TÉCNICAS DE APRENDIZADO DE MÁQUINA APLICADAS À 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 À 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 À 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

À minha mãe engenheira mecânica que tanto amo no meu coração.

AGRADECIMENTO

Agradeço ao meu país, que esconde um povo tão sofrido e ao mesmo tempo tão amoroso. Este trabalho é apenas um pedacinho do pagamento da minha dívida.

RESUMO

Todos os dias, diversas negociações são realizadas nas bolsas de valores no 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 aumento de performance de uma estratégia de swing trade no mercado de ações brasileiro (B3). Para isso, é realizada a reprodução aproximada da estratégia, seguida pela substituição dos critérios de decisão de entrada nas operações e preços alvos de venda por um modelo de aprendizado de máquina.

Palavras-Chave: Machine Learning, 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

JSON - *JavaScript Object Notation*

k-NN - *K Nearest Neighbors*

MACD - *Moving Average Convergence/Divergence*

ML - *Machine Learning*

MME - Média Móvel Exponencial

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

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Tema

O tema deste trabalho se resume no aperfeiçoamento de uma estratégia de swing trade na bolsa de valores 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, cujos dados diários são de domínio público e 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 e 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, portanto só há lucro em variações positivas 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 [7] 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 [8], sejam 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 [9].

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 [10].

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 estratégias de trading de forma a trazer uma melhor performance, colaborando assim para uma maior variedade de opções de investimentos à população brasileira.

1.4 Objetivos

O objetivo geral deste trabalho é implementar um software capaz de simular uma estratégia de swing trade e gerar uma nova estratégia baseada na anterior utilizando aprendizado de máquina a fim de melhorar sua performance. 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) Simular a estratégia de swing trade do trader André Moraes da forma mais fidedigna que a janela de dados diária permita; (3) Criar e simular um novo algoritmo baseado na estratégia anterior utilizando aprendizado de máquina; (4) Criar e simular uma estratégia de baseline, referente à estratégia de aprendizado de máquina; (5) Analisar os modelos gerados.

1.5 Metodologia

O trabalho teve início na criação de um ambiente propício à simulação de estratégias, bem como sua configuração e manutenção. Consequentemente, 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 geração de candles semanais a partir de candles diários, a identificação de picos, os momentos de tendência de alta do mercado e as médias móveis exponenciais dos preços de fechamento.

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. Aqui são criadas e executadas: a estratégia base, que é uma adaptação do André Moraes; a estratégia aprimorada, que utiliza aprendizado de máquina; e

a estratégia de baseline, respectivamente.

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, especialmente de Data Sciente, 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 básico do Mercado Financeiro e de Aprendizado de Máquina. Bem como uma revisão dos Trabalhos Relacionados ao tema.

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 tem capital de investimento, e agentes deficitários, que buscam capital para rentabilizá-lo, através da compra e venda valores mobiliários (i.e., ativos financeiros) [11]. Consequentemente, gera-se uma maior liquidez destes ativos e também uma melhora no fluxo de capitais entre os agentes econômicos, sejam eles 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 [12].

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) [13], 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 [14] 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 [15], ou IPO (Initial Public Offering), onde as ações que compõe o capital social [16] de uma companhia são vendidas pela primeira vez ao público geral. Uma vez encerrado o IPO, estas mesmas ações passam para o mercado secundário [17], onde investidores as negociam entre si. Em retorno ao capital adquirido pela companhia, surgem algumas responsabilidades, dentre elas a publicação de demonstrações financeiras [18], auditadas pela própria CVM [19].

Para o acionista de uma sociedade anônima, existem duas formas de se obter lucro: através de proventos (dividendos e juros sobre capital próprio) [20]; 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 (i.e., ação) posteriormente à sua aquisição a um preço de compra inferior. No entando, também é possível trabalhar com posições vendidas (short selling) [21], 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 expectativa de queda de um ativo se mostra verdadeira.

2.1.1 Hipótese do Mercado Eficiente

A Hipótese do Mercado Eficiente, definida por FAMA [22], 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 o todo o histórico de informações disponibilizados publicamente.

- HME semi-forte: Engloba a HME fraca, acrescentando a existência de uma mudança instantânea que os preços sofrem ao surgirem novas informações.

- HME forte: Engloba a HME semi-forte, porém entende 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 o 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 significativa [23].

A HME fraca foi verificada devido à 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 dividendos das companhias se refletem, em média, no preços das ações [24].

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 no mercado financeiro. No entanto, é importante ressaltar que outros autores questionam ao menos parcialmente os estudos realizados por FAMA, sejam por resultados inconclusivos ou por anomalias detectadas no comportamento do mercado. Por exemplo, SHOSTAK [25] 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 [26]. Já o investidor Jack Schwager afirma que a HME está correta pelos motivos errados [27], pois é muito difícil bater a média do mercado de forma consis-

tente 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 [28] 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: Um dos mais conhecidos no mercado. É a média ponderada pelo capital social das 500 maiores companhias do mercado americano.
- Dow Jones Industrial Average: É a média ponderada pelo preço da ação das 30 maiores blue-chips industriais e financeiras do mercado americano (i.e., companhias bem conhecidas, bem estabelecidas e com grande capital social).
- Ibovespa: Principal indicador de desempenho do mercado brasileiro. Possui alguns critérios específicos, mas basicamente é composto pelas ações com maior volume de negociação na B3 [29].

Índices não são negociáveis pois não passam de métricas de mercado. Para isso existem fundos de investimentos chamados ETFs (Exchange-Traded Funds) [30], 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 [31] surge 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 [32]. 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) [33, 34]. 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 [35]; e o MACD (Moving Average Convergence-Divergence) [36].

Leitura de Gráficos de Candlesticks

Gráficos de Candlesticks[1] são bastante utilizados na AT. A leitura é padronizada de acordo com a Figura 2.1. Neste tipo de gráfico as cores importam, pois indicam se o balanço do período foi positivo ou negativo.

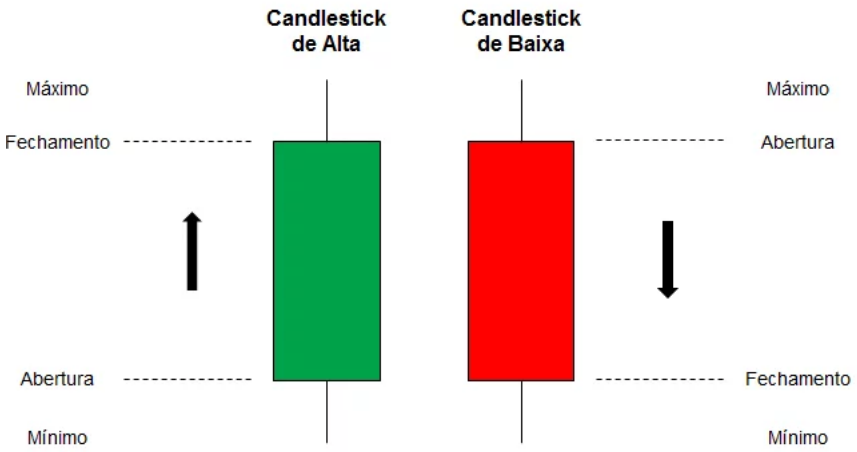

Figura 2.1: Leitura de um gráfico de *candlestick* [1]

[1]Em português: Gráfico de Velas.

Teoria de Dow

A Teoria de Dow, criada pelo americano Charles Henry Dow em 1884 é considerada a base da AT moderna [2]. Embora não tivesse sido formalizada explicitamente pelo autor enquanto estava vivo, amigos e profissionais da época tiveram o trabalho de divulgar e fazer alguns ajustes. Baseada na HME, a ideia central por trás da Teoria de Dow é que a lógica econômica deve ser usada para explicar os movimentos do mercado, que em condições ideais segue o padrão de: tendência de alta[2]; topo; tendência de baixa[3]; e fundo, intercalados com períodos de consolidação[4]. A Figura 2.2 ilustra esse comportamento.

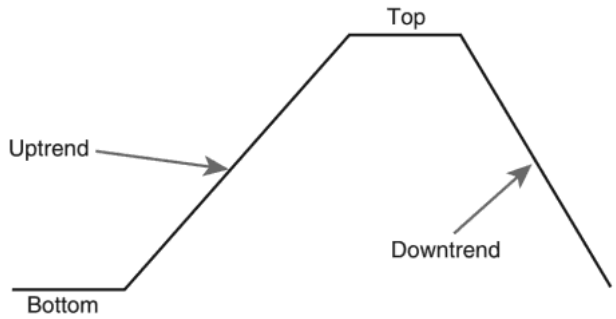

Figura 2.2: Comportamento do mercado ideal segundo a Teoria de Dow [2]

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 $n > 0$.

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

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

[2]Topos e fundos ascendentes.

[3]Topos e fundos descendentes.

[4]Topos e fundos lateralizados.

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.

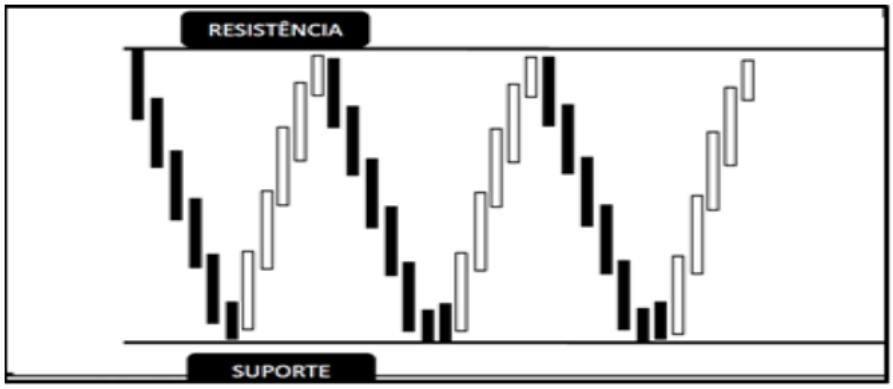

Figura 2.3: Formação de linhas de Suporte e de Resistência [3]

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.

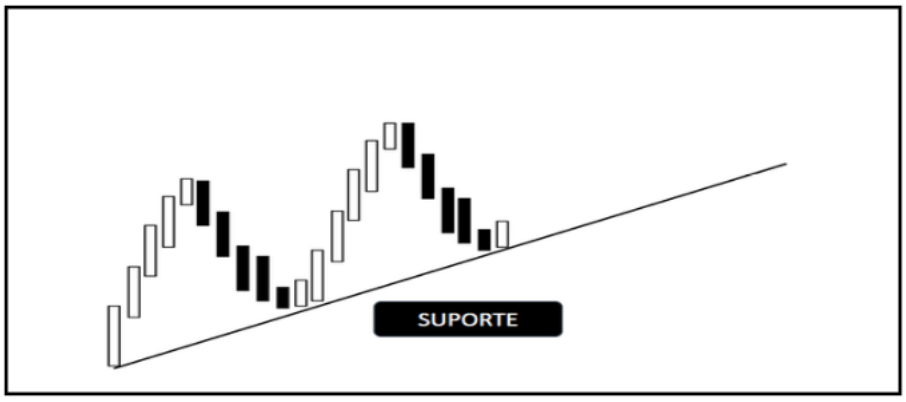

Figura 2.4: Formação de uma Linha de Tendência de Alta [3]

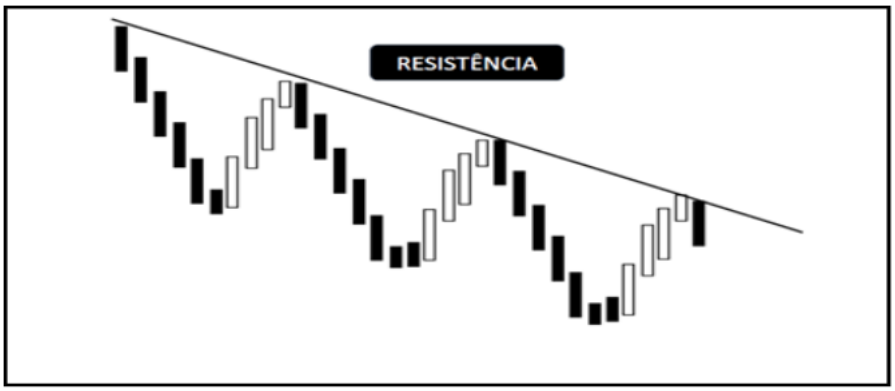

Figura 2.5: Formação de uma Linha de Tendência de Baixa [3]

2.3 Aprendizado de Máquina

Aprendizado de Máquina (Machine Learning) é um campo de estudo dentro de Ingeligência Artificial [37] 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 [38] para um estudo de caso do jogo de damas [39].

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.

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

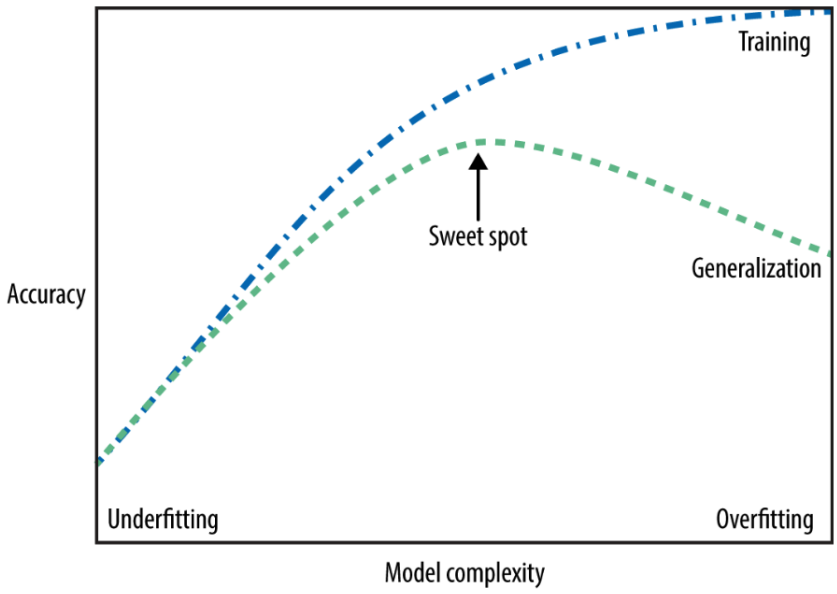

Figura 2.6: Relação entre complexidade e acurácia de um modelo [4]

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 [40]. 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) [40].

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.

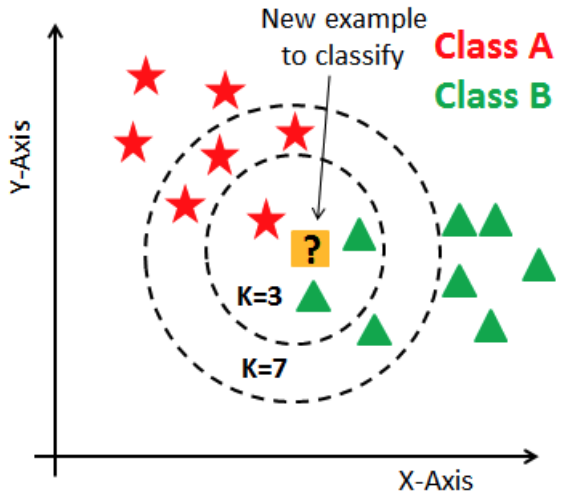

Figura 2.8: Funcionamento de um algoritmo k-NN para o problema de classificação [6]. Para K=3 a classe é B e para K=7 a classe é A.

Decision Tree

Em essência, uma Árvore de Decisão[5] é uma sequência hierárquica de estruturas de decisão *if/else*[6] acerca das características do conjunto de dados. Tecnicamente, pode-se construir uma Árvore de Decisão até que todas as suas folhas[7] 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 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

[5]Em inglês: *Decision Tree*.

[6]Em português: se/senão.

[7]Em inglês: leafs.

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.

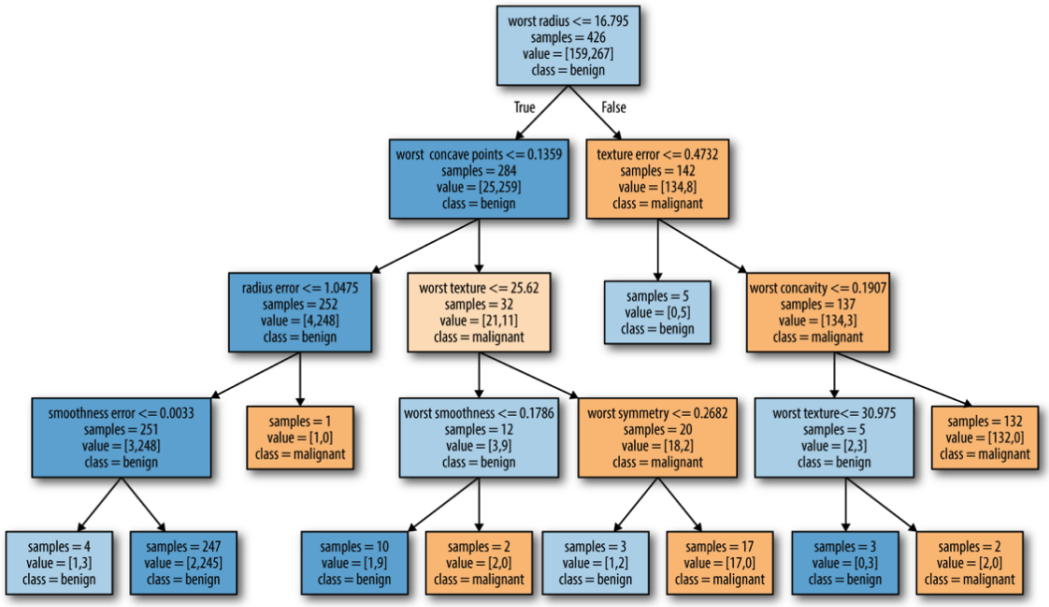

Figura 2.9: Visualização de uma Árvose de Decisão para um *dataset* de câncer de mama [4].

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*[8] é a combinação de diversas Árvores de Decisão ligeiramente diferentes entre si [4]. A

[8]Em português: Floresta Aleatória.

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[9].

2.4 Considerações para Análise de Resultados

2.4.1 Índice de Sharpe

Criado pelo americano William F. Sharpe em 1966 e revisado em 1994, o Índice de Sharpe[10] 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) [41]. 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.

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

[9]Do inglês: *Central Process Unit*.

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

2.4.2 Índice de Sortino

2.4.3 Correlação de Spearman

2.5 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 [42], 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.

2.5.1 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*[11] 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.

2.5.2 Modelos Baseados em Processos Estocásticos

De acordo com GODFREY [43], a hipótese de que a flutuação de preços no mercado de ações poderia ser explicada por uma Random Walk[12] foi feita por BACHELIER [44]. A partir da década de 60, muitos trabalhos acadêmicos foram realizados nessa linha na tentativa de entender o comportamento e a previsibilidade do mercado [22, 45, 46], assim como estratégias [47]. Nota-se que até hoje utiliza-se Random Walks para testar a hipótese de eficiência de mercados [48].

Outra abordagem utilizada são os Modelos Ocultos de Markov (do inglês Hidden Markov Model, ou HMM) [49]. 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 [50] utiliza um modelo HMM para previsão do preço de fechamento do dia seguinte para ações FAANG[13]. A partir da série histórica de preços OHLC[14], seu modelo atinge uma eficiência de 97%-99%, calculado a partir do erro percentual absoluto médio[15].

---

[11] 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.

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

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

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

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

Uma outra aplicação de modelos HMM é dada por DE ANGELIS et al [51], que criou uma metodologia a partir de índices da bolsa americana capaz de identificar períodos estáveis e instáveis (i.e. crises econômicas), assim como as probabilidades de transição entre um estado e o outro.

Por fim, pode-ser mencionar o uso de modelos ARCH[16] (Autoregressive Conditional Heteroskedasticity). A ideia central está na modelagem de uma variância condicional, ou seja, que muda de acordo o instante da série [52]. 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 2.4.

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-1}^2 \tag{2.4}$$

O modelo ARCH foi proposto por ENGLE em 1982 para estimar a variância da inflação do Reino Unido [53]. A partir daí, várias derivações surgiram, como por exemplo: GARCH [17] por BOLLERSLEV [54] em 1986, EGARCH [18] por NELSON em 1991, NGARCH [19] por HIGGINS e BERA[55] em 1992, TGARCH [20] por ZAKOIAN e RABEMANANJARA [56] em 1993, dentre outros. Alguns dos modelos da família ARCH podem ser encontrados nos trabalhos de FRANSES e DIJK [57], de MARCUCCI [58] e de ALBERG et al [59].

2.5.3 Modelos Baseados em Aprendizado de Máquina

Existem registros de estudos sobre inteligência artificial aplicados ao mercado financeiro por volta da década de 70 [60], porém ainda em um estágio embrionário devido às dificuldades de processamento computacional e de acesso a dados na época.

---

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

[17] Generalised ARCH.

[18] Exponential Generalised ARCH.

[19] Non-linear Generalised ARCH.

[20] Threshold Generalised ARCH.

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 [61] relata que dos 122 trabalhos mais relevantes publicados entre 2007 e 2018 com o tema de predição do mercado financeiro usando ML, 66% são baseados em AT, 23% são baseados em AF e 11% usam análises mistas. Além disso, os algoritmos mais utilizados são ANN[21] (*Artificial Neural Networks*) e SVM[22] (*Support Vector Machine*).

De forma semelhante, GANDHMAL e KUMAR [62] 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*[23]

É possível encontrar também modelos híbridos, com uma combinação de GARCH com ANN feita por BILDIRCI e ERSIN [63].

[21]Em português: Redes Neurais Artificiais.

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

[23]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.

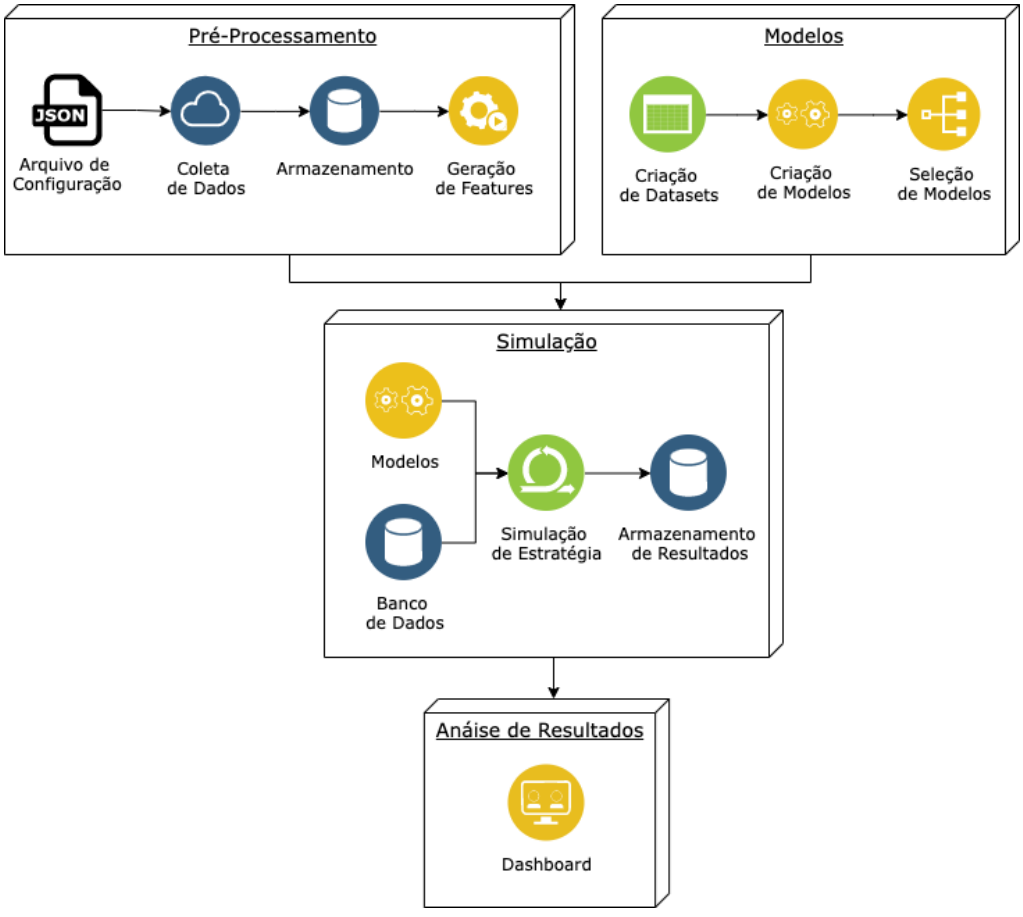

Figura 3.1: Estrutura do técnica do projeto

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 (re)calculadas 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 3.1 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

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.

Continuação da Tabela 3.1		
Nome do Parâmetro	Escopo	Descrição
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.
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.

Continuação da Tabela 3.1		
Nome do Parâmetro	Escopo	Descrição
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 3.4.1. Tipo: *Boolean*. *Default*: *False*. Listável: Sim.
enable_profit _compensation	Local	Uso de compensação por lucratividade acumulada. Ver Seção 3.4.2. 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 3.4.3. Tipo: *Boolean*. *Default*: *False*. Listável: Sim.
dynamic_rcc_reference	Local	Valor de referência de uso de capital médio no controle do RCC dinâmico. Ver Seção 3.4.3. Tipo: *Float*. *Default*: 0,80. Listável: Sim.
dynamic_rcc_k	Local	Valor do ganho proporcional K no controle do RCC dinâmico. Ver Seção 3.4.3. 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.

Continuação da Tabela 3.1		
Nome do Parâmetro	Escopo	Descrição
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.
max_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 3.1		

Tabela 3.1: Lista de parâmetros detalhados.

3.2.2 Coleta de Dados

A Coleta de Dados ocorre através da biblioteca *open-source yfinance* [64], uma ferramenta não oficial que transmite dados públicos da plataforma *Yahoo! Finance* [65], 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].

3.2.3 Armazenamento de Dados

O Armazenamento de Dados é feito 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 por causa da 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[2] do banco. O *script* de criação e população inicial do banco de dados pode ser encontrado em [66].

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

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

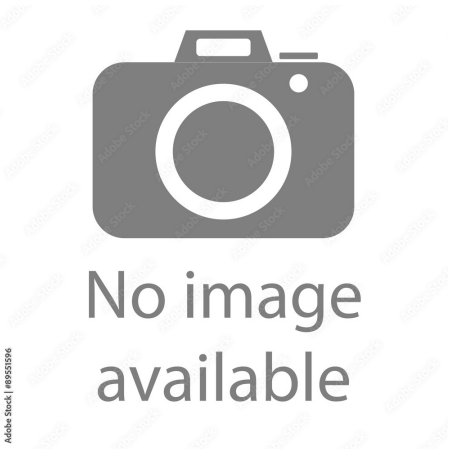

Figura 3.4:

3.2.4 Geração de *Features* de Uso Geral

EXCLUIR - Quais features; Cuidados com não-causalidade.

3.3 Simulação de Estratégia

3.3.1 Estrutura

EXCLUIR - Carteira com N ativos de datas distintas; Regra de 3 para 1 entre stop e alvo; 1 operação por ativo.

3.3.2 Premissas

EXCLUIR - Compra na abertura do merdado; Sem venda no dia da compra; Prioridades durante venda (stop primeiro).

3.3.3 Período Máximo de Dias por Operação

EXCLUIR - Motivação da escolha dos 45 dias; Gráfico entre ABEV e MGLU.

3.3.4 Gerenciamento de Risco

EXCLUIR - Coeficiente de Risco-Capital

3.3.5 Risco de Entrada por Operação

EXCLUIR - Cálculo do risco mínimo; Cálculo do risco Máximo.

3.3.6 Descanso por Tendência de Baixa

3.3.7 Descanso por Identificação de Crises

3.3.8 Lista de Parâmetros de Configuração

EXCLUIR - Lista todos e explicar o que fazem.

3.3.9 Ensaios Paralelos

EXCLUIR - Parâmetros que estão implementados e não trouxeram resultados expressivos.

3.4 Otimizações de Gerenciamento de Carteira

3.4.1 Normalização por Frequência de Operações

3.4.2 Compensação por Lucratividade

3.4.3 Controle Proporcional para Uso de Capital

3.5 Criação de Modelos

3.5.1 Resumo

3.5.2 *Feature Selection*

3.5.3 Geração de *Datasets*

3.5.4 *Walk Forward Optimization*

3.5.5 Critérios de Escolha

3.6 Análise de Resultados

EXCLUIR - Dashboard; Baseline

Capítulo 4

Conclusão

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] 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).

[8] 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).

[9] 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).

[10] FERNÁNDEZ, A., “Artificial intelligence in financial services”, *Banco de Espana Article*, v. 3, pp. 19, 2019.

[11] 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).

[12] 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.

[13] 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).

[14] B3, “Ações”, https://www.b3.com.br/pt_br/produtos-e-servicos/negociacao/renda-variavel/acoes.htm, (Acessado em 24 de Março de 2022).

[15] 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).

[16] 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).

[17] 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).

[18] 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).

[19] 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).

[20] 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).

[21] 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).

[22] 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.

[23] 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).

[24] 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.

[25] 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.

[26] JUNG, J., SHILLER, R. J., "Samuelson's dictum and the stock market", *Economic Inquiry*, v. 43, n. 2, pp. 221–228, 2005.

[27] SCHWAGER, J. D., *Market Sense and Nonsense: How the Markets Really Work (and how They Don't)*. John Wiley & Sons, 2012.

[28] FORBES, "Investing Basics: What Is A Market Index?", https://www.forbes.com/advisor/investing/stock-market-index/, (Acessado em 28 de Março de 2022).

[29] 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).

[30] 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).

[31] INVESTOPEDIA, "Fractional Share", https://www.investopedia.com/terms/f/fractionalshare.a (Acessado em 28 de Março de 2022).

[32] BULKOWSKI, T. N., *Fundamental Analysis and Position Trading: Evolution of a Trader*, v. 605. John Wiley & Sons, 2012.

[33] MURPHY, J. J., *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin, 1999.

[34] EDWARDS, R. D., MAGEE, J., BASSETTI, W. C., *Technical analysis of stock trends*. CRC press, 2018.

[35] BOLLINGER, J., *Bollinger on Bollinger bands*. McGraw Hill Professional, 2002.

[36] APPEL, G., DOBSON, E., *Understanding MACD*. Traders Press, 2007.

[37] IBM, "Artificial Intelligence (AI)", https://www.ibm.com/cloud/learn/what-is-artificial-intelligence, (Acessado em 4 de Abril de 2022).

[38] IBM, "Machine Learning", https://www.ibm.com/cloud/learn/machine-learning#: :text=IBM(Acessado em 4 de Abril de 2022).

[39] 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.

[40] 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.

[41] SHARPE, W. F., "The sharpe ratio", *Streetwise–the Best of the Journal of Portfolio Management*, pp. 169–185, 1998.

[42] KIM, K., *Electronic and algorithmic trading technology: the complete guide*. Academic Press, 2010.

[43] 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.

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

[45] 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.

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

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

[48] 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.

[49] 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.

[50] 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.

[51] 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.

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

[53] 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.

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

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

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

[57] 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.

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

[59] 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.

[60] 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.

[61] 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.

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

[63] 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.

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

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

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