



TÉCNICAS DE APRENDIZADO DE MÁQUINA APLICADAS A ESTRATÉGIA DE SWING TRADE NO MERCADO FINANCEIRO

Pedro Henrique Barbosa Nori

Projeto de Graduação apresentado ao Curso de Engenharia Eletrônica e de Computação da Escola Politécnica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Engenheiro.

Orientador: Heraldo Luis Silveira de Almeida

Rio de Janeiro
Setembro de 2022

TÉCNICAS DE APRENDIZADO DE MÁQUINA APLICADAS A
ESTRATÉGIA DE SWING TRADE NO MERCADO
FINANCEIRO

Pedro Henrique Barbosa Nori

PROJETO DE GRADUAÇÃO SUBMETIDO AO CORPO DOCENTE DO CURSO
DE ENGENHARIA ELETRÔNICA E DE COMPUTAÇÃO DA ESCOLA PO-
LITÉCNICA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO
PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU
DE ENGENHEIRO ELETRÔNICO E DE COMPUTAÇÃO

Autor:

Pedro Henrique Barbosa Nori

Orientador:

Prof. Heraldo Luis Silveira de Almeida, D. Sc.

Examinador:

Prof. Flávio Luis de Mello, D. Sc.

Examinador:

Prof. Natanael Nunes de Moura Junior, D. Sc.

Rio de Janeiro
Setembro de 2022

Declaração de Autoria e de Direitos

Eu, *Pedro Henrique Barbosa Nori* CPF 134.129.077-82, autor da monografia *TÉCNICAS DE APRENDIZADO DE MÁQUINA APLICADAS A ESTRATÉGIA DE SWING TRADE NO MERCADO FINANCEIRO*, subscrevo para os devidos fins, as seguintes informações:

1. O autor declara que o trabalho apresentado na disciplina de Projeto de Graduação da Escola Politécnica da UFRJ é de sua autoria, sendo original em forma e conteúdo.
2. Excetua-se do item 1. eventuais transcrições de texto, figuras, tabelas, conceitos e idéias, que identifiquem claramente a fonte original, explicitando as autorizações obtidas dos respectivos proprietários, quando necessárias.
3. O autor permite que a UFRJ, por um prazo indeterminado, efetue em qualquer mídia de divulgação, a publicação do trabalho acadêmico em sua totalidade, ou em parte. Essa autorização não envolve ônus de qualquer natureza à UFRJ, ou aos seus representantes.
4. O autor pode, excepcionalmente, encaminhar à Comissão de Projeto de Graduação, a não divulgação do material, por um prazo máximo de 01 (um) ano, improrrogável, a contar da data de defesa, desde que o pedido seja justificado, e solicitado antecipadamente, por escrito, à Congregação da Escola Politécnica.
5. O autor declara, ainda, ter a capacidade jurídica para a prática do presente ato, assim como ter conhecimento do teor da presente Declaração, estando ciente das sanções e punições legais, no que tange a cópia parcial, ou total, de obra intelectual, o que se configura como violação do direito autoral previsto no Código Penal Brasileiro no art.184 e art.299, bem como na Lei 9.610.
6. O autor é o único responsável pelo conteúdo apresentado nos trabalhos acadêmicos publicados, não cabendo à UFRJ, aos seus representantes, ou ao(s) orientador(es), qualquer responsabilização/ indenização nesse sentido.
7. Por ser verdade, firmo a presente declaração.

Pedro Henrique Barbosa Nori

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Escola Politécnica - Departamento de Eletrônica e de Computação

Centro de Tecnologia, bloco H, sala H-217, Cidade Universitária

Rio de Janeiro - RJ CEP 21949-900

Este exemplar é de propriedade da Universidade Federal do Rio de Janeiro, que poderá incluí-lo em base de dados, armazenar em computador, microfilmear ou adotar qualquer forma de arquivamento.

É permitida a menção, reprodução parcial ou integral e a transmissão entre bibliotecas deste trabalho, sem modificação de seu texto, em qualquer meio que esteja ou venha a ser fixado, para pesquisa acadêmica, comentários e citações, desde que sem finalidade comercial e que seja feita a referência bibliográfica completa.

Os conceitos expressos neste trabalho são de responsabilidade do(s) autor(es).

DEDICATÓRIA

À minha mãe, engenheira mecânica.

AGRADECIMENTO

Agradeço à minha mãe Ana Christina e ao meu pai Adilson Nori por todo apoio e paciência que pude receber durante todos esses anos. Agradeço também ao meu irmão João por estar comigo nessa grande jornada da vida.

Agradeço aos professores que tive a oportunidade de conhecer, a começar pelo Ariedio Schiappacassa do CEFET/RJ por toda a paciência em me ajudar na montagem do meu primeira rádio FM e nos primeiros passos com PIC. Agradeço também aos meus professores da UFRJ, os quais guardo enorme respeito, carinho e admiração. Em especial: Casé, Luiz Wagner, Pino, Brafman, Teodósio, Wallace e Jomar.

Deixo um grande abraço a todos os meus companheiros da equipe de robótica MinervaBots, onde tanto aprendi e tanto amei pertencer.

À todas as experiências que puder compartilhar com meus amigos da vida e da faculdade. Deixo também um abraço para o meu amigo de infância Daniel Iunes Monteiro que infelizmente não se encontra mais nesta vida.

Por fim, agradeço ao meu orientador Heraldo por me receber de braços abertos e por todo o suporte no cumprimento deste projeto.

RESUMO

Todos os dias, diversas negociações são realizadas nas bolsas de valores do mundo inteiro. Com os mais diversos objetivos, investidores buscam um aumento crescente de patrimônio de forma consistente. Paralelamente, inteligências artificiais vem substituindo cada vez mais atividades antes desempenhadas pelo homem.

Nesse sentido, este trabalho visa a aplicação de técnicas de aprendizado de máquina para a elaboração de uma estratégia de *swing trade* no mercado acionário brasileiro. Para isso, é concebida uma estrutura de regras e premissas que criam uma base ao modelo de aprendizado de máquina, responsável por decidir o momento de entrada em operações a partir de um conjunto de dados.

Palavras-Chave: *Machine Learning*, *Random Forest*, Análise Técnica, *Swing Trade*, Mercado Financeiro.

ABSTRACT

Every day, several negotiations are carried out on stock exchanges around the world. With the most diverse objectives, investors seek a consistently growing increase in equity. At the same time, artificial intelligences are increasingly replacing activities previously performed by man.

In this context, this work aims at the application of machine learning techniques for the elaboration of a swing trade strategy in the Brazilian stock market. So, a structure of rules and assumptions is conceived to create a basis for the machine learning model, responsible for deciding when to enter into operations given a set of data.

Key-words: Machine Learning, Random Forest, Technical Analysis, Swing Trade, Stock Market.

SIGLAS

AF - Análise Fundamentalista

API - *Application Programming Interface*

ANN - *Artificial Neural Networks*

ARCH - *Autoregressive Conditional Heteroskedasticity*

AS - Aprendizado Supervisionado

AT - Análise Técnica

B3 - Bolsa, Brasil, Balção

CPU - *Central Process Unit*

CSL - *Cost Sensitive Learning*

CSV - *Comma-separated values*

CVM - Comissão de Valores Mobiliários

DT - *Decision Tree*

EGARCH - *Exponential Generalised ARCH*

EMA - *Exponential Moving Average*

ETF - *Exchange-Traded Funds*

GARCH - *Generalised ARCH*

HME - Hipótese do Mercado Eficiente

HMM - *Hidden Markov Model*

iBovespa - Índice Bovespa

IPO - *Initial Public Offering*

IIR - *Infinite Impulse Reponse*

IL - Índice de Lucratividade

JSON - *JavaScript Object Notation*

k-NN - *K Nearest Neighbors*

MACD - *Moving Average Convergence/Divergence*

ML - *Machine Learning*

MME - Média Móvel Exponencial

NFO - Normalização por Frequência de Operações

NGARCH - *Non-linear Generalised ARCH*

RCC - *Risk-Capital Coefficient*

RF - *Random Forest*

RI - Relações com Investidores

SVM - *Support Vector Machine*

TGARCH - *Threshold Generalised ARCH*

UFRJ - Universidade Federal do Rio de Janeiro

WFA - *Walk-Forward Analysis*

Sumário

1	Introdução	1
1.1	Tema	1
1.2	Delimitação	1
1.3	Justificativa	2
1.4	Objetivos	3
1.5	Metodologia	3
1.6	Descrição	4
2	Fundamentação Teórica	5
2.1	Mercado de Capitais, Bolsa de Valores e Ações	5
2.1.1	Hipótese do Mercado Eficiente	6
2.1.2	Índice de Bolsa de Valores	8
2.1.3	Mercado Fracionário	8
2.2	Tipos de Análises	9
2.2.1	Análise Fundamentalista	9
2.2.2	Análise Técnica	9
2.3	Aprendizado de Máquina	13
2.3.1	Aprendizado Supervisionado	13
2.3.2	Problema de Regressão	15
2.3.3	Problema de Classificação	15
2.3.4	Algoritmos de Aprendizado Supervisionado	16
2.4	<i>Walk-Forward Analysis</i>	19
2.5	Considerações para Análise de Resultados	20
2.5.1	Índice de Sharpe	20
2.5.2	Índice de Sortino	20

2.5.3	Correlação de Spearman	21
2.6	Trabalhos Relacionados	22
2.6.1	Modelos Baseados em Indicadores Técnicos	23
2.6.2	Modelos Baseados em Processos Estocásticos	23
2.6.3	Modelos Baseados em Aprendizado de Máquina	25
3	Metodologia	26
3.1	Resumo	26
3.2	Pré-Processamento	28
3.2.1	Arquivo de Configuração	28
3.2.2	Coleta de Dados	29
3.2.3	Armazenamento de Dados	31
3.2.4	Geração de <i>Features</i> de Uso Geral	33
3.3	Simulação de Estratégia	39
3.3.1	Estrutura	39
3.3.2	Premissas	40
3.3.3	Período Máximo de Dias por Operação	41
3.3.4	Risco de Entrada por Operação	45
3.3.5	Gerenciamento de Risco	46
3.3.6	Controle Proporcional para Uso de Capital	49
3.3.7	Lista de Parâmetros de Configuração	54
3.3.8	<i>Dashboard</i>	55
3.4	Modelos de Aprendizado Supervisionado	59
3.4.1	Resumo	59
3.4.2	<i>Datasets</i> e <i>Feature Selection</i>	59
3.4.3	Índice de Lucratividade	60
3.4.4	Balanceamento de Classes	61
3.4.5	Geração de Modelos	62
3.4.6	Modelo <i>Baseline</i>	65
4	Resultados	67
5	Considerações Finais	72
5.1	Conclusão	72

5.2	Trabalhos Futuros	73
	Bibliografia	74
A	Inconsistência de Proventos na Biblioteca <i>yfinance</i>	82

Lista de Figuras

2.1	Leitura de um gráfico de <i>candlestick</i> [1]	10
2.2	Comportamento do mercado ideal segundo a Teoria de Dow [2]	11
2.3	Formação de linhas de Suporte e de Resistência [3]	12
2.4	Formação de uma Linha de Tendência de Alta [3]	12
2.5	Formação de uma Linha de Tendência de Baixa [3]	12
2.6	Relação entre complexidade e acurácia de um modelo [4]	14
2.7	<i>Oversampling</i> e <i>Undersampling</i> de classes desbalanceadas [5]	16
2.8	Funcionamento de um algoritmo k-NN para o problema de classi- ficação [6]. Para K=3 a classe é B e para K=7 a classe é A.	17
2.9	Visualização de uma Árvore de Decisão para um <i>dataset</i> de câncer de mama [4].	18
2.10	<i>Walk-Forward Analysis</i> (Não-Ancorado e Ancorado) [7].	19
3.1	Estrutura do técnica do projeto	26
3.2	Estrutura do Arquivo de Configuração	28
3.3	Arquivo de Configuração para Execuções Múltiplas	29
3.4	ERD do Banco de Dados	32
3.5	MGLU3 - Risco Mínimo (01/01/2019 a 31/12/2019)	35
3.6	ABEV3 - Risco Mínimo (01/01/2019 a 31/12/2019)	36
3.7	CPLE6 - Algoritmo de identificação de picos (01/01/2019 a 31/12/2019)	37
3.8	CPLE6 - Subidas de preços entre picos (01/01/2019 a 31/12/2019)	37
3.9	MGLU3 - Riscos Máximo e Mínimo (01/01/2019 a 31/12/2019)	38
3.10	ABEV3 - Riscos Máximo e Mínimo (01/01/2019 a 31/12/2019)	38
3.11	MGLU3 - Histograma de dias com risco mínimo em operações de sucesso (01/01/2016 a 31/12/2018)	42

3.12	ABEV3 - Histograma de dias com risco mínimo em operações de sucesso (01/01/2016 a 31/12/2018)	43
3.13	MGLU3 - Histograma de dias com risco ótimo em operações de sucesso (01/01/2016 a 31/12/2018)	43
3.14	ABEV3 - Histograma de dias com risco ótimo em operações de sucesso (01/01/2016 a 31/12/2018)	44
3.15	MGLU3 - Histograma de todas as operações de sucesso (01/01/2016 a 31/12/2018)	44
3.16	Indicadores de performance em função do risco de entrada (71 tickers, 01/01/2019 a 31/03/2020)	46
3.17	Indicadores de performance em função do RCC (71 tickers: 01/01/2019 a 31/12/2021)	48
3.18	Uso de Capital (71 tickers, 01/01/2019 a 31/12/2021, RCC = 0,29%)	48
3.19	Uso de Capital (71 tickers, 01/01/2019 a 31/12/2021, RCC = 1,88%)	49
3.20	Rendimento sob uso de RCC dinâmico	52
3.21	Índice de Sharpe sob uso de RCC dinâmico	52
3.22	Uso médio de capital sob uso de RCC dinâmico	53
3.23	Total de operações sob uso de RCC dinâmico	53
3.24	<i>Dashboard</i> - Performance	56
3.25	<i>Dashboard</i> - Parâmetros de entrada	56
3.26	<i>Dashboard</i> - Resultados e estatísticas	57
3.27	<i>Dashboard</i> - Gráfico de uso de capital	57
3.28	<i>Dashboard</i> - Gráficos de análise individual de ações	58
3.29	Diagrama de criação de modelos	64
3.30	<i>Baseline</i> para o intervalo de 01/01/2019 a 31/12/2021	66
4.1	Performance final	68

Lista de Tabelas

2.1	Amostras das variáveis aleatórias X e Y	22
2.2	Postos rgX e rgY	22
3.1	Ações Escolhidas	28
3.2	Período de dias que engloba 90% das contagens dos histogramas . . .	45
3.3	Lista de parâmetros detalhados	55
3.4	Comparação de Resultados	60
3.5	Balanceamento via CSL	62
3.6	Parâmetros fixos	63
3.7	Parâmetros variáveis	63
3.8	WFA - Intervalos de treinamento, teste e validade dos modelos	65
3.9	<i>Baseline</i> - Indicadores de Performance	66
4.1	Resultado final	69
A.1	Análise de Consistência de Proventos: MGLU3	83

Capítulo 1

Introdução

1.1 Tema

O tema deste trabalho se resume na elaboração de uma estratégia de *swing trade* na bolsa de valores brasileira através de métodos de aprendizado de máquina.

Nesse contexto, o problema a ser abordado é a identificação do momento apropriado para compra de um determinado ativo, como também os preços alvos determinantes para venda, tendo em vista uma variação positiva de seu preço.

1.2 Delimitação

Este trabalho se limita aos ativos negociados na Bolsa de Valores de São Paulo, a B3, de cujos dados diários de domínio público foram adquiridos através da plataforma *Yahoo Finance* pela API *open-source yfinance*, disponível em Python. Não são levadas em consideração informações sobre proventos (dividendos e juros sobre capital próprio) devido à inconsistência dos mesmos na API supracitada, aliada à dificuldade técnica para automatização da busca de tais dados.

A duração das operações tem em vista um horizonte mínimo de um dia, sendo portanto operações de *swing trade*. Não são realizadas vendas a descoberto¹, portanto

¹Venda a descoberto (*short selling*): Venda de ações anterior a compra das mesmas através de um contrato de aluguel. O lucro ocorre na queda de preço do mercado.

só há lucro em variações positivas de preço dos ativos. Apenas uma operação por ativo pode existir em um determinado instante de tempo para uma estratégia. Em outras palavras, só é possível comprar mais ações de uma empresa após a venda completa das ações da mesma, caso existam.

A incidência de impostos devidos (e.g., imposto de renda) está fora do escopo. Assim como a utilização de critérios baseados em análise fundamentalista, por causa da dificuldade de obtenção dessas informações de maneira automatizada e estruturada.

1.3 Justificativa

O crescimento do número de investidores na bolsa de valores brasileira [8] demonstra um maior interesse da população na busca por um complemento da renda familiar ou até na substituição da fonte de renda principal.

No cenário global, o aumento do uso de robôs de *trading* (ou algoritmos) tem se mostrando expressivo [9], seja por pessoas físicas ou fundos de investimento, de forma total ou parcial em suas estratégias. Por outro lado, tal crescimento não vem sendo igualmente representado no Brasil devido às peculiaridades do mercado de capitais nacional, como a alta volatilidade e a alta sensibilidade a notícias [10].

Paralelamente, estudos relacionados a aprendizado de máquina vem trazendo resultados práticos no dia-a-dia das pessoas, desde o clássico exemplo de reconhecimento de mensagens de *spam* em um caixa de email à identificação do perfil de consumo de clientes em uma loja. Da mesma forma, instituições financeiras e bancos centrais também estão, com cautela, incorporando aplicações de aprendizado de máquina em tarefas internas [11].

Apesar das dificuldades inerentes ao cenário atual do mercado de capitais brasileiro, não se pode ignorar o potencial que os algoritmos podem trazer. Desta forma, o presente trabalho visa a união de técnicas de aprendizado de máquina a práticas de *trading*, consolidando uma estratégia que sirva de suporte a uma maior variedade de opções de investimento à população brasileira.

1.4 Objetivos

O objetivo geral deste trabalho é implementar um *software* capaz de simular uma estratégia de *swing trade* utilizando aprendizado de máquina. Especificamente, o software deve: (1) Criar um ambiente automatizado que permita buscar, atualizar e armazenar dados diários da bolsa brasileira de forma simples e conforme necessidade do usuário da aplicação; (2) Criar a estrutura de uma estratégia por meio de um conjunto de regras e premissas baseadas em práticas de *trading*; (3) Gerar os modelos de aprendizado de máquina e acoplá-los à estrutura criada; (4) Simular a estratégia obtida; (5) Criar um mecanismo de fácil visualização dos resultados das simulações; (6) Analisar os resultados gerados.

1.5 Metodologia

O trabalho tem início na criação de um ambiente propício à simulação de estratégias, bem como sua configuração e manutenção. Para isso, a fim de: otimizar o tráfego de dados pela internet; minimizar o processamento necessário para a geração de dados derivados (pré-processamento); e armazenar os resultados das estratégias de forma organizada, foi utilizado um banco de dados PostgreSQL. Dentre as atividades realizadas durante o pré-processamento dos dados, anteriores à simulação, é possível citar a identificação de momentos de tendência de baixa e crises do mercado, a identificação de picos na série histórica e os valores de risco mínimo e máximo a serem utilizados nas operações.

Em seguida, a etapa de simulação começa na leitura de um arquivo JSON contendo todos parâmetros necessários para a execução das estratégias. Nesta etapa, o programa itera dia após dia para cada estratégia configurada verificando os momentos e os valores de compra e de venda para cada ativo que compõe as carteiras. Ao final, registram-se no banco todas as operações executadas, independente da obtenção de lucro, junto com as informações estatísticas necessárias para a avaliação da performance.

Por fim, com o objetivo de facilitar a análise dos resultados gerados, criou-se um *dashboard* responsável por centralizar todas as informações pertinentes a uma execução de estratégia em uma única página web.

Observa-se que além do uso de estruturas do banco de dados PostgreSQL, como *triggers* e *functions*, o código foi construído em Python devido à ampla variedade de bibliotecas e ao suporte da comunidade, apesar da desvantagem de desempenho por ser uma linguagem interpretada. Bastante foco foi dado à escalabilidade e à manutenção do código, que contou com as bibliotecas e as APIs *yfinance*, *pandas*, *numpy*, *scikit-learn*, *multiprocessing*, *matplotlib* e *dash*. Também utilizou-se *containers* Docker para simplificar a execução.

1.6 Descrição

No capítulo 2 é desenvolvida a fundamentação teórica acerca de temas relevantes ao entendimento geral de mercado financeiro e de aprendizado de máquina. Também é realizada uma revisão dos trabalhos relacionados ao tema, em outras palavras, a revisão bibliográfica.

No capítulo 3, toda a metodologia é descrita, o que envolve o pré-processamento dos dados as simulações e a criação dos modelos de ML.

Por fim, o capítulo 4 encerra com a conclusão e uma recomendação de trabalhos futuros.

Capítulo 2

Fundamentação Teórica

2.1 Mercado de Capitais, Bolsa de Valores e Ações

O Mercado de Capitais, também conhecido como Mercado de Valores Mobiliários, é um dos segmentos do sistema financeiro responsável por fazer o intermédio entre agentes superávitaros, que detém capital de investimento, e agentes deficitários, que buscam capital para rentabilizá-lo, através da compra e venda valores mobiliários (i.e., ativos financeiros) [12]. Consequentemente, gera-se uma maior liquidez destes ativos e também uma melhora no fluxo de capitais entre os agentes econômicos, seja os governos por meio dos bancos centrais, os bancos privados, as instituições financeiras ou até mesmo as pessoas físicas.

No Brasil, o Mercado de Capitais é regulado e fiscalizado pela CVM, uma autarquia federal vinculada ao Ministério da Fazenda e criada em 1976 através da Lei nº 6.385 [13].

A Bolsa de Valores é uma plataforma onde se negociam os valores mobiliários do Mercado de Capitais, dentre eles ações (i.e., fatias, pedaços) de sociedades anônimas (ou companhias). No Brasil, a única Bolsa de Valores oficial existente é a B3 [14], que administra os sistemas de negociação, compensação, liquidação, depósito e registro para todas as principais classes de ativos.

O processo de abertura de capital de uma empresa é uma iniciativa que possui vantagens estratégicas [15] como: o aumento da confiança na perspectiva do

mercado, seja para o consumidor final ou para parceiros comerciais; a solução de problemas decorrentes de processos sucessórios; e também a captação de capital de investimento, a fim de contribuir para o crescimento ou para a consolidação da companhia. Esse processo acontece através de uma oferta pública [16], ou IPO, onde as ações que compõem o capital social [17] de uma companhia são vendidas pela primeira vez ao público geral. Uma vez encerrado o IPO, estas mesmas ações passam para o mercado secundário [18], onde investidores as negociam entre si. Em retorno ao capital adquirido pela companhia, surgem algumas responsabilidades, dentre elas a publicação de demonstrações financeiras [19], auditadas pela própria CVM [20].

Para o acionista de uma sociedade anônima, existem duas formas de se obter lucro: através de proventos (dividendos e juros sobre capital próprio) [21]; ou através de operações de compra e de venda de ações, mediante oscilações de seu valor de mercado. Conforme a expectativa corretamente induz, o lucro é comumente aferido durante a venda de um determinado papel posteriormente à sua aquisição a um preço de compra inferior. No entanto, também é possível trabalhar com posições vendidas (short selling) [22], onde um investidor aluga ações de outro investidor por meio de um contrato. Em seguida as vende para posteriormente recomprá-las a um preço inferior, devolvendo-as assim ao respectivo dono. Neste caso, o lucro é obtido quando a expectativa de queda de um ativo se prova verdadeira.

2.1.1 Hipótese do Mercado Eficiente

A Hipótese do Mercado Eficiente, definida por Eugene Fama [23], afirma que idealmente o preço de um ativo reflete toda a informação disponível sobre seu valor intrínseco. Em outras palavras, quanto menor o efeito de fatores que contribuam para uma inércia no fluxo de capital de investidores e na transmissão de informações, mais o mercado tende a ser eficiente. São estudados os três níveis de hipóteses:

- HME fraca: Os preços atuais refletem todo o histórico de informações disponibilizados publicamente.
- HME semi-forte: Engloba a HME fraca, acrescentando-se a existência de uma mudança instantânea que os preços sofrem ao surgirem novas informações.

- HME forte: Engloba a HME semi-forte, porém entende-se que a mudança instantânea dos preços acompanha toda e qualquer informação existente sobre o ativo. Assim, absolutamente nenhum investidor conseguiria obter lucro superior à média do mercado, pois não há como acessar nenhuma informação privilegiada, uma vez que ela já estaria refletido no preço corrente do ativo.

O autor menciona que a HME forte não é estritamente válida na realidade, o que é uma afirmação coerente quando se verifica a existência de casos em que o vazamento de informações confidenciais trouxe aos acusados uma lucratividade muito acima da média [24].

A HME fraca foi verificada pela consistência da correlação dos preços dia após dia de determinadas ações, mesmo que esta fosse baixa.

A hipótese semi-forte também foi sustentada por alguns fatores, dentre eles a verificação de que os futuros pagamentos de proventos das companhias se refletem em média no preços das ações [25].

Em resumo, o estudo das Hipóteses de Mercado Eficiente traz informações relevantes quanto se avalia a teoria por trás da possibilidade de aplicação de estratégias de *trading* ao mercado financeiro. No entanto, é importante ressaltar que outros autores questionam ao menos parcialmente os estudos realizados por Eugene Fama, seja por resultados inconclusivos ou por anomalias detectadas no comportamento do mercado. Por exemplo, Frank Shostak [26] critica abertamente a premissa de que todos os investidores teriam a mesma expectativa sobre os retornos da empresa. O ganhador do prêmio Nobel em ciências econômicas Paul Samuelson, que afirma que o a HME funciona muito melhor para ações individuais do que para o mercado como um todo [27]. Já o investidor Jack Schwager afirma que a HME está correta pelos motivos errados [28], pois é muito difícil bater a média do mercado de forma consistente ao mesmo tempo que investidores possuem habilidades diferentes, portanto a informação não é interpretada e aplicada por todos da mesma forma.

2.1.2 Índice de Bolsa de Valores

Índices de Bolsas de Valores [29] são métricas criadas para avaliar a saúde de um determinado grupo de ações negociadas na bolsa. Cada índice possui uma regra própria de criação que define quais ações são englobadas e com quais pesos, como por exemplo:

- S&P 500 (*Standard and Poor's 500*): Um dos mais conhecidos no mercado global. É a média ponderada pelo capital social das 500 maiores companhias do mercado americano.
- DJIA (*Dow Jones Industrial Average*): É a média ponderada pelo preço das ações das 30 maiores *blue-chips*¹ industriais e financeiras do mercado americano.
- Ibovespa (Índice Bovespa): Principal indicador de desempenho do mercado brasileiro. Possui alguns critérios específicos, mas basicamente é composto pelas ações com maior volume de negociação na B3 [30].

Índices não são negociáveis pois não passam de métricas de mercado. Para isso existem fundos de investimentos chamados ETFs [31], especializados em seguir um determinado índice.

No Brasil, um investidor que deseja que uma parte de seu capital acompanhe um rendimento equivalente ao Ibovespa deverá investir no ETF cujo código de negociação é BOVA11.

2.1.3 Mercado Fracionário

Ações são negociadas em múltiplos de um lote, que representa uma quantidade mínima de papéis. Nesse contexto, o Mercado Fracionário [32] surge com o objetivo de facilitar negociações de volumes menores que o lote mínimo permitido. No Brasil, o lote é de 100 ações e o Mercado Fracionário permite a compra de no mínimo 1 ação.

¹Companhias bem conhecidas, bem estabelecidas e com grande capital social.

Na prática, ações fracionárias são agrupadas até formarem um lote para então serem negociadas. Elas podem ser criadas devido: a um desdobramento de ações que não gera resultado par (e.g., 3 para 2); ou a fusões e aquisições de empresas que combinam suas ações em uma razão predeterminada.

Normalmente o Mercado Fracionário possui menor liquidez e maior volatilidade, mas sempre acompanha o preço do ativo negociado no mercado aberto.

Grandes investidores e fundos de investimentos não possuem problemas quanto ao capital mínimo necessário para a compra de um lote de ações, visto que negociam em quantidades muito maiores. O problema surge quando um investidor com pouco aporte financeiro deseja entrar no mercado e não consegue encontrar ativos cujo lote mínimo esteja dentro de seu orçamento.

2.2 Tipos de Análises

2.2.1 Análise Fundamentalista

A Análise Fundamentalista é um bastante utilizada para identificar tendências de flutuação no preço de ações tendo em vista um horizonte de longo prazo [33]. Ela se baseia em fatores econômicos relacionados à companhia, como: o quadro de diretores e dirigentes maiores; o fluxo de caixa; a saúde e a situação financeira; o contexto político do país; os concorrentes de mercado; as circunstâncias e os desastres climáticos, naturais ou não, dentre outros fatores.

Devido à natureza desorganizada e desestruturada do acesso e da leitura dos dados que representam os fatores mencionados, torna-se difícil implementar uma automação eficaz.

2.2.2 Análise Técnica

A Análise Técnica busca identificar tendências de curto prazo na série temporal de preços de ações através da identificação de padrões e da criação de informações derivadas (indicadores técnicos) [34, 35]. Segundo a Teoria de Dow, o preço das ações

é consequência de todos os acontecimentos relacionados direta ou indiretamente a uma companhia [2].

Diferentemente da AF, a automação desta análise é mais fácil pois os dados normalmente são organizados e estruturados. No entanto, como são obtidos *a posteriori*, a dificuldade desta análise se dá na separação entre o que é ruído e o que é de fato tendência de mercado, e também na criação de informações derivadas que se mostrem relativamente úteis.

Dentre os indicadores mais famosos e portanto utilizados, podemos citar: o volume financeiro; a identificação de tendências de alta, de baixa e de consolidação de acordo com a Teoria de Dow; as linhas de suporte e de resistência do mercado; as médias móveis; as bandas de Bollinger [36]; e o MACD [37].

Leitura de Gráficos de Candlesticks

Gráficos de Candlesticks² são bastante utilizados na AT. A leitura é padronizada de acordo com a Figura 2.1. Neste tipo de gráfico as cores importam, pois indicam se o balanço do período foi positivo ou negativo.

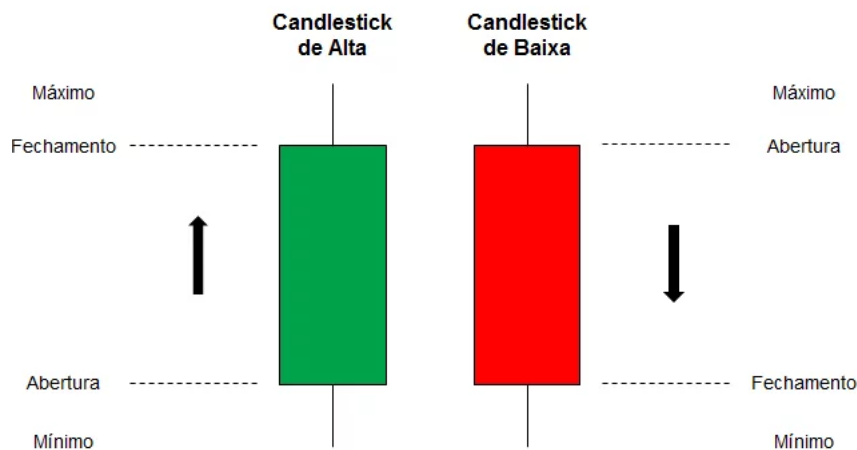


Figura 2.1: Leitura de um gráfico de *candlestick* [1]

²Em português: Gráfico de Velas.

Teoria de Dow

A Teoria de Dow, criada pelo americano Charles Henry Dow em 1884 é considerada a base da AT moderna [2]. Embora não tivesse sido formalizada explicitamente pelo autor enquanto estava vivo, amigos e profissionais da época tiveram o trabalho de divulgar e fazer alguns ajustes. Baseada na HME, a ideia central por trás da Teoria de Dow é que a lógica econômica deve ser usada para explicar os movimentos do mercado, que em condições ideais segue o padrão de: tendência de alta³; topo; tendência de baixa⁴; e fundo, intercalados com períodos de consolidação⁵. A Figura 2.2 ilustra esse comportamento.



Figura 2.2: Comportamento do mercado ideal segundo a Teoria de Dow [2]

Suporte, Resistência e Linhas de Tendência

Suporte e Resistência são regiões em um gráfico de *candlestick* onde existe um efeito memória associado a grandes ganhos ou perdas históricas [3]. Normalmente estão associadas a eventos econômicos relevantes. A Figura 2.3 ilustra essas regiões, comumente chamadas de Linhas de Suporte e de Resistência.

De maneira semelhante, as Linhas de Tendência oferecem uma inspeção gráfica do quanto o preço de um ativo está crescendo ou diminuindo. Portanto, estão necessariamente atreladas a movimentos de tendência de alta ou de tendência de baixa. Em essência, não deixam de ser linhas de Suporte e de Resistência. As Figuras 2.4 e 2.5 exemplificam esses indicadores.

³Topos e fundos ascendentes.

⁴Topos e fundos descendentes.

⁵Topos e fundos lateralizados.

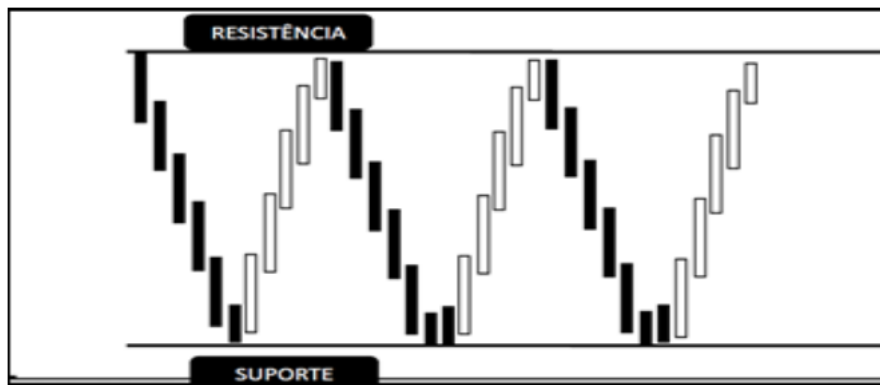


Figura 2.3: Formação de linhas de Suporte e de Resistência [3]

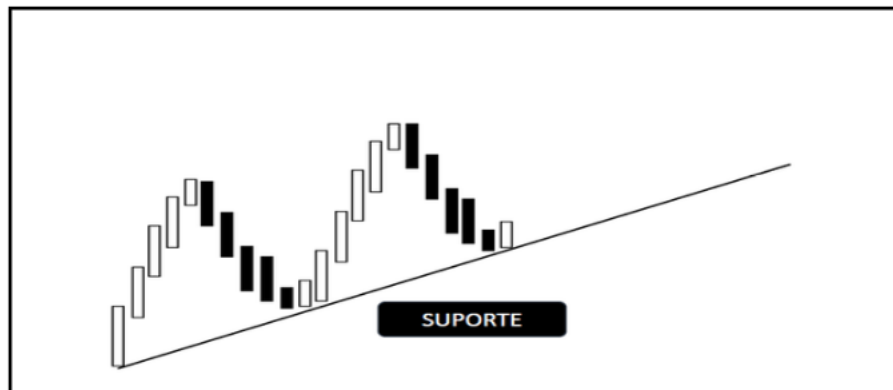


Figura 2.4: Formação de uma Linha de Tendência de Alta [3]

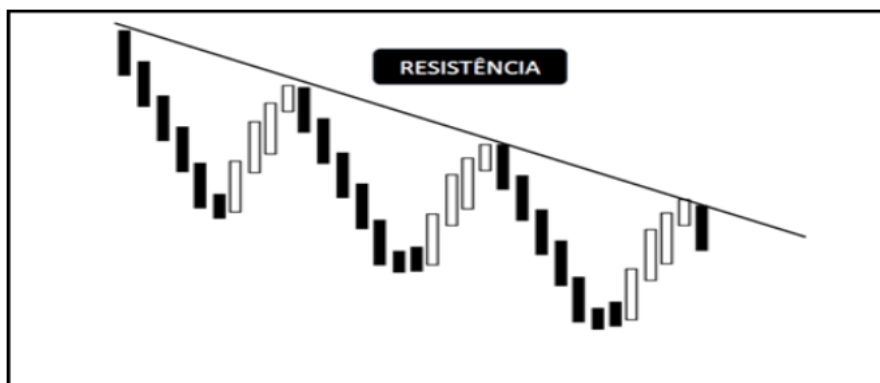


Figura 2.5: Formação de uma Linha de Tendência de Baixa [3]

2.3 Aprendizado de Máquina

Aprendizado de Máquina é um campo de estudo dentro de Inteligência Artificial [38] que engloba estatística e ciência da computação. O objetivo é extrair conhecimento a partir de um conjunto de dados [4]. A terminologia foi criada por um pesquisador da IBM chamado Arthur Samuel em 1959 [39] para um estudo de caso do jogo de damas [40].

Em geral, algoritmos de ML buscam realizar tarefas extremamente complexas computacionalmente sem serem explicitamente programadas caso a caso. Alguns exemplos de aplicações que deixam evidente os benefícios deste método são: visão computacional, identificação de rosto, recomendação de produtos em plataformas de *e-commerce*, identificação de transações financeiras fraudulentas, suporte a diagnósticos médicos, dentre diversos outros.

Algoritmos de ML podem ser baseados em Aprendizado Supervisionado, Aprendizado Não Supervisionado ou até mesmo em um aprendizado híbrido. Este trabalho utiliza apenas AS para a criação de modelos.

2.3.1 Aprendizado Supervisionado

Uma das metodologias mais comuns de ML, seu objetivo é a predição de um resultado a partir de um conjunto de dados de entrada, com a condição de que o modelo tem acesso a vários exemplos de entrada e saída de dados para obter uma melhor performance [4].

O conjunto de dados (*dataset*) com exemplos de entrada e saída utilizado para criação do modelo é chamado de conjunto de dados de treinamento (*training set*). Existe um outro conjunto de dados utilizado para testar a performance do modelo. Este segundo conjunto, chamado de dados de teste (*test set*), precisa ser necessariamente diferente dos dados de treinamento para evitar que o efeito memória se sobreponha à qualidade de generalização do modelo (explicado a seguir). Como regra geral de uso, é aconselhável separar 75% dos dados para o conjunto de treinamento e 25% para o conjunto de teste, ou algo próximo desta proporção [4].

Todo modelo pode ser avaliado sob o ponto de vista da generalização. Essa característica indica a capacidade de realizar previsões acuradas no conjuntos de dados de teste. Quanto maior a taxa de acerto no conjunto de teste, melhor tende a ser a capacidade de generalização.

Outras características importantes são conhecidas como *overfitting* e *underfitting*. Quando um modelos está muito complexo a ponto de ser sensível demais aos ruídos do conjunto de treinamento, trazendo dificuldades de generalização, diz-se que ocorreu um *overfitting*. De forma análoga, quando a complexidade do modelo é baixa de forma a não aproveitar devidamente as característica importantes do conjunto de treinamento, implicando também em perda de generalização, diz-se que ocorreu um *underfitting*. O objetivo do projetista de um modelo por AS é encontrar um ponto de equilíbrio entre essas características, chamada de “*sweet spot*” na Figura 2.6, que mostra a relação entre generalização, *overfitting* e *underfitting*.

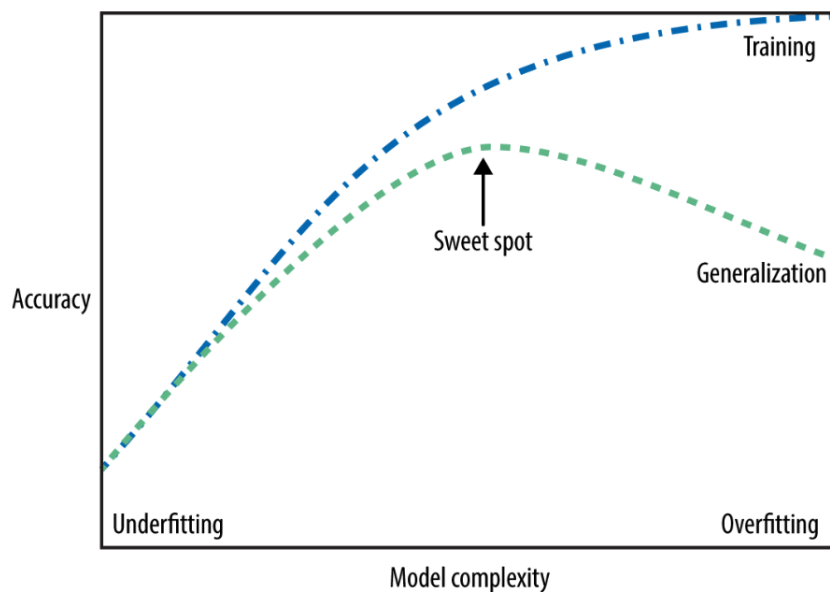


Figura 2.6: Relação entre complexidade e acurácia de um modelo [4]

Existem dois tipos de problemas associados ao AS, os problemas de Regressão e os problemas de Classificação.

2.3.2 Problema de Regressão

Este problema envolve a predição de um número contínuo a partir dos dados de entrada [4]. Para exemplificar, pode-se citar a probabilidade de uma pessoa desenvolver uma doença auto-imune a partir de indicadores médicos específicos. Ou também um índice que traz uma expectativa de quantos kilogramas de milho serão colhidos em uma safra a partir de dados geológicos e meteorológicos.

2.3.3 Problema de Classificação

Os problemas de classificação buscam escolher um rótulo (ou classe) mais provável dentre uma lista de possibilidades finitas e pré-estabelecidas [4]. Como aplicações, pode-se citar: a previsão de escolha eleitoral de pessoas a partir de indicadores socioeconômicos; o diagnóstico de câncer em pacientes a partir de informações médicas; ou mesmo a presença e ausência de animais catalogados em um conjunto de imagens.

É importante mencionar que problemas de classificação precisam de atenção ao balanceamento das classes (i.e. mesma relevância para cada classe durante o treinamento). Em outras palavras, um conjunto de dados não balanceado pode gerar um modelo pouco complexo para uma aplicação não trivial, o que implica em um índice ilusório de acurácia nos dados de teste. Isso acontece porque o modelo tende a quase sempre escolher a classe com maior frequência em seu treinamento, independentemente da composição dos dados. Para corrigir este efeito, deve-se deixar todas as classes com a mesma relevância durante o treinamento do modelo, o que pode ser feito através dos seguintes métodos:

- *Undersampling*: Diminuição de amostras pertencentes à classe mais presente. É aconselhável quando o *dataset* é grande o suficiente para suportar a perda de dados sem perda significativa de generalização. Como vantagem, diminui o tempo de treinamento de um modelo. Ver Figura 2.7.
- *Oversampling*: Replica ou gera sinteticamente amostras pertencentes à classe menos presente. Como consequência, não há perda de informação potencialmente relevante, porém pode gerar *overfitting*. Pode ser uma boa opção em *datasets* pequenos [41]. Ver Figura 2.7.

- *Cost Sensitive Learning* (CSL): Ao invés de alterar o tamanho do *dataset*, criam-se pesos diferentes para um erro de classificação durante o treinamento. Portanto, um erro em uma classe menos frequente deve ser mais penalizado do que o contrário. É aconselhável em *datasets* grandes (> 10000) [41].

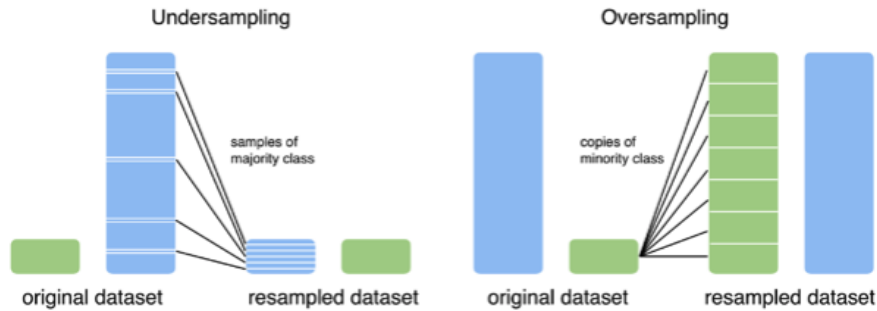


Figura 2.7: *Oversampling* e *Undersampling* de classes desbalanceadas [5]

2.3.4 Algoritmos de Aprendizado Supervisionado

Esta seção trará uma visão simplificada sobre os algoritmos de AS mais pertinentes ao presente trabalho, em ordem crescente de complexidade. Os exemplos citados serão focados em problemas de classificação apenas para entendimento do raciocínio por detrás dos modelos, porém todos possuem variantes para problemas de regressão.

k-Nearest Neighbors

k-NN é talvez o algoritmo mais simples de todos. Consiste na memorização dos dados de treinamento para prever a classe ou o valor a partir da média dos K registros mais próximos encontrados. A Figura 2.8 mostra como funciona o critério de seleção da classe de uma amostra de teste a partir dos dados de treinamento e do parâmetro K de vizinhos selecionados.

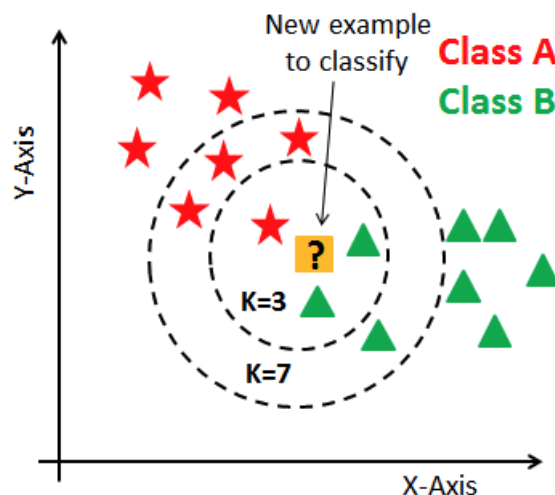


Figura 2.8: Funcionamento de um algoritmo k-NN para o problema de classificação [6]. Para $K=3$ a classe é B e para $K=7$ a classe é A.

Decision Tree

Em essência, uma Árvore de Decisão⁶ é uma sequência hierárquica de estruturas de decisão *if/else*⁷ acerca das características do conjunto de dados. Tecnicamente, pode-se construir uma Árvore de Decisão até que todas as suas folhas estejam totalmente puras, ou seja, as sequências de decisão que levam a um resultado só englobam amostras de um tipo de classe. Por outro lado, folhas impuras contém a presença de mais de uma classe, onde se escolhe a classe de maior número de amostras como resultado. O problema é que a presença excessiva de folhas totalmente puras normalmente é acompanhado de um *overfitting* do modelo, portanto precisa ser controlado. Para isso, é possível ajustar alguns parâmetros, como por exemplo: a profundidade, que define a quantidade máxima de camadas que a árvore atingirá qualquer que seja o ramo; o número mínimo de amostras necessário para criação de uma nova ramificação; dentre outros.

Algumas das vantagens deste modelo estão no fácil entendimento e visualização dos critérios de decisão em árvores pequenas aos olhos do projetista. O tempo de processamento computacional envolvido na criação deste modelo é razoavelmente

⁶Em inglês: *Decision Tree*.

⁷Em português: se/senão.

curto. Não é necessário um pré-processamento dos dados, uma vez que cada característica é processada separadamente. A Figura 2.9 exemplifica a estrutura por trás de uma Árvore de Decisão.

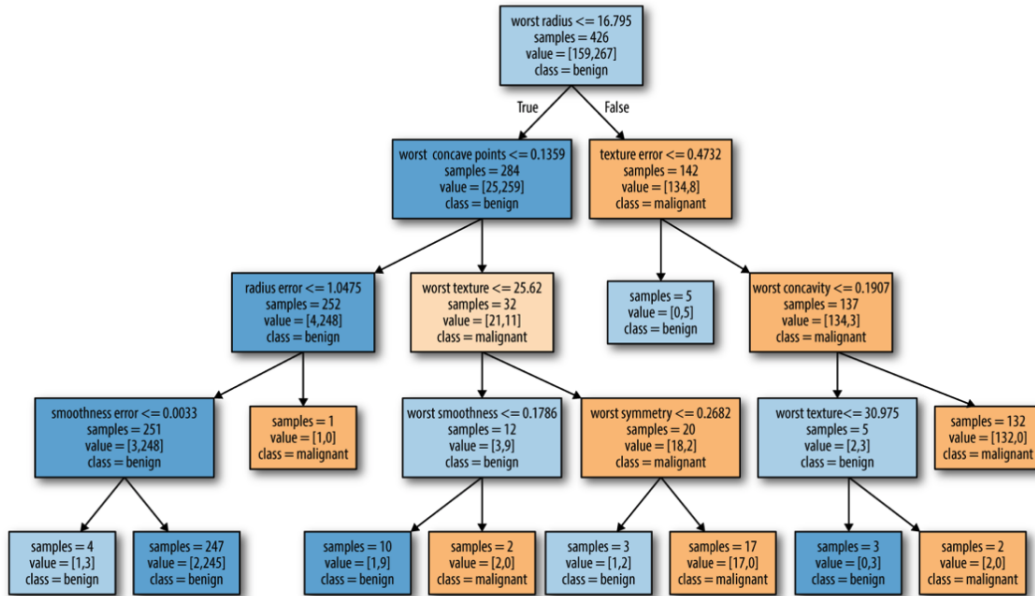


Figura 2.9: Visualização de uma Árvore de Decisão para um *dataset* de câncer de mama [4].

Por outro lado, uma desvantagem eminente é a tendência *overfitting* e a baixa capacidade de generalização, que podem ser mitigados através de um algoritmo derivado chamado *Random Forest*.

Random Forest

Um dos modelos mais utilizado atualmente, o algoritmo *Random Forest*⁸ é a combinação de diversas Árvores de Decisão ligeiramente diferentes entre si [4]. A ideia é que apesar da tendência de *overfitting* existente, a média dos resultados de cada árvore tende a diminuir esse fator. Além dos parâmetros responsáveis por configurar as árvores individualmente, este modelo também precisa do número de árvores que serão utilizadas.

⁸Em português: Floresta Aleatória.

Normalmente é preferível utilizar *Random Forests* ao invés de Árvores de Decisão, salvo casos em que o entendimento e a visualização clara do modelo se torna um fator importante. É possível compensar o aumento do tempo de processamento envolvido na criação de uma *Random Forest* com a paralelização dos núcleos de processamento da CPU⁹.

2.4 Walk-Forward Analysis

Walk-Forward Analysis [42] é um processo de otimização mais voltado para séries temporais no contexto de finanças. Para isso, o *dataset* é dividido em múltiplos segmentos consecutivos, que são iterados progressivamente a fim de se obter os parâmetros ou modelos desejados. A Figura 2.10 ilustra o processo. Observa-se que são utilizados os termos *in-sample data* (IS) como sinônimo de *training set* e *out-of-sample data* (OOS) como sinônimo de *test set*. A imagem à esquerda representa um processo não-ancorado, onde o início do IS caminha com o decorrer do processo, já a imagem à direita mantém o início fixo por ser um processo ancorado.

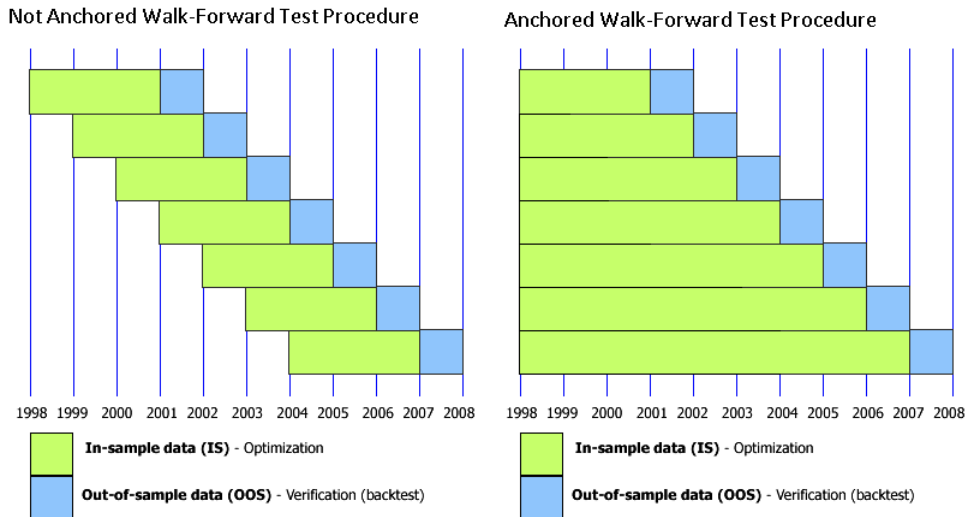


Figura 2.10: *Walk-Forward Analysis* (Não-Ancorado e Ancorado) [7].

A natureza do WFA ancorado permite uma maior adaptação dos modelos de ML de acordo com as tendências de mercado, que mudam significativamente com

⁹Do inglês: *Central Process Unit*.

o tempo. Durante o treinamento de um modelo, considerar informações temporalmente muito distantes do período de aplicação do mesmo pode comprometer sua acurácia, pois os padrões que guiavam os preços anteriormente não necessariamente são iguais aos padrões atuais.

2.5 Considerações para Análise de Resultados

2.5.1 Índice de Sharpe

Criado pelo americano William F. Sharpe em 1966 e revisado em 1994, o Índice de Sharpe¹⁰ tem como objetivo medir a performance de um investimento em relação a sua volatilidade, levando também em consideração o rendimento e a volatilidade de um investimento relativamente livre de risco (e.g., título público) [43]. Seja R_a o retorno do investimento alvo, R_b o retorno do investimento livre de risco e σ_a seu respectivo desvio padrão, pode-se calcular o Índice de Sharpe através da Equação 2.1.

$$S_a = \frac{E[R_a - R_b]}{\sigma_a} \quad (2.1)$$

2.5.2 Índice de Sortino

O Índice de Sortino, criado pelo americano Frank Sortino [44] é uma variante do Índice de Sharpe que considera o desvio padrão apenas dos rendimentos abaixo da média. As Equações 2.2 e 2.3 mostram seu cálculo, onde assim como no Índice de Sharpe, R_a é o retorno do investimento alvo, R_b é o retorno do investimento livre de risco e σ_a seu respectivo desvio padrão. Considere também X_i o i -ésimo retorno e T o retorno médio do investimento.

$$S_a = \frac{E[R_a - R_b]}{\sigma_a} \quad (2.2)$$

$$\sigma_a = \sqrt{\frac{1}{N} \sum_{i=1}^N (\text{Min}(0, X_i - T))^2} \quad (2.3)$$

¹⁰Também conhecido como *Sharpe Index*, *Sharpe Ratio* ou até *Sharpe Measure*.

2.5.3 Correlação de Spearman

A Correlação de Postos de Spearman ou simplesmente Correlação de Spearman foi criada pelo psicólogo inglês Charles Edward Spearman e revelada em 1904 [45]. Em resumo, a Correlação avalia o grau de proximidade que duas variáveis aleatórias possuem em relação a uma função monotônica. Matematicamente, é o mesmo que a correlação de Pearson aplicada aos postos¹¹ das duas variáveis envolvidas.

Para duas variáveis aleatórias X_i e Y_i , são criados os postos rgX_i e rgY_i para as N amostras presentes. Existem duas formas de se calcular o coeficiente: a primeira, mostrada pela Equação 2.4, é para o caso em que há apenas postos inteiros distintos, sem presença de nós (i.e., valores iguais em cada uma das variáveis); já a segunda, ilustrada pela Equação 2.5, é para o caso em que há presença de nós.

$$\rho_s = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)} \quad (2.4)$$

$$\rho_s = \rho_{rgX, rgY} = \frac{cov(rgX, rgY)}{\sigma_{rgX}, \sigma_{rgY}} \quad (2.5)$$

Na Equação 2.4, d_i é diferença entre os dois postos de cada variáveis aleatória, mostrado através da Equação 2.6.

$$d_i = rgX_i - rgY_i \quad (2.6)$$

Um exemplo prático do cálculo da correlação pode ser analisado a partir da Tabela 2.1, onde são ilustradas amostras para duas variáveis aleatórias X e Y . A Tabela 2.2 acrescenta a informação dos postos rgX e rgY , que ordenam as amostras em ordem decrescente. Observa-se em rgX a presença de dois postos com valores de 6.5, causados pelo nó em X de dois valores repetidos (61). Neste caso, a regra é a escolha valor médio dos postos que seriam ocupados, no caso 6 e 7.

Após o cálculo dos postos, deve-se utilizar a Equação 2.5, encontrando-se o valor aproximado de 0.6687.

¹¹Classificação ordenada das amostras em escala ordinal. Do inglês: *ranks*.

	1	2	3	4	5	6	7	8	9	10
X	56	75	45	71	61	64	58	80	76	61
Y	66	70	40	60	65	56	59	77	67	63

Tabela 2.1: Amostras das variáveis aleatórias X e Y

	1	2	3	4	5	6	7	8	9	10
X	56	75	45	71	61	64	58	80	76	61
Y	66	70	40	60	65	56	59	77	67	63
rgX	9	3	10	4	6.5	5	8	1	2	6.5
rgY	4	2	10	7	5	9	8	1	3	6

Tabela 2.2: Postos rgX e rgY

2.6 Trabalhos Relacionados

Tendo em vista o conflito de interesses existente por trás de trabalhos de cujo tema está relacionado à previsibilidade do mercado financeiro, pode-se questionar se as estratégias mais promissoras de fato são encontradas em domínio público. Isso ocorre pois a democratização de uma estratégia lucrativa poderia implicar na redução das lucratividades individuais, especialmente se for utilizada em escala.

Segundo Kendall Kim [46], somente a partir dos anos 80 que as corretoras começaram a utilizar protocolos de comunicação eletrônica para substituir a corretagem por voz. Essa inovação permitiu o desenvolvimento do *Algorithmic Trading*, que é a automação da tomada de decisões de estratégias por um computador capaz de enviar ordens de compra e venda diretamente ao mercado.

A partir do trabalho de Danilo Pereira [47], pode-se simplificar os modelos de AT aplicados ao mercado financeiro em três metodologias centrais: modelos baseados em indicadores técnicos; modelos baseados em processos estocásticos; e modelos baseados em aprendizado de máquina.

2.6.1 Modelos Baseados em Indicadores Técnicos

Este tipo de abordagem utiliza informações derivadas da série temporal de preços para criar uma combinação de indicadores que possuam algum poder de previsibilidade de tendência de mercado. Quando comparada aos outros tipos, é a metodologia mais simples e democrática, uma vez que investidores com um conhecimento superficial sobre estatística e inteligência artificial já podem operar em estratégias próprias.

Diversos *traders*¹² e investidores utilizam este tipo de abordagem. Dentre eles podemos citar André Moraes [3], de cujas contribuições servirão como base neste trabalho para um aperfeiçoamento via aprendizado de máquina.

2.6.2 Modelos Baseados em Processos Estocásticos

De acordo com Michael Godfrey *et al.* [48], a hipótese de que a flutuação de preços no mercado de ações poderia ser explicada por uma *Random Walk*¹³ foi feita por Louis Bachelier [49]. A partir da década de 60, muitos trabalhos acadêmicos foram realizados nessa linha na tentativa de entender o comportamento e a previsibilidade do mercado [23, 50, 51], assim como estratégias [52]. Nota-se que até hoje utiliza-se *Random Walks* para testar a hipótese de eficiência de mercados [53].

Outra abordagem utilizada são os Modelos Ocultos de Markov (do inglês *Hidden Markov Model*) [54]. Uma Cadeia de Markov é um processo estocástico que modela um sistema por meio de uma sequência finita de estados. A mudança ou a permanência em cada estado é determinada por probabilidades que dependem somente do estado atual. Em uma Cadeia de Markov, pressupõe-se que seus estados sejam observáveis, o que para algumas aplicações, pode não ser verdade. Nesse sentido surge o modelo HMM, que busca aprender sobre um processo não observável

¹²Em português: negociantes. Pessoas que compram e vendem bens, moedas ou ações com o objetivo de lucrar, mas não necessariamente com foco em investimento, podendo até assumir um papel especulativo.

¹³Processo aleatório definido pela equação $y_t = y_{t-1} + X$, onde X é uma variável aleatória e y é a variável resultante.

(oculto) a partir de um processo observável.

Em sua pesquisa, Aishwary Jadhav *et al.* [55] utiliza um modelo HMM para previsão do preço de fechamento do dia seguinte para ações FAANG¹⁴. A partir da série histórica de preços OHLC¹⁵, seu modelo atinge uma eficiência de 97%-99%, calculado a partir do erro percentual absoluto médio¹⁶.

Uma outra aplicação de modelos HMM é dada por Luca De Angelis *et al.* [56], que criou uma metodologia a partir de índices da bolsa americana capaz de identificar períodos estáveis e instáveis (i.e. crises econômicas), assim como as probabilidades de transição entre um estado e o outro.

Por fim, pode-ser mencionar o uso de modelos ARCH¹⁷. A ideia central está na modelagem de uma variância condicional, ou seja, que muda de acordo o instante da série [57]. Essa característica se faz muito útil em séries que possuem períodos de alta volatilidade se alternando com períodos de baixa volatilidade. Para um modelo genérico ARCH(q), seja ϵ_t o erro (resíduo) no instante t e α_0 um ruído branco, pode-se descrever a variância condicional de acordo com a Equação 2.7.

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^q \alpha_i \epsilon_{t-i}^2 \quad (2.7)$$

O modelo ARCH foi proposto por Robert Engle em 1982 para estimar a variância da inflação do Reino Unido [58]. A partir daí, várias derivações surgiram, como por exemplo: GARCH¹⁸ por Tim Bollerslev [59] em 1986, EGARCH¹⁹ por Daniel Nelson [60] em 1991, NGARCH²⁰ por Matthew Higgins *et al.* [61] em 1992, TGARCH²¹ por Roger Rabemananjara *et al.* [62] em 1993, dentre outros. Alguns dos modelos

¹⁴Facebook, Amazon, Apple, Netflix, Google.

¹⁵Open, High, Low, Close. Em português: Abertura, Máximo, Mínimo, Fechamento.

¹⁶Mean Absolute Percentage Error (MAPE): $\frac{1}{N} \sum_{i=1}^N \frac{|Predicted(i) - Actual(i)|}{Actual(i)}$

¹⁷Em português: Heteroscedasticidade Condicional Auto-regressiva.

¹⁸Generalised ARCH.

¹⁹Exponential Generalised ARCH.

²⁰Non-linear Generalised ARCH.

²¹Threshold Generalised ARCH.

da família ARCH podem ser encontrados nos trabalhos de Philip Franses *et al.* [63], de Juri Marcucci [64] e de Dima Alberg *et al.* [65].

2.6.3 Modelos Baseados em Aprendizado de Máquina

Existem registros de estudos sobre inteligência artificial aplicados ao mercado financeiro por volta da década de 70 [66], porém ainda em um estágio embrionário devido às dificuldades de processamento computacional e de acesso a dados na época. Por ser uma área de estudo extremamente dependente de ambas as questões, conforme elas foram evoluindo, mais trabalhos puderam ser realizados sobre o tema.

Isaac Nti *et al.* [67] relata que dos 122 trabalhos mais relevantes publicados entre 2007 e 2018 com o tema de predição do mercado financeiro usando ML, 66% são baseados em AT, 23% são baseados em AF e 11% usam análises mistas. Além disso, os algoritmos mais utilizados são ANN²² (*Artificial Neural Networks*) e SVM²³ (*Support Vector Machine*).

De forma semelhante, Dattatray Gandhmal *et al.* [68] verificou que a partir de uma análise detalhada de 50 trabalhos com o tema de predição do mercado financeiro, os algoritmos que mais costumam trazer resultados efetivos são ANN e técnicas baseadas em lógica *Fuzzy*²⁴

É possível encontrar também modelos híbridos, com uma combinação de GARCH com ANN feita por Melike Bildirici *et al.* [69].

²²Em português: Redes Neurais Artificiais.

²³Em português: Máquina de Vetor de Suporte.

²⁴Em português: Difuso.

Capítulo 3

Metodologia

3.1 Resumo

As seções a seguir trazem detalhes quanto a estrutura técnica do projeto. A Figura 3.1 apresenta um diagrama geral de como essas estruturas se conectam.



Figura 3.1: Estrutura do técnica do projeto

Antes da execução do código principal, é necessário garantir que os modelos estão devidamente criados e acessíveis. Para isso, é importante a elaboração dos *datasets* de cada ação a ser simulada, pois servem de entrada de dados para a criação e seleção dos modelos. A biblioteca *multiprocessing* foi utilizada para minimizar o tempo total gasto nestas etapas.

Após a criação dos modelos, tem-se início a etapa de pré-processamento de dados, onde ocorre a leitura e interpretação do arquivo de configuração para se obter o número de estratégias a executar, além dos ativos envolvidos e seus respectivos intervalos de tempo. Uma vez verificado no banco os dados já existentes, faz-se um *download* apenas dos dados necessários. Se houver alguma atualização, as *features* de uso geral são recalculadas e armazenadas no banco a fim de servir de insumo para as estratégias que estarão por vir.

Completa a etapa de pré-processamento, inicia-se a simulação das estratégias. O arquivo de configuração foi projetado para ser capaz de designar diversas estratégias de parâmetros distintos a uma mesma ordem de execução de programa. Também fez-se uso da biblioteca *multiprocessing* para paralelizar as simulações, cujos resultados e estatísticas são salvas no banco para posterior análise.

Por fim, é possível visualizar os resultados de forma clara através de uma aplicação secundária responsável por criar um *dashboard* interativo.

Em relação às tecnologias utilizadas, a aplicação foi desenvolvida em *Python* com o apoio das bibliotecas *yfinance*, *pandas*, *numpy*, *scikit-learn*, *multiprocessing*, *matplotlib* e *dash*. Foi estruturado um banco de dados PostgreSQL para armazenamento dos *candlesticks* obtidos, das *features* geradas e das estratégias simuladas. Também foi incorporado o uso de *Docker* especificamente para a execução de estratégias sem a necessidade de configuração de ambiente.

A Tabela 3.1 lista todos os 71 ativos escolhidos para simulação no escopo deste trabalho. Os critérios de escolha envolvem as seguintes preferências: diversidade de segmentos; disponibilidade da série temporal de dados a partir de 2013; e presença na composição do IBovespa em qualquer data.

Ações Escolhidas (71)

ABEV3	ALPA4	AMER3	B3SA3	BBAS3	BBDC3	BBDC4	BBSE3
BEEF3	BPAN4	BRAP4	BRFS3	BRKM5	BRML3	CCRO3	CIEL3
CMIG4	COGN3	CPFE3	CPLE6	CSAN3	CSNA3	CVCB3	CYRE3
DXCO3	ECOR3	EGIE3	ELET3	ELET6	EMBR3	ENBR3	ENEV3
ENGI11	EQTL3	EZTC3	FLRY3	GGBR4	GOAU4	GOLL4	HYPE3
ITSA4	ITUB4	JBSS3	JHSF3	LAME4	LCAM3	LREN3	MGLU3
MRFG3	MRVE3	MULT3	PETR3	PETR4	POSI3	PRI03	QUAL3
RADL3	RENT3	SANB11	SBSP3	SULA11	TAE11	TIMS3	TOTS3
UGPA3	USIM5	VALE3	VIAA3	VIVT3	WEGE3	YDUQ3	

Tabela 3.1: Ações Escolhidas

3.2 Pré-Processamento

3.2.1 Arquivo de Configuração

O Arquivo de Configuração é um arquivo no formato JSON responsável por configurar detalhadamente cada parâmetro da sequência de estratégias que se deseja executar. Uma ordem de execução do programa pode conter diversas simulações de estratégias, que são configuradas neste Arquivo. A Figura 3.2 mostra sua estrutura.



Figura 3.2: Estrutura do Arquivo de Configuração

Nota-se que no topo são listados os parâmetros de uso geral, ou variáveis de escopo global, de cujos valores precedem quaisquer outros listados a seguir, em caso de sobreposição. Em seguida abre-se o vetor de tipos de estratégias, onde o campo *name* representa o nome da classe selecionada, sendo este o elemento que conecta o usuário ao tipo de estratégia desejada. Ressalta-se que este trabalho compreende apenas uma estratégia. Após a seleção do nome, são configurados os parâmetros internos da estratégia. A Tabela 3.3 da Seção 3.3.7 descreve todos os parâmetros disponíveis.

Para se criar mais de um perfil de simulação, é necessário modificar o Arquivo conforme a Figura 3.3. Automaticamente, o código interpreta que existe mais de uma simulação a executar, com todos os parâmetros em comum exceto aqueles em formato de listas. Caso haja mais de um parâmetro no formato de lista, seus comprimentos precisam ser iguais. No caso da Figura 3.3, a primeira simulação utilizará os valores (100, 0.01) para o par (variável_local_1, variável_local_2), a segunda utilizará (200, 0.02) e assim sucessivamente.

```
{
  "variável_global_1": false,
  "variável_global_2": 1.0,
  "strategies": [
    {
      "name": "Estratégia",
      "comment": "Maximização de Ganhos.",
      "variável_local_1": [100, 200, 300],
      "variável_local_2": [0.01, 0.02, 0.03],
      "stock_targets": [
        {
          "name": "XYZW1",
          "start_date": "01/01/2019",
          "end_date": "31/03/2021"
        },
        {
          "name": "XYZW2",
          "start_date": "01/01/2019",
          "end_date": "31/03/2021"
        }
      ]
    }
  ]
}
```

3 Estratégias

Figura 3.3: Arquivo de Configuração para Execuções Múltiplas

3.2.2 Coleta de Dados

A Coleta de Dados ocorre através da biblioteca *open-source yfinance* [70], uma ferramenta não oficial que transmite dados públicos da plataforma *Yahoo! Finance*

[71], um subsistema da rede *Yahoo!*.

A escolha desta biblioteca como fonte primária de dados se deve principalmente pela facilidade de uso associada à ausência de custos. Contudo, alguns testes e verificações com outras fontes de dados evidenciaram destantagens relevantes, porém não impeditivas para uso. São elas:

- Os valores de proventos que a biblioteca disponibiliza não são consistentes com as declarações dos sites das próprias companhias, portanto não podem ser utilizados por este projeto. Testes internos confirmaram a presença de diversos proventos corretamente apresentados e ajustados pelos respectivos desdobramentos acumulados. O problema é que os mesmos estavam misturados com alguns *outliers* inexistentes na realidade, o suficiente para questionar o uso em escala (i.e., para vários ativos sem verificação individual). O Apêndice A evidencia os problemas encontrados em mais detalhes.
- Os volumes de negociação disponibilizados coincidem em valores relativos com os volumes da plataforma *TradingView* [72] (i.e., variação de volume comparativa para um mesmo ativo), o que é suficiente para este trabalho. Não foram encontradas evidências do contrário.
- *Candlesticks* de janelas temporais inferiores à diária (*intraday*) são disponibilizados, porém as limitações envolvidas inviabilizam seu uso, como: o limite de 730 dias para a busca dos dados; a inconsistência com os dados diários quanto ao volume; e a alguns *bugs* como a ausência de *candlesticks* em todo dia de parcial do pregão da B3 (Quarta-feira de Cinzas).

Apenas os dados não existentes no banco são baixados via *yfinance*. Para isso, um *trigger*¹ é acoplado às tabelas de *candlesticks* e acionado sempre que operações de *insert*, *update*, *delete* e *truncate* são realizadas. Quando ativado, ele chama uma função responsável por atualizar a tabela de *status*, que registra o intervalo de tempo representado nas tabelas de *candlesticks* para cada *ticker* envolvido. Deve-se

¹Procedimento armazenado em um banco de dados que é chamado automaticamente sempre que ocorre um evento.

ressaltar que os devidos cuidados foram tomados para evitar buracos entre intervalos de tempo não adjacentes. Portanto, apenas uma consulta à tabela de *status* é executada para se verificar a necessidade de *download* de novos dados.

3.2.3 Armazenamento de Dados

O Armazenamento de Dados é realizado por um banco de dados *PostgreSQL*, criado com o objetivo de salvar: os resultados das simulações; as *features* de uso geral; e os *candlesticks* obtidos. As vantagens de um banco de dados em relação a um arquivo CSV ou a uma planilha de Excel dispensam comentários. Contudo, quanto ao escopo deste trabalho, pode-se mencionar os seguintes pontos:

- Fácil acesso aos resultados das simulações de forma estruturada e consistente, recurso este utilizado pela aplicação que gera o *dashboard*.
- Economia de processamento devido ao armazenamento das *features* de uso geral, uma vez que estratégias simuladas não necessitam recalculá-las a cada execução.
- Independência da plataforma *Yahoo! Finance* para o caso de não continuidade dos dados ou qualquer alteração repentina.
- Diminuição do tráfego na rede pela persistência dos *candlesticks* já obtidos.

A figura 3.4 mostra o ERD² do banco de dados. Os *scripts* de criação e população inicial, bem como as *constraints* envolvidas podem ser encontrados em [73].

²*Entity-Relationship Diagram*. Em português: Diagrama de Entidade Relacionamento.

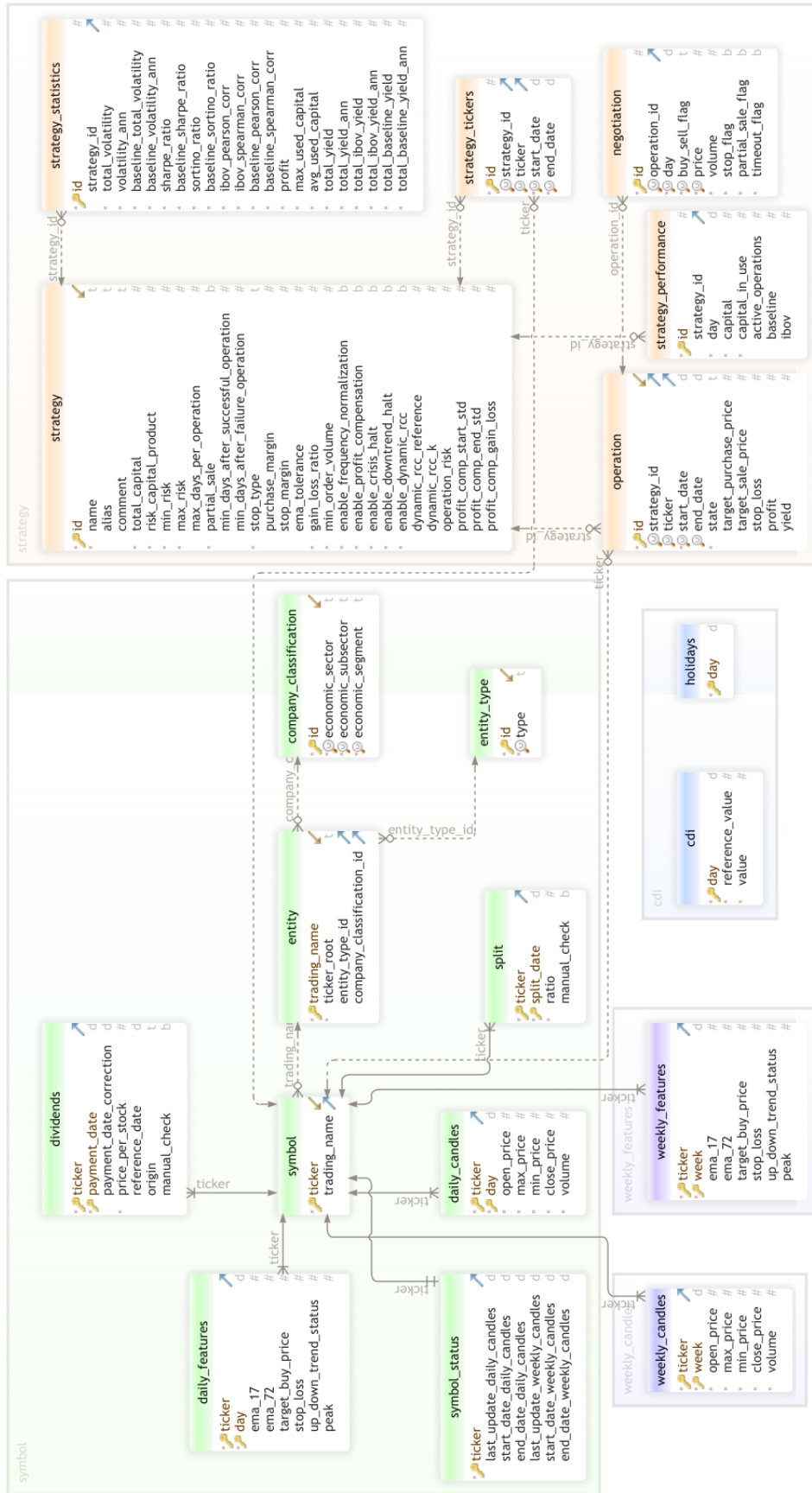


Figura 3.4: ERD do Banco de Dados

3.2.4 Geração de *Features* de Uso Geral

As *Features* de Uso Geral são características derivadas dos *candlesticks* que podem auxiliar qualquer decisão interna de uma estratégia, porém seu objetivo principal está no suporte à escolha do momento de entrada apropriado nas operações, o que é fundamentalmente a responsabilidade do modelo de *Machine Learning*. Como podem ser utilizadas por diversas estratégias, são calculadas antes do início das simulações e somente quando há necessidade, ou seja, quando os *candlesticks* são inseridos pela primeira vez no banco ou quando são atualizados. Ao final dos cálculos, são armazenadas nas tabelas de *features* para posteriores consultas durante as simulações.

Devido à natureza do problema, que se trata da análise de dados do passado, é necessário garantir que as *features* geradas estão sempre apontando para um intervalo de tempo anterior ao dia no qual elas serão consumidas, o que será feito pelo modelo durante a simulação. Do contrário, um erro de não-causalidade poderia surgir, corrompendo a performance dos modelos. Analogamente, durante a geração dos modelos (Seção 3.4.5) também é necessário se garantir que o período de treinamento é anterior ao período de apuração, diretamente via a configuração das datas limites de cada período e indiretamente via remoção de datas cuja *features* possuem efeito memória, evitando assim que qualquer informação esteja presente simultaneamente dos dados de treinamento e de apuração.

As *features* de Uso Geral utilizadas são:

- **Risco Mínimo**

O Risco Mínimo é uma *feature* de suporte à escolha do risco de entrada em uma operação, não sendo assim consumido diretamente pelo modelo de ML. Ressalta-se que o conceito de risco no escopo deste trabalho está relacionado à diferença de valor no qual o *stop loss* é colocado abaixo do preço de compra (Equação 3.10). A escolha do risco também implica no valor do preço alvo de uma operação, pois o mesmo é definido como 3 vezes a magnitude do risco escolhido, percentualmente acima do preço de compra (ver Seção 3.3.1).

A fórmula é composta por uma parcela fixa somada a uma parcela variável, conforme mostrado pela Equação 3.1.

$$Risk_{min} = Risk_{min_f} + Risk_{min_v} \quad (3.1)$$

Seja P_{Δ} a diferença entre o preço máximo e mínimo de um *candle* (Equação 3.2), pode-se definir $Risk_{min_f}$ como o valor mínimo de risco necessário para superar as oscilações diárias dos preços dos últimos 20 dias úteis (Equação 3.3). Nota-se que $\sigma_{P_{\Delta}}$ é o desvio padrão relativo aos últimos 20 dias úteis.

$$P_{\Delta} = P_{high} - P_{low} \quad (3.2)$$

$$Risk_{min_f} = \frac{\sigma_{P_{\Delta}}}{P_{mid}} \quad (3.3)$$

A parcela variável $Risk_{min_v}$ está associada à tendência de queda de preço no curto prazo. Seu cálculo é realizado a partir da derivada de preços médios ajustada por um filtro digital IIR passa-baixas (Equações 3.4 e 3.5), onde α é o coeficiente de amortecimento. O sinal negativo indica que quanto maior a tendência de queda, maior precisa ser o risco associado.

$$P_{mid_LPF(i)} = \alpha P_{mid(i)} + (1 - \alpha) P_{mid_LPF(i-1)}, \quad \text{onde } 0 \leq \alpha \leq 1 \quad (3.4)$$

$$Risk_{min_v} = -P_{mid_LPF(i)} \quad (3.5)$$

Foi utilizado $\alpha = 0.30$, uma vez que neste caso é mais interessante uma resposta rápida a um baixo ruído.

Por fim, adicionou-se um segundo filtro de passa-baixas de $\alpha = 0.10$ apenas aos movimentos de descida dos valores de $Risk_{min}$ com o objetivo de aumentar a cautela durante momentos mais turbulentos do mercado.

As Figuras 3.5 e 3.6 e mostram os resultados do algoritmo para dois papéis de comportamentos distintos: MGLU3 representando um companhia com foco

em alto crescimento, portanto mais volátil; e ABEV3 representando uma companhia já bem consolidada no mercado, portanto menos volátil.

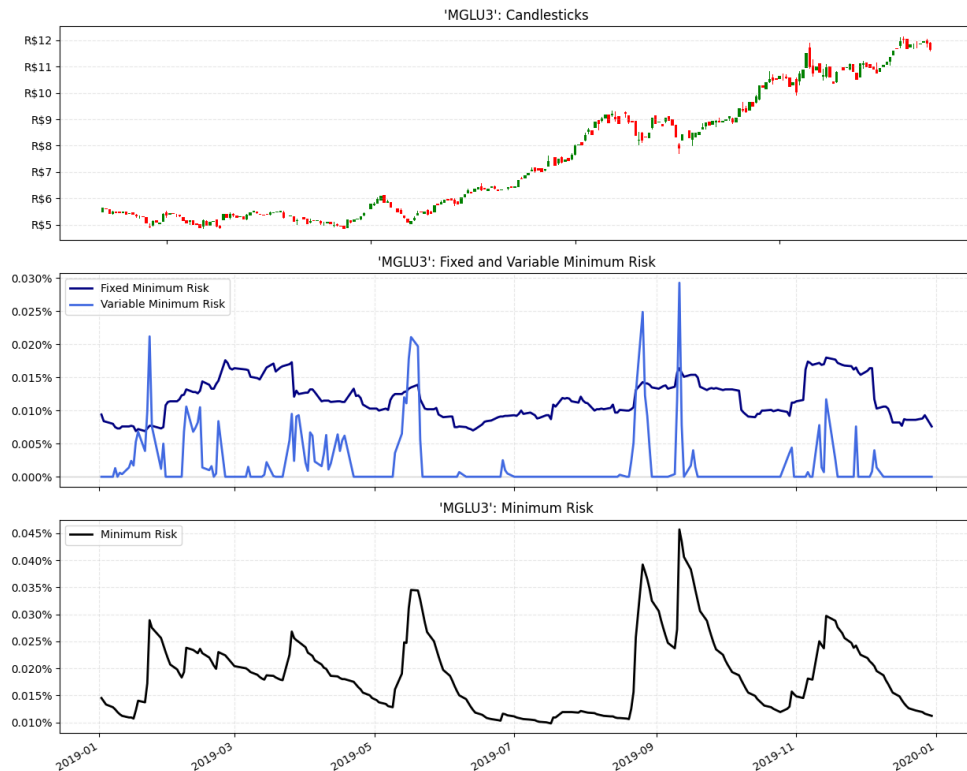


Figura 3.5: MGLU3 - Risco Mínimo (01/01/2019 a 31/12/2019)

• Risco Máximo

O Risco Máximo é uma *feature* de suporte à escolha do risco de entrada em uma operação, não sendo assim consumido diretamente pelo modelo de ML. Ressalta-se que o conceito de risco no escopo deste trabalho está relacionado à diferença de valor no qual o *stop loss* é colocado abaixo do preço de compra (Equação 3.10). A escolha do risco também implica no valor do preço alvo de uma operação, pois o mesmo é definido como 3 vezes a magnitude do risco escolhido, percentualmente acima do preço de compra (ver Seção 3.3.1). A ideia central está na análise estatística das subidas de preços entre os últimos picos identificados dentro do intervalo de 80 dias úteis.

Para se iniciar o cálculo, primeiro é necessário a criação de um algoritmo de identificação de picos, conforme mostrado pela Figura 3.7. O método usa uma janela móvel de $W = 5$ *candles* que corre dia após dia até a data corrente

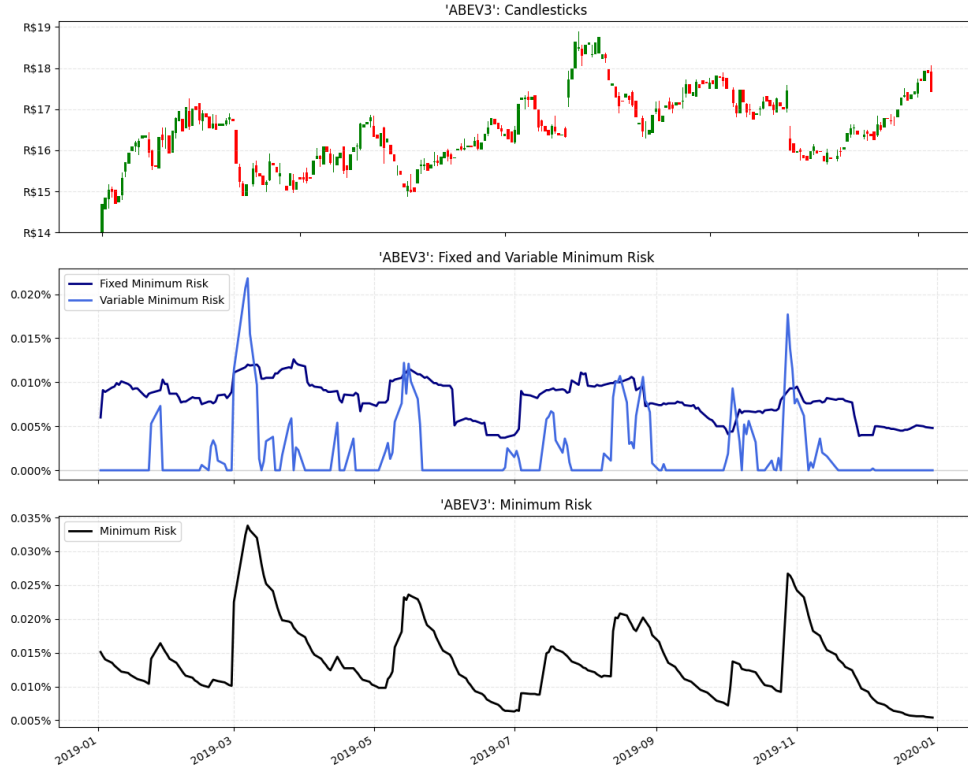


Figura 3.6: ABEV3 - Risco Mínimo (01/01/2019 a 31/12/2019)

e atribui 1 voto de máximo e 1 voto de mínimo aos preços de máximo e preços de mínimo encontrado na janela, respectivamente. Em todos os passos, o primeiro e o último *candle* da janela nunca recebem votos devido à falta de um *candle* adjacente. São elegíveis à picos apenas os *candles* que obtiveram um mínimo de $\text{floor}(W/2) = 2$ votos. Ao final, garante-se a alternância entre máximos e mínimos locais através da remoção de picos consecutivos de um mesmo tipo.

Depois da identificação de picos, extraem-se as n subidas de preços de cada mínimo para o máximo consecutivo no período designado, normalizados pelo pico de mínimo (Figura 3.8 e Equação 3.6). Por fim, o Risco máximo é obtido pelo cálculo da média $\overline{C_{(i)}}$ com um filtro digital IIR passa-baixas (Equações 3.7 e 3.8).

$$c_k = (P_{\max(k)} - P_{\min(k)})/P_{\min(k)}, \quad \text{onde } 0 < k \leq n \quad (3.6)$$

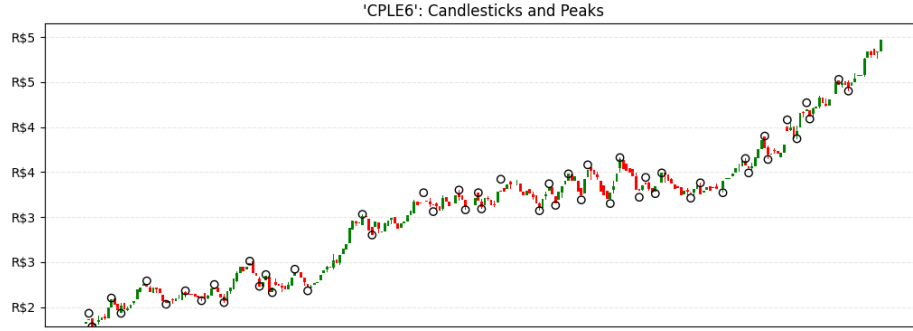


Figura 3.7: CPLE6 - Algoritmo de identificação de picos (01/01/2019 a 31/12/2019)

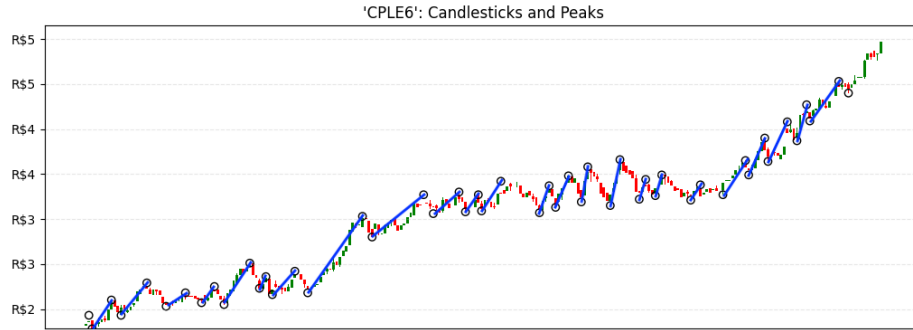


Figura 3.8: CPLE6 - Subidas de preços entre picos (01/01/2019 a 31/12/2019)

$$\overline{C}_{(i)} = \frac{1}{n} \sum_{k=1}^n c_k \quad (3.7)$$

$$Risk_{max(i)} = \alpha \overline{C}_{(i)} + (1 - \alpha) Risk_{max(i-1)} \quad (3.8)$$

Foi utilizado $\alpha = 0.50$.

As Figuras 3.9 e 3.10 mostram o Risco Máximo para os ativos MGLU3 e ABEV3, respectivamente.

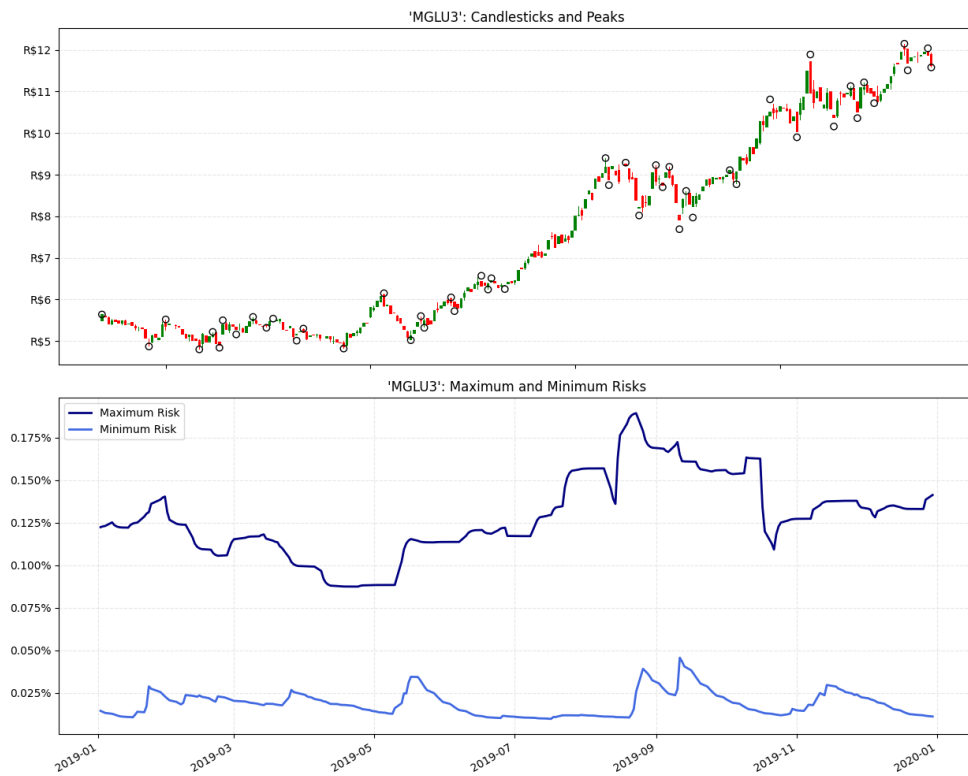


Figura 3.9: MGLU3 - Riscos Máximo e Mínimo (01/01/2019 a 31/12/2019)



Figura 3.10: ABEV3 - Riscos Máximo e Mínimo (01/01/2019 a 31/12/2019)

3.3 Simulação de Estratégia

3.3.1 Estrutura

O tema escolhido pelo presente trabalho permite uma enorme quantidade possíveis implementações, onde muitas se mostram como promissoras e interessantes de se explorar. No entanto, dar vida a um projeto de engenharia envolve a delimitação de um escopo, que necessariamente restringe as possibilidades. Dessa forma, a Estrutura na qual as estratégias são simuladas se baseia nas seguintes declarações:

- Toda estratégia possui um **capital inicial**, que representa uma quantidade de capital pré-alocado para compra dos ativos financeiros. Essa quantia deve ser sempre respeitada ao longo da simulação de forma a não representar nunca um valor negativo.
- Toda estratégia deve possuir uma **carteira de ativos** (ou lista de ativos) com as respectivas datas iniciais e finais de validade, isto é, intervalos de tempo onde as operações podem ser realizadas. Embora sejam permitidos intervalos diferentes, é convencionado a mesma data de início e de fim para todos os papéis.
- Define-se uma **operação** como o processo de compra única de um volume de ações de um ativo seguido pela venda de todo o volume comprado, independentemente do tempo, mesmo que esta ocorra em estágios. Nota-se que apenas a venda é cabível de ocorrer em estágios (i.e., venda parcial).
- Toda operação possui um **preço alvo** e um **stop loss**. O preço alvo é um valor acima do preço de compra e o *stop loss* é um valor abaixo do preço de compra. Quando o mercado atinge qualquer um dos dois valores, uma venda é disparada, encerrando a operação em vigor. No entanto, considera-se uma operação de sucesso aquela que encerrou por atingir o preço alvo e uma operação de falha aquela que encerrou por atingir o *stop loss*.
- Uma estratégia pode possuir no máximo **uma operação em vigência** para cada *ticker* em sua bolsa de ativos, portanto para que uma segunda compra

ocorra no momento em que já existem papéis adquiridos, é necessários vendê-los primeiro.

- A **razão entre ganho e perda**, também denominada relação risco/ganho, é definida por André Moraes [3] e predetermina a relação entre o preço alvo e o *stop loss* em qualquer operação. Ela indica a razão entre a diferença do preço alvo P_{target} para o preço de compra P_{buy} sobre a a diferença do preço de compra para o *stop loss* P_{stop} (Equação 3.9). O valor recomendado é 3, portanto este é o valor que será utilizado em todo o escopo deste trabalho como uma constante.

$$G = \frac{P_{target} - P_{buy}}{P_{buy} - P_{stop}} = 3 \quad (3.9)$$

Utiliza-se o termo “risco de uma operação” ou simplesmente “risco” como sendo a diferença de valor no qual o *stop loss* é colocado abaixo do preço de compra (Equação 3.10). Por exemplo, se o preço de compra de uma operação é de R\$10,00 e o seu risco é de 5%, então o *stop loss* se encontra em R\$9,50 e o preço alvo em R\$11,50 necessariamente.

$$Risk = \frac{P_{buy} - P_{stop}}{P_{buy}} \quad (3.10)$$

- Não há **operações a descoberto**.
- Não há **operações alavancadas**.

3.3.2 Premissas

As Premissas são um conjunto de afirmações que visam complementar a Estrutura das simulações ao mesmo tempo que garantir a integridade dos resultados, muitas vezes optando pelo pior cenário em situações inconclusivas. São elas:

- O momento de decisão de **entrada em uma operação** por uma estratégia ocorre durante a abertura de mercado do dia corrente, mais precisamente no instante em que o preço de abertura é definido.

- No dia que houver a compra de um ativo, não pode haver a venda do mesmo. Em outras palavras, o **período mínimo de duração de uma operação é de 1 dia útil**.
- A **venda por *timeout*** ocorre quando o número máximo de dias de uma operações extrapola um valor definido (ver Seção 3.3.3)
- Devido a ausência de informações mais detalhadas que a janela de tempo diária, a seguinte ordem é priorizada durante a **venda de um ativo**:
 1. Venda por *stop loss*
 2. Venda parcial (caso habilitada)
 3. Venda por preço alvo
 4. Venda por *timeout*
- Caso algum **preço de venda seja pulado**, ou seja, a descontinuidade entre o preço de abertura do dia corrente e o preço de fechamento do dia anterior não englobe o valor de venda, utiliza-se o preço de abertura do dia corrente. A única exceção acontece para a venda por *timeout*, já que se trata de uma venda compulsória que sempre ocorre no preço de fechamento do dia designado.

3.3.3 Período Máximo de Dias por Operação

Em teoria, poderia-se permitir que operações não tivessem um período máximo de dias para serem encerradas. Contudo, isso facilmente se prova uma decisão ruim de alocação de capital em ativos que passam por uma fase de consolidação, ou seja, sem qualquer tendência. Além do ativo em questão não encerrar a operação e finalizar seu lucro ou seu prejuízo na carteira, o capital alocado nele não pode ser utilizado por outros ativos que eventualmente venham a lucrar, ou seja, gera-se um efeito de inércia ao aumento da performance geral. Portanto, foi imposto um limite do período de tempo de todas as operação em um valor máximo.

Nesta linha, um algoritmo auxiliar foi criado para varrer um período de dias passados e criar operações com diversos valores de risco, observando quais riscos levariam a operações de sucesso e quais levariam a operações de falha. Também

analisou-se a distribuição de operações de sucesso de acordo com os valores de risco e o intervalo de dias corridos.

De início, foi fixado um intervalo máximo de 90 dias para cada operação hipotética que o algoritmo gerou. O valor é propositalmente excessivo, pois sua função é apenas não forçar *timeout* na maioria das operações. As Figuras 3.11 e 3.12 mostram dois histogramas dos dias das operações de sucesso que consideram o menor risco possível, isto é, o menor valor de risco que se pode utilizar a cada dia da série temporal de forma a tornar a operação um sucesso, caso ele exista. Se não existir, é considerado operação de falha, portanto está fora dos histogramas. A legenda indica faixas onde, no caso da linha tracejada em verde na Figura 3.11, 50% das contagens se encontram dentro dos 12 primeiros dias, e assim por diante.

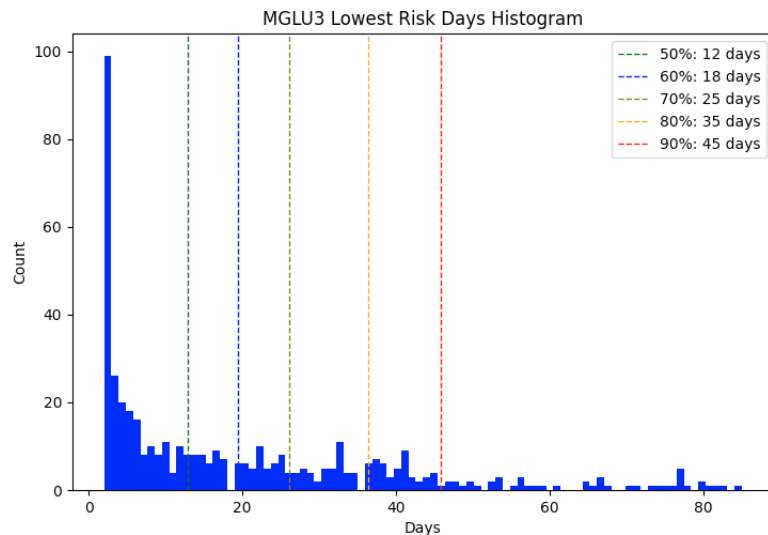


Figura 3.11: MGLU3 - Histograma de dias com risco mínimo em operações de sucesso (01/01/2016 a 31/12/2018)

Também foram analisados os histogramas de risco ótimo por operação, ou seja, o valor de risco que traz o melhor rendimento por operação considerando os dias corridos (Figuras 3.13 e 3.14).

A Figura 3.15 mostra a distribuição de todas as operações de sucesso possíveis no intervalo selecionado. Obseva-se que valores baixos de risco não costumam estar associados a longos períodos de operação e que a maior densidade de operações de sucesso se encontra em baixo risco e em baixa duração de operação.

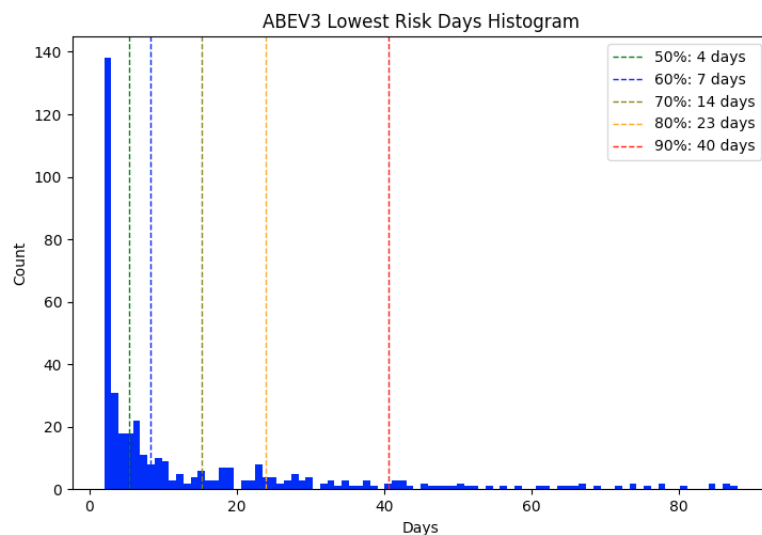


Figura 3.12: ABEV3 - Histograma de dias com risco mínimo em operações de sucesso (01/01/2016 a 31/12/2018)

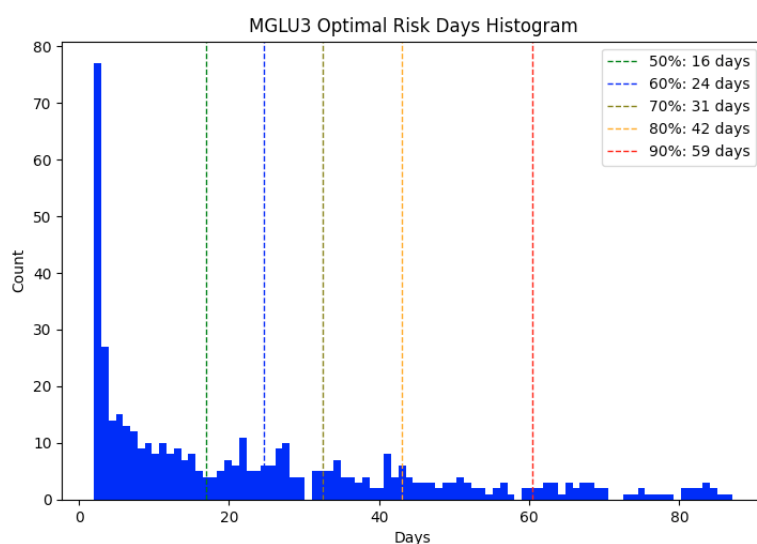


Figura 3.13: MGLU3 - Histograma de dias com risco ótimo em operações de sucesso (01/01/2016 a 31/12/2018)

A Tabela 3.2 resume o período de dias que engloba 90% das contagens dos histogramas conforme indicado nas Figuras 3.11, 3.12, 3.13 e 3.14.

Com base nos valores encontrados, foi escolhido o período máximo de **45 dias** para qualquer operação.

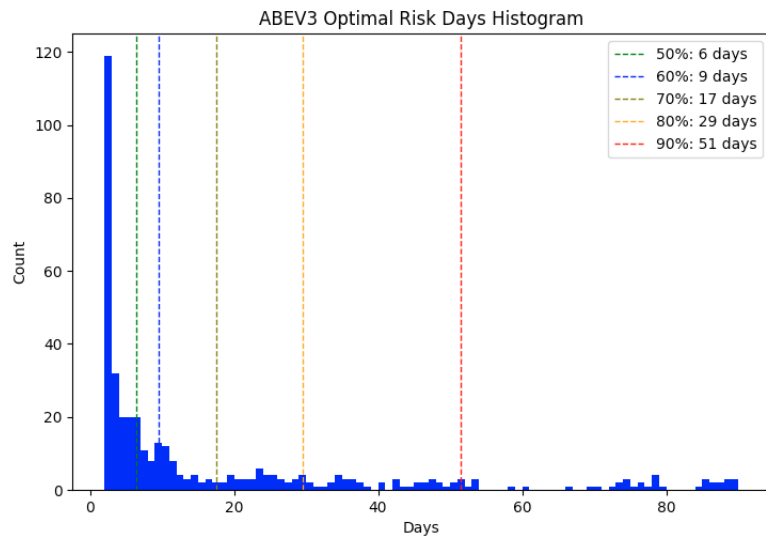


Figura 3.14: ABEV3 - Histograma de dias com risco ótimo em operações de sucesso (01/01/2016 a 31/12/2018)

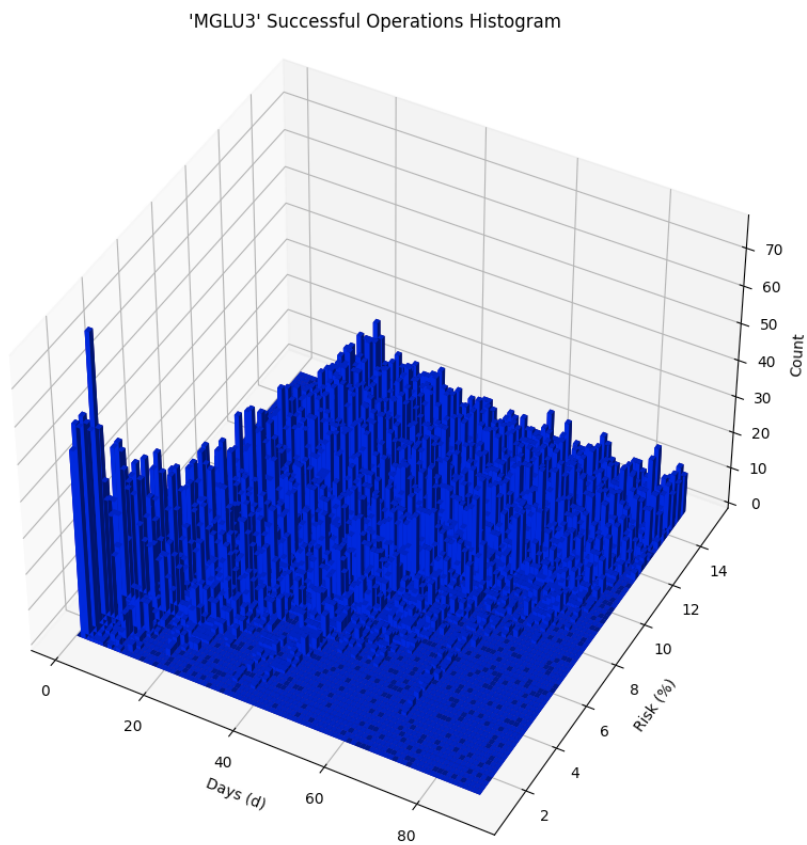


Figura 3.15: MGLU3 - Histograma de todas as operações de sucesso (01/01/2016 a 31/12/2018)

	Menor Risco	Risco Ótimo
MGLU3	45 dias	59 dias
ABEV3	40 dias	51 dias

Tabela 3.2: Período de dias que engloba 90% das contagens dos histogramas

3.3.4 Risco de Entrada por Operação

O Risco de Entrada por Operação é encontrado a partir de um valor intermediário entre o Risco Mínimo e o Risco Máximo.

Pelo fato da metodologia de cálculo do Risco Máximo e do Risco Mínimo seguirem raciocínios diferentes (Seção 3.2.4), podem ocorrer momentos nos quais a equação $Risk_{max} < Risk_{min}$ seja verdadeira, em outras palavras, as ondas de subida de preço entre picos no gráfico diário não compensem as oscilações inerentes ao ruído diário dos *candlesticks*. Enquanto este evento ocorrer, não haverá entrada em operações para o ativo envolvido.

Para a maioria dos casos, tem-se $Risk_{max} > Risk_{min}$. O valor do Risco de Entrada por Operação ($Risk_{operation}$) é definido pelo parâmetro $Risk_{coef}$ de acordo com a Equação 3.11, onde $0 \leq Risk_{coef} \leq 1$.

$$Risk_{operation} = Risk_{min} + Risk_{coef}(Risk_{max} - Risk_{min}) \quad (3.11)$$

A Figura 3.16 resume as simulações cujos valores de Risco de Entrada estão no intervalo fechado $[0.10, 0.70]$. Foram utilizados os 71 *tickers* da Tabela 3.1 no período de 01/01/2019 a 31/03/2020 sem qualquer otimização. Utilizou-se um RCC de 0,10% para evitar saturação de capital. Nota-se que o ponto de máximo para os três indicadores de performance coincidiu para o Risco de Entrada de 0,43.

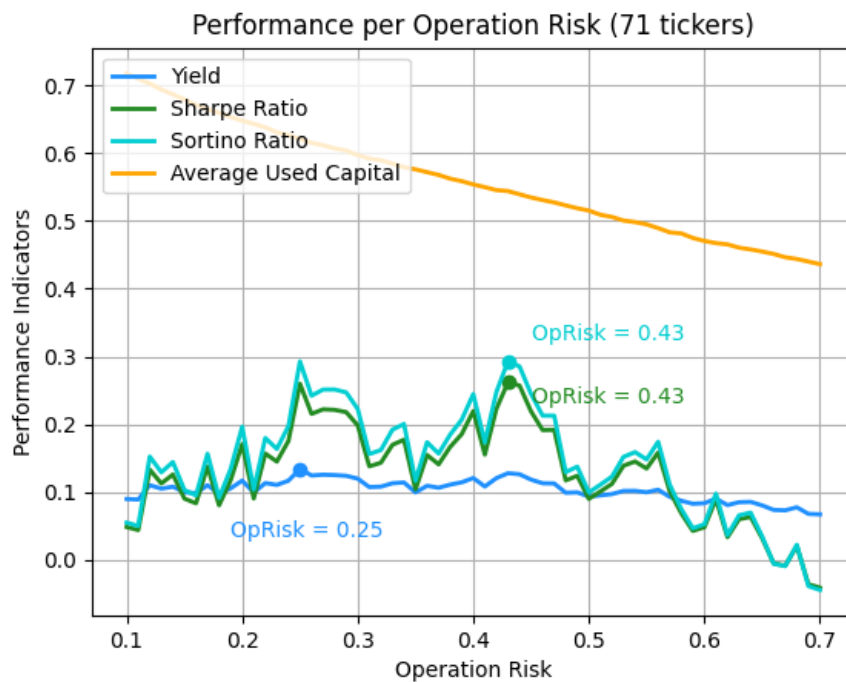


Figura 3.16: Indicadores de performance em função do risco de entrada (71 tickers, 01/01/2019 a 31/03/2020)

3.3.5 Gerenciamento de Risco

Segundo André Moraes [3], um bom Gerenciamento de Risco é essencial para a performance de uma estratégia. Afinal, não adianta obter uma alta taxa de acerto em operações de cujo lucro médio não compense as perdas acumuladas pelas operações que falham. Além disso, estar com o capital muito alocado em ativos de um único segmento é perigoso devido à exposição à fatores como falta de insumos industriais, mudanças na legislação, crises internas, instabilidade política, dentre outros.

Para mitigar as questões levantadas, algumas medidas foram tomadas inspiradas no trabalho de André Moraes [3]. São elas:

- Diversificação de ativos em segmentos de mercado variados através da escolha de um alto número de *tickers* na carteira, mais especificamente 71.
- Criação do Coeficiente de Risco-Capital³

³ou *Risk-Capital Coefficient* (RCC)

O Coeficiente de Risco-Capital, definido pela Equação 3.12, é uma constante que equilibra a relação entre o capital de entrada em uma operação e o risco escolhido. Seu valor é configurado previamente no Arquivo de Configuração (ver Tabela 3.3) e vale para todos os ativos da carteira.

$$RCC = Capital \times Risk \quad (3.12)$$

Durante uma simulação, a estratégia primeiro encontra o valor do risco desejado para entrar na operação, depois escolhe o capital a ser alocado. Dessa forma, a Equação 3.13 mostra de fato a aplicação do RCC. É evidente que quanto maior o risco envolvido, menor o capital a ser alocado e vice-versa.

$$Capital = \frac{RCC}{Risk} \quad (3.13)$$

O RCC influencia diretamente no uso médio de capital de uma estratégia. Por um lado, um uso de capital baixo significa um mau aproveitamento do capital, o que leva a uma performance ruim. Por outro lado, muito uso de capital implica em pouco capital disponível para entrada em novas operações, o que também leva a uma performance ruim e instável, visto que a operação que iniciar imediatamente antes das demais leva quase todos o capital consigo.

A Figura 3.17 mostra o gráfico da relação entre os indicadores de performance em função do RCC para os 71 *tickers* da Tabela 3.1. Nenhuma otimização foi adicionada. Nota-se no início do gráfico uma região de máximos locais em cada indicador, seguido por uma pequena queda até retomar novamente a subida dos valores. Percebe-se também que por volta do RCC de 1,88%, o uso de capital está quase saturando.

Dentre as execuções que compõem a Figura 3.17, extraiu-se os gráficos de uso de capital para as simulações de RCC=0.29% e RCC=1,88% (Figuras 3.18 e 3.19, respectivamente). Nota-se o efeito de saturação causado pelo aumento significativo do RCC, que reduziu o número de operações totais de 3972 para 3390 pela falta de capital disponível em alguns intervalos. Por outro lado, operações longe da saturação

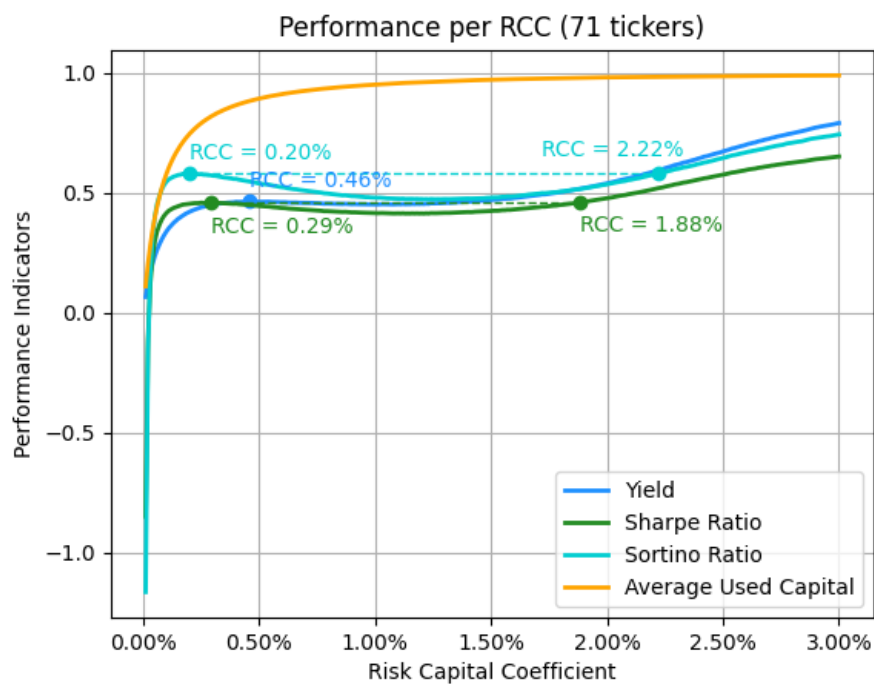


Figura 3.17: Indicadores de performance em função do RCC (71 tickers: 01/01/2019 a 31/12/2021)

receberam mais capital e como estas são mais expressivas, a performance geral foi maior.

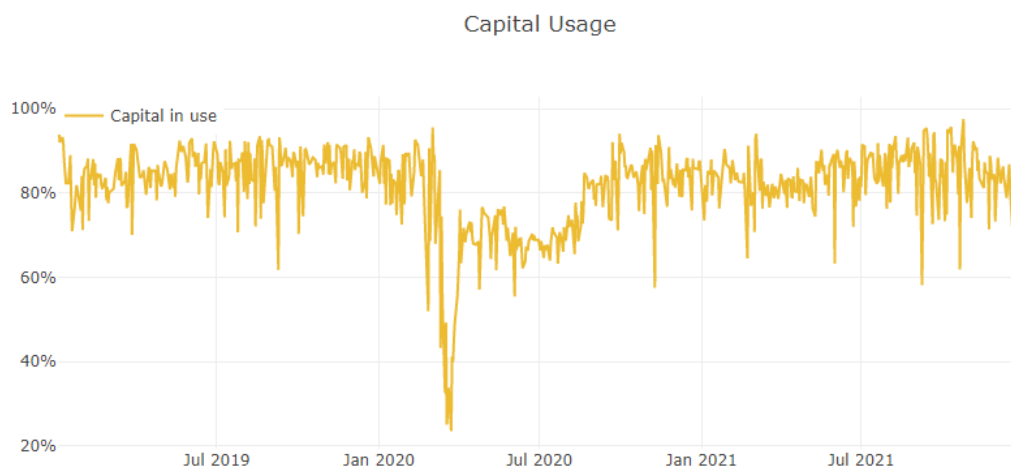


Figura 3.18: Uso de Capital (71 tickers, 01/01/2019 a 31/12/2021, RCC = 0,29%)

Um problema geral e inerente à abordagem do RCC é a necessidade do conhecimento *a priori* de um valor razoável. Ou seja, sem algumas simulações prévias, não

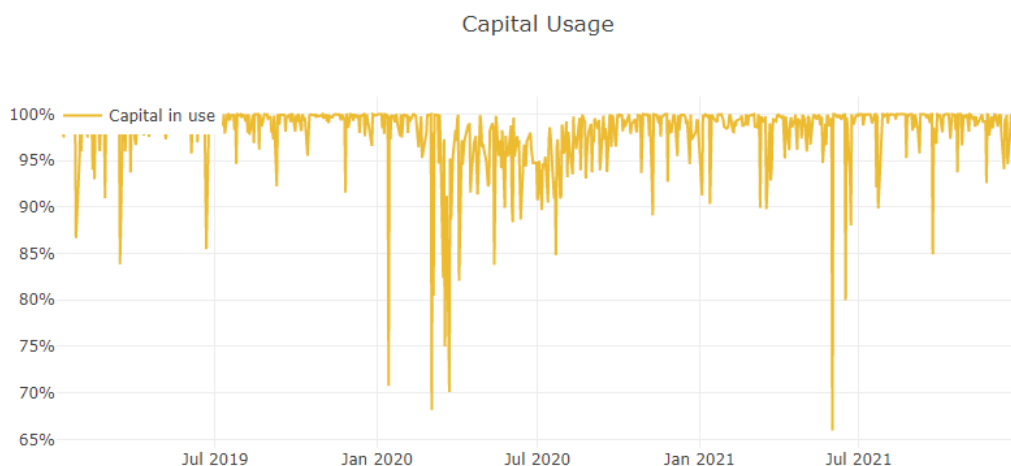


Figura 3.19: Uso de Capital (71 tickers, 01/01/2019 a 31/12/2021, RCC = 1,88%)

há como se saber um valor ótimo ou pelo menos próximo do ótimo. Uma forma de se atenuar esse problema é através da criação de um controle proporcional que aumente o RCC geral em função do baixo aproveitamento do uso de capital ao longo dos dias e vice-versa. Essa alternativa também é chamada de RCC Dinâmico e abordada em mais detalhes na Seção 3.3.6.

3.3.6 Controle Proporcional para Uso de Capital

A criação de um RCC fixo pode ser interessante do ponto de vista de Gerenciamento de Risco (Seção 3.3.5), mas na prática deixa um pouco a desejar por requerer uma noção prévia de um valor adequado. Esse valor só pode ser obtido através de simulações anteriores ao período desejado, onde o comportamento do mercado pode ser significativamente diferente a ponto de requerer um novo RCC, dificultando um bom ajuste. Em outras palavras, um RCC fixo pode levar a problemas de subaproveitamento do Uso de Capital da carteira.

Uma forma de se atenuar esse problema é através da criação de um RCC Dinâmico, configurado através de um Controle Proporcional. A vantagem dessa abordagem está na diminuição da sensibilidade do RCC em relação à performance geral, permitindo um ajuste menos preciso sem grande impacto de performance. O Controle atua no rebalanceamento de capital em função do uso médio de capital vigente, ou seja, períodos com mais disponibilidade de capital terão uma maior alavancagem.

As Equações 3.14 e 3.15 mostram o cálculo do erro e do RCC dinâmico (RCC_{din}) a partir do valor de referência para o uso médio de capital (C_{ref}), do uso médio de capital dos últimos 10 dias de simulação ($\overline{C_{10d}}$), do RCC fixo (RCC_{fix} , definido pela Equação 3.12) e da constante de ganho proporcional (K).

$$e = C_{ref} - \overline{C_{10d}}, \quad \text{para } 0 \leq C_{ref}, \overline{C_{10d}} \leq 1 \quad (3.14)$$

$$RCC_{din} = RCC_{fix}(1 + Ke) \quad (3.15)$$

As Figuras 3.20, 3.21 e 3.22 apresentam indicadores de performance para valores de RCC e de K diferentes. Foram utilizados na simulação os 71 *tickers* indicados na Tabela 3.1 no intervalo de 01/01/2019 a 31/12/2021. Cada curva é composta por um conjunto de simulações diferentes que compartilham entre si o mesmo produto $RCC \times K$. Dessa forma, é possível acoplar em um único parâmetro duas variáveis que influem diretamente no rendimento e no uso de capital. Consequentemente, também facilita a avaliação do grau de importância que cada uma dessas variáveis apresenta.

Dando prosseguimento à análise, primeiro pode-se observar nas Figuras 3.20 e 3.21 que a curva que apresentou o maior rendimento foi a de $RCC \times K = 0,75$. Isso mostra que forçar a estratégia além desse valor não é vantajoso. As curvas imediatamente próximas (0,85; 0,80; 0,70; 0,65) também se distribuíram consistentemente logo no entorno da anterior, indicando que não se trata de uma anormalidade e sim uma região de máximo (aproximadamente, devido à discretização).

A segunda observação importante é que o aumento do valor de K, mesmo que linearmente balanceado pela diminuição do RCC, traz uma melhoria de performance enquanto o contrário piora os resultados.

Já observando a Figura 3.22, nota-se que o produto $RCC \times K$ possui relação direta com o uso médio de capital. O fato da curva mais performática ($RCC \times K = 0,75$) não ser a de maior uso médio de capital indica que também há a necessidade de uma

pequena parcela de capital ocioso para uma boa fluidez da estratégia, neste caso um valor próximo de 2,7%.

O objetivo central desta análise é encontrar um conjunto de valores otimizados para RCC e K . No entanto, as Figuras 3.20 e 3.21 dão a entender que um K tendendo ao infinito, aliado a um RCC tendendo a zero seria a solução. O problema que ocorre em decorrência desta linha de raciocínio é mostrado através da Figura 3.23, onde se pode ver uma queda acentuada do número de operações totais efetuadas pelas estratégias. Em um regime longe de saturação, as mesmas estratégias simuladas possuem todas 3972 operações totais. No entanto, a estratégia com maior índice de Sharpe da Figura 3.21, onde $RCC \times K = 0,75$, possui 3145 operações. Isso significa que 20,8% de operações foram perdidas por ausência de capital de entrada devido à saturação. Como não há nenhum critério envolvido, tanto as operações de sucesso quando as de falha são perdidas em igual proporção, não afetando muito a performance final neste caso.

Contudo, a escolha de um valor de K excessivamente grande pode levar o número de operações totais para um valor baixo o suficiente para gerar instabilidades indesejáveis na performance. Isso ocorreria pois um *ticker* que em determinado momento comprar um papel monopolizará praticamente todo o capital disponível até que ele decida vender. Tal comportamento é válido e até poderia se mostrar bastante lucrativo e confiável em determinados contextos através de um estudo mais aprofundado. Porém, isto foge ao escopo deste projeto, que foca na utilização de aprendizado de máquina como o elemento central por trás da performance das estratégias.

Desta forma, escolheu-se o valor de $RCC = 0,000025$ e $K = 30000$, representados pelo maior índice de Sharpe na Figura 3.21. Conforme mencionado anteriormente, esta escolha admite uma perda de 20,8% no número de operações.

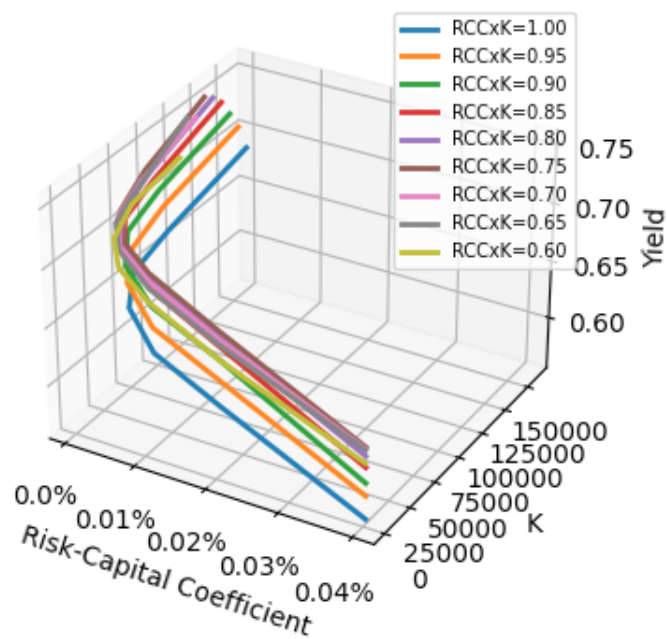


Figura 3.20: Rendimento sob uso de RCC dinâmico

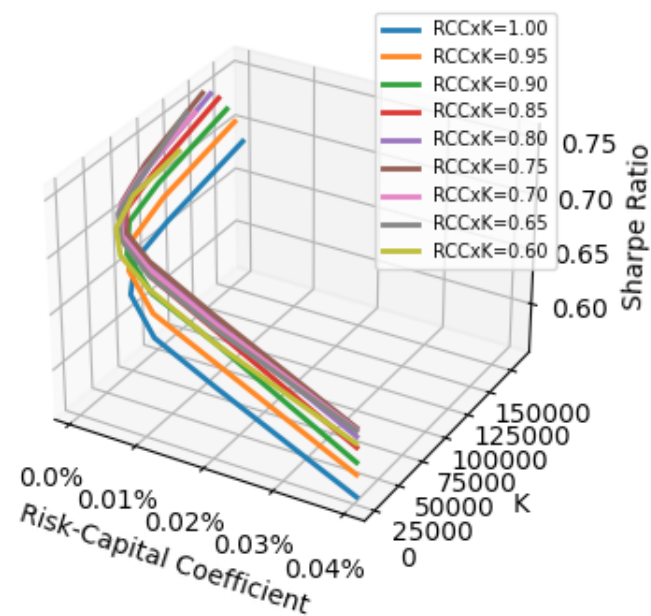


Figura 3.21: Índice de Sharpe sob uso de RCC dinâmico

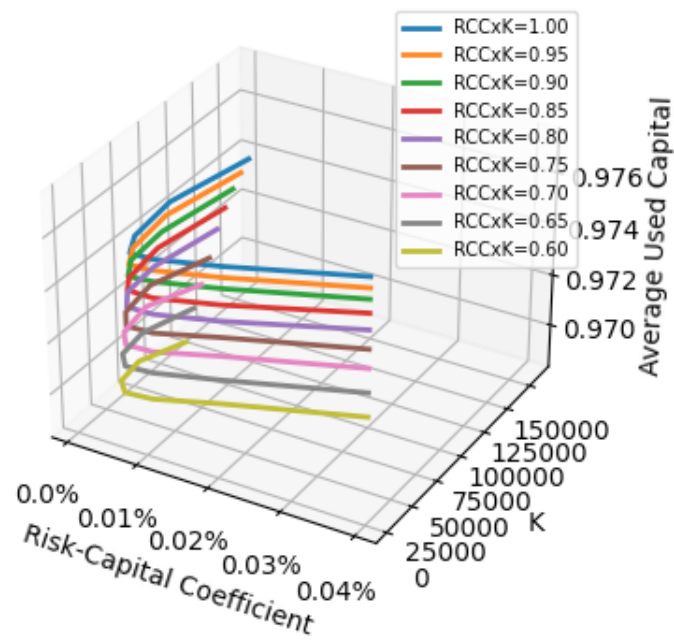


Figura 3.22: Uso médio de capital sob uso de RCC dinâmico

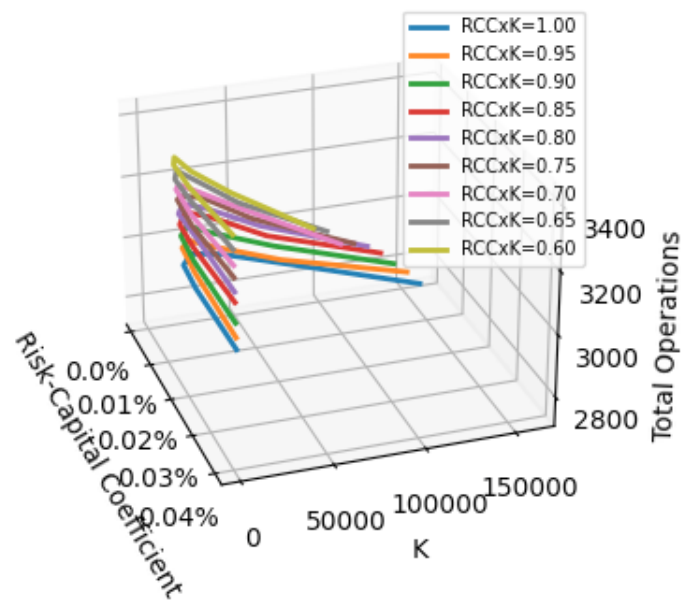


Figura 3.23: Total de operações sob uso de RCC dinâmico

3.3.7 Lista de Parâmetros de Configuração

A Tabela 3.3 mostra uma lista de todos os parâmetros configuráveis em uma simulação. Nota-se que as variáveis de escopo geral são aplicáveis a toda e qualquer estratégia presente no Arquivo de Configuração enquanto as variáveis de escopo local dizem respeito apenas a um grupo de estratégias em particular (ver Seção 3.2.1).

Lista de Parâmetros	
Nome do Parâmetro	Descrição
name	(OBRIGATÓRIO) Nome da estratégia a ser executada. Único valor válido: “ML”. Tipo: <i>String</i> . Listável: Não.
alias	(OBRIGATÓRIO) Rótulo de Identificação. Tipo: <i>String</i> . <i>Default</i> : <i>String</i> vazia. Listável: Não.
stock_targets	(OBRIGATÓRIO) <i>Array</i> de ações a incluir na carteira. Formato indicado pela Figura 3.2.
comment	Comentário. Tipo: <i>String</i> . <i>Default</i> : <i>String</i> vazia. Listável: Não.
capital	Capital total da carteira em reais (R\$). Tipo: <i>Float</i> . <i>Default</i> : 100000. Listável: Sim.
risk_capital_coefficient	Coefficiente de risco-capital (RCC) geral (Seção 3.3.5). Tipo: <i>Float</i> . <i>Default</i> : 0,001. Listável: Sim.
tickers_number	Número de ativos a escolher dentro de “stock_targets” em ordem de listagem. Tipo: <i>Int</i> . <i>Default</i> : 0 (todos). Listável: Sim.
min_order_volume	Volume mínimo por operação. Tipo: <i>Int</i> . <i>Default</i> : 1. Listável: Sim.
max_days_per_operation	Número máximo de dias por operação (Seção 3.3.3). Inclui o dia de compra. Caso excedido, ocorre venda compulsória pelo preço de fechamento no último dia da contagem. Tipo: <i>Int</i> . <i>Default</i> : 45. Listável: Não.

Continuação da Tabela 3.3	
Nome do Parâmetro	Descrição
min_risk	Risco mínimo por operação. Tipo: <i>Float</i> . <i>Default</i> : 0,003. Listável: Sim.
max_risk	Risco máximo por operação. Tipo: <i>Float</i> . <i>Default</i> : 0,15. Listável: Sim.
operation_risk	Valor percentual de escolha do risco de entrada em operação (Seção 3.3.4). Tipo: <i>Float</i> . <i>Default</i> : 0,5. Listável: Sim.
enable_dynamic_rcc	Uso de Coeficiente de Risco-Capital dinâmico (Seção 3.3.6). Tipo: <i>Boolean</i> . <i>Default</i> : <i>False</i> . Listável: Sim.
dynamic_rcc_reference	Valor de referência de uso de capital médio no controle do RCC dinâmico (Seção 3.3.6). Tipo: <i>Float</i> . <i>Default</i> : 1,0. Listável: Sim.
dynamic_rcc_k	Valor do ganho proporcional K no controle do RCC dinâmico (Seção 3.3.6). Tipo: <i>Float</i> . <i>Default</i> : 10. Listável: Sim.
Fim da Tabela 3.3	

Tabela 3.3: Lista de parâmetros detalhados

3.3.8 *Dashboard*

Um *Dashboard* interativo é gerado por aplicação secundária a fim de auxiliar a análise dos resultados obtidos em cada simulação. O *framework Dash* [74] foi utilizado para criar uma interface web resumindo todas as informações pertinentes a uma simulação executada. As Figuras 3.24, 3.25, 3.26, 3.27 e 3.28 mostram em partes as seções de uma simulação genérica.

Strategy Analytics

Analyze Stock Market swing trade strategies

Strategy: ML Derivation

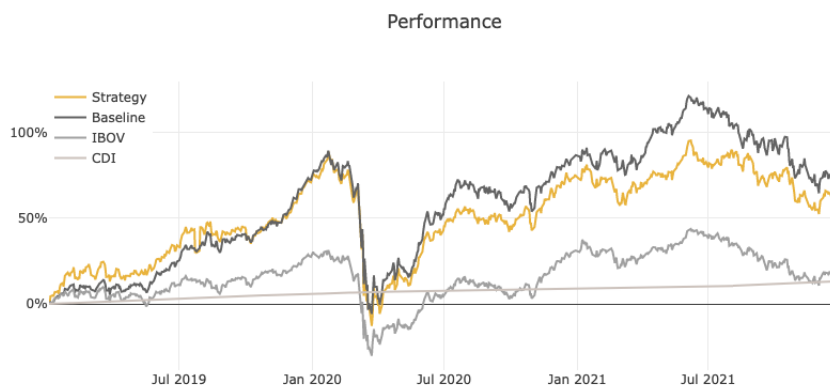


Figura 3.24: *Dashboard* - Performance

Parameters

Alias	2019-1 to 2021_4. All tickers. K analysis.
Total Tickers	71
Start Date	01/01/2019
End Date	31/12/2021
Capital (R\$)	100000
Risk-Capital Coefficient - RCC (%)	0.02
Gain-Loss Ratio	3
Minimum Order Volume	1
Minimum Operation Risk (%)	0.3
Maximum Operation Risk (%)	15
Partial Sale	False
Stop Loss Type	Normal
Min Days after Successfull Operation (days)	0
Min Days after Failure Operation (days)	0
Maximum Days per Operation (days)	45
Enable Frequency Normalization	False
Enable Profit Compensation	False
Enable Crisis Halt	False
Enable Downtrend Halt	False
Enable Dynamic RCC	True
Dynamic RCC Reference (%)	100
Dynamic RCC K	5000
Operation Risk	0.29
Profit Compensation Std Dev Range	[0.2, 2.0]
Profit Compensation Gain/Loss	0.6

Figura 3.25: *Dashboard* - Parâmetros de entrada

Results and Statistics

Strategy Total Yield (%)	61.34
Baseline Total Yield (%)	73.94
IBOVESPA Total Yield (%)	15.17
CDI Total Yield (%)	13.66
Strategy Total Volatility (%)	59.09
Baseline Total Volatility (%)	55.48
Strategy Sharpe Ratio (-)	0.56
Baseline Sharpe Ratio (-)	0.68
Strategy Sortino Ratio (-)	0.64
Baseline Sortino Ratio (-)	0.74
Strategy-Baseline Spearman Correlation (-)	0.97
Strategy-IBOV Spearman Correlation (-)	0.94
Maximum Used Capital (%)	100.00
Average Used Capital (%)	97.62
Maximum Active Operations	71
Average Active Operations	65.97
Active Operations Standard Deviation	6.13
Profit (R\$)	61344.67
Total Operations	3054
---Successful Operations (hit 3:1 target)	901 (29.5%)
---Partial Sale Successfull Operations (hit 1:1 or 2:1 target)	0 (0.0%)
---Failed Operations	1954 (64.0%)
---Timed Out Operations	145 (4.7%)
---Unfinished Operations	54 (1.8%)
Strategy Yield (% ann)	17.61
Baseline Yield (% ann)	20.65
IBOVESPA Yield (% ann)	4.91
CDI Yield (% ann)	4.37
Strategy Volatility (%ann)	34.41
Baseline Volatility (%ann)	32.31

Figura 3.26: *Dashboard* - Resultados e estatísticas

Capital Usage

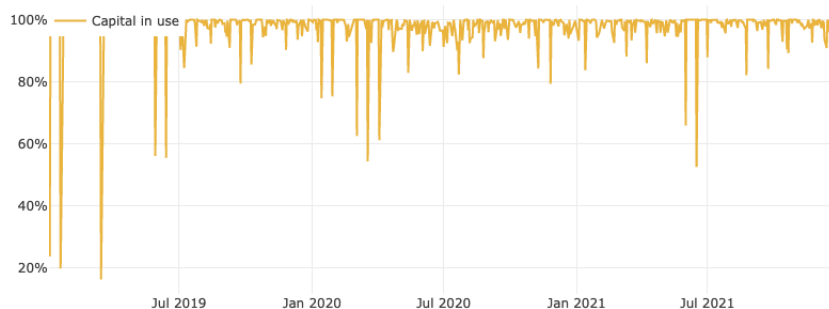


Figura 3.27: *Dashboard* - Gráfico de uso de capital

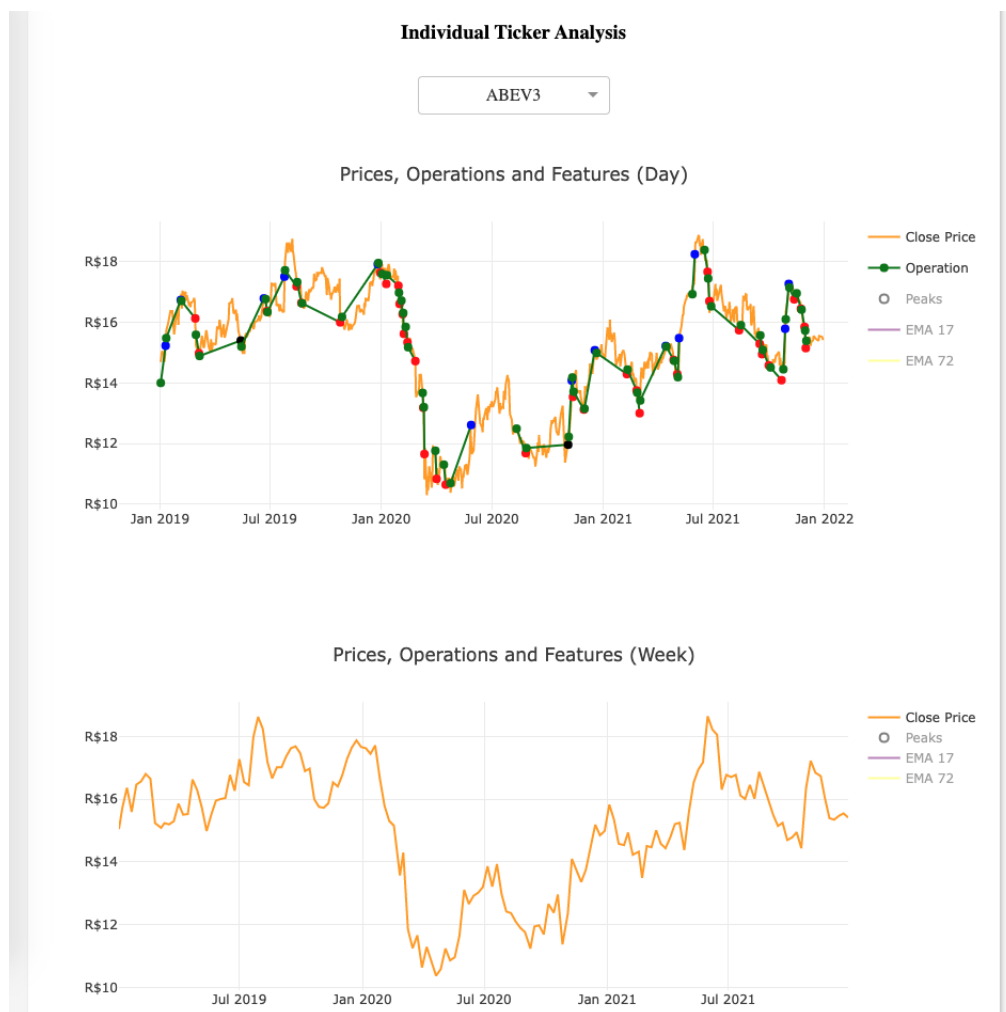


Figura 3.28: *Dashboard* - Gráficos de análise individual de ações

3.4 Modelos de Aprendizado Supervisionado

3.4.1 Resumo

A partir de *datasets* previamente populados, modelos do tipo *Random Forest* são gerados para cada ação a cada intervalo de 3 meses de simulação. Um critério particular de performance foi criado para auxiliar na escolha do melhor modelo, que é filtrado por uma varredura de diversos parâmetros.

3.4.2 *Datasets e Feature Selection*

Os *datasets* são arquivos CSV criados para cada *ticker* através de uma varredura da série histórica. Analisa-se dia após dia as *features* acumuladas e o resultado de uma operação hipotética iniciada no dia corrente. A Tabela 3.4 mostra a lista de *features* relevantes do arquivo, onde as linhas marcadas em negrito indicam as colunas utilizadas na entrada de dados dos modelos. A coluna Resultado da Operação indica a saída observada para o treinamento supervisionado.

O termo preço médio se refere ao definido pela Equação ???. Da mesma forma, a derivada do preço médio é indicada pela Equação ??. As colunas de cujos nomes se iniciam com *Spearman* são na verdade a correlação entre o vetor de preços médios dos últimos N dias acumulados e uma função puramente monotônica crescente $f(x) = x$. Isso permite a extração de uma medida para intensidade de subida dos preços que independe da normalização pelo preço da ação. Como o que importa na correlação de Spearman são os postos, o valor numérico do vetor utilizado para representar a função $f(x) = x$ não tem relevância, desde que seja monotônico crescente.

O *flag* de fim de intervalo indica que, pelo fato do *dataset* ter chegado ao final, não é possível dizer se a operação foi de sucesso ou foi de falha, portanto a mesma é desconsiderada do treinamento.

Por fim, para cada dia de operação, foram cruzadas diversas opções de risco a fim de enriquecer o *dataset* com mais diversidade, permitindo modelos mais robustos. Foram utilizadas 119 opções de risco: de 0,2% a 12% em passos de 0,1%.

Nome	Coluna	Tipo
<i>Ticker</i>	<i>ticker</i>	<i>string</i>
Início da Operação	<i>day</i>	<i>datetime</i>
Risco da Operação	<i>risk</i>	<i>float</i>
Resultado da Operação	<i>success_oper_flag</i>	<i>boolean</i>
<i>Flag</i> de Fim de Intervalo	<i>end_of_interval_flag</i>	<i>boolean</i>
Derivada Preço Médio	<i>mid_prices_dot</i>	<i>float</i>
<i>Spearman</i> (5 dias)	<i>spearman_corr_5_day</i>	<i>float</i> : Preço Médio, $f(x)=x$
<i>Spearman</i> (10 dias)	<i>spearman_corr_10_day</i>	<i>float</i> : Preço Médio, $f(x)=x$
<i>Spearman</i> (15 dias)	<i>spearman_corr_15_day</i>	<i>float</i> : Preço Médio, $f(x)=x$
<i>Spearman</i> (20 dias)	<i>spearman_corr_20_day</i>	<i>float</i> : Preço Médio, $f(x)=x$
<i>Spearman</i> (25 dias)	<i>spearman_corr_25_day</i>	<i>float</i> : Preço Médio, $f(x)=x$
<i>Spearman</i> (30 dias)	<i>spearman_corr_30_day</i>	<i>float</i> : Preço Médio, $f(x)=x$
<i>Spearman</i> (35 dias)	<i>spearman_corr_35_day</i>	<i>float</i> : Preço Médio, $f(x)=x$
<i>Spearman</i> (40 dias)	<i>spearman_corr_40_day</i>	<i>float</i> : Preço Médio, $f(x)=x$
<i>Spearman</i> (50 dias)	<i>spearman_corr_50_day</i>	<i>float</i> : Preço Médio, $f(x)=x$
<i>Spearman</i> (60 dias)	<i>spearman_corr_60_day</i>	<i>float</i> : Preço Médio, $f(x)=x$

Tabela 3.4: Comparação de Resultados

3.4.3 Índice de Lucratividade

Durante a etapa de criação dos modelos, é necessário uma métrica para ranqueamento das performances que acompanhe o contexto. O impacto de um acerto por parte do modelo implica em um ganho de $3X$ para a carteira, onde X é o valor de entrada da operação. Assim como uma falha implica em uma perda de $-X$ para a carteira.

O Índice de Lucratividade é um coeficiente entre 0 e 1 onde 0 significa o pior resultado possível, isto é, aquele no qual o modelo errou todas as operações no *dataset* de forma a trazer o maior prejuízo possível. Por outro lado, 1 significa o maior lucro que o modelo pode trazer se acertar todas as operações que o *dataset* permite. Nota-se que o valor do Índice que representa o lucro zero não é necessariamente 0.5, e sim algum valor intermediário que precisa ser fornecido durante o cálculo. Também é

importante ressaltar que como os *datasets* variam entre si, os valores de lucro zero também mudam, portanto a métrica é útil para criação de modelos que utilizem exatamente a mesma fonte de dados.

Seja A o número de operações de sucesso no *dataset* (classe 1) e B o número de operações de falha (classe 0), pode-se definir S_a como a soma dos riscos de todas as operações da classe 1 e S_b a soma dos riscos de todas as operações da classe 0. Assim, a Equação 3.16 representa uma função linear responsável por mapear o lucro L_m de um modelo no Índice de Lucratividade.

$$I_L = \frac{1}{(S_a - S_b)} L_m - \frac{S_b}{(S_a - S_b)} \quad (3.16)$$

O lucro L_m pode ser encontrado através da Equação 3.17, onde TP é a contagem de verdadeiros positivos do modelos e FN é a contagem de falsos negativos. Falsos positivos e verdadeiros negativos são desconsiderados do cálculo pois como não geram operações, não trazem ganho ou perda alguma.

$$L_m = 3 \times TP - FN \quad (3.17)$$

Por fim, o valor do Índice que representa o lucro zero pode ser encontrado a partir da Equação 3.16, basta impor a condição $L_m = 0$ (Equação 3.18).

$$I_{L_0} = -\frac{S_b}{(S_a - S_b)} \quad (3.18)$$

3.4.4 Balanceamento de Classes

Devido ao desbalanceamento dos *datasets* utilizados, que pode chegar até cerca de 90% de operações de falha contra 10% de operações de sucesso, faz-se imprescindível alguma técnica de balanceamento mencionada na Seção 2.3.3: *undersampling*, *oversampling* ou CSL. A natureza do problema também deve ser levada em consideração, pois um simples balanceamento via CSL daria a mesma importância à predição dos acertos e das falhas, no entanto o impacto de um acerto por parte do modelo gera um ganho de 3X para a carteira e uma falha gera uma perda de -X para os mesmos X apostados.

Levando em consideração as questões levantadas, optou-se por um balanceamento em cascada de 2 níveis:

- *Undersampling* via *Tomek Links*⁴, de cujo impacto na proporção das classes é muito pequeno.
- CSL levando em considerações o número de amostras e o impacto das classes para a carteira (Tabela 3.5).

Classe	Amostras	Peso p/ Carteira	Balanceamento
Op. de Falha (Classe 0)	N_0	1	$\frac{(1/4)N_1}{(N_0 + (1/4)N_1)}$
Op. de Sucesso (Classe 1)	N_1	4	$\frac{N_0}{(N_0 + (1/4)N_1)}$

Tabela 3.5: Balanceamento via CSL

3.4.5 Geração de Modelos

Os modelos *Random Forest* foram criados a partir da biblioteca *Scikit-Learn* [76]. Para a escolha do modelo a ser utilizado na simulação final, diversos outros modelos temporários foram criados e excluídos a fim de se selecionar apenas o melhor. Todos modelos temporários compartilham os parâmetros fixos indicados pela Tabela 3.6, porém diferem quanto aos parâmetros variáveis da Tabela 3.7 ou quanto às sementes aleatórias utilizadas. A lógica de criação dos modelos pode ser interpretada da seguinte forma:

- Para cada um dos 71 *tickers* da carteira (Tabela 3.1), o período completo de simulação vai de 01/01/2019 a 31/12/2021 e é dividido em 12 intervalos de 3 meses (WFA).
- Cada intervalo de 3 meses de simulação é governado pelo melhor modelo de ML criado previamente.

⁴Algoritmo que remove amostras da classe majoritária via identificação de *links*. Um *link* entre a amostra A da classe 1 e a amostra B da classe 2 é encontrado se o vizinho mais próximo de A é B e vice-versa [75].

- Por trás de cada modelo final selecionado existem 100 modelos temporários organizados em dois níveis: parâmetros variáveis e sementes aleatórias.
- O primeiro nível de diferenciação é dado parâmetros variáveis da Tabela 3.7, ou seja, as combinações entre *max_depth* e *max_features*, que geram os 10 pares (3, 3), (3, 4), (4, 3), (4, 4), ..., (7, 4).
- Para cada par criado, são criados 10 modelos com 10 sementes aleatórias diferentes.
- O modelo final escolhido possui os mesmos parâmetros do grupo de 10 modelos de sementes aleatórias diferentes que possuir o maior I_L/σ_{I_L} .

Parâmetro	Valor
n_estimators	200
criterion	gini
min_samples_split	24
min_samples_leaf	12
min_weight_fraction_leaf	0.0
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	True
oob_score	False
warm_start	False
ccp_alpha	0.0
max_samples	None

Tabela 3.6: Parâmetros fixos

Parâmetro	Valor
max_depth	[3, 4, 5, 6, 7]
max_features	[3, 4]

Tabela 3.7: Parâmetros variáveis

A Figura 3.29 ilustra a lógica de criação dos modelos apresentada através de um diagrama. O período de treinamento e de teste de cada modelo foi separado utilizando WFA e pode ser verificado pela Tabela 3.8. Entende-se por validade o período de tempo durante a simulação no qual o algoritmo criado pode atuar. Desta forma, cada ação na carteira possui 12 modelos com validade de 3 meses cada.

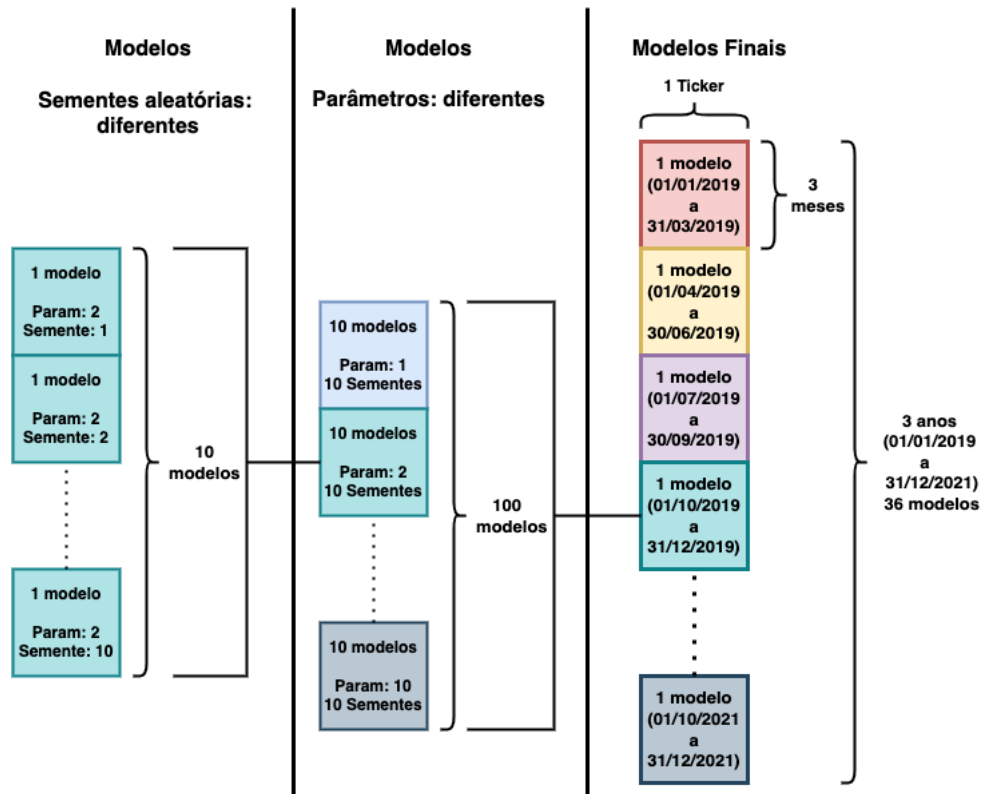


Figura 3.29: Diagrama de criação de modelos

Treinamento		Teste		Validade	
Início	Fim	Início	Fim	Início	Fim
01/01/2013	31/03/2018	01/04/2018	31/12/2018	01/01/2019	31/03/2019
01/04/2013	30/06/2018	01/07/2018	31/03/2019	01/04/2019	30/06/2019
01/07/2013	30/09/2018	01/10/2018	30/06/2019	01/07/2019	30/09/2019
01/10/2013	31/12/2018	01/01/2019	30/09/2019	01/10/2019	31/12/2019
01/01/2014	31/03/2019	01/04/2019	31/12/2019	01/01/2020	31/03/2020
01/04/2014	30/06/2019	01/07/2019	31/03/2020	01/04/2020	30/06/2020
01/07/2014	30/09/2019	01/10/2019	30/06/2020	01/07/2020	30/09/2020
01/10/2014	31/12/2019	01/01/2020	30/09/2020	01/10/2020	31/12/2020
01/01/2015	31/03/2020	01/04/2020	31/12/2020	01/01/2021	31/03/2021
01/04/2015	30/06/2020	01/07/2020	31/03/2020	01/04/2021	30/06/2021
01/07/2015	30/09/2020	01/10/2020	30/06/2020	01/07/2021	30/09/2021
01/10/2015	31/12/2020	01/01/2021	30/09/2021	01/10/2021	31/12/2021

Tabela 3.8: WFA - Intervalos de treinamento, teste e validade dos modelos

3.4.6 Modelo *Baseline*

No contexto de ML, entende-se como *baseline* a linha base de comparação de um modelo. Em outras palavras, é uma estratégia simples e de fácil implementação que traz uma performance razoável de se obter na realidade. Neste caso, utilizou-se a média de performance das ações da carteira, ou seja, supondo-se que o capital inicial fosse igualmente distribuído em cada ação disponível, o rendimento médio destas ações ao longo do tempo é o *baseline*.

A Figura 3.30 mostra o *baseline* calculado para os 71 *tickers* indicados na Tabela 3.1 no intervalo de 01/01/2019 a 31/12/2021. Adicionou-se o iBovespa (Seção 2.1.2) e o CDI⁵ acumulado para fins de comparação. Nota-se a diferença de performance do *baseline* para o iBovespa, o que é razoável já que o *baseline* tem composição fixa e o iBovespa tem uma composição variável tanto na escolha dos ativos quanto em seus

⁵Certificado de Depósito Interbancário. Indexador cujo valor é numericamente muito próximo à taxa básica de juros da economia, a Taxa Selic.

respectivos pesos. Além disso, as 71 ações escolhidas são de empresas em maioria presentes no mercado desde 2013, portanto possuem algum grau de consolidação e resiliência. A Tabela 3.9 mostra os indicadores de performance para o modelo *baseline*.

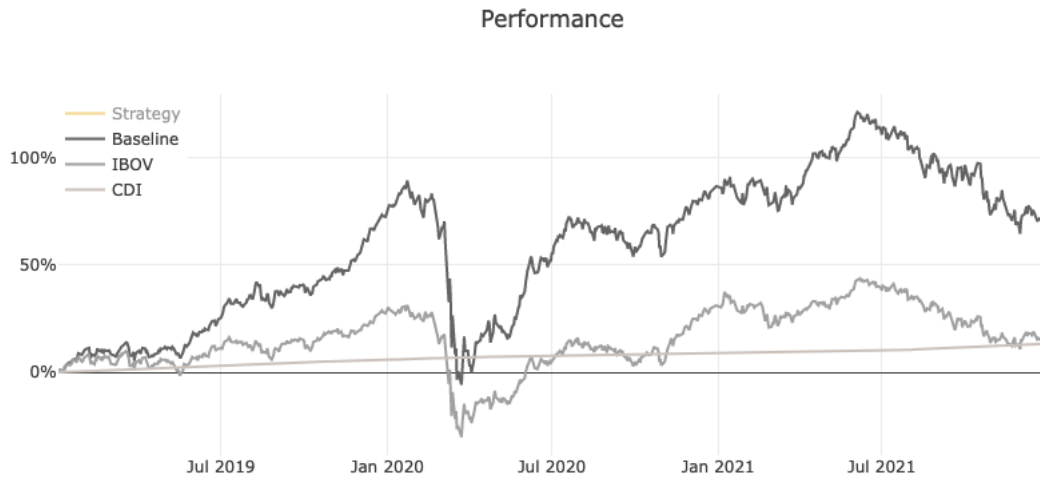


Figura 3.30: *Baseline* para o intervalo de 01/01/2019 a 31/12/2021

Rend. Final	Volatilidade	Sharpe	Sortino
73,94%	55,48%	0,68	0,74

Tabela 3.9: *Baseline* - Indicadores de Performance

Capítulo 4

Resultados

A partir dos parâmetros descritos e refinados na Seção 3.3, utilizou-se a seguinte configuração para encontrar a simulação com os melhores índices de performance:

- 71 *tickers* (Tabela 3.1)
- Período de simulação: 01/01/2019 a 31/12/2021
- Capital: R\$ 100000,00
- Volume mínimo de ações por negociação: 1
- Risco de entrada por operação: 0,29
- Período máximo de dias por operação: 45
- RCC: 0,000025
- Controle Proporcional para Uso de Capital (RCC Dinâmico): Sim
- Valor de Referência: 100%
- Constante K de Ganho Proporcional: 30000

Conforme descrito na Seção 3.3.5, a escolha dos 71 *tickers* foi realizada levando em consideração as seguintes preferências: diversidade de segmentos de mercado; disponibilidade da série temporal de dados a partir de 2013; e presença na composição do Ibovespa em qualquer data. O período de simulação escolhido engloba um intervalo recente dos últimos 3 anos, já que a escolha de momentos anteriores

a este pode estar associado a padrões de mercado muito diferentes do atual. Como os modelos criados utilizam *walk-forward analysis* e portanto atualizados a cada 3 meses de simulação para cada *ticker*, o comprimento do intervalo de 3 anos não é um problema. O capital e o volume mínimo por negociação foram escolhidos visando evitar potenciais problemas quanto à granularidade dos montantes de negociação.

Os demais parâmetros revelam valores refinados através de suas respectivas seções, a mencionar: Seção 3.3.4 para o Risco de entrada por operação; Seção 3.3.3 para o Período máximo de dias por operação; e Seção 3.3.6 para o RCC, o Valor de referência e a Constante K.

Os resultados encontrados podem ser verificados pela Figura 4.1 e pela Tabela 4.1.

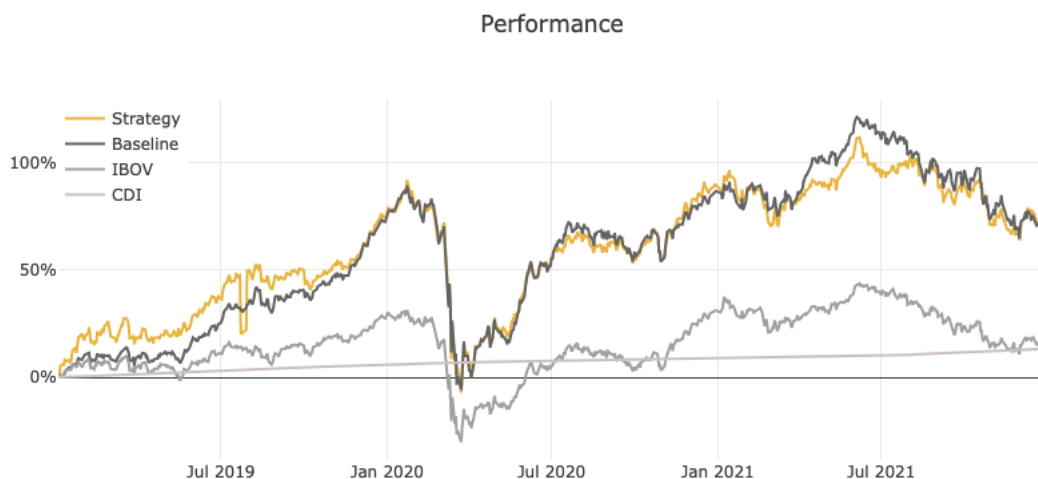


Figura 4.1: Performance final

Deve-se lembrar que a estratégia ou o modelo *baseline* é uma representação interna da média de rendimento do mercado para a mesma cesta de ações da estratégia a ser simulada (Seção 3.4.6). Com isso em mente, observa-se pela Figura 4.1 que o rendimento da estratégia final fica em torno do *baseline*, com exceção do primeiro ano de simulação, onde o desempenho é significativamente maior.

A Tabela 4.1 também mostra que ambas as estratégias obtiveram resultados bem próximos, em particular a estratégia final teve performance ligeiramente inferior,

Parâmetro	Estratégia	<i>Baseline</i>
Rendimento Final	72,79%	73,94%
Volatilidade	63,20%	55,48%
Índice de Sharpe	0,64	0,68
Índice de Sortino	0,74	0,74
Correlação de Spearman (c/ <i>Baseline</i>)	0,99	-
Correlação de Spearman (c /Ibovespa)	0,92	-
Uso Máximo de Capital	100%	100%
Uso Médio de Capital	97,22%	100%
Máximo de Operações Ativas	71	-
Média de Operações Ativas	65,97	-
Desvio Padrão de Operações Ativas	6,13	-
Operações Totais	3145	71
Operações de Sucesso	891 (28,3%)	-
Operações de Falha	2048 (65,1%)	-
Operações de <i>Timeout</i>	147 (4,7%)	-
Operações Incompletas	59 (1,9%)	-

Tabela 4.1: Resultado final

conforme esperado. Analisando individualmente os resultados, o rendimento da estratégia final está apenas 1,15% abaixo do *baseline*, o que é relevante, mas não tanto quando se tem em mente que este é apenas o retrato de um dos diversos dias de simulação volátil.

Apesar dos rendimentos finais equivalentes, não se pode deixar de notar o aumento de volatilidade da estratégia final em 7,72%, o que prejudica sua qualidade. Contudo, a volatilidade não deve ser analisada isoladamente, mas sim através do índice de Sharpe, que está 0,04 pontos abaixo da referência. Um desvio relevante, porém ainda sutil, mostrando que uma parcela da alta de volatilidade foi compensada por um rendimento médio maior. Já o índice de Sortino está igual em 0,74, mostrando que ambas as estratégias possuem o mesmo grau de confiança quando se analisa a rentabilidade em relação às oscilações de capital abaixo da média.

Dando sequência à análise, as altas correlações de Spearman mostram uma forte dependência da estratégia em relação ao *baseline* e ao iBovespa, o que pode ser facilmente verificado pela Figura 4.1. No entanto, aqui valem algumas ressalvas. Como a correlação de Spearman se baseia apenas nos ranques das funções e não propriamente na magnitude dos valores, é um cenário possível uma estratégia obter uma correlação com o *baseline* muito próxima de 1,0 ao mesmo tempo que um rendimento superior. Isso porque o ganho de capital com cada oscilação positiva seria maior do que a perda em uma oscilação negativa. Por outro lado, também é possível um cenário onde uma estratégia com uma baixa correlação de Spearman com o *baseline* apresente um rendimento superior. Este caso em particular seria um caminho mais interessante, pois mostraria uma versatilidade maior da estratégia criada para diferentes tipos de cenários de mercado.

O uso máximo de capital em 100% indica apenas que em algum momento todo capital esteve alocado em ativos. Contudo, o uso médio de capital em 97,22% traz a informação de que como quase todo capital esteve alocado o tempo todo, há pouco espaço para uma melhora de performance de maneira indiscriminada daqui para frente. Em outras palavras, as melhoras de rendimento precisarão passar por um aprimoramento dos modelos no que diz respeito a quais operações abrir mão para que outras possam prosperar mais. Nota-se que um uso médio de capital em 100% não seria vantajoso, uma vez que diminuiria-se a disponibilidade de capital para aporte em novas operações. Por outro lado, um uso médio muito baixo indicaria que a estratégia estaria longe de alcançar o seu máximo potencial de performance. Portanto, uma pequena folga é importante e necessária.

Assim como o uso médio de capital está alto, a média de operações ativas também está, o que é esperado. O mesmo raciocínio vale para o máximo de operações ativas e o uso máximo de capital. **HERALDO: Não mencionei o Desvio Padrão de Operações Ativas por não achar relevante. Devo removê-lo da tabela? Ou deixo assim mesmo?**

Por fim, deve-se ressaltar que o número total de operações de 3145 está saturado, conforme explicado ao final da Seção 3.3.6. Resumidamente, os parâmetros

de simulação, em especial o RCC e o K, estão configurados de forma a estressar a estratégia a ponto dela desobedecer em alguns momentos as ordens de compra vindas dos modelos de ML por falta de capital disponível. Tal abordagem trouxe um aumento de performance, mas deve ser usada com cautela. Independentemente desta questão, a taxa de acerto de 28,3% contra os 65,1% de falha mostra que apesar da estratégia acertar pouco, o lucro obtido é proporcionalmente maior que as perdas acumuladas.

Capítulo 5

Considerações Finais

5.1 Conclusão

Neste trabalho, foi proposta a criação de uma estratégia de *swing trade* com a utilização de técnicas de aprendizado de máquina. A partir dos resultados apresentados no Capítulo 4, foi possível observar que a estratégia proposta possui indicadores de performance muito próximos da estratégia *baseline*, que representa a média de rendimento do mercado.

Como não há consideração de proventos nos dados utilizados, a tendência é que ambas as estratégias apresentem uma performance real melhor do que a simulada. No entanto, a estratégia *baseline* deve revelar um aumento de performance relativamente maior, pois os papéis adquiridos ficam sempre em posse durante todo o período de simulação, não deixando intervalos de tempo descobertos. Também se deve ressaltar que o presente trabalho considerou premissas pessimistas durante as simulações, o que traz perspectivas ligeiramente mais promissoras.

Por fim, os resultados encontrados mostram que o uso de modelos de aprendizado de máquina pode auxiliar investidores no mercado de ações, no entanto para superar a média do mercado são necessários estudos mais aprofundados antes de uma implementação real.

5.2 Trabalhos Futuros

Este trabalho se baseou em premissas e regras que moldaram uma estrutura na qual os modelos pudessem operar. No entanto, alguns valores escolhidos podem ser revisitados a fim de serem aprimorados. Dentre eles, podemos citar a escolha do valor de 45 dias para o período máximo de dias por operação. A criação dos modelos de ML necessita de um valor prévio deste parâmetro. Assim, para se otimizar o valor do período, é necessário criar vários conjuntos de modelos, um para cada dia de período máximo que se deseja considerar, seguido pela execução da simulação. O problema desta abordagem é o tempo total de criação de apenas um conjunto de modelos, pois devido ao WFA, um conjunto requer 12 modelos para cada um dos 71 *tickers* da listagem geral (Figura 3.29 e Tabela 3.1), o que demora mais de um dia de execução em um computador de uso doméstico nos padrões atuais.

Deve-se mencionar também o parâmetro de razão entre ganho e perda, já que foi utilizado o valor constante de 3 a partir da recomendação do André Moraes. É possível abrir uma frente de trabalho visando encontrar um valor mais otimizado, o que necessitaria, assim como o caso anterior do período máximo de dias por operação, a criação de conjuntos de modelos com valores diferentes, o que também não é trivial.

Referências Bibliográficas

- [1] INVESTIDOR, B. D., “Como Interpretar o Gráfico de Candlestick”, <https://www.bussoladoinvestidor.com.br/grafico-de-candlestick/>, (Acessado em 5 de Abril de 2022).
- [2] KIRKPATRICK II, C. D., DAHLQUIST, J. A., *Technical analysis: the complete resource for financial market technicians*. FT press, 2010.
- [3] MORAES, A., *Se Afastando da Manada: Estratégias para vencer no Mercado de Ações*. Infomoney, 2016.
- [4] MÜLLER, A. C., GUIDO, S., *Introduction to machine learning with Python: a guide for data scientists*. ”O’Reilly Media, Inc.”, 2016.
- [5] STRANDS, “Unbalanced Datasets & What To Do About Them”, <https://blog.strands.com/unbalanced-datasets>, (Acessado em 5 de Abril de 2022).
- [6] DATACAMP, “KNN Classification Tutorial using Scikit-learn”, <https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn>, (Acessado em 5 de Abril de 2022).
- [7] MULTICHARTS, “Walk Forward Optimization”, https://www.multicharts.com/trading-software/index.php/Walk_Forward_Optimization, (Acessado em 28 de Junho de 2022).
- [8] B3, “B3 atinge 5 milhões de contas de investidores em renda variável em janeiro”, <https://www.b3.com.br/pt.br/noticias/5-milhoes-de-contas-de-investidores.htm>, (Acessado em 21 de Março de 2022).

- [9] INFOMONEY, “Robôs de investimentos já controlam mais de US\$ 200 bilhões ao redor do mundo”, <https://www.infomoney.com.br/onde-investir/robos-de-investimentos-ja-controlam-mais-de-us-200-bilhoes-ao-redor-do-mundo>, (Acessado em 22 de Março de 2022).
- [10] INFOMONEY, “No Brasil, robôs de investimento não conseguem bater melhores fundos”, <https://www.infomoney.com.br/onde-investir/no-brasil-robos-de-investimento-nao-conseguem-bater-melhores-fundos>, (Acessado em 22 de Março de 2022).
- [11] FERNÁNDEZ, A., “Artificial intelligence in financial services”, *Banco de Espana Article*, v. 3, pp. 19, 2019.
- [12] CVM, “Entendendo o Mercado de Valores Mobiliários”, <https://www.investidor.gov.br/menu/primeiros-passos/entendendo-mercado-valores.html>, (Acessado em 24 de Março de 2022).
- [13] BRASIL, “Lei nº 6.385, de 7 de dezembro de 1976. Dispõe sobre o mercado de valores mobiliários e cria a Comissão de Valores Mobiliários.”, http://www.planalto.gov.br/ccivil_03/leis/l6385.htm.
- [14] B3, “Uma das principais empresas de infraestrutura de mercado financeiro do mundo”, https://www.b3.com.br/pt_br/b3/institucional/quem-somos/, (Acessado em 24 de Março de 2022).
- [15] B3, “Ações”, https://www.b3.com.br/pt_br/produtos-e-servicos/negociacao/renda-variavel/acoes.htm, (Acessado em 24 de Março de 2022).
- [16] BRASIL, “Lei nº 6.404, de 15 de dezembro de 1976. Dispõe sobre as Sociedades por Ações.”, Capítulo XC, Seção VII, http://www.planalto.gov.br/ccivil_03/leis/l6404consol.htm, (Acessado em 24 de Março de 2022).
- [17] CVM, “Lei 6.404/76: Exposição de Motivos”, Capítulo II, Seção I, <https://www.gov.br/cvm/pt-br/acesso-a-informacao-cvm/institucional/sobre-a-cvm/>, (Acessado em 24 de Março de 2022).

- [18] INVESTIMENTOS, X., “Mercado secundário: entenda as diferenças com o mercado primário”, <https://conteudos.xpi.com.br/aprenda-a-investir/relatorios/mercado-secundario/>, (Acessado em 24 de Março de 2022).
- [19] BRASIL, “Lei nº 6.404, de 15 de dezembro de 1976. Dispõe sobre as Sociedades por Ações.”, Capítulo XV, Seção II, Art. 176, http://www.planalto.gov.br/ccivil_03/leis/l6404consol.htm, (Acessado em 24 de Março de 2022).
- [20] BRASIL, “Lei nº 6.404, de 15 de dezembro de 1976. Dispõe sobre as Sociedades por Ações.”, Capítulo XXI, Seção IV, Art. 275, § 4º, http://www.planalto.gov.br/ccivil_03/leis/l6404consol.htm, (Acessado em 24 de Março de 2022).
- [21] INFOMONEY, “Proventos: O que são, como funcionam e como ganhar dinheiro com eles?”, <https://www.infomoney.com.br/guias/proventos/>, (Acessado em 24 de Março de 2022).
- [22] B3, “Posições vendidas no mercado de ações”, https://www.b3.com.br/pt_br/noticias/short-selling.htm, (Acessado em 24 de Março de 2022).
- [23] FAMA, E. F., “Efficient capital markets: A review of theory and empirical work”, *The journal of Finance*, v. 25, n. 2, pp. 383–417, 1970.
- [24] INVESTOPEDIA, “Four Scandalous Insider Trading Incidents”, <https://www.investopedia.com/articles/stocks/09/insider-trading.asp#:~:text=Four>(Acessado em 25 de Março de 2022).
- [25] FAMA, E. F., FISHER, L., JENSEN, M., *et al.*, “The adjustment of stock prices to new information”, *International economic review*, v. 10, n. 1, 1969.
- [26] SHOSTAK, F., “In defense of fundamental analysis: A critique of the efficient market hypothesis”, *The Review of Austrian Economics*, v. 10, n. 2, pp. 27–45, 1997.

- [27] JUNG, J., SHILLER, R. J., “Samuelson’s dictum and the stock market”, *Economic Inquiry*, v. 43, n. 2, pp. 221–228, 2005.
- [28] SCHWAGER, J. D., *Market Sense and Nonsense: How the Markets Really Work (and how They Don’t)*. John Wiley & Sons, 2012.
- [29] FORBES, “Investing Basics: What Is A Market Index?”, <https://www.forbes.com/advisor/investing/stock-market-index/>, (Acessado em 28 de Março de 2022).
- [30] B3, “Ibovespa B3”, https://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/ibovespa.htm, (Acessado em 28 de Março de 2022).
- [31] B3, “ETF de Renda Variável”, https://www.b3.com.br/pt_br/produtos-e-servicos/negociacao/renda-variavel/etf-de-renda-variavel.htm, (Acessado em 28 de Março de 2022).
- [32] INVESTOPEDIA, “Fractional Share”, <https://www.investopedia.com/terms/f/fractionalshare.a>, (Acessado em 28 de Março de 2022).
- [33] BULKOWSKI, T. N., *Fundamental Analysis and Position Trading: Evolution of a Trader*, v. 605. John Wiley & Sons, 2012.
- [34] MURPHY, J. J., *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin, 1999.
- [35] EDWARDS, R. D., MAGEE, J., BASSETTI, W. C., *Technical analysis of stock trends*. CRC press, 2018.
- [36] BOLLINGER, J., *Bollinger on Bollinger bands*. McGraw Hill Professional, 2002.
- [37] APPEL, G., DOBSON, E., *Understanding MACD*. Traders Press, 2007.
- [38] IBM, “Artificial Intelligence (AI)”, <https://www.ibm.com/cloud/learn/what-is-artificial-intelligence>, (Acessado em 4 de Abril de 2022).

- [39] IBM, “Machine Learning”, [https://www.ibm.com/cloud/learn/machine-learning#:~:text=IBM\(Acessado em 4 de Abril de 2022\)](https://www.ibm.com/cloud/learn/machine-learning#:~:text=IBM(Acessado em 4 de Abril de 2022).).
- [40] ARTHUR, S., OTHERS, “Some studies in machine learning using the game of checkers”, *IBM Journal of research and development*, v. 3, n. 3, pp. 210–229, 1959.
- [41] WEISS, G. M., MCCARTHY, K., ZABAR, B., “Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?”, *Dmin*, v. 7, n. 35-41, pp. 24, 2007.
- [42] PARDO, R., *The evaluation and optimization of trading strategies*. John Wiley & Sons, 2011.
- [43] SHARPE, W. F., “The sharpe ratio”, *Streetwise—the Best of the Journal of Portfolio Management*, pp. 169–185, 1998.
- [44] ROLLINGER, T. N., HOFFMAN, S. T., “Sortino: a ‘sharper’ ratio”, *Chicago, Illinois: Red Rock Capital*, , 2013.
- [45] SPEARMAN, C., “The proof and measurement of association between two things.”, , 1961.
- [46] KIM, K., *Electronic and algorithmic trading technology: the complete guide*. Academic Press, 2010.
- [47] PEREIRA, D. F. R., *Aprendizado de máquina e aprendizado profundo para apoio à decisão no mercado financeiro*. Dissertação de graduação, Escola Politécnica - Universidade Federal do Rio de Janeiro, <https://monografias.poli.ufrj.br/download.php?arquivo=monopoli10025708.pdf&fcodigo=3659>, 2018.
- [48] GODFREY, M. D., GRANGER, C. W., MORGENSTERN, O., “THE RANDOM-WALK HYPOTHESIS OF STOCK MARKET BEHAVIOR a”, *Kyklos*, v. 17, n. 1, pp. 1–30, 1964.
- [49] BACHELIER, L., “Théorie de la spéculation”. In: *Annales scientifiques de l’École normale supérieure*, v. 17, pp. 21–86, 1900.

- [50] SOLNIK, B. H., “Note on the validity of the random walk for European stock prices”, *The journal of Finance*, v. 28, n. 5, pp. 1151–1159, 1973.
- [51] COOPER, J. C., “World stock markets: Some random walk tests”, *Applied Economics*, v. 14, n. 5, pp. 515–531, 1982.
- [52] MALKIEL, B. G., *A random walk down Wall Street: the time-tested strategy for successful investing*. WW Norton & Company, 2019.
- [53] SAID, A., HARPER, A., “The efficiency of the Russian stock market: A revisit of the random walk hypothesis”, *Academy of Accounting and Financial Studies Journal*, v. 19, n. 1, pp. 42–48, 2015.
- [54] RABINER, L. R., “A tutorial on hidden Markov models and selected applications in speech recognition”, *Proceedings of the IEEE*, v. 77, n. 2, pp. 257–286, 1989.
- [55] JADHAV, A., KALE, J., RANE, C., *et al.*, “Forecasting FAANG Stocks using Hidden Markov Model”. In: *2021 6th International Conference for Convergence in Technology (I2CT)*, pp. 1–4, IEEE, 2021.
- [56] DE ANGELIS, L., PAAS, L. J., “A dynamic analysis of stock markets using a hidden Markov model”, *Journal of Applied Statistics*, v. 40, n. 8, pp. 1682–1700, 2013.
- [57] ENDERS, W., *Applied econometric time series*. John Wiley & Sons, 2008.
- [58] ENGLE, R. F., “Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation”, *Econometrica: Journal of the econometric society*, pp. 987–1007, 1982.
- [59] BOLLERSLEV, T., “Generalized autoregressive conditional heteroskedasticity”, *Journal of econometrics*, v. 31, n. 3, pp. 307–327, 1986.
- [60] NELSON, D. B., “Conditional heteroskedasticity in asset returns: A new approach”, *Econometrica: Journal of the econometric society*, pp. 347–370, 1991.
- [61] HIGGINS, M. L., BERA, A. K., “A class of nonlinear ARCH models”, *International Economic Review*, pp. 137–158, 1992.

- [62] RABEMANANJARA, R., ZAKOIAN, J.-M., “Threshold ARCH models and asymmetries in volatility”, *Journal of applied econometrics*, v. 8, n. 1, pp. 31–49, 1993.
- [63] FRANCES, P. H., VAN DIJK, D., “Forecasting stock market volatility using (non-linear) Garch models”, *Journal of forecasting*, v. 15, n. 3, pp. 229–235, 1996.
- [64] MARCUCCI, J., “Forecasting stock market volatility with regime-switching GARCH models”, *Studies in Nonlinear Dynamics & Econometrics*, v. 9, n. 4, 2005.
- [65] ALBERG, D., SHALIT, H., YOSEF, R., “Estimating stock market volatility using asymmetric GARCH models”, *Applied Financial Economics*, v. 18, n. 15, pp. 1201–1208, 2008.
- [66] FELSEN, J., “Artificial intelligence techniques applied to reduction of uncertainty in decision analysis through learning”, *Journal of the Operational Research Society*, v. 26, n. 3, pp. 581–598, 1975.
- [67] NTI, I. K., ADEKOYA, A. F., WEYORI, B. A., “A systematic review of fundamental and technical analysis of stock market predictions”, *Artificial Intelligence Review*, v. 53, n. 4, pp. 3007–3057, 2020.
- [68] GANDHMAL, D. P., KUMAR, K., “Systematic analysis and review of stock market prediction techniques”, *Computer Science Review*, v. 34, pp. 100190, 2019.
- [69] BILDIRICI, M., ERSIN, Ö. Ö., “Improving forecasts of GARCH family models with the artificial neural networks: An application to the daily returns in Istanbul Stock Exchange”, *Expert Systems with Applications*, v. 36, n. 4, pp. 7355–7362, 2009.
- [70] YFINANCE, “yfinance”, <https://pypi.org/project/yfinance/>, (Acessado em 1 de Junho de 2022).
- [71] YAHOO!, “Yahoo! Finance”, <https://finance.yahoo.com>, (Acessado em 1 de Junho de 2022).

- [72] VIEW, T., “Trading View”, <https://www.tradingview.com/about/>, (Acessado em 20 de Setembro de 2022).
- [73] NORI, P., “Database Files in Project Github Page”, <https://github.com/Nori12/Projeto-Final/tree/master/database>.
- [74] PLOTLY, “Dash”, <https://dash.plotly.com/>, (Acessado em 28 de Junho de 2022).
- [75] HE, H., MA, Y., *Imbalanced learning: foundations, algorithms, and applications*. Wiley-IEEE Press, 2013.
- [76] SCIKIT-LEARN, “Random Forest”, <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>, (Acessado em 5 de Julho de 2022).
- [77] LUIZA, M., “Relações com Investidores - Magazine Luiza”, <https://ri.magazineluiza.com.br/>, (Acessado em 20 de Setembro de 2022).

Apêndice A

Inconsistência de Proventos na Biblioteca *yfinance*

Apesar da praticidade de obtenção de *candlesticks* diários que a biblioteca *yfinance* (Python) traz, seus valores de proventos (dividendos e juros sobre capital próprio) não são totalmente confiáveis. O estudo em questão mostra inconsistências tanto por duplicação quanto por inserção incorreta de proventos. Para isso, uma análise de caso foi realizada para a companhia Magazine Luiza (*ticker* MGLU3), onde foram comparados os dados obtidos do *yfinance* via *script* com o site de relações com investidores [77] da mesma. Além disso, utilizou-se a plataforma *TradingView* [72] para confirmação dos valores de preço de fechamento.

A Tabela A.1 mostra o histórico dos preços de fechamento para alguns dias específicos e datas importantes, como distribuição de proventos e desdobramentos, além de outras datas. A Tabela está ordenada do *candle* mais recente para o mais antigo e as marcações: em vermelho indicam valores incorretos; em azul indicam valores corretos; e em laranja indicam valores que progridam os erros dos valores incorretos. A data de execução do *script* é de 19/09/2020, o que é relevante, uma vez que a plataforma sempre retorna os preços dos *candles* já normalizado por todos os desdobramentos ¹ acumulados.

¹Em inglês: *split*

Data	Preço Fch <i>yfinance</i> (R\$)	Preço Fch Site RI (R\$)	Preço Fch <i>TradingView</i> (R\$)	Provento/Ação <i>yfinance</i> (R\$)	Provento/Ação Site RI (R\$/ação)	<i>Split</i> <i>yfinance</i>	<i>Split</i> Site RI
16/09/2022	4,46	4,46	4,46	-	-	-	-
01/07/2022	2,20	2,20	2,20	-	-	-	-
03/01/2022	6,72	6,72	6,72	-	-	-	-
07/07/2021	22,01	22,01	22,01	-	-	-	-
06/07/2021	21,07	21,07	21,07	0,015494	0,0154942583	-	-
05/07/2021	21,3645	21,37	21,36	-	-	-	-
04/01/2021	25,1817	25,18	25,18	-	-	-	-
30/12/2020	24,9319	24,93	24,93	0,026301	0,0263019985	-	-
29/12/2020	25,2354	25,24	25,24	-	-	-	-
15/10/2020	25,4650	25,47	25,46	-	-	-	-
14/10/2020	25,6347	25,64	25,54	-	-	1:4	1:4
13/10/2020	25,9541	25,96	25,95	-	-	-	-
03/08/2020	20,6061	20,61	20,61	0,094176	-	-	-
31/07/2020	20,0479	20,15	20,14	0,023541	0,094165968	-	-
30/07/2020	20,6654	20,77	20,76	-	-	-	-
29/07/2020	19,9012	20,00	19,99	-	-	-	-
15/04/2020	10,8798	10,93	10,93	-	-	-	-
14/04/2020	10,6441	10,70	10,69	0,179508	-	-	-
13/04/2020	10,2203	10,45	10,45	-	-	-	-
09/04/2020	10,1569	10,39	10,38	-	-	-	-
03/01/2020	11,9224	12,19	12,19	-	-	-	-
02/01/2020	12,0297	12,30	12,30	0,008947	0,0357891574	-	-
30/12/2019	11,6235	11,89	11,88	-	-	-	-
09/10/2019	9,6619	9,88	9,88	-	-	-	-
08/10/2019	9,2598	9,47	9,47	0,018402	0,0736066061	-	-
07/10/2019	9,3394	9,55	9,55	-	-	-	-
06/08/2019	8,9016	9,11	9,10	-	-	1:8	1:8
17/04/2019	4,8588	4,97	4,97	-	-	-	-
16/04/2019	4,9463	5,06	5,06	0,011571	0,370259884	-	-
15/04/2019	4,9594	5,07	5,07	-	-	-	-
03/01/2019	5,5812	5,71	5,71	-	-	-	-
02/01/2019	5,6416	5,77	5,77	0,018522	0,59270489	-	-
28/12/2018	5,4744	5,60	5,60	-	-	-	-

Tabela A.1: Análise de Consistência de Proventos: MGLU3

Analisando a Tabela A.1 de cima para baixo, nota-se que a primeira irregularidade notável ocorre no dia 14/10/2020, onde a plataforma *TradingView* apresenta um

preço de fechamento discrepante em relação ao site de RI da própria companhia e do *yfinance*. Como se trata de um evento singular e não é o foco deste estudo, ele foi desconsiderado.

Em seguida, nos dias 03/08/2020 e 31/07/2020, o *yfinance* registrou a presença dos proventos de R\$0,094176/ação e R\$0,023541/ação, respectivamente. O problema aqui é que além de ser muito improvável que qualquer empresa na bolsa brasileira distribuía proventos duas vezes em dois dias úteis seguidos, pode-se notar que o valor de R\$0,023541/ação equivale ao de R\$0,094176/ação quando multiplicado por 4. Em outras palavras, normalizando pelo desdobramento de 1:4 ocorrido em 14/10/2020, conclui-se que um dos proventos é duplicado. Como confirmação, o site de RI da Magazine Luiza dispõe de um comunicado sobre a distribuição de proventos de R\$0,094165968/ação em 31/07/2020.

Continuando a análise, é possível verificar que a partir da duplicata encontrada, os preços de fechamento do *yfinance* vão acumulando o erro. Em 14/04/2020, o *yfinance* contabilizou proventos de R\$0,179508/ação, no entanto, nada foi encontrado no site de RI, o que evidencia um lançamento incorreto. Nota-se também que o valor é relativamente alto quando comparado aos outros proventos de outras datas.

Os restantes dos valores de proventos do *yfinance* em azul equivalem aos comunicados pelo site de RI da companhia, porém deve-se levar em consideração os desdobramentos acumulados.

Por fim, pode-se concluir que o uso da plataforma *yfinance* no que diz respeito à disponibilização de proventos no contexto deste projeto não pode ser deferida, uma vez que a presença e a magnitude dos valores incorretos não é desprezível.