Universidade Federal
do Rio de Janeiro
Escola Politécnica

TÉCNICAS DE APRENDIZADO DE MÁQUINA APLICADAS A ESTRATÉGIA DE SWING TRADE DO 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

Julho de 2021

TÉCNICAS DE APRENDIZADO DE MÁQUINA APLICADAS A ESTRATÉGIA DE SWING TRADE DO MERCADO FINANCEIRO

Pedro Henrique Barbosa Nori

PROJETO DE GRADUAÇÃO SUBMETIDO AO CORPO DOCENTE DO CURSO DE ENGENHARIA ELETRÔNICA E DE COMPUTAÇÃO DA ESCOLA POLITÉCNICA 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:

Heraldo Luis Silveira de Almeida, D. Sc.

Examinador:

Prof xxxxx

Examinador:

Prof xxxx

Rio de Janeiro

Julho de 2021

Declaração de Autoria e de Direitos

Eu, *Pedro Henrque Barbosa Nori* CPF *134.129.077-82*, autor da monografia *TÉCNICAS DE APRENDIZADO DE MÁQUINA APLICADAS A ESTRATÉGIA DE SWING TRADE DO 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. Excetuam-se do item 1. eventuais transcrições de texto, figuras, tabelas, conceitos e idéias, que identifiquem claramente a fonte original, explicitando as autorizaçõees 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 violaçõo 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

DEDICATÓRIA

AGRADECIMENTO

INCLUIR - Pais, Professores, Aridio, MinervaBots, Amigos, Heraldo

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 artificias 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 pela decisão de entrada nas operações e os respectivos preços alvos de venda.

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

ABSTRACT

Insert your abstract here. Insert your abstract here. Insert your abstract here. Insert your abstract here. Insert your abstract here.

Key-words: word, word, word.

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*

DT - *Decision Tree*

EGARCH - *Exponential Generalised ARCH*

EMA - *Exponential Moving Average*

GARCH - *Generalised ARCH*

HME - Hipótese do Mercado Eficiente

HMM - *Hidden Markov Model*

iBovespa - Índice Bovespa

IIR - *Infinite Impulse Reponse*

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*

SVM - *Support Vector Machine*

TGARCH - *Threshold Generalised ARCH*

UFRJ - Universidade Federal do Rio de Janeiro

WFA - *Walk-Forward Analysis*

Sumário

Lista de Figuras

Lista de Tabelas

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 realizas vendas a descoberto[1], portanto

[1]Venda a descoberto (*short selling*): Venda de ações anterior a compra das mesmas através de um contrato de aluguél. 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 pecualiaridades 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 algortimos 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* resposável 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

Neste capítulo, são introduzidos alguns conceitos chave para o entendimento do projeto. Nas próximas seções, são feitas contextualizações sobre o Mercado de Capitais, Bolsa de Valores, Ações e Aprendizado de Máquina.

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 superávitarios, 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 insituições financeiras ou até mesmo as pessoas físicas.

No Brasil, o Mercado de Capitais é regulado e fiscalizado pela CVM (Comissão de Valores Mobiliários), 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 (Brasil, Bolsa, Balcão) [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 vantanges 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 (Initial Public Offering), onde as ações que compõe 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 expectavida corretamente induz, o lucro é comumente aferido durante a venda de um determinado papél posteriormente à sua aquisição a um preço de compra inferior. No entando, 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 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 FAMA, seja por resultados inconclusivos ou por anomalias detectadas no comportamento do mercado. Por exemplo, 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 (*Exchange-Traded Funds*) [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 a transacionar. Nesse contexto, o Mercado Fracionário [32] surge

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

com o objetivo de facilitar negiociações de volumes menores que o lote mínimo permitido. Na prática, ações fracionárias são agrupadas até formarem um lote para então serem negociadas. Normalmente o Mercado Fracionário possui menor liquidez e maior volatilidade, mas sempre acompanha o preço do ativo negociado no mercado aberto.

Ações fracionárias 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.

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.

No Brasil, o lote mínimo é de 100 ações e o Mercado Fracionário permite a compra de no mínimo 1 ação.

2.2 Tipos de Análises

2.2.1 Análise Fundamentalista

A Análise Fundamentalista (AF) é um muito 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 climáticas; os desastres climáticos, naturais ou não, dentre outros fatores.

Devido à natureza desorganizada e desestruturada dos dados que representam os fatores mencionados, torna-se muito difícil implementar uma automação.

2.2.2 Análise Técnica

A Análise Técnica (AT) 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 muito 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, além da criação de informações derivadas que se mostram 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 (Moving Average Convergence-Divergence) [37].

Leitura de Gráficos de Candlesticks

Gráficos de Candlesticks[2] 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.

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

[2]Em português: Gráfico de Velas.

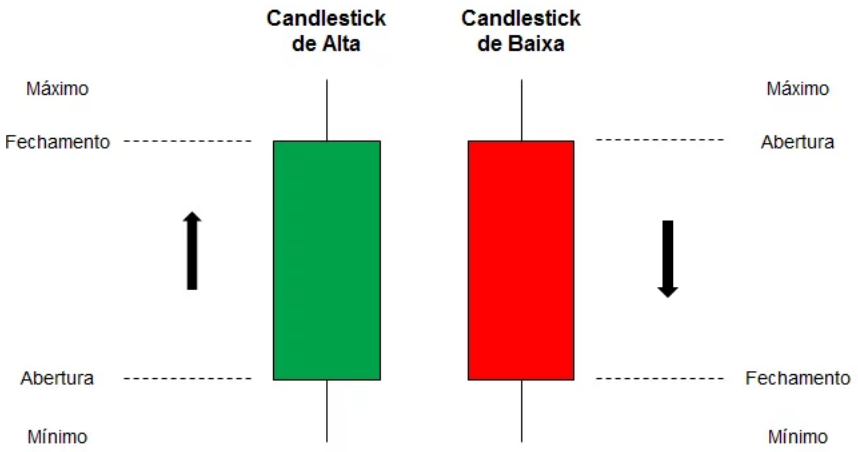

Figura 2.1: Leitura de um gráfico de *candlestick* [1]

mercado, que em condições ideais segue o padrão de: tendência de alta[3]; topo; tendência de baixa[4]; e fundo, intercalados com períodos de consolidação[5]. A Figura 2.2 ilustra esse comportamento.

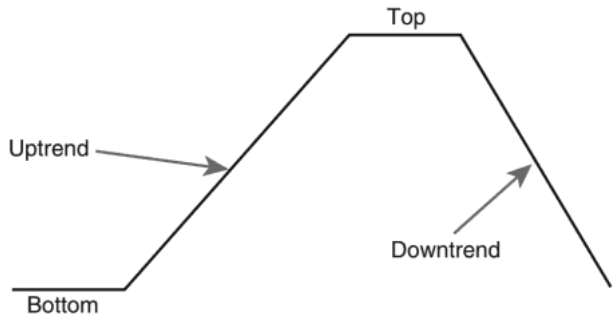

Figura 2.2: Comportamento do mercado ideal segundo a Teoria de Dow [2]

Média Móvel Exponencial

A Média Móvel Exponencial (MME) possui uma característica que a torna relevante para estratégias de AT. Ela dá um maior peso relativo às amostras mais recentes dentro de uma série temporal. As Equações 2.1 e 2.2 mostram o seu cálculo, onde P_t representa o preço atual, MME_{t-1} é a média acumulada até o instante anterior e K é uma constante definida pela quantidade de amostras desejadas $n > 0$.

[3]Topos e fundos ascendentes.

[4]Topos e fundos descendentes.

[5]Topos e fundos lateralizados.

$$MME_t = (P_t - MME_{t-1}) * K + MME_{t-1} \tag{2.1}$$

$$K = \frac{2}{n+1} \tag{2.2}$$

Suporte, Resistência e Linhas de Tendência

Suporte e Resistência são regiões em um gráfico de *candlestick* onde existe um grande 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.

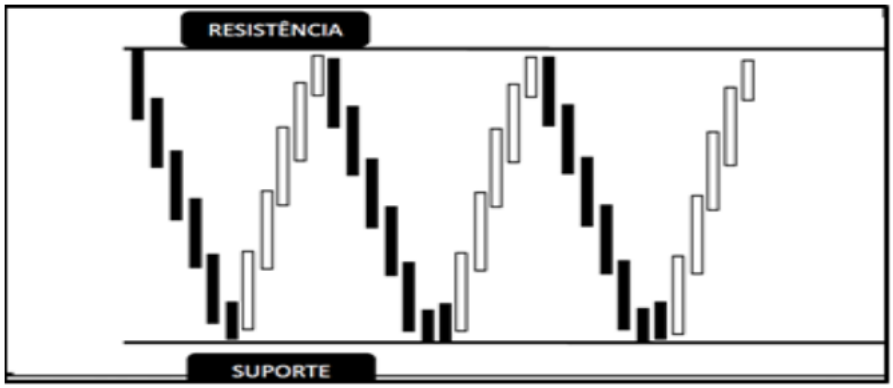

Figura 2.3: Formação de linhas de Suporte e de Resistência [3]

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

2.3 Aprendizado de Máquina

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

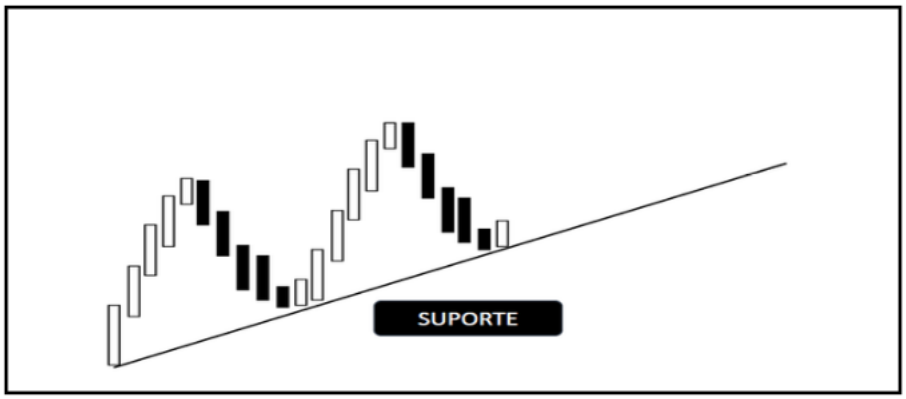

Figura 2.4: Formação de uma Linha de Tendência de Alta [3]

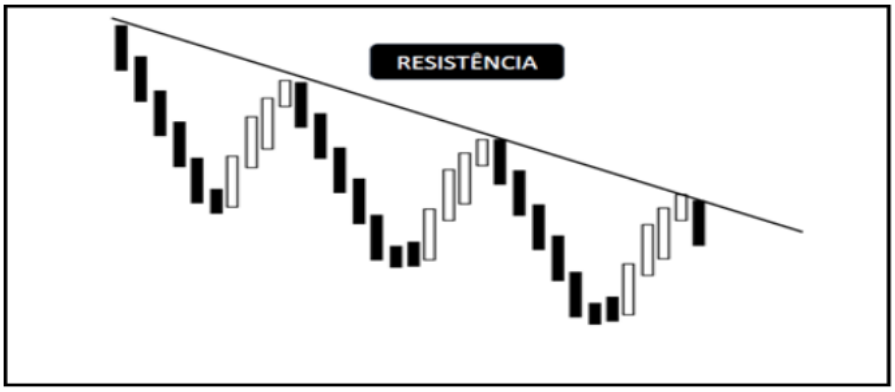

Figura 2.5: Formação de uma Linha de Tendência de Baixa [3]

Em geral, algoritmos de ML buscam realizar tarefas extremamente complexas computacionalmente sem serem explitamente 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 um modelo 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 uma melhor performance [4].

O conjunto de dados (*dataset*) com exemplos de entrada e saída utilizado para criação do modelo é chamado de dados de treinamento (*training set*). Existe um outro conjunto de dados utilzado 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 sobreponho à qualidade de generalização do modelo (explicado a seguir). Como regra geral de uso, é aconselhável separar 75% dos dados para os dados de treinamento e 25% para os dados 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 predições acuradas em conjuntos de dados semelhantes ao de treinamento, porém jamais vistos (dados de teste). Quanto maior a taxa de acerto nos dados 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 dos dados 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 dos dados de treinamento, implicado 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*.

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

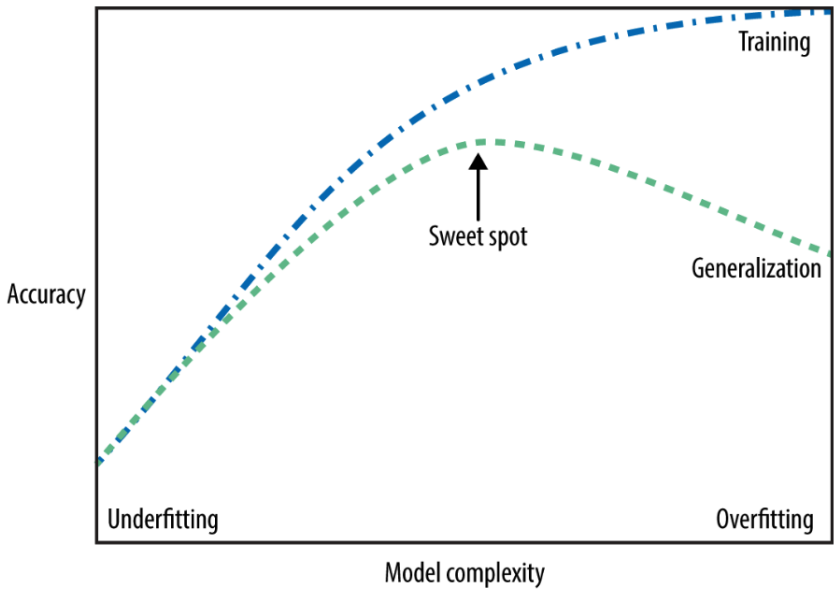

Figura 2.6: Relação entre complexidade e acurácia de um modelo [4]

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 espectativa 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 ilusório indíce de acurácia nos dados de teste. Isso acontece porque o modelo tende

a quase sempre escolher a classe com maior frequencia 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 numa 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 simplificado sobre os algoritmos de AS mais pertinentes ao presente trabalho, em ordem crescente de complexidade. O 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 predizer 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.

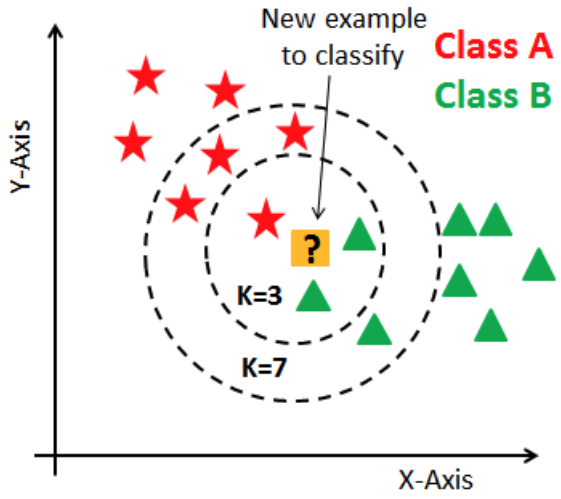

Figura 2.8: Funcionamento de um algoritmo k-NN para o problema de classificação [6]. Para K=3 a classe é B e para K=7 a classe é A.

Decision Tree

Em essência, uma Árvore de Decisão[6] 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[8] estejam totalmente puras, ou seja, as sequências de decisão que levam a um resultado só englobam amostras de um tipo de classe. Ao contrário de folhas impuras, que contém a presença de mais de uma classe, onde normalmente se escolhe a de maior número de amostras como resultado. O problema é que a presença excessiva de folhas

[6]Em inglês: *Decision Tree*.

[7]Em português: se/senão.

[8]Em inglês: leafs.

totalmente puras 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 para se criar uma nova ramificação; dentre outros.

Algumas vantagens deste modelo estão no relativamente fácil entendimento e visualização dos critérios de decisão para o projetista em árvores pequenas. O tempo de processamento computacional envolvido na criação deste modelo é razoavelmente curto. Não é necessário um pré-processamento dos dados, uma vez que cada característica é processada separadamente. A Figura 2.9 mostra a estrutura por trás de uma Árvode de Decisão.

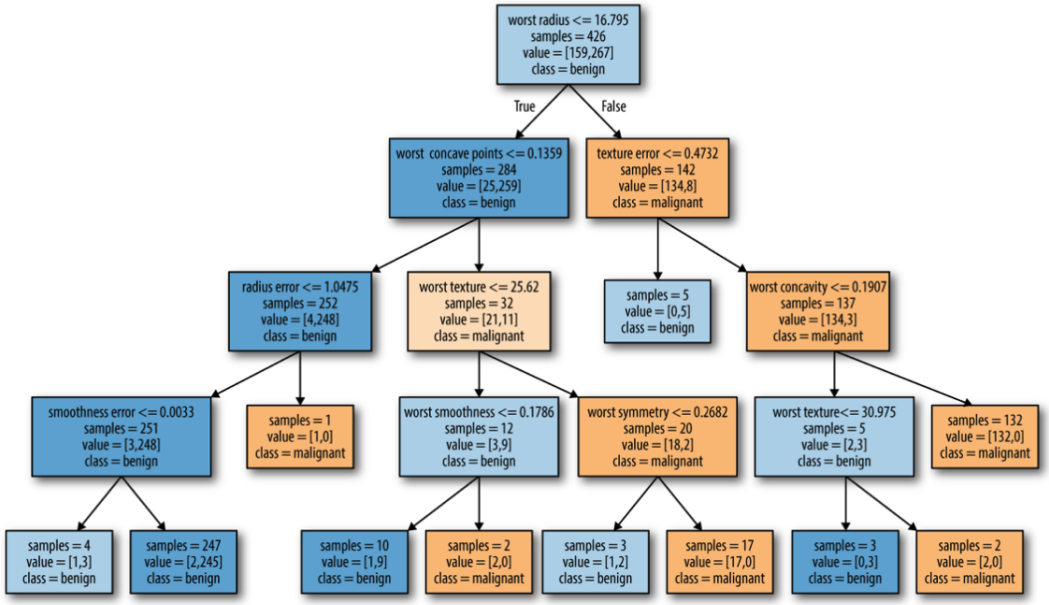

Figura 2.9: Visualização de uma Árvose de Decisão para um *dataset* de câncer de mama [4].

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*[9] 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 no número de árvores que serão utilizadas.

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, o que difícil de ser analisado quando existem muitras árvores. É possível compensar o aumento do tempo de processamente envolvido na criação deste modelo com paralelização em núcleos de processamento da CPU[10].

2.4 *Walk-Forward Analysis*

Walk-Forward Analysis [42] é um processo de otimização mais voltado para séries temporais no contexto de finanças. O conjunto de dados diponível (*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 mostra o processo para uma série temporal. 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.

A natureza do WFA ancorado permite uma maior adaptação dos modelos de ML de acordo com as tendências de mercado, que mudam significativamente com o tempo. Durante o treinamento de um modelo, considerar informações temporalmente muito distantes do período de aplicação do mesmo pode comprometer

[9]Em português: Floresta Aleatória.

[10]Do inglês: *Central Process Unit*.

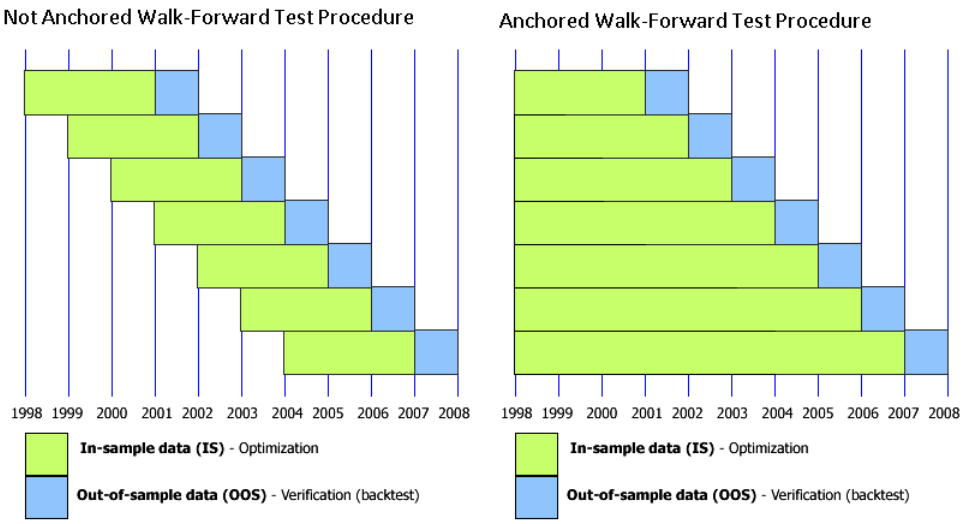

Figura 2.10: *Walk-Forward Analysis* (Não-Ancorado e Ancorado) [7].

significavamente 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 Sharpe[11] 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 2.3.

$$S_a = \frac{E[R_a - R_b]}{\sigma_a} \tag{2.3}$$

[11] Também conhecido como *Sharpe Index*, *Sharpe Ratio* ou até *Sharpe Measure*.

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 2.4 e 2.5 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 do investimento.

$$S_a = \frac{E[R_a - R_b]}{\sigma_a} \tag{2.4}$$

$$\sigma_a = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Min(0, X_i - T))^2} \tag{2.5}$$

2.5.3 Correlação de Spearman

A Correlação de Postos de Spearman ou simplesmente Correlação de Spearman for 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 postos[12] 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 2.6, 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 2.7, é para o caso em que há presença de nós.

$$\rho_s = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2 - 1)} \tag{2.6}$$

$$\rho_s = \rho_{rgX,rgY} = \frac{cov(rgX, rgY)}{\sigma_{rgX}, \sigma_{rgY}} \tag{2.7}$$

[12]Classificação ordenada das amostras em escala ordinal. Do inglês: *ranks*.

Na Equação 2.6, d_i é diferença entre os dois postos de cada variáveis aleatória, mostrado através da Equação 2.8.

$$d_i = rgX_i - rgY_i \tag{2.8}$$

Um exemplo prático do cálculo da correlação pode ser mostrado 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 amostram 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.

	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

Após o cálculo dos postos, deve-se utilizar a Equação 2.7, encontrado o valor 0.6687.

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 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 automatização da tomada de decisões de estratégias por um computador capaz de enviar ordens de compra e venda diretamente ao mercado.

Para efeito de simplificação, os modelos de AT aplicados ao mercado financeiro serão agrupados 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 da tendência de mercado. Quando comparada aos outros tipos, é a metodologia mais simples e democrática, uma vez que pessoas com pouco ou nenhum conhecimento sobre estatística e inteligência artificial podem operar em estratégias próprias.

Diversos *traders*[13] e investidores utilizam este tipo de abordagem. Dentre eles podemos citar MORAES [3], de cujas contribuições servirão como base neste trabalho para um aperfeiçoamento via aprendizado de máquina.

[13]Em 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 viés especulativo.

2.6.2 Modelos Baseados em Processos Estocásticos

De acordo com GODFREY [47], a hipótese de que a flutuação de preços no mercado de ações poderia ser explicada por uma Random Walk[14] foi feita por BACHELIER [48]. 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, 49, 50], assim como estratégias [51]. Nota-se que até hoje utiliza-se Random Walks para testar a hipótese de eficiência de mercados [52].

Outra abordagem utilizada são os Modelos Ocultos de Markov (do inglês Hidden Markov Model, ou HMM) [53]. 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 (oculto) a partir de um processo observável.

Em sua pesquisa, JADHAV et al [54] utiliza um modelo HMM para previsão do preço de fechamento do dia seguinte para ações FAANG[15]. A partir da série histórica de preços OHLC[16], seu modelo atinge uma eficiência de 97%-99%, calculado a partir do erro percentual absoluto médio[17].

Uma outra aplicação de modelos HMM é dada por DE ANGELIS et al [55], 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.

---

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

[15]Facebook, Amazon, Apple, Netflix, Google.

[16]Open, High, Low, Close. Em português: Abertura, Máximo, Mínimo, Fechamento.

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

Por fim, pode-ser mencionar o uso de modelos ARCH[18] (Autoregressive Conditional Heteroskedasticity). A ideia central está na modelagem de uma variância condicional, ou seja, que muda de acordo o instante da série [56]. 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 2.9.

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-1}^2 \tag{2.9}$$

O modelo ARCH foi proposto por ENGLE em 1982 para estimar a variância da inflação do Reino Unido [57]. A partir daí, várias derivações surgiram, como por exemplo: GARCH [19] por BOLLERSLEV [58] em 1986, EGARCH [20] por NELSON em 1991, NGARCH [21] por HIGGINS e BERA[59] em 1992, TGARCH [22] por ZAKOIAN e RABEMANANJARA [60] em 1993, dentre outros. Alguns dos modelos da família ARCH podem ser encontrados nos trabalhos de FRANSES e DIJK [61], de MARCUCCI [62] e de ALBERG et al [63].

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 [64], 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.

NTI et al [65] 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

[18] Em português: Heteroscedasticidade Condicional Auto-regressiva.

[19] Generalised ARCH.

[20] Exponential Generalised ARCH.

[21] Non-linear Generalised ARCH.

[22] Threshold Generalised ARCH.

em AT, 23% são baseados em AF e 11% usam análises mistas. Além disso, os algoritmos mais utilizados são ANN[23] (*Artificial Neural Networks*) e SVM[24] (*Support Vector Machine*).

De forma semelhante, GANDHMAL e KUMAR [66] verificaram 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*[25]

É possível encontrar também modelos híbridos, com uma combinação de GARCH com ANN feita por BILDIRCI e ERSIN [67].

[23]Em português: Redes Neurais Artificiais.

[24]Em português: Máquina de Vetor de Suporte.

[25]Em português: Difuso.

Capítulo 3

Metodologia

3.1 Resumo

As seções a seguir trazem detalhes quanto a estrutura técnica do projeto. Portanto, a Figura 3.1 apresenta uma noção geral de como as estrutras se conectam.

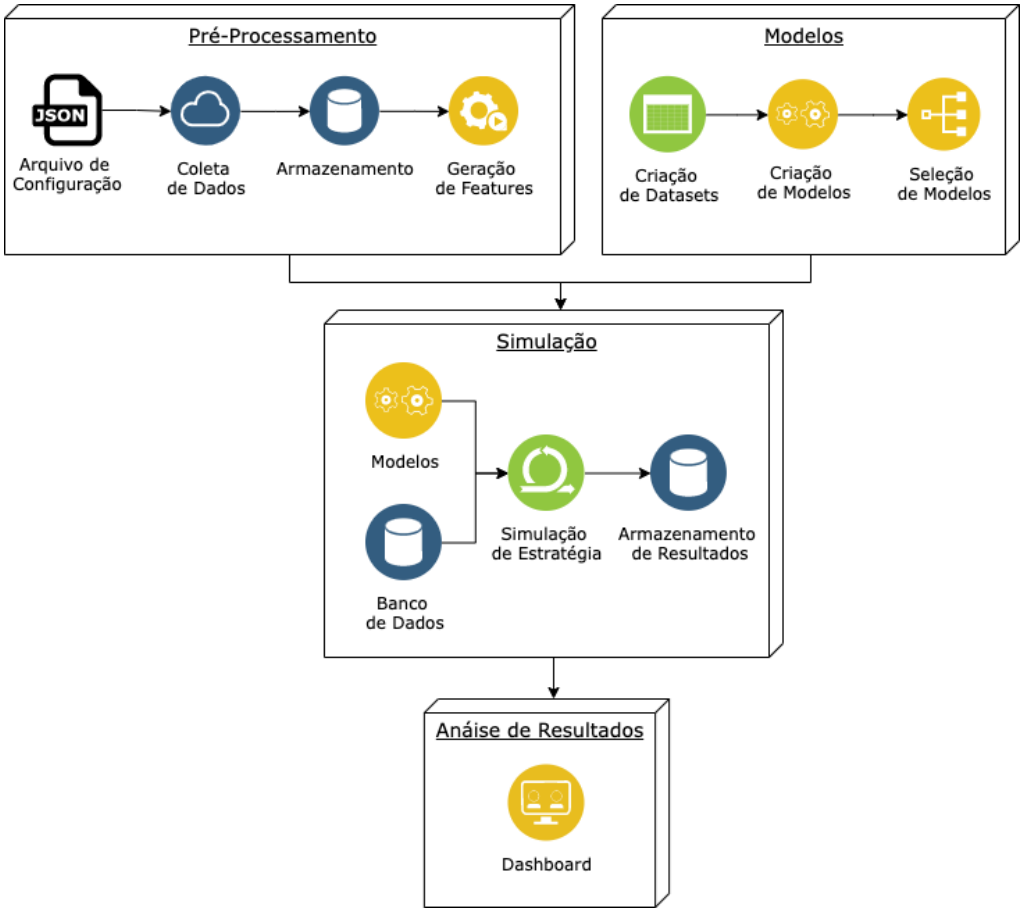

Figura 3.1: Estrutura do técnica do projeto

Primeiro, antes da execução do código principal, é necessário garantir que os modelos estão devidamente localizados em pasta apropriada. Para isso, faz-se imprescindível a criação dos *datasets* para cada ação a ser simulada, pois servem de entrada de dados para a criação e seleção dos modelos, etapa esta que deve ser executada logo em sequência. 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, quais os ativos envolvidos e seus recpectivos intervalos de tempo. Uma vez verificado no banco os dados já existentes, faz-se um *download* apenas dos dados necessários. Se houver alguma atualização de dados, as *features* de uso geral são recalculadas e armazenadas no banco a fim de servir de insumo para as estratégias que estarão por vir.

Completada 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*, *dash* e *multiprocessing*. 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.

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. Após a seleção do nome, são configurados os parâmetros internos da estratégia. A Tabela 3.3 da Seção 3.3.8 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 compri-

mentos 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.

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* [68], uma ferramenta não oficial que transmite dados públicos da plataforma *Yahoo! Finance* [69], um subsistema da rede *Yahoo!*.

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

- Os valores de proventos (i.e., dividendos e juros sobre capital próprio) que a biblioteca disponibiliza não são consistentes para a B3, portanto não são utilizados por este projeto. Testes internos confirmaram a presença de diversos proventos corretamente apresentados e ajustados pelos respectivos desdobramentos acumulados, porém somados a alguns *outliers* inexistentes na realidade, o suficiente para questionar seu uso em escala (i.e., para vários ativos

sem verificação individual). HERALDO: Devo mostrar evidências do teste que corrobora esta afirmação?

- Os volumes de negociação disponibilizados não necessariamente coincidem com a plataforma TradingView em valores absolutos, porém coincidem em valores relativos (i.e., variação de volume dia após dia para um mesmo ativo), o que é suficiente para este trabalho. HERALDO: (1) Na verdade, encontrei algumas evidências de que os valores relativos conferem, mas nenhum evidência de que não conferem. (2) Será que posso citar a plataforma TradingView? Ou melhor, devo tomar algum cuidado?

- *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).

Os dados obtidos são *candlesticks* diários (OHLCV). Com a mesma facilidade, é possível adquirir janelas de tempo semanais, no entanto para evitar potenciais problemas de consistência de dados, as mesmas são calculadas internamente a partir da janela diária via comandos SQL[1].

Apenas os dados não existentes no banco são baixados via *yfinance*. Para isso, um *trigger*[2] é acoplado às tabelas de *candlesticks* e acionado sempre que operações de *insert*, *update*, *delete* e *truncate* são realizadas. Quando ativado, chama uma função responsável por atualizar a tabela de *status*, que registra o intervalo de tempo representado nas tabelas de *candlesticks* para cada *ticker* envolvidos. Deve-se 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.

[1] *Structured Query Language*: Linguagem usada para administrar bancos de dados relacionais.

[2] Procedimento armazenado em um banco de dados que é chamado automaticamente sempre que ocorre um evento.

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* de resultados.
- 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 repentida.
- Diminuição do tráfego na rede pela persistência dos *candlesticks* já obtidos.

Os *candlesticks* semanais são calculados via *query* SQL para garantir a consistência dos dados, já que a possibilidade de inconsistência se fez presente entre dados *intraday* e diários, conforme mencionado na Seção 3.2.2.

A figura 3.4 mostra o ERD[3] do banco. Os *scripts* de criação e população inicial do banco de dados pode ser encontrado em [70]. HERALDO: (1) Vale a pena mostrar o ERD do banco? (2) Devo mencionar as constraints de banco que garantem consistência, como por exemplo a comparação do preço máximo de um candle com seus outros valores a fim de garantir que este de fato é máximo? Preços não negativos, valores não nulos, etc.

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

[3] *Entity-Relationship Diagram*. Em português: Diagrama de Entidade Relacionamento.

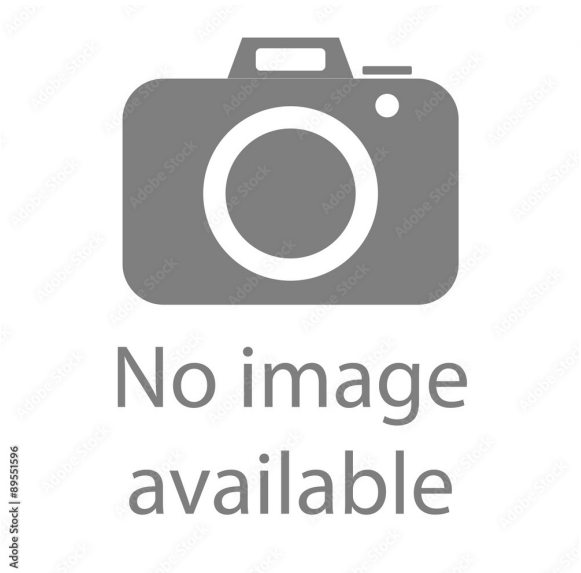

Figura 3.4: ERD do Banco de Dados

está no suporte à escolha do momento de entrada apropriado nas operações, o que é fundalmentalmente a responsabilidade do modelo de *Machine Learning*. Como podem ser utilizadas por qualquer estratégia, 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.

Nesta etapa do projeto, assim como em diversas outras, faz-se necessário atenção e cuidados quanto a erros de não-causalidade, que mesmo sendo sutis, podem influenciar drasticamente os resultados finais. HERALDO: Devo omitir esse parágrafo? A intenção é dizer que o autor não subestima essa problema, portanto tomou cuidados a nível de implementação para evitá-lo. Por outro lado, talvez essa preocupação já esteja subentendida, sendo desnecessário enfatizá-la.

As *features* de Uso Geral utilizadas são:

- ~~Média Móvel Exponencial de 17 períodos no gráfico diário.~~
  HERALDO: Feature original do André Moraes. Não utilizado nos modelos.
- ~~Média Móvel Exponencial de 72 períodos no gráfico diário.~~
  HERALDO: Feature original do André Moraes. Não utilizado nos modelos.

- ~~Média Móvel Exponencial de 72 períodos no gráfico semanal.~~

  HERALDO: Feature original do André Moraes. Não utilizado nos modelos.

- ~~*Flag* de Tendência de Alta.~~

  HERALDO: Feature original do André Moraes. Não utilizado nos modelos DA FORMA QUE O ANDRÉ USA, por isso o risco no texto. Ele se baseia em picos e vales ascendentes para justificar uma tendência de alta (de acordo com a teoria de Dow). A nível de código, a estratégia mais simples é comparar os últimos 2 picos de mínimo e 2 picos de máximo, verificar se um é maior que o outro com alguma margem de tolerância, que se caracteriza tendência de alta. Porém essa informação é ruidosa, principalmente em mercados em consolidação (=muito lateralizados). O que ocorre na prática do André, é que não são só os ultimos 4 picos que são analisados. Primeiro, ele já tira uma noção visual se o mercado anda em consolidação, isso requer olhar uma quantidade variável de picos que possuem algum grau de proximidade nas magnitudes, mas se comportam entre linhas de suporte e resistência. Quantos dias passados devo olhar para medir a consolidação? (retórica rs). Se uma ação vem em tendência de baixa, por exemplo, ele não espera exatamente o par perfeito de 4 picos ascendentes para dizer se o mercado está em alta, muitas vezes porque o quarto pico ainda nem se formou consistentemente. Tudo isso se baseia em uma boa identificação de picos, que felizmente foi implementada, porém tem um atraso mínimo de 9 dias para identificar qualquer pico. Enfim, embora possível, não é trivial uma boa métrica fidedigna ao André, seguindo esta lógica.

- ~~*Stop Loss* no último pico de rompimento/reversão de tendência (Pressupõe preço de compra definido).~~

  HERALDO: Não utilizado mais pois era feature do André Moraes. Não utilizado nos modelos. Em particular, essa não é trivial de se calcular e foi uma das quais distanciou a simulação feita da estratégia real do André. Os picos relevantes que retratam a memória do mercado muitas vezes estão relacionados a acontecimentos notáveis, como crises financeiras, acidentes industriais, relatórios jurícos, escândalos, aquisições novas, etc. O André pode não avaliar os acontecimentos menores na escolha do stop por ser grafista e não estar

inteiramente ligado nas notícias, mas leva em consideração os mais marcantes. Como análise de notícias está completamente fora do escopo, obter "grau de importância"de um pico com alguma precisão requer no mínimo olhar os dados intraday (seção de coleta de dados explica porque não usei) e avaliar o volume de negociações na regiões. Contudo, idealmente deve-se olhar o livro de ofertas para tirar métricas das forças de compra e de venda, talvez aliadas ao volume de negociação e assim obter um valor razoável. Quando implementei na tentativa de simular o André, usei simplemente o pico mais próximo abaixo do preço de compra, dado uma distância mínima de 1%.

- ***Flag* de Identificação de Crises**

Flag criado para prever crises financeiras através da identificação de anomalias nos volumes de negociação. Seu objetivo é impedir que o modelo de ML entre em qualquer operação durante sua presença. Para isso, utiliza-se a média $\overline{V}$ e o desvio padrão σ_V do volume de negociações dos últimos 60 dias úteis, junto com o volume V do dia corrente. Adiciona-se um efeito de inércia de 2 dias úteis consecutivos para ativar uma anomalia de volume e outra inércia de 8 dias úteis para persistência do *flag*. As Equações 3.1 e 3.1 mostram o cálculo e a Figura 3.5 ilustra a eficácia do *flag*.

$$V_{anomaly(i)} = \begin{cases} 1, & \text{se } V_{(i)} \geq \overline{V} + \sigma_V \quad \text{e} \quad V_{(i-1)} \geq \overline{V} + \sigma_V \\ 0, & \text{caso contrário} \end{cases} \tag{3.1}$$

$$F_{crisis(i)} = \begin{cases} 1, & \text{se } V_{anomaly(i)} = 1 \\ 1, & \text{se } V_{anomaly(i)} = 0 \quad e \quad F_{crisis(i-1)} = 1 \quad \text{(Até 8 vezes consecutivas)} \\ 0, & \text{caso contrário} \end{cases} \tag{3.2}$$

- ***Flag* de Tendência de Baixa**

Semelhante ao *Flag* de Identificação de Crises, este *Flag* também tem o objetivo de impedir entrada em operações pelo modelo de ML durante sua pre-

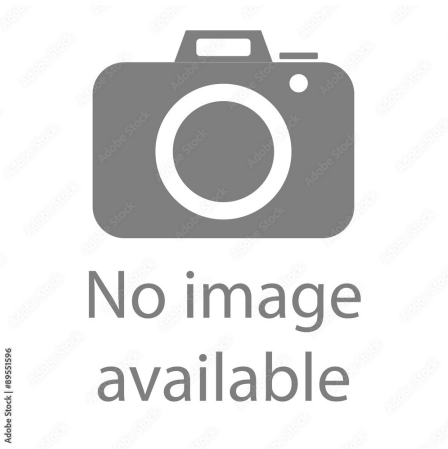

Figura 3.5: *Flag* de Identificação de Crises para a ação XYZ no período de XX/YY/ZZZZ a XX/YY/ZZZZ

sença. No entanto, o critério é diferente. Primeiro, calcula-se a derivada dos preços médios $\dot{P_{mid}}$ entre o dia corrente e o anterior normalizado pela média dos preços médios (Equações 3.3 e 3.4). Em seguida, ajusta-se um filtro digital IIR passa-baixas de coeficiente de amortecimento α (Equação 3.5). Por fim, acrescenta-se um efeito de inércia de 3 dias úteis consecutivos para persistência do *flag* em caso de ocorrência (Equação 3.6). HERALDO: Devo adicionar na fundamentação um tópico falando sobre filtro digital IIR passa-baixas e mostrando de onde vem essa fórmula?

$$P_{mid} = \frac{P_{open} + P_{close}}{2} \tag{3.3}$$

$$\dot{P}_{mid(i)} = \frac{P_{mid(i)} - P_{mid(i-1)}}{\frac{1}{2}(P_{mid(i)} + P_{mid(i-1)})} \tag{3.4}$$

$$\dot{P}_{mid\_LPF(i)} = \alpha \dot{P}_{mid(i)} + (1-\alpha)\dot{P}_{mid\_LPF(i-1)}, \quad \text{onde} \quad 0 \leq \alpha \leq 1 \tag{3.5}$$

$$F_{downtrend(i)} = \begin{cases} 1, & \text{se } \dot{P}_{mid\_LPF(i)} < 0 \\ 1, & \text{se } \dot{P}_{mid\_LPF(i)} \geq 0 \quad e \quad F_{downtrend(i-1)} = 1 \quad \text{(Até 3 vezes consecutivas)} \\ 0, & \text{caso contrário} \end{cases} \tag{3.6}$$

Foi utilizado $\alpha = 0.10$, pois trantando-se de *flag* que pode impedir diretamente o fluxo de negociações, faz-se mais necessário um baixo ruído à inércia da medida. A Figura 3.6 ilustra a eficácia do *flag*.

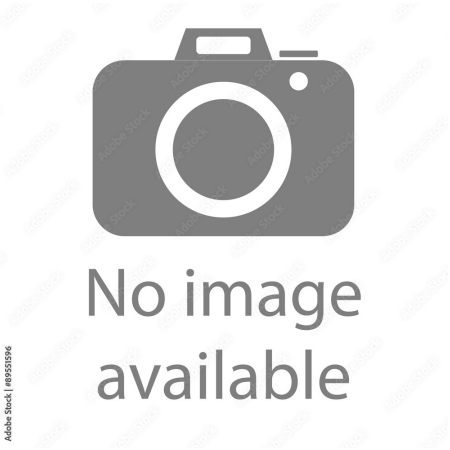

Figura 3.6: *Flag* de Identificação de Crises para a ação XYZ no período de XX/YY/ZZZZ a XX/YY/ZZZZ

- **Risco Mínimo**

O Risco Mínimo serve 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 3.17). Sua fórmula é composta por uma parcela fixa somada a uma parcela variável, conforme mostrado pela Equação 3.7.

$$Risk_{min} = Risk_{min_f} + Risk_{min_v} \tag{3.7}$$

Seja Δ a diferença entre o preço máximo e mínimo de um *candlestick* (Equação 3.8), 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 3.9).

$$\Delta = P_{high} - P_{low} \tag{3.8}$$

$$Risk_{min_f} = \frac{\sigma_\Delta}{P_{mid}} \tag{3.9}$$

A parcela variável $Risk_{min_v}$ representa o inverso da derivada de preços ajustada por um filtro digital IIR passa-baixas (Equações 3.5 e 3.10), onde α é o coeficiente de amortecimento.

$$Risk_{min_v} = -\dot{P}_{mid\_LPF(i)} \tag{3.10}$$

Diferentemente do *flag* de Tendência de Baixa, foi utilizado $\alpha = 0.30$, uma vez que neste caso é muito mais interessante uma resposta rápida do que 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 $Risk_{min}$ a fim de se aumentar a cautela apenas durante momentos mais turbulentos do mercado.

A Figura 3.7 mostra o resultado do algoritmo para dois papéis de comportamentos distintos: MGLU3 representando um companhia com foco em alto crescimento, portanto mais instável; e ABEV3 representando uma companhia já bem consolidada no mercado, com menos oportunidades de crescimento. HERALDO: Precisa de alguma citação aqui para suportar as afirmações feitas?

Figura 3.7: Risco Mínimo para ABEV3 e MGLU3

- **Risco Máximo**

O Risco Máximo serve 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 3.17). A escolha do risco também implica diretamente 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.

A ideia central está na análise estatística das subidas de preços entre os últimos picos identificados dentro do intervalo de 80 dias úteis. Portanto, primeiro se faz necessário a criação de um um algoritmo de identificação de picos, conforme mostrado pela Figura 3.8. O método usa uma janela móvel de $W = 17$ *candles* que corre dia após dia até a data corrente e atribui votos de máximo e votos de mínimo ao maior e menor valor encontrado na janela, respectivamente. São elegíveis à picos apenas os *candles* que obtiveram um mínimo de $floor(W/2) = 8$ votos. Ao final, garante-se a alternância entre máximos e mínimos locais através da remoção de picos consecutivos de um mesmo tipo.

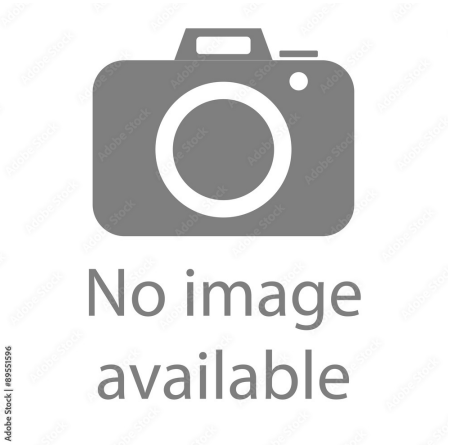

Figura 3.8: Algoritmo de Identificação de Picos

O algoritmo implementado cumpre seu propósito pois se assemelha ao método grafista de identificação de picos [3]. O valor da janela de 17 *candles* foi escolhido devido a teoria do Phi Cube [3]. HERALDO: O estômago até embrulha quando cito essa teoria do Phi Cube, que parece mais uma tentativa deses-

perada de trazer algum critério a um processo caótico. Faz lembrar filme de superherói que peca por tentar justificar demais a origem de um poder com uma base científica. Talvez a teoria do "É 17 porque dá certo" seja tão científica quanto o Phi Cube.

Depois da identificação de picos, extraem-se as n subidas de preços de cada mínimo para o máximo consecutivo, normalizados pelo pico de mínimo (Figura 3.9 e Equação 3.11). Em seguida, calcula-se a média $\overline{C_{LPF(i)}}$ e o desvio padrão $\sigma_{C\_LPF(i)}$ com um filtro digital IIR passa-baixas (Equações 3.12, 3.13 e 3.14). Finalmente, o Risco Máximo $Risk_{max}$ pode ser encontrado segundo a Equação 3.15, onde G é a constante de razão entre ganho e perda.

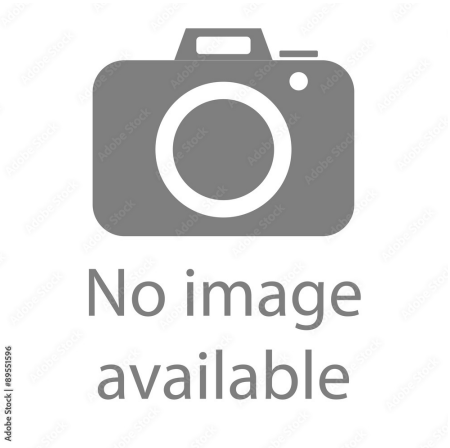

Figura 3.9: Algoritmo de Identificação de Picos

$$c_k = (P_{max(k)} - P_{min(k)})/P_{min(k)}, \quad \text{onde } 0 < k \leq n \tag{3.11}$$

$$\overline{C_{(i)}} = \frac{1}{n}\sum_{k=1}^{n} c_k \tag{3.12}$$

$$\overline{C_{LPF(i)}} = \alpha\overline{C_{(i)}} + (1-\alpha)\overline{C_{LPF(i-1)}} \tag{3.13}$$

$$\sigma_{C\_LPF(i)} = \alpha\sigma_{C(i)} + (1-\alpha)\sigma_{C\_LPF(i-1)} \tag{3.14}$$

$$Risk_{max} = \frac{1}{G}(\overline{C_{LPF(i)}} - 0.5\sigma_{C\_LPF(i)}), \quad \text{onde } G = 3 \tag{3.15}$$

Foi utilizado $\alpha = 0.50$.

Analogamente à Figura 3.7, a Figura 3.10 mostra o Risco Máximo para os ativos MGLU3 e ABEV3.

Figura 3.10: Risco Máximo para ABEV3 e MGLU3

3.3 Simulação de Estratégia

3.3.1 Estrutura

O tema escolhido pelo presente trabalho abrange uma quantidade gigantesca de trajetos possíveis de implementação. Facilmente a imaginação mostra uma possibilidade diferente onde a intuição por detrás da pesquisa vê campo fértil para exploração. 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 serão 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 datas iniciais e finais de validade, sendo estes intervalos de tempo onde as operações podem ser realizadas. Embora se permitam 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 aquele 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** predetermina a relação entre o preço alvo de venda 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 3.16). Seu valor é constante e igual a 3. HERALDO: Devo citar o André aqui? Não achei uma referência disso no livro dele, porém tem nos vídeos.

$$G = \frac{P_{target} - P_{buy}}{P_{buy} - P_{stop}} = 3 \tag{3.16}$$

  Utiliza-se o termo "risco de uma operação"como sendo a diferença de valor no qual o *stop loss* é colocado abaixo do preço de compra (Equação 3.17).

$$Risk = \frac{P_{buy} - P_{stop}}{P_{buy}} \tag{3.17}$$

- 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 garantindo 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 deve ocorrer está na abertura de mercado do dia corrente, mais precisamente no instante em que o preço de abertura é definido.
- No dia que ocorrer 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 2 dias úteis**.
- 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**:
  - Venda por *stop loss*
  - Venda parcial (caso habilitada)
  - Venda por preço alvo
  - Venda por *timeout*
- Caso um **preço de venda seja pulado**, ou seja, a descontinuidade entre o preço de fechamento do dia anterior e o preço de abertura do dia corrente não englobe o valor de venda, utiliza-se o preço de abertura de mercado. 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 péssima decisão 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 trazer seu lucro ou prejuízo para a carteira, o capital alocado nele não pode ser utilizado por outros ativos que eventualmente vem a lucrar. Quanto mais lucrativa tende a ser uma estratégia, maior a inércia que esta decisão pode oferecer. Portanto, optou-se por um limite finito de dias por operação.

A escolha de um número adequado para o limite de dias possui alguns caminhos, resumidos entre os extremos de: um valor fixo geral; ou um valor dinâmico para cada ação. Nota-se que criar um algoritmo que escolha o valor dinâmico a partir da análise dos dados dos últimos meses para cada ação não é trivial. Assim, optou-se por um valor fixo geral, onde a base do estudo se deu na análise ações de perfis opostos.

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 levaram a operações de sucesso e quais levaram 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 grande, 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 de sucesso, caso exista este valor. Se não houver, é considerado operação de falha, portanto está fora dos histogramas. A legenda indica faixas onde, no caso la linha tracejada verde na Figura 3.11, 50% das contagens se encontram dentro dos 12 primeiros dias, e assim por diante. Foi considerado o período de 01/01/2016 a 31/12/2018.

Também foi analisado os histogramas para o caso de risco ótimo por operação, ou seja, o valor de risco que traz o maior rendimento por operação considerando os dias corridos (Figuras 3.13 e 3.14).

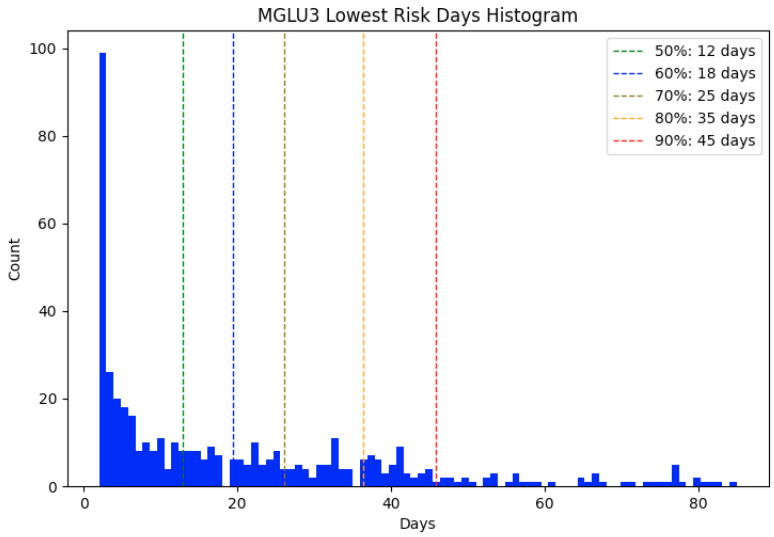

Figura 3.11: MGLU3 - Histograma de Dias com Risco Mínimo em Operações de Sucesso

A Tabela 3.1 resume o período de dias que engloba 90% das contagens dos histogramas.

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

Tabela 3.1: Período de dias que engloba 90% das contagens dos histogramas

Com base nos valores encontrados, arbitrou-se o período máximo fixo de **45 dias** para qualquer operação. HERALDO: É muita ousadia afirmar que escolhi arbitrariamente o valor de 45 dias com base na tabela apresentada?

3.3.4 Gerenciamento de Risco

Segundo MORAES [3], o Gerenciamento de Risco é imprescindível para um bom rendimento 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 falhadas. Além disso, estar com o capital muito alocado em ativos de um único segmento é perigoso devido à alta exposição à fatores prejudicias como falta de

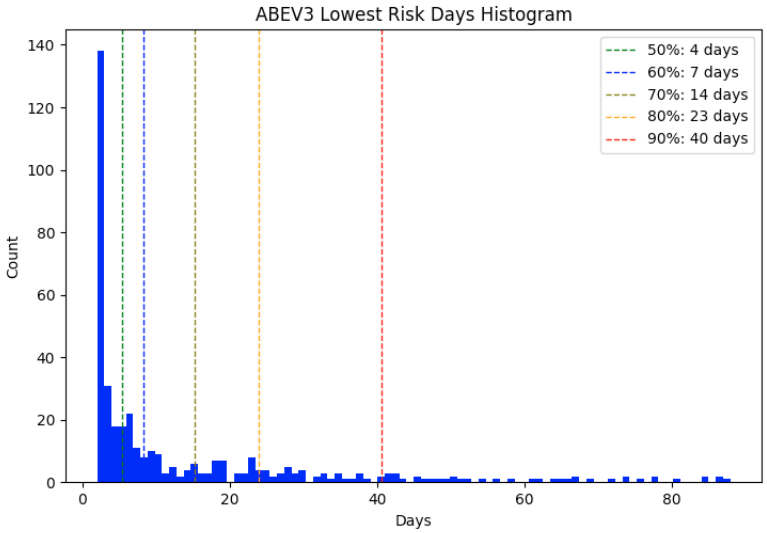

Figura 3.12: ABEV3 - Histograma de Dias com Risco Mínimo em Operações de Sucesso

insumos industriais, mudanças na legislação, crises internas, instabilidade política, dentre outros.

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

- Diversificação de ativos em segmentos de mercado variados através da escolha de um alto número de *tickers*, mais especificamente 71.
- Criação do Coeficiente de Risco-Capital[4]

O Coeficiente de Risco-Capital, definido pela Equação 3.18, 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.

$$RCC = Capital \times Risk \tag{3.18}$$

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

[4]ou *Risk-Capital Coefficient* (RCC)

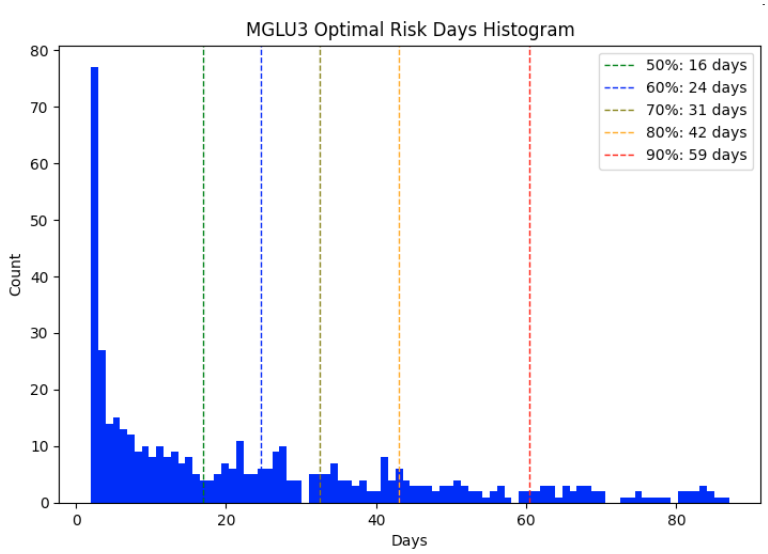

Figura 3.13: MGLU3 - Histograma de Dias com Risco Ótimo em Operações de Sucesso

Equação 3.19 mostra de fato a aplicação do RCC. É evidente que quanto maior o risco envolvido, menor o capital a ser alocado e vice-versa.

$$Capital = \frac{RCC}{Risk} \tag{3.19}$$

A necessidade de um melhor uso médio de capital ao longo do período de simulação inspirou a criação de um RCC Dinâmico, que é tratado em detalhes na Seção 3.4.4 por se tratar de uma otimização.

Por fim, a Tabela 3.2 lista todos os 71 ativos escolhidos para simulação. 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. HERALDO: Será que não tem outro lugar melhor para colocar essa tabela? Coloquei aqui pois foi o primeiro momento no qual foi relevante mencionar as 71 ações escolhidas, portanto aproveitei o gancho.

3.3.5 Risco de Entrada por Operação

O Risco de Entrada por Operação é encontrado a partir da média aritmética entre as *features* Risco Mínimo e Risco Máximo (ver Seção 3.2.4). A Equação 3.20 mostra

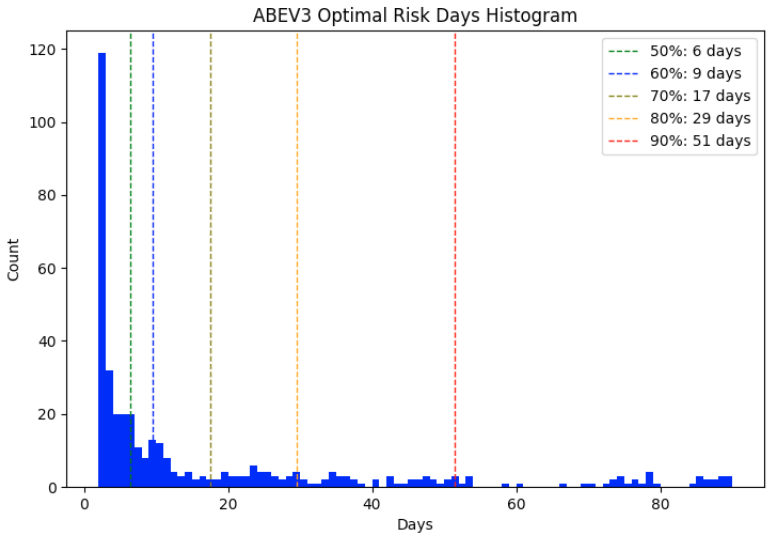

Figura 3.14: ABEV3 - Histograma de Dias com Risco Ótimo em Operações de Sucesso

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	PRIO3	QUAL3
RADL3	RENT3	SANB11	SBSP3	SULA11	TAEE11	TIMS3	TOTS3
UGPA3	USIM5	VALE3	VIIA3	VIVT3	WEGE3	YDUQ3	

Tabela 3.2: Ações Escolhidas

essa relação:

$$Risk_{operation} = \frac{1}{2}(Risk_{max} + Risk_{min}), \quad \text{se } Risk_{max} \geq Risk_{min} \tag{3.20}$$

Pelo fato da metodologia de cálculo do Risco Máximo e do Risco Mínimo seguirem raciocínios diferentes, podem ocorrer momentos nos quais a condição expressa na Equação 3.20 não seja verdadeira, em outras palavras, as ondas de subida de preço

entre picos no gráfico diário não compensam o risco inerente ao ruído diário dos *candlesticks*. Enquanto este evento ocorrer, não haverá entrada em operações para o ativo envolvido. Por outro lado, é comum uma interseção entre os períodos de Descanso por Tendência de Baixa ou até mesmo de Descanso por Identificação de Crises (Seções 3.3.6 e 3.3.7, respectivamente). HERALDO: Algo me diz que um gráfico aqui confirmando essa interseção não seria uma má ideia, certo?

3.3.6 Descanso por Tendência de Baixa

O Descanso por Tendência de Baixa é um intervalo que impede qualquer nova operação durante a ativação do *Flag* de Tendência de Baixa (ver Seção 3.2.4). O objetivo é esperar o mercado entrar em uma nova tendência de alta ou pelo menos se estabilizar para que uma nova operação se justifique, mesmo que esta decisão implique em uma pequena inércia. Operações em vigor não são canceladas. HERALDO: Optei por não colocar uma imagem mostrando a eficácia desse flag porque em princípio a Seção de Features (3.2.4) já faz isso.

3.3.7 Descanso por Identificação de Crises

O Descanso por Identificação de Crises é um intervalo que impede qualquer nova operação durante a ativação do *Flag* de Identificação de Crises (ver Seção 3.2.4). O objetivo é esperar o mercado se estabilizar de uma crise em potencial para que uma nova operação se justifique, mesmo que esta decisão implique em uma pequena inércia. Operações em vigor não são canceladas. HERALDO: Optei por não colocar uma imagem mostrando a eficácia desse flag porque em princípio a Seção de Features (3.2.4) já faz isso.

3.3.8 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 parciluar (ver Seção 3.2.1 para mais esclarecimentos).

Lista de Parâmetros		
Nome do Parâmetro	Escopo	Descrição
show_results	Geral	Exibe *dashboard* da última simulação completada ao final. Tipo: *Boolean*. *Default*: *True*. Listável: Não.
min_risk_features	Geral	Risco mínimo para o cálculo de *features*. Tipo: *Float*. *Default*: 0,01. Listável: Não.
max_risk_features	Geral	Risco máximo para o cálculo de *features*. Tipo: *Float*. *Default*: 0,10. Listável: Não.
name	Local	**(OBRIGATÓRIO)** Nome da estratégia a ser executada. Valores válidos: "ML Derivation". Tipo: *String*. Listável: Não.
comment	Local	Comentário. Tipo: *String*. *Default*: *String* vazia. Listável: Não.
capital	Local	Capital total da carteira em reais (R$). Tipo: *Float*. Listável: Sim.
risk_capital_coefficient	Local	Coeficiente de risco-capital (RCC) geral. Tipo: *Float*. *Default*: 0,001. Listável: Sim.
tickers_bag	Local	Grupo de ativos a escolher dentro de "stock_targets". Valores aceitos: "listed_first"(ordem de listagem); "random"(ordem aleatória). *Default*: "listed_first". Listável: Sim.
tickers_number	Local	Número de ativos a escolher dentro de "stock_targets", de acordo com "tickers_bag". Tipo: *Int*. *Default*: 0 (todos). Listável: Sim.
min_order_volume	Local	Volume mínimo por operação. Tipo: *Int*. *Default*: 1. Listável: Sim.
gain_loss_ratio	Local	Razão entre ganho e perda. Para uma unidade de risco (delta pencentual entre preço de compra e *stop loss*) são utilizadas N unidades de risco acima no preço preço de compra para definir o preço alvo. Tipo: *Float*. *Default*: 3. Listável: Sim.

Continuação da Tabela 3.3		
Nome do Parâmetro	Escopo	Descrição
max_days_per_operation	Local	Número máximo de dias por operação. Inclui o dia de compra. Caso excedido, ocorre venda compulsória pelo preço de fechamento no último dia da contagem. Tipo: *Int.* *Default*: 45. Listável: Não.
min_risk	Local	Risco mínimo por operação. Tipo: *Float.* *Default*: 0,003. Listável: Sim.
max_risk	Local	Risco máximo por operação. Tipo: *Float.* *Default*: 0,10. Listável: Sim.
max_risk	Local	Risco máximo por operação. Tipo: *Float.* *Default*: 0,10. Listável: Sim.
enable_frequency _normalization	Local	Uso de normalização por frequência de operações. Ativos com N vezes mais operações que a média receberão N vezes menos capital. Ver Seção 3.4.2. Tipo: *Boolean.* *Default*: *False.* Listável: Sim.
enable_profit _compensation	Local	Uso de compensação por lucratividade acumulada. Ver Seção 3.4.3. Tipo: *Boolean.* *Default*: *False.* Listável: Sim.
enable_crisis_halt	Local	Bloqueio de novas aquisições em caso de identificação de potenciais crises financeiras (para ativo). Ver Seção 3.3.7. Tipo: *Boolean.* *Default*: *False.* Listável: Sim.
enable_downtrend_halt	Local	Bloqueio de novas aquisições em caso de identificação de tendências de baixo nos preços (para ativo). Ver Seção 3.3.6. Tipo: *Boolean.* *Default*: *False.* Listável: Sim.
enable_dynamic_rcc	Local	Uso de Coeficiente de Risco-Capital dinâmico (para carteira). Ver Seção 3.4.4. Tipo: *Boolean.* *Default*: *False.* Listável: Sim.

Continuação da Tabela 3.3		
Nome do Parâmetro	Escopo	Descrição
dynamic_rcc_reference	Local	Valor de referência de uso de capital médio no controle do RCC dinâmico. Ver Seção 3.4.4. Tipo: *Float*. *Default*: 0,80. Listável: Sim.
dynamic_rcc_k	Local	Valor do ganho proporcional K no controle do RCC dinâmico. Ver Seção 3.4.4. Tipo: *Float*. *Default*: 3. Listável: Sim.
purchase_margin	Local	Margem percentual aplicada ao valor de compra. Ex: Se o alvo de compra estiver configurado para R$100, uma margem de 1% permitirá a compra antecipada em R$99. Tipo: *Float*. *Default*: 0. Listável: Sim.
stop_margin	Local	Margem percentual aplicada ao valor do *stop loss*. Ex: Se o *stop* estiver configurado para R$100, uma margem de 1% permitirá a compra antecipada em R$101. Tipo: *Float*. *Default*: 0. Listável: Sim.
partial_sale	Local	Uso de saídas parciais. Tipo: *Boolean*. *Default*: *False*. Listável: Sim.
stop_type	Local	Tipo de *stop loss* utilizado. Valores aceitos: "normal"; "staircase"(para cada patamar de unidade de risco que o preço atinge acima do valor de compra, o *stop* sobe igualmente, até uma unidade de risco abaixo do preço alvo). Ver "gain_loss_ratio". *Default*: "normal". Listável: Sim.
min_days_after_successful_operation	Local	Mínimo de dias sem novas aquisições após operação de sucesso, para cada ação. Ex: para 1 dia mínimo, se a última venda de sucesso ocorreu durante o dia X, a próxima compra só ocorrerá a partir do dia X+2, inclusive. Tipo: *Int*. *Default*: 0. Listável: Sim.

Continuação da Tabela 3.3		
Nome do Parâmetro	Escopo	Descrição
min_days_after_failure _operation	Local	Mínimo de dias sem novas aquisições após operação de falha, para cada ação. Ex: para 1 dia mínimo, se a última venda de falha ocorreu durante o dia X, a próxima compra só ocorrerá a partir do dia X+2, inclusive. Tipo: *Int.* *Default*: 0. Listável: Sim.
stock_targets	Local	**(OBRIGATÓRIO)** *Array* de ações a incluir na carteira. Formato indicado pela Figura 3.2. Atenção ao parâmetro "tickers_bag".
Fim da Tabela 3.3		

Tabela 3.3: Lista de parâmetros detalhados

3.3.9 Ensaios Paralelos

HERALDO: Vou fazer apenas uma introdução para te mostrar o caminho que seguirei caso você julgue que valha a pena mencionar essa seção. Se não for necessário, removo as referências dos parâmentros mencionados ao longo desta monografia.

Alguns parâmetros não se mostraram eficazes em melhorar a performance das simulações, muito embora tenham se apresentado como alternativas plausíveis na resolução de problemas ao longo deste trabalho. São eles os parâmetros:

- Venda Parcial (partial_sale) HERALDO: A motivação da criação desse parâmetro vem de uma prática do André como uma forma de auxiliar o fator psicológico do trader para os casos frustantes em que o mercado quase chega no preço de venda, mas depois cai e bate no stop. Ele não usa isso constantemente, mas deixa como uma opção. Venda parcial em uma operação é vender 50% do volume de compra adquirido quando o preço bater uma unidade de risco (1un Risco = Pcompra - StopLoss) acima do preço de compra. Dessa forma,

ao invés da operação ter um ganho de 3X para cada X de possível perda, ela teria um ganho de 2X para os mesmos X, mais os casos de saída sem prejuízo. A grande questão é que essa conta não compensa, seja quando eu estava tentando replicar a estratégia do André, seja com os modelos de ML agora, nunca trouxe um rendimento maior.

- Saídas Parcias (stop_type = "staircase") HERALDO: É uma extensão que criei derivada da Venda Parcial para tentar granularizar um pouco mais o critério anterior e verificar se o problema da ineficácia estava na escolha do limiar de venda. Igualmente não traz melhora alguma. A ideia é: cada vez que o preço de marcado toca o limiar de uma unidade de risco (1un Risco = Pcompra - StopLoss) acima do preço de compra, o stop loss sobre igualmente 1 un de risco. Exemplo: se o preço de compra é Pcom e o Stop Loss é Pcom - X, quando o preço de mercado atingir Pcom + X, o Stop Loss sobe para Pcom. Quando subir de novo para Pcom + 2X, o Stop sobe para Pcom + X. O alvo continua sendo Pcom + 3X como sempre.

- Dias Mínimos de Espera após Operação de Falha (min_days_after_failure_operation)

- Dias Mínimos de Espera após Operação de Sucesso (min_days_after_successful_operation) HERALDO: A motivação de ambos os parâmetros aqui era diminuir o ruído de entrada e saída em operações sequenciais que o modelo de ML arrisca, mas toda hora a operação é stopada. São regiões do gráfico que apresentam 2, 3, 4 operações de falha curtas e seguidas, eventualmente com uma de sucesso e já volta pra falha. Na minha interpretação, os motivos da existência dessas regiões tem a ver com a qualidade do modelo, porém mais especificamente com: (1) a semelhança forte do momento presente com um passado de treinamento onde houve muito sucesso, muito embora possa ser apenas coincidência; e (2) a escolha de um valor de risco de entrada muito baixo para a volatilidade atual do mercado, porém sufiente para o modelo arriscar a operação (pode se somar aqui o efeito indicado em (1)). Moral da história: algumas vezes funcionou (com modelos anteriores que já discartei), porém mesmo QUANDO funciona, é difícil ter clareza do número ótimo de dias para ambos os parâmetros. Digo isso pois acontece do valor de dias escolhido estar em uma região da curva

instável onde a melhora do rendimento geral foi coincidência. A prova disso se dá quando você escolhe valores imediatamente próximos e verifica novos resultados discrepantes (simulando sempre com os 71 tickers em uma carteira de 100000 reais). Enfim, eles estão sempre aqui como opções, mas são são consistentes.

3.4 Otimizações de Gerenciamento de Carteira

3.4.1 Resumo

As otimizações apresentadas nesta Seção independem de um modelo de ML específico, sendo portanto algoritmos gerais que, motivados ou não por problemas oriundos dos modelos, buscam uma abordagem geral para aumento da performance da carteira.

3.4.2 Normalização por Frequência de Operações

Cada ativo de uma carteira possui um critério próprio de análise das condições de mercado que o auxilia na decisão de entrada nas operações. Muitas vezes, ativos diferentes acumulam um número bastante variado de operações concluídas ao longo da simulação. Por outro lado, esse número não posui ligação direta com a performance individual dos mesmos. Em outras palavras, facilmente ocorre a situação de um *ticker* monopolizar grande parte do capital total da carteira ao longo do tempo, simplesmente por ter uma frequência de operações maior que os outros, sem qualquer fator meritocrático que embase uma justificativa.

A fim de se endereçar essa questão, foi criado o critério de Normalização por Frequência de Operações, onde cada ativo receberá de capital para uma determinada operação um valor inversamente proporcional a frequência de operações acumulada até o momento.

$$Capital_{norm} = Capital \times \frac{\overline{f_{total}}}{f_{stock}} \tag{3.21}$$

$$\overline{f_{total}} = \frac{N_{total\_operations}}{N_{total\_stocks}} \tag{3.22}$$

As Equações 3.21 e 3.22 mostram a obtenção do novo Capital Normalizado $Capital_{norm}$ a partir do capital que em princípio seria alocado (ver Equação 3.19), a frequência média de operações totais acumulada $\overline{f_{total}}$ e a frequência média de operações do ativo envolvido f_{stock}, igualmente acumulada.

Para a Normalização começar a ser aplicada a um ativo, é necessário que haja pelo menos uma operação concluída do mesmo, assim como um mínimo de duas vezes o número total de ativos na carteira em operações concluídas. A Equação 3.23 mostra as condições citadas.

$$f_{stock} > 0, \quad N_{total\_operations} \geq 2 \times N_{total\_stocks} \tag{3.23}$$

As Figuras 3.15 e 3.16 mostram eficácia do critério criado para a simulação dos 71 *tickers* listados na Tabela 3.2, durante o período de 01/01/2019 a 31/12/2021, onde a Figura 3.15 desabilita a Normalização e a Figura 3.16 habilita.

Figura 3.15: Simulação sem uso da Normalização por Frequência de Operações

A Tabela 3.4 traz um comparativo dos resultados das simulações. Observa-se que o critério de Normalização criado se auto-compensa, ou seja, realoca capital dentro da própria estratégia sem alterar o Uso Médio de Capital da carteira. A vantagem de um critério auto-compensado é que ele não traz uma potencial ilusão de melhora

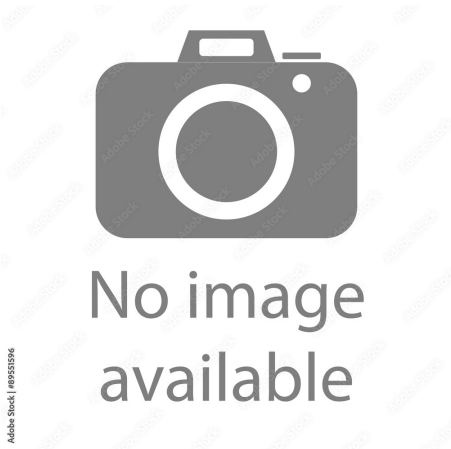

Figura 3.16: Simulação com uso da Normalização por Frequência de Operações

de performance, já que a comparação de resultados entre estratégias precisa ter em vista Usos Médios de Capital razoavelmente próximos entre si para que a escolha dos RCCs individuais não influencie a análise.

	hue	hue
hue	hue	hue
hue	hue	hue

Tabela 3.4: Comparação de Resultados

(Mencionar mais comentários sobre os outros parâmetros quando tiver as imagens e a tabela).

3.4.3 Compensação por Lucratividade

A Compensacão por Lucratividade é um ajuste auto-compensado[5] que aumenta o capital em operações de *tickers* que possuem um lucro acumulado acima da média da carteira. Da mesma forma, também diminui o capital daqueles que estão com o lucro acumulado abaixo da média da carteira.

A Equação 3.24 mostra a primeira etapa do cálculo da Compensação, onde σ_{eq} é o valor em unidades de desvio padrão do quanto o lucro acumulado p_t do ativo está

[5]Realoca capital dentro da própria estratégia sem alterar o Uso Médio de Capital da carteira

em relação à média $\overline{P_w}$ e desvio padrão σ_w da carteira.

$$\sigma_{eq} = \frac{p_t - \overline{P_w}}{\sigma_w} \tag{3.24}$$

Em seguida, a Equações 3.25 e 3.26 dão sequência ao cálculo criando as constantes que serão utilizadas diretamente na Compensação. Nota-se que C_{max} é a variação máxima positiva que a Compensação pode alcançar. σ_s e σ_e são limiares de σ_{eq} que definem lugares geométricos diferentes.

$$m_1 = \frac{C_{max}}{\sigma_e - \sigma_s}, \quad n_1 = 1 - \sigma_s m_1 \tag{3.25}$$

$$m_2 = \frac{C_{max}}{\sigma_e - \sigma_s}, \quad n_2 = 1 + \sigma_s m_2 \tag{3.26}$$

Finalmente, a Equação 3.27 mostra o cálculo final, que pode ser facilmente visualizado pela Figura 3.17.

$$C = \begin{cases} m_1\sigma_{eq} + n_1, & \text{se } \sigma_{eq} \geq \sigma_s \quad \text{e} \quad \sigma_{eq} \leq \sigma_e \\ m_2\sigma_{eq} + n_2, & \text{se } \sigma_{eq} \leq -\sigma_s \quad \text{e} \quad \sigma_{eq} \geq -\sigma_e \\ 1 + C_{max}, & \text{se } \sigma_{eq} > \sigma_e \\ 1 - C_{max}, & \text{se } \sigma_{eq} < -\sigma_e \\ 0, & \text{se } |\sigma eq| < \sigma_s \end{cases} \tag{3.27}$$

Utilizou-se $C_{max} = 0.60$, $\sigma_s = 0.2$ e $\sigma_e = 2.0$.

A Figura 3.18 mostra o ganho de performance obtido para a simulação dos 71 *tickers* listados na Tabela 3.2, durante o período de 01/01/2019 a 31/12/2021. Ressalta-se que nenhuma outra Otimização de Gerenciamento de Carteira foi utilizada a fim de se isolar o efeito desta Compensação.

A Figura 3.19 também mostra o ganho de performance nas mesmas condições da Figura 3.18, com a exceção do uso comum da Normalização por Frequência de Operações, o que mostra que ambas as otimizações de carteira não se exluem, muito pelo contrário.

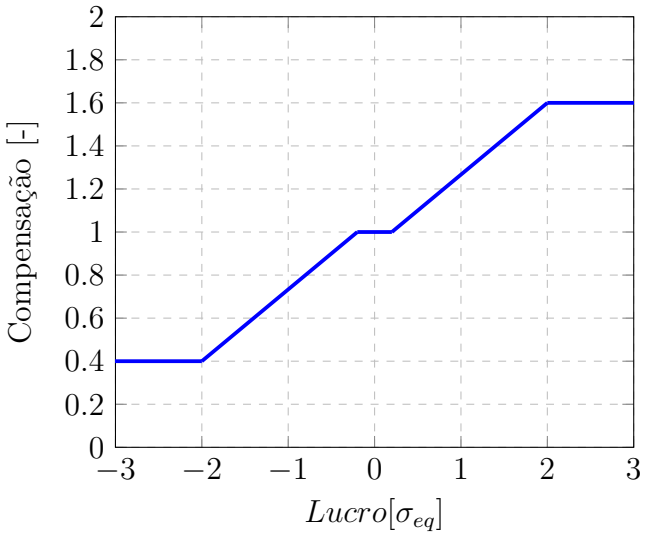

Figura 3.17: Gráfico da Função de Compensação por Lucratividade

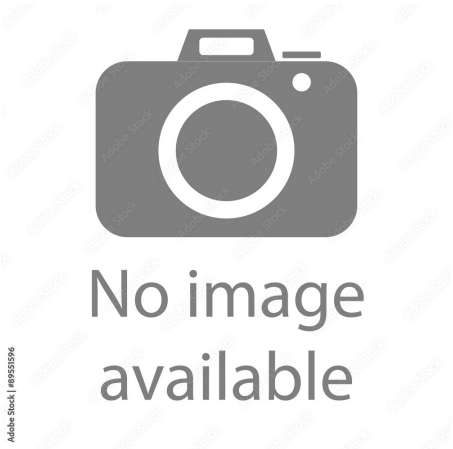

Figura 3.18: Ganho de performance pelo uso da Compensação por Lucratividade

3.4.4 Controle Proporcional para Uso de Capital (Risco Dinâmico)

A criação de um RCC fixo pode ser interessante do ponto de vista de Gerenciamento de Risco (ver Seção 3.3.4), 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 suficientemente diferente a ponto de requerer um novo RCC, dificultando um bom ajuste. Em outras palavras, um RCC fixo leva a problemas de subaproveitamento do Uso de Capital da carteira.

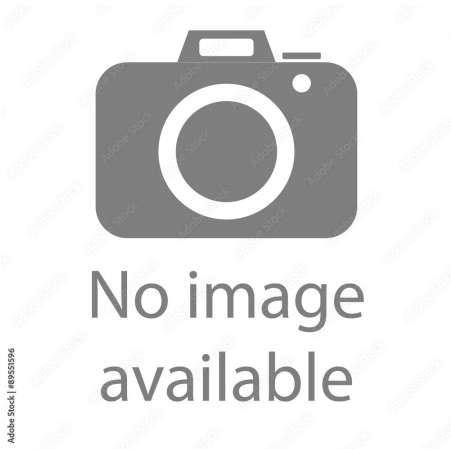

Figura 3.19: Ganho de performance pelo uso da Compensação por Lucratividade (com NFO)

A solução desse problema se dá pela 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. Tudo isso aliado a um rebalanceamento dinâmico de capital em função do uso médio de capital vigente, ou seja, períodos com menores oportunidades de operações terão mais disponibilidade de capital e vice-versa.

As Equações 3.28 e 3.29 mostram o cálculo do RCC dinâmico (RCC_{din}) a partir do RCC fixo (RCC_{fix}, definido pela Equação 3.18), do valor de referência para o uso médio de capital (C_{ref}), da constante de ganho proporcional (K) e do uso médio de capital dos últimos 10 dias de simulação ($\overline{C_{10d}}$).

$$e = C_{ref} - \overline{C_{10d}}, \quad \text{para } 0 \leq C_{ref}, \overline{C_{10d}} \leq 1 \tag{3.28}$$

$$RCC_{din} = RCC_{fix}(1 + Ke) \tag{3.29}$$

Foi utilizado $C_{ref} = 1.0$, $K = 5$ e $RCC_{fix} = 0.003$

A Figura 3.20 mostra o ganho de performance obtido pela implementação do RCC Dinâmico com os valores mencionados.

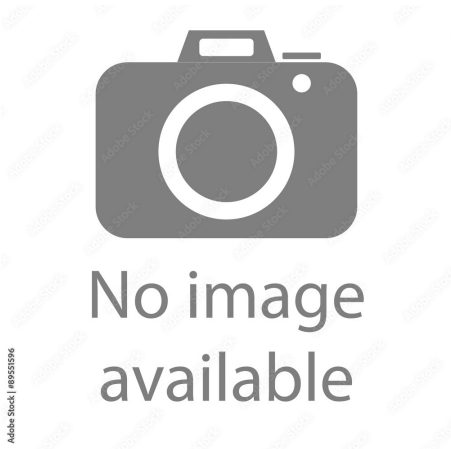

Figura 3.20: Ganho de performance pelo uso do RCC Dinâmico

(Falar que essa medida aumenta o uso de capital bem alocado, mas também o mal alocado, por isso tem um limite)

(Mostrar que RCC dinâmico não tem problema com otimizações anteriores)

3.5 Criação de Modelos

3.5.1 Resumo

A partir de *datasets* previamente populados, modelos do tipo *Random Forest* são gerados para cada ação e para cada intervalo de 3 meses de simulação. Um critério particular de performance foi criado para auxiliar na escolha do melhor modelo, que é filtrado tanto por uma varredura de parâmetros variados quanto por uma análise estatística das sementes aleatórias utilizadas.

3.5.2 Geração de *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 3.5 mostra a lista de colunas presentes no arquivo, onde as linhas marcadas em negrito indicam o uso efetivo da coluna durante a criação dos modelos. A coluna Resultado da Operação não é utilizada na entrada dos dados, mas sim indica a saída observada para o treinamento supervisionado.

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

Tabela 3.5: Comparação de Resultados

O termo preço médio se refere ao definido pela Equação 3.3. Da mesma forma, a derivada do preço médio é indicada pela Equação 3.4. 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 númerico 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 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. Por isso foram utilizadas 56 opções de risco: de 1% a 12% em passos de 0,2%.

3.5.3 Índice de Lucratividade

3.5.4 Critérios de Escolha

Os modelos *Random Forest* foram criados a partir da biblioteca *Scikit-Learn* [71] com a configuração indicada pela Tabela 3.6.

Parâmetro	Valor
n_estimators	200
criterion	gini
min_samples_split	12
min_samples_leaf	6
min_weight_fraction_leaf	0.0
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	True
oob_score	False
warm_start	False
class_weight	balanced_subsample
ccp_alpha	0.0
max_samples	None

Tabela 3.6: Comparação de Resultados

3.6 Análise de Resultados

3.6.1 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 performace razoável de se obter na realidade. Neste caso, utilizou-se a média da performance das ações da carteira, ou seja, supondo-se que o capital inicial fosse igualmente distribuido em cada ação disponível, o rendimento médio destas ações ao longo do tempo é o *baseline*.

A Figura 3.21 mostra o *baseline* calculado para os 71 *tickers* indicados na Tabela 3.2 no intervalo de 01/01/2019 a 31/12/2021. Adicionou-se o iBovespa (ver Seção 2.1.2) e o CDI[6] acumulado para fins de comparação. Nota-se que a performance do *baseline* é significativamente maior que o iBovespa, o que é razoável já que o iBovespa é uma composição variável tanto na escolha dos ativos quanto em seus 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.

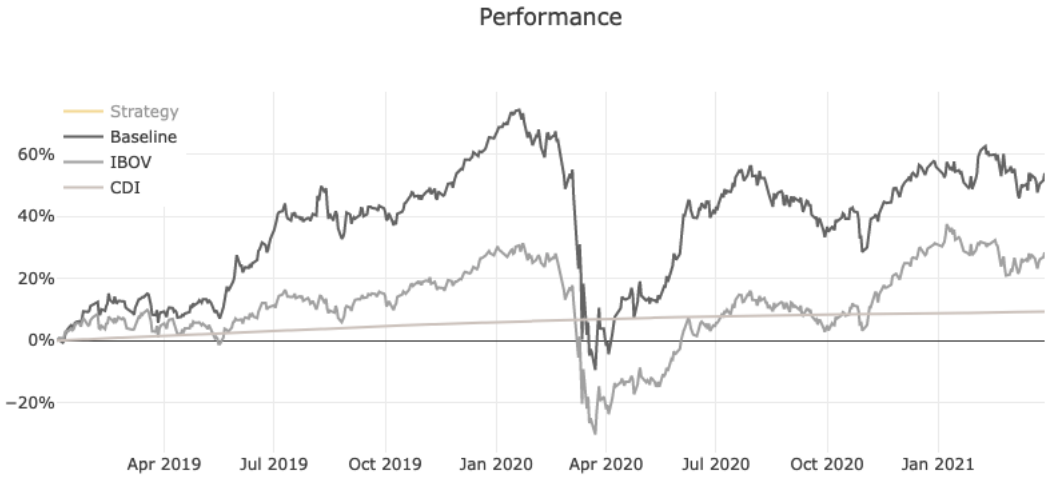

Figura 3.21: *Baseline* para o intervalo de 01/01/2019 a 31/12/2021 (CORRIGIR)

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

3.6.2 *Dashboard*

Um *Dashboard* interativo é gerado por aplicação secundária a fim de auxiliar a análise dos resultados obtidos em cada simulação. O *framework Dash* [72] foi utilizado para criar uma interface web resumindo todas as informações pertinentes a uma simulação executada. As Figuras 3.22, 3.23, 3.24, 3.25 e 3.26 mostram em partes as seções de uma simulação genérica.

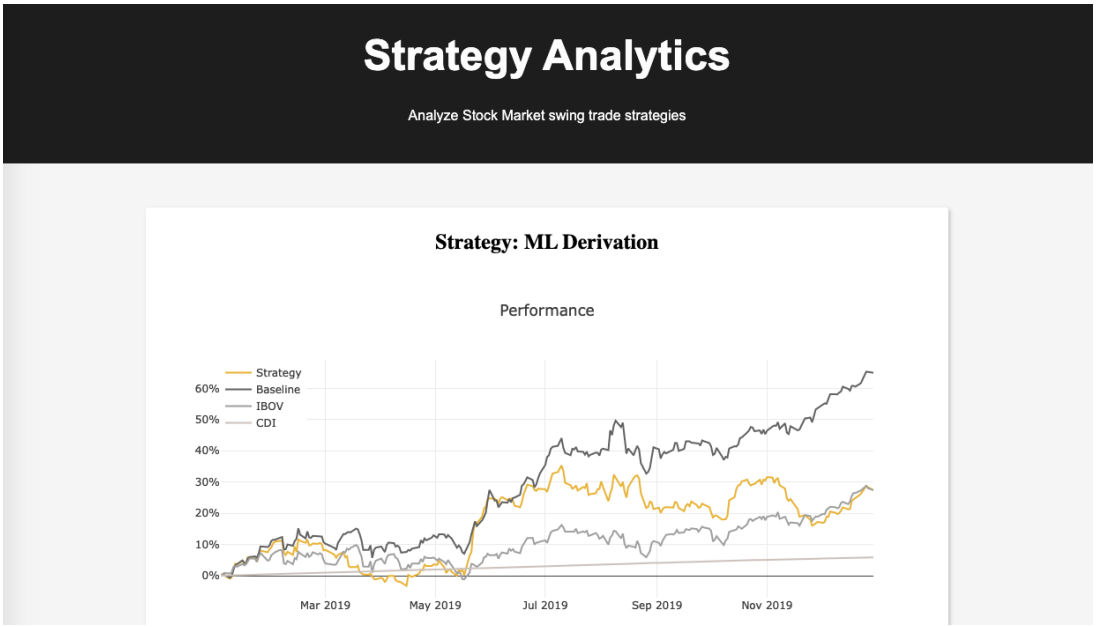

Figura 3.22: Dashboard - Performance

Parameters

Alias	2019-1 to 2019-4. 20 tickers. BAG_2.
Total Tickers	20
Start Date	01/01/2019
End Date	30/12/2019
Capital (R$)	100000
Risk-Capital Coefficient - RCC (%)	0.3
Gain-Loss Ratio	3
Minimum Order Volume	1
Minimum Operation Risk (%)	0.3
Maximum Operation Risk (%)	10
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	True
Enable Profit Compensation	True
Enable Crisis Halt	True
Enable Downtrend Halt	True
Enable Dynamic RCC	True
Dynamic RCC Reference (%)	100
Dynamic RCC K	5

Figura 3.23: Dashboard - Parâmetros de entrada

Results and Statistics

Strategy Total Yield (%)	**29.72**
Baseline Total Yield (%)	64.98
IBOVESPA Total Yield (%)	27.42
CDI Total Yield (%)	5.92
Strategy Total Volatility (%)	**21.95**
Baseline Total Volatility (%)	20.03
Strategy Sharpe Ratio (-)	**1.07**
Baseline Sharpe Ratio (-)	3
Strategy Sortino Ratio (-)	**1.72**
Baseline Sortino Ratio (-)	4.24
Strategy-Baseline Spearman Correlation (-)	**0.74**
Strategy-IBOV Spearman Correlation (-)	0.72
Maximum Used Capital (%)	100.00
Average Used Capital (%)	**80.22**
Maximum Active Operations	13
Average Active Operations	7.36
Active Operations Standard Deviation	2.58
Profit (R$)	29720.37
Total Operations	356
---Successful Operations (hit 3:1 target)	**109 (30.6%)**
---Partial Sale Successfull Operations (hit 1:1 or 2:1 target)	0 (0.0%)
---Failed Operations	**239 (67.1%)**
---Timed Out Operations	0 (0.0%)
---Unfinished Operations	8 (2.2%)
Strategy Yield (% ann)	**30.27**
Baseline Yield (% ann)	66.32
IBOVESPA Yield (% ann)	27.92
CDI Yield (% ann)	5.95
Strategy Volatility (%ann)	**22.13**
Baseline Volatility (%ann)	20.19

Figura 3.24: Dashboard - Resultados e Estatísticas

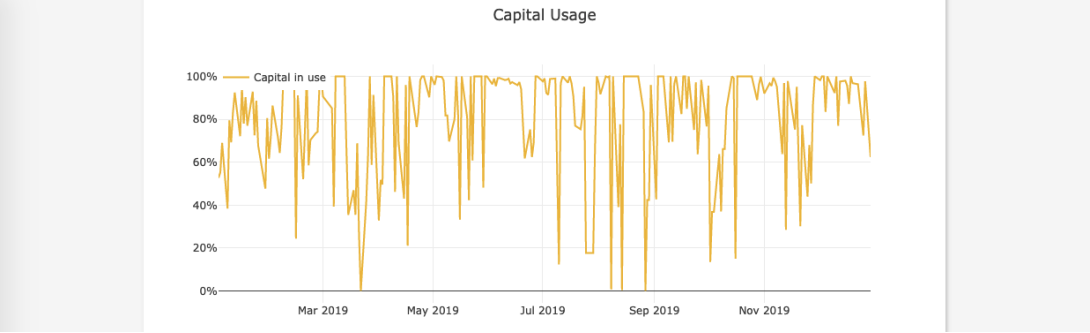

Figura 3.25: Dashboard - Gráfico de uso de capital

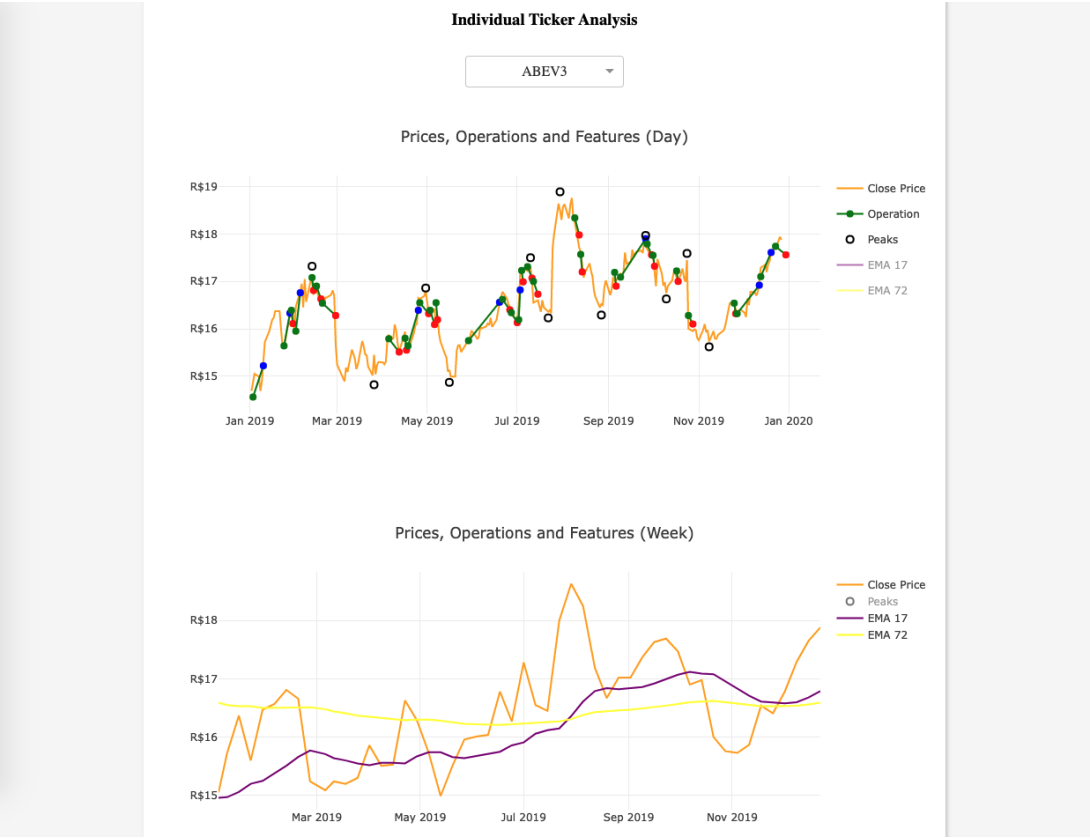

Figura 3.26: Dashboard - Gráficos de análise individual de ações

Capítulo 4

Conclusão

Foram realizadas diversas simulações para os 71 *Tickers* apresentados na Tabela 3.2, durante o período de simulação de 01/01/2019 e 31/12/2021, que inclusive engloba a Crise do Coronavírus ocorrida em março de 2020. O melhor resultado obtido levou em consideração os parâmetros indicados pela Tabela 4.1:

As Figuras 4.1 e 4.2 mostram a performance da carteira em comparação ao Baseline, iBovespa e CDI acumulado no período, assim como os valores métricas resultantes dessa performance.

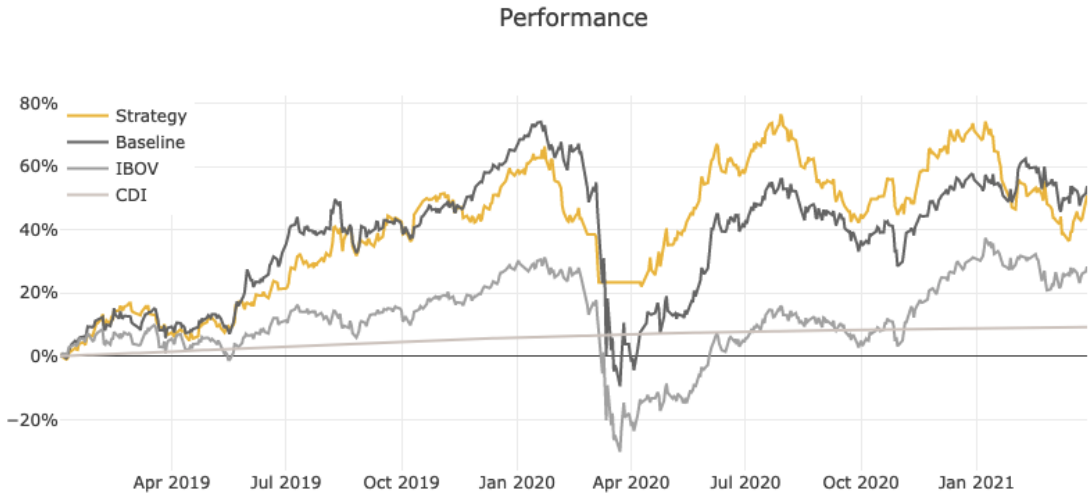

Figura 4.1: Performance da Carteira

Analisando a Figura 4.1, pode-se perceber que (FAZER COMENTÁRIOS). Quanto aos resultados mostrados na Figura 4.2, (FAZER COMENTÁRIOS).

Parâmetro	Valor
Capital	100000
RCC	0.3%
Gain Loss Ratio	3
Menor Risco por Operação	0.3%
Maior Risco por Operação	10%
Volume Mínimo por Operação	1
Venda Parcial	Não
Stop Loss	Normal
Descanso após Operação de Sucesso	Não
Descanso após Operação de Falha	Não
Normalização por Frequência de Operações	Sim
Compensação por Lucratividade	Sim
Descanso por Identificação de Crises	Sim
Descanso por Tendência de Baixa	Sim
RCC Dinâmico	Sim
RCC Dinâmico (Referência)	100%
RCC Dinâmico (K)	5

Tabela 4.1: Configurações de Simulação

(Mostrar gráficos de operações por ticker).

Strategy Total Yield (%)	**54.06**
Baseline Total Yield (%)	53.82
IBOVESPA Total Yield (%)	28.39
CDI Total Yield (%)	9.39
Strategy Total Volatility (%)	**33.77**
Baseline Total Volatility (%)	52.92
Strategy Sharpe Ratio (-)	**0.83**
Baseline Sharpe Ratio (-)	0.69
Strategy Sortino Ratio (-)	**1.24**
Baseline Sortino Ratio (-)	0.77
Strategy-Baseline Spearman Correlation (-)	**0.78**
Strategy-IBOV Spearman Correlation (-)	0.58
Maximum Used Capital (%)	100.00
Average Used Capital (%)	**77.37**
Maximum Active Operations	18
Average Active Operations	8.96
Active Operations Standard Deviation	4.28
Profit (R$)	54060.98
Total Operations	1073
---Successful Operations (hit 3:1 target)	**328 (30.6%)**
---Partial Sale Successfull Operations (hit 1:1 or 2:1 target)	0 (0.0%)
---Failed Operations	**737 (68.7%)**
---Timed Out Operations	1 (0.1%)
---Unfinished Operations	7 (0.7%)
Strategy Yield (% ann)	**21.68**
Baseline Yield (% ann)	21.59
IBOVESPA Yield (% ann)	12.02
CDI Yield (% ann)	4.10
Strategy Volatility (%ann)	**22.76**
Baseline Volatility (%ann)	35.66

Figura 4.2: Resultados

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, Seçao 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).

[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] 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.

[48] BACHELIER, L., "Théorie de la spéculation". In: *Annales scientifiques de l'École normale supérieure*, v. 17, pp. 21–86, 1900.

[49] 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.

[50] COOPER, J. C., "World stock markets: Some random walk tests", *Applied Economics*, v. 14, n. 5, pp. 515–531, 1982.

[51] MALKIEL, B. G., *A random walk down Wall Street: the time-tested strategy for successful investing*. WW Norton & Company, 2019.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] ENDERS, W., *Applied econometric time series*. John Wiley & Sons, 2008.

[57] 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.

[58] BOLLERSLEV, T., "Generalized autoregressive conditional heteroskedasticity", *Journal of econometrics*, v. 31, n. 3, pp. 307–327, 1986.

[59] HIGGINS, M. L., BERA, A. K., "A class of nonlinear ARCH models", *International Economic Review*, pp. 137–158, 1992.

[60] RABEMANANJARA, R., ZAKOIAN, J.-M., "Threshold ARCH models and asymmetries in volatility", *Journal of applied econometrics*, v. 8, n. 1, pp. 31–49, 1993.

[61] FRANSES, 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.

[62] MARCUCCI, J., "Forecasting stock market volatility with regime-switching GARCH models", *Studies in Nonlinear Dynamics & Econometrics*, v. 9, n. 4, 2005.

[63] 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.

[64] 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.

[65] 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.

[66] GANDHMAL, D. P., KUMAR, K., "Systematic analysis and review of stock market prediction techniques", *Computer Science Review*, v. 34, pp. 100190, 2019.

[67] 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.

[68] YFINANCE, "yfinance", https://pypi.org/project/yfinance/, (Acessado em 1 de Junho de 2022).

[69] YAHOO!, "Yahoo! Finance", https://finance.yahoo.com, (Acessado em 1 de Junho de 2022).

[70] NORI, P., "Project Github Page", https://github.com/Nori12/Projeto-Final.

[71] SCIKIT-LEARN, "Random Forest", https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, (Acessado em 5 de Julho de 2022).

[72] PLOTLY, “Dash”, https://dash.plotly.com/, (Acessado em 28 de Junho de 2022).