligência computacional*. 140 p. Doutorado em Engenharia Elétrica — Universidade Federal de Minas Gerais, Belo Horizonte, 2017.

PINHEIRO, J. L. *Mercado de Capitais*. 8. ed. São Paulo: Atlas, 2016.

PRADO, M. L. de. *Advances in Financial Machine Learning*. 1. ed. [S.l.]: Wiley, 2018. ISBN 9781119482086.

QIAN, B.; RASHEED, K. Stock market prediction with multiple classifiers. v. 26, p. 25–33, 02 2007.

RAZA, N. et al. Asymmetric impact of gold, oil prices and their volatilities on stock prices of emerging markets. *Resources Policy*, Elsevier, v. 49, p. 290–301, 2016.

RESENDE, F. S. de. *Ciência de Dados aplicada à BM&FBovespa*. Curitiba: Appris Editora, 2016.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3rd. ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009. ISBN 0136042597, 9780136042594.

SAFFI, P. A. C. Análise técnica: Sorte ou realidade? *Revista Brasileira de Economia*, v. 57, n. 4, p. 954–974, 2003. ISSN 1806-9134.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, JSTOR, v. 52, n. 3/4, p. 591–611, 1965.

- SHARPE, W. F. Mutual fund performance. *The Journal of Business*, University of Chicago Press, v. 39, n. 1, p. 119–138, 1966. ISSN 00219398, 15375374. Disponível em: <<http://www.jstor.org/stable/2351741>>.
- SHUMWAY, R. H.; STOFFER, D. S. *Time series analysis and its applications: with R examples*. [S.l.]: Springer, 2017.
- SILVA, L. A. da. *Introdução à mineração de dados com aplicações em R*. [S.l.]: Editora Elsevier, 2016.
- SIMÕES, A. R. C. R. A. *STOCKS: Computação Inteligente Aplicada ao Mercado Accionista*. 94 p. Mestrado — Universidade de Lisboa, Portugal, 2010.
- SLOVIC, P. Psychological study of human judgment: Implications for investment decision making. *The Journal of Finance*, [American Finance Association, Wiley], v. 27, n. 4, p. 779–799, 1972. ISSN 00221082, 15406261. Disponível em: <<http://www.jstor.org/stable/2978668>>.
- STOCK, J. H.; WATSON, M. W. *Econometria*. [S.l.]: Pearson Education do Brasil, 2004.
- SZYSZKA, P. X. Random forest: uma investigação da eficiência de mercado de ações da ações da vale. 2017.
- TAYLOR, S. J. *Asset price dynamics, volatility, and prediction*. [S.l.]: Princeton university press, 2011.
- THENMOZHI, M.; CHAND, G. S. Forecasting stock returns based on information transmission across global markets using support vector machines. *Neural Computing and Applications*, Springer, v. 27, n. 4, p. 805–824, 2016.
- THODE, H. C. *Testing for normality*. [S.l.]: CRC press, 2002.
- TILLY, E. T.; MONTESANO, A.; SMITH, E. C. *Automated trading system for routing and matching orders*. [S.l.]: Google Patents, dez. 8 2016. US Patent App. 14/991,688.
- TUKEY, J. W. Exploratory data analysis. Addison-Wesley, 1977.

VARIAN, H. A portfolio of nobel laureates: Markowitz, miller and sharpe. *Journal of Economic Perspectives*, v. 7, n. 1, p. 159–169, 1993.

VERIKAS, A.; GELZINIS, A.; BACAUSKIENE, M. Mining data with random forests: A survey and results of new tests. *Pattern recognition*, Elsevier, v. 44, n. 2, p. 330–349, 2011.

VERMA, R.; VERMA, P. Noise trading and stock market volatility. *Journal of Multinational Financial Management*, Elsevier, v. 17, n. 3, p. 231–243, 2007.

WONG, W. K.; DU, J.; CHONG, T. T. L. Do technical indicators reward chartists? a study of stock market of china. *Review of Applied Economics*, v. 1, n. 2, p. 183–205, 2005.

WOOLDRIDGE, J. M. *Introdução à Econometria: Uma Abordagem Moderna*. 1. ed. [S.l.: s.n.], 2006.

YOSHINAGA, C. et al. Finanças comportamentais: uma introdução. *Revista de Gestão*, v. 15, p. 25–35, 01 2008.

ZHAO, L.; WANG, L. Price trend prediction of stock market using outlier data mining algorithm. In: IEEE. *2015 IEEE Fifth International Conference on Big Data and Cloud Computing*. [S.l.], 2015. p. 93–98.