Universidade Federal
do Rio de Janeiro
Escola Politécnica

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

Pedro Henrique Barbosa Nori

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

Orientador: Heraldo Luis Silveira de Almeida

Rio de Janeiro

Setembro de 2022

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

Pedro Henrique Barbosa Nori

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

_____

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

Examinador:

_____

Prof. Flávio Luis de Mello, D. Sc.

Examinador:

_____

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

Rio de Janeiro

Setembro de 2022

Declaração de Autoria e de Direitos

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

1. O autor declara que o trabalho apresentado na disciplina de Projeto de Graduação da Escola Politécnica da UFRJ é de sua autoria, sendo original em forma e conteúdo.

2. 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.

AGRADECIMENTO

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

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

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

A todas as experiências que puder compartilhar com meus amigos. Deixo um abraço especial para o meu amigo de infância Daniel Iunes Monteiro, que infelizmente não se encontra mais nesta vida.

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

RESUMO

Todos os dias, diversas negociações são realizadas nas bolsas de valores do mundo inteiro. Com os mais diversos objetivos, investidores buscam um aumento crescente de patrimônio de forma consistente. Paralelamente, inteligências artificias estão substituindo cada vez mais atividades antes desempenhadas pelo homem. Nesse sentido, este trabalho visa a aplicação de técnicas de aprendizado de máquina para a elaboração de uma estratégia de *swing trade* no mercado acionário brasileiro. Para isso, é concebida uma estrutura de regras e premissas que criam uma base ao modelo de aprendizado de máquina, responsável por decidir o momento de entrada em operações a partir de um conjunto de dados. Ao final, algumas estratégias são simuladas e suas performances analisadas com um modelo *baseline*.

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

ABSTRACT

Every day, several negotiations are carried out on stock exchanges around the world. With the most diverse objectives, investors seek a consistently growing increase in equity. At the same time, artificial intelligences are increasingly replacing activities previously performed by man. In this context, this work aims at the application of machine learning techniques for the elaboration of a swing trade strategy in the Brazilian stock market. So, a structure of rules and assumptions is conceived to create a basis for the machine learning model, responsible for deciding when to enter into operations given a set of data. In the end, some strategies are simulated and its results compared to a *baseline* model.

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

SIGLAS

AF - Análise Fundamentalista

API - *Application Programming Interface*

ANN - *Artificial Neural Networks*

ARCH - *Autoregressive Conditional Heteroskedasticity*

AS - Aprendizado Supervisionado

AT - Análise Técnica

B3 - Brasil, Bolsa, Balção

CPU - *Central Process Unit*

CSL - *Cost Sensitive Learning*

CSV - *Comma-separated values*

CVM - Comissão de Valores Mobiliários

DT - *Decision Tree*

EGARCH - *Exponential Generalised ARCH*

EMA - *Exponential Moving Average*

ETF - *Exchange-Traded Fund*

GARCH - *Generalised ARCH*

HME - Hipótese do Mercado Eficiente

HMM - *Hidden Markov Model*

iBovespa - Índice Bovespa

IPO - *Initial Public Offering*

IIR - *Infinite Impulse Reponse*

IL - Índice de Lucratividade

JSON - *JavaScript Object Notation*

k-NN - *K Nearest Neighbors*

MACD - *Moving Average Convergence/Divergence*

ML - *Machine Learning*

MME - Média Móvel Exponencial

NFO - Normalização por Frequência de Operações

NGARCH - *Non-linear Generalised ARCH*

RCC - *Risk-Capital Coefficient*

RF - *Random Forest*

RI - Relações com Investidores

SVM - *Support Vector Machine*

TGARCH - *Threshold Generalised ARCH*

UFRJ - Universidade Federal do Rio de Janeiro

WFA - *Walk-Forward Analysis*

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Tema

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 estão substituindo cada vez mais atividades antes desempenhadas pelo homem.

Nesse contexto, este trabalho se resume na elaboração de uma estratégia de *swing trade* [1] na bolsa de valores brasileira através de métodos de *machine learning* (ML).

Desta forma, o problema a ser abordado é a identificação do momento apropriado para compra de um determinado ativo, como também os preços 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 API *open-source yfinance*, disponível em Python [2]. 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. O Apêndice A evidencia os problemas encontrados em mais detalhes.

Na estratégia de negociação desenvolvida neste trabalho, a duração das operações tem em vista um horizonte mínimo de um dia, sendo portanto operações de *swing trade*. Também não são realizas vendas a descoberto[1] [3], portanto só há lucro em movimentos crescentes de preço. 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 [4], por causa da dificuldade técnica de obtenção dessas informações de maneira automatizada e estruturada.

1.3 Justificativa

O crescimento do número de investidores na bolsa de valores brasileira [5] pode ser parcialmente justificado por 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 investimento tem se mostrado expressivo, seja por pessoas físicas ou fundos de investimento, de forma total ou parcial em suas estratégias [6, 7]. Por outro lado, tal crescimento não vem sendo igualmente representado no Brasil devido às peculiaridades do mercado de capitais nacional, como a alta volatilidade e a alta sensibilidade a notícias [8].

Paralelamente, estudos relacionados a aprendizado de máquina (ML) vêm trazendo resultados práticos no dia-a-dia das pessoas, desde o clássico exemplo de reconhecimento de mensagens de *spam* em um caixa de email [9] à identificação do perfil de consumo de clientes em uma loja [10]. Da mesma forma, instituições financeiras e bancos centrais também estão, com cautela, incorporando aplicações de aprendizado de máquina em tarefas internas [11].

[1] Venda a descoberto (*short selling*): Processo no qual a venda de ações ocorre antes da compra, fazendo com que o lucro seja obtido na queda do preço de mercado.

Apesar das dificuldades inerentes ao cenário atual do mercado de capitais brasileiro [12], não se pode ignorar o potencial de ganhos e de economia de tempo que os algortimos podem trazer aos investidores, uma vez que se bem configurados, operam diretamente conectados à plataforma da bolsa de valores e de forma automatizada. 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) treinar 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; e (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. Dentre as atividades realizadas durante o pré-processamento dos dados, anteriores às simulações, é possível citar a identificação de picos na série histórica e a criação de *features* de suporte às decisões de entrada e saída das operações.

Em seguida, foi construído o *software* principal, que além de se comunicar com o banco de dados, também busca os dados históricos quando há necessidade e simula

as estratégias solicitadas via um arquivo de configuração.

Paralelamente, ocorreu o treinamento dos modelos de ML por meio de aprendizado supervisionado. Estes por sua vez, são acoplados ao programa principal para a tomada da decisão do momento de entrada nas operações.

Como atividades secundárias, porém importantes para os resultados finais, foram concebidos e refinados alguns parâmetros de simulações para aperfeiçoamento das estratégias. Nota-se que alguns parâmetros não tem relação alguma com os modelos gerados e sim com o gerenciamento de capital da carteira de ativos utilizada durante as simulações.

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

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, a Metodologia é descrita junto da Implementação. São desenvolvidas as etapas de: pré-processamento de dados; simulações e refinamento de seus parâmetros; assim como o treinamento dos modelos de ML.

No Capítulo 4, são apresentados os resultados das simulações a partir dos modelos criados e dos parâmetros otimizados no Capítulo 3.

Por fim, o Capítulo 5 encerra com a conclusão e as recomendações de trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo busca fornecer ao leitor alguns insumos para uma melhor contextualização do trabalho. Serão abordadas as dinâmicas básicas de funcionamento do mercado, seguida por uma introdução de ML, um tipo específico de análise e alguns índices relavantes a este estudo. Por fim, uma revisão bibliográfica é realizada.

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) [13]. Consequentemente, gera-se uma maior liquidez destes ativos e também uma melhora no fluxo de capitais entre os agentes econômicos, seja os governos por meio dos bancos centrais, os bancos privados, as instituições financeiras ou até mesmo as pessoas físicas.

No Brasil, o Mercado de Capitais é regulado e fiscalizado pela Comissão de Valores Mobiliários (CVM), uma autarquia federal vinculada ao Ministério da Fazenda e criada em 1976 através da Lei nº 6.385 [14].

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 Brasil,

Bolsa, Balcão (B3) [15], 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 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 [16]. Esse processo acontece através de uma oferta pública inicial (IPO) [17], onde as ações que compõe o capital social [18] de uma companhia são vendidas pela primeira vez ao público geral. Uma vez encerrado o IPO, estas mesmas ações passam para o mercado secundário [19], onde investidores as negociam entre si. Em retorno ao capital adquirido pela companhia, surgem algumas responsabilidades, dentre elas a publicação de demonstrações financeiras [20], auditadas pela própria CVM [21].

Para o acionista de uma sociedade anônima, existem duas formas de se obter lucro: através de proventos (dividendos e juros sobre capital próprio) [22]; ou através de operações de compra e de venda de ações, mediante oscilações de seu valor de mercado.

2.1.1 Hipótese do Mercado Eficiente

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

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

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

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

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

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

Em resumo, o estudo da HME traz informações relevantes quanto se avalia a teoria por trás da possibilidade de aplicação de estratégias de *trading* ao mercado financeiro. No entanto, é importante ressaltar que outros autores questionam ao menos parcialmente os estudos realizados por Eugene Fama, seja por resultados inconclusivos ou por anomalias detectadas no comportamento do mercado. Por exemplo, Frank Shostak critica abertamente a premissa de que todos os investidores teriam a mesma expectativa sobre os retornos da empresa [26]. O ganhador do prêmio Nobel em ciências econômicas Paul Samuelson, que afirma que 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, 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 [28].

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 existe um tipo de fundo de investimento chamado *Exchange-Traded Fund* (ETF) [31], que é especializado 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 [32].

2.1.3 Mercado Fracionário

Ações são negociadas em múltiplos de um lote, que representa uma quantidade mínima de papéis. Nesse contexto, o Mercado Fracionário surge com o objetivo de facilitar negociações de volumes menores que o lote mínimo permitido. No Brasil, o lote é de 100 ações e o Mercado Fracionário permite a compra de no mínimo 1 ação [33]. No entanto, este mercado possui menor liquidez e maior volatilidade, mas sempre acompanha o preço do ativo negociado no mercado aberto.

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

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 [33].

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

2.2 Tipos de Análises

Nesta seção, será abordado as duas formas utilizadas para análise do comportamento das companhias ao longo do tempo.

2.2.1 Análise Fundamentalista

A Análise Fundamentalista (AF) é um bastante utilizada para identificar tendências de flutuação no preço de ações tendo em vista um horizonte de longo prazo [4]. Ela se baseia em fatores econômicos relacionados à companhia, como: o quadro de diretores e dirigentes maiores; o fluxo de caixa; a saúde e a situação financeira; o contexto político do país; os concorrentes de mercado; as circunstâncias e os desastres climáticos, naturais ou não, dentre outros fatores.

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

2.2.2 Análise Técnica

A Análise Técnica (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 [36].

Diferentemente da AF, a automação desta análise é mais fácil pois os dados normalmente são organizados e estruturados. No entanto, uma das dificuldades desta análise está na separação entre o que é ruído e o que é de fato tendência de mercado.

Para facilitar a análise, são utilizados indicadores, dentre os quais pode-se 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 [37]; e o MACD [38].

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.

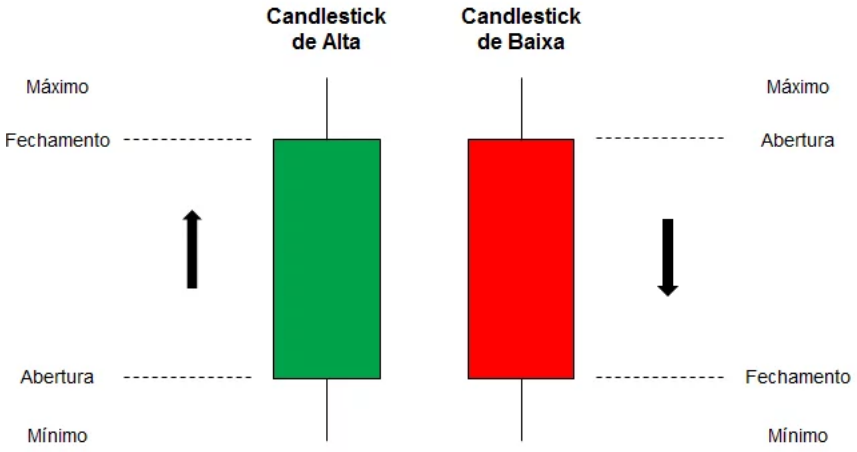

Figura 2.1: Leitura de um gráfico de *candlestick* [39]

Teoria de Dow

A Teoria de Dow, criada pelo americano Charles Henry Dow em 1884 é considerada a base da AT moderna [36]. Embora não tivesse sido formalizada explicitamente

[2]Em português: Gráfico de Velas.

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; topo; tendência de baixa; e fundo, intercalados com períodos de consolidação. A Figura 2.2 ilustra esse comportamento.

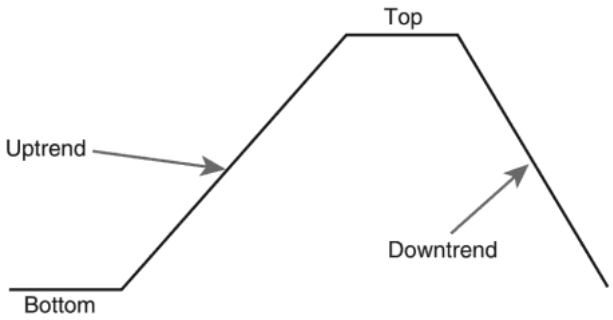

Figura 2.2: Comportamento do mercado ideal segundo a Teoria de Dow [36]

Suporte, Resistência e Linhas de Tendência

Suporte e Resistência são regiões em um gráfico de *candlestick* onde existe um efeito memória associado a grandes ganhos ou perdas históricas [40]. Normalmente estão associadas a eventos econômicos relevantes. A Figura 2.3 ilustra essas regiões, comumente chamadas de Linhas de Suporte e de Resistência.

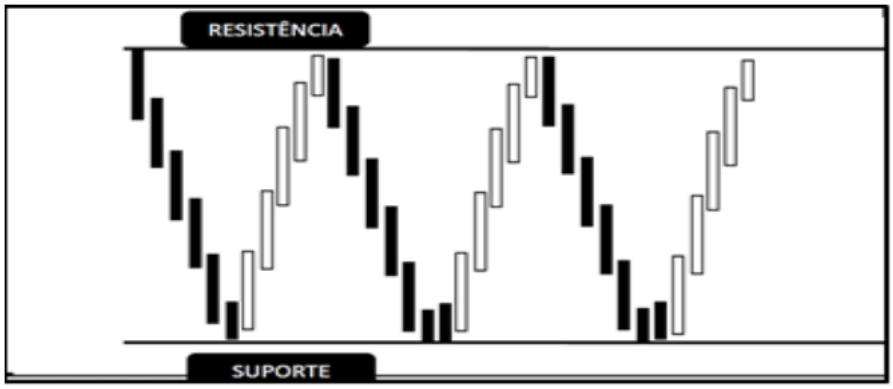

Figura 2.3: Formação de linhas de Suporte e de Resistência [40]

De maneira semelhante, as Linhas de Tendência oferecem uma inspeção gráfica do quanto o preço de um ativo está crescendo o diminuindo. Portanto, estão neces-

sariamente atreladas a movimentos de tendência de alta ou de tendência de baixa. Em essência, não deixam de ser linhas de Supote e de Resistência. As Figuras 2.4 e 2.5 exemplificam esses indicadores.

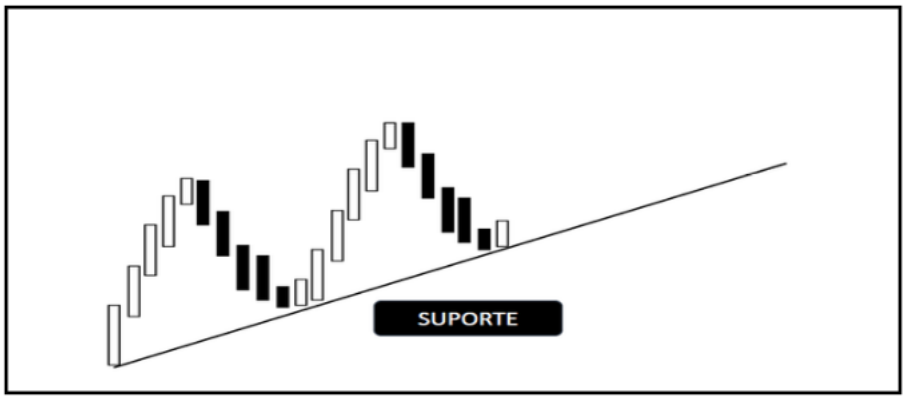

Figura 2.4: Formação de uma Linha de Tendência de Alta [40]

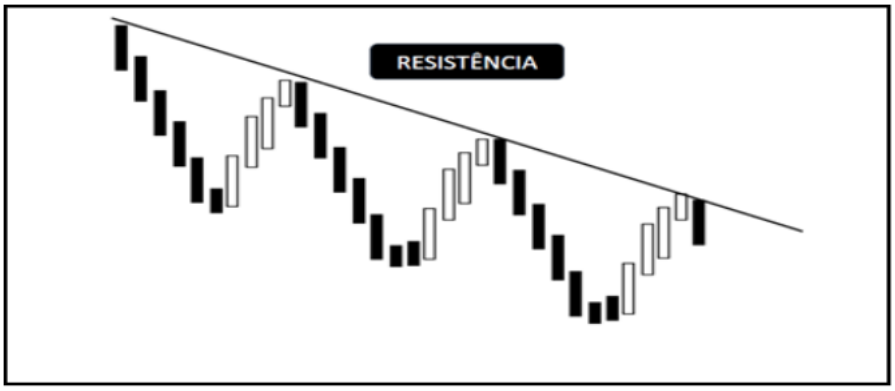

Figura 2.5: Formação de uma Linha de Tendência de Baixa [40]

2.3 Aprendizado de Máquina

Aprendizado de Máquina é um campo de estudo dentro de Inteligência Artificial [41]. O objetivo é extrair conhecimento a partir de uma conjunto de dados [42]. A terminologia foi criada por um pesquisador da IBM chamado Arthur Samuel em 1959 [43] para um estudo de caso do jogo de damas [44].

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 [45], reconhecimento de faces [46], recomendação de produtos em plataformas de *e-commerce* [47], identificação de transações financeiras fraudulentas [48] e suporte a diagnósticos médicos [49].

Dentre as diferentes abordagens de ML que podem ser utilizadas, este trabalho utiliza apenas Aprendizado Supervisionado (AS) [42].

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 entradas e de saídas de dados para assim obter uma melhor performance [42].

O conjunto de dados (*dataset*) com exemplos de entrada e saída utilizado para criação do modelo é chamado de conjunto de dados de treinamento (*training set*). Existe um outro conjunto de dados utilzado para testar a performance do modelo. Este por sua vez é chamado de conjunto de dados de teste (*test set*) e precisa ser diferente dos dados de treinamento para evitar que o efeito memória se sobreponha à qualidade de generalização do modelo, explicado a seguir.

Todo modelo pode ser avaliado sob o ponto de vista da generalização. Essa característica indica a capacidade de realizar predições acuradas no conjuntos de dados de teste. Um dos métodos de avaliação de performance de generalização é chamado de validação cruzada (*cross-validation* [42]. Neste método, quanto maior a taxa de acerto no conjunto de teste, melhor tende a ser a capacidade de generalização.

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

de equilíbrio entre essas características, chamada de "*sweet spot*" na Figura 2.6, que mostra a relação entre generalização, *overfitting* e *underfitting*.

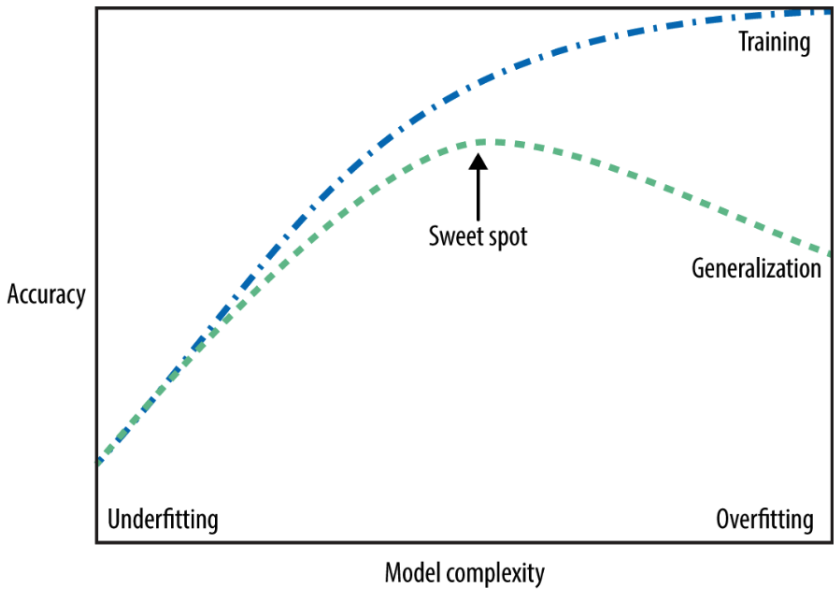

Figura 2.6: Relação entre complexidade e acurácia de um modelo [42]

Existem dois tipos principais de problemas de AS, os problemas de Regressão e os problemas de Classificação [42].

2.3.2 Problema de Regressão

Este problema envolve a predição de um número contínuo a partir dos dados de entrada [42]. 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 [42]. 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 apresentar altos valores de acurácia a partir de um modelo extremamente simples para uma determinada aplicação. Isso acontece porque o modelo aprende que é mais fácil escolher a classe com maior frequencia em seu treinamento do que tentar se aperfeiçoar [50]. 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 [51]. Ver Figura 2.7.

- *Cost Sensitive Learning* (CSL): Ao invés de alterar o tamanho do *dataset*, criam-se pesos diferentes para um erro de classificação durante o treinamento. Portanto, um erro em uma classe menos frequente deve ser mais penalizado do que o contrário. É aconselhável em *datasets* grandes (> 10000) [51].

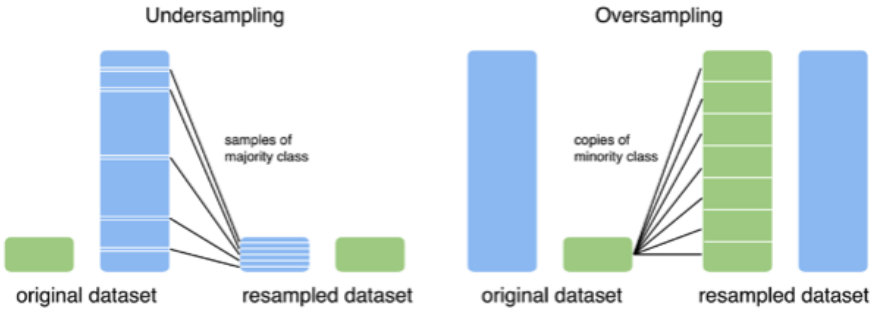

Figura 2.7: *Oversampling* e *Undersampling* de classes desbalanceadas [52]

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 [42, 53]. Pode-se citar o uso deste algoritmo por Wolberg e Mangasarian para identificação de malignidade de amostras citológicas de mamas [54].

A simplicidade deste algoritmo é uma grande vantagem, mas também é possível citar a facilidade de treinamento e a robustez a dados ruidosos [55]. Por outro lado, a limitação de memória, a execução demorada devido ao *lazy learning* [56] e a alta sensibilidade a características irrelevantes são seus pontos negativos. 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.

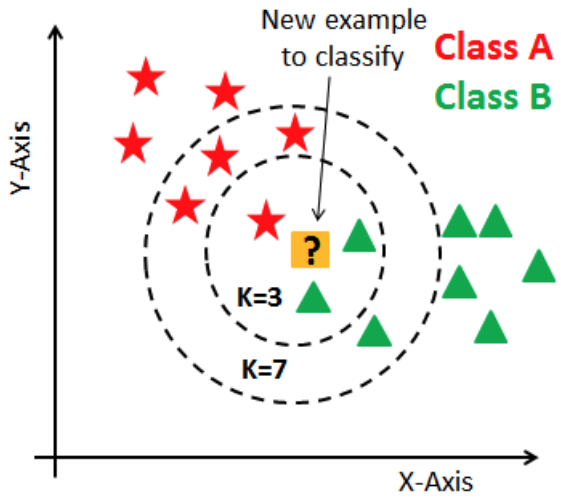

Figura 2.8: Funcionamento de um algoritmo k-NN para o problema de classificação. Para K=3 a classe é B e para K=7 a classe é A [57].

Decision Tree

Em essência, uma Árvore de Decisão é uma sequência hierárquica de estruturas de decisão *if/else* acerca das características do conjunto de dados. Pode-se mencionar o uso deste algoritmo por Lobato *et al* nos sistemas de energia espanhóis [58].

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

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

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*.

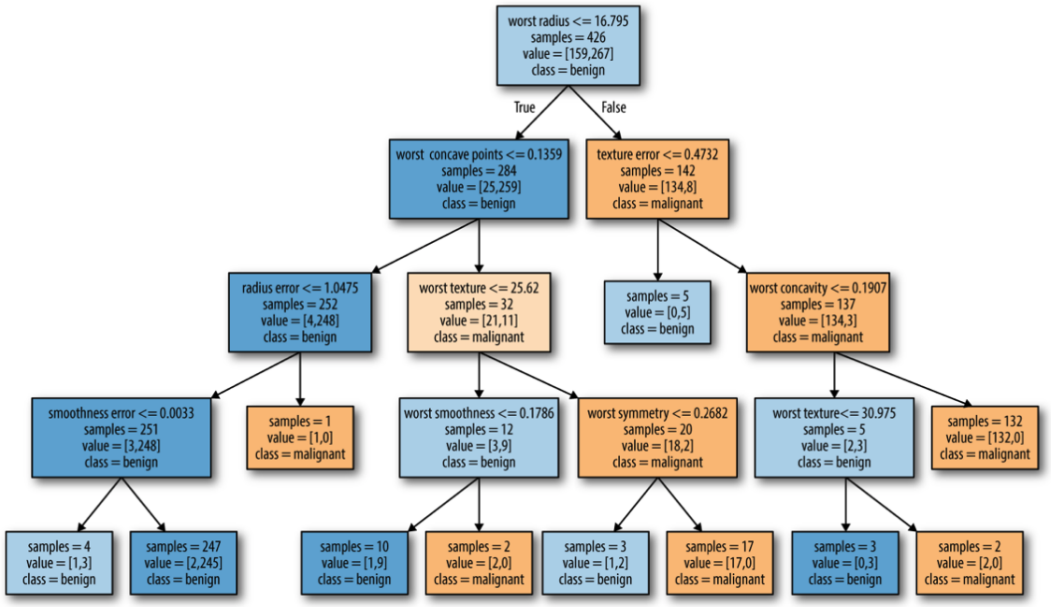

Figura 2.9: Visualização de uma Árvose de Decisão para um *dataset* de câncer de mama [42].

Random Forest

Um dos modelos mais utilizados atualmente, o algoritmo *Random Forest* é a combinação de diversas Árvores de Decisão ligeiramente diferentes entre si [42]. A ideia é que apesar da tendência de *overfitting* existente, a média dos resultados de cada árvore tende a diminuir esse fator. Além dos parâmetros responsáveis por configurar as árvores individualmente, este modelo também precisa do número de árvores que serão utilizadas.

Normalmente é preferível utilizar *Random Forests* ao invés de Árvores de Decisão, salvo casos em que o entendimento e a visualização clara do modelo se torna um fator importante. É possível compensar o aumento do tempo de processamente envolvido na criação de uma *Random Forest* com a paralelização dos núcleos de processamento da CPU[3].

[3]Do inglês: *Central Process Unit*.

2.4 *Walk-Forward Analysis*

Walk-Forward Analysis (WFA) [59] é um processo de otimização mais voltado para séries temporais no contexto de finanças. Para isso, o *dataset* é dividido em múltiplos segmentos consecutivos, que são iterados progressivamente a fim de se obter os parâmetros ou modelos desejados. A Figura 2.10 ilustra o processo. Observa-se que são utilizados os termos *in-sample data* (IS) como sinônimo de *training set* e *out-of-sample data* (OOS) como sinônimo de *test set*. A imagem à esquerda representa um processo não-ancorado, onde o início do IS caminha com o decorrer do processo, já a imagem à direita mantém o início fixo por ser um processo ancorado.

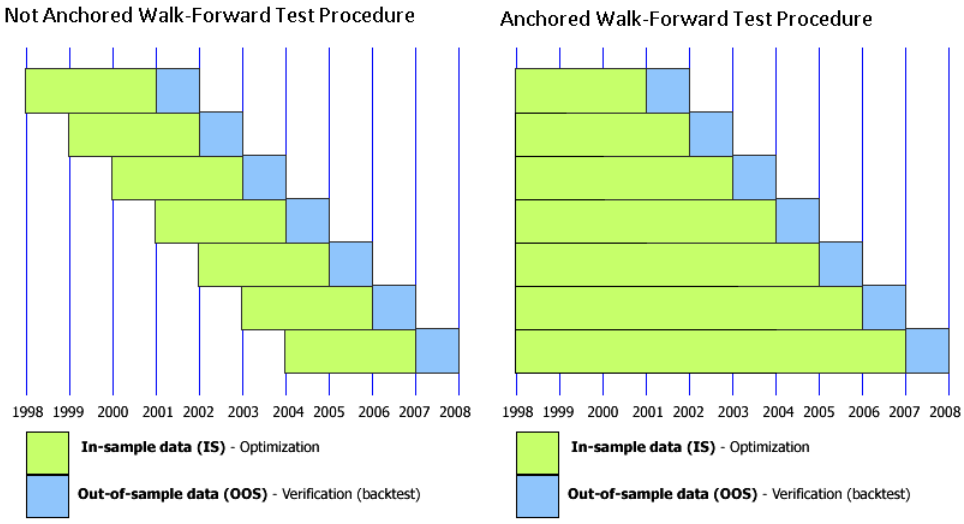

Figura 2.10: *Walk-Forward Analysis* Não-Ancorado e Ancorado [60].

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 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 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) [61]. Seja R_a o retorno do investimento alvo, R_b o retorno do investimento livre de risco e σ_a seu respectivo desvio padrão, pode-se calcular o Índice de Sharpe através da Equação 2.1.

$$S_a = \frac{E[R_a - R_b]}{\sigma_a} \tag{2.1}$$

2.5.2 Índice de Sortino

O Índice de Sortino, criado pelo americano Frank Sortino [62] é uma variante do Índice de Sharpe que considera o desvio padrão apenas dos rendimentos abaixo da média. As Equações 2.2 e 2.3 mostram seu cálculo, onde assim como no Índice de Sharpe, R_a é o retorno do investimento alvo, R_b é o retorno do investimento livre de risco e σ_a seu respectivo desvio padrão. Considere também X_i o i-ésimo retorno e T o retorno médio do investimento.

$$S_a = \frac{E[R_a - R_b]}{\sigma_a} \tag{2.2}$$

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

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 [63]. 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[4] das duas variáveis envolvidas.

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

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

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

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

$$d_i = rgX_i - rgY_i \tag{2.6}$$

Um exemplo prático do cálculo da correlação pode ser analisado a partir da Tabela 2.1, onde são ilustradas amostras para duas variáveis aleatórias X e Y. A Tabela 2.2 acrescenta a informação dos postos rgX e rgY, que ordenam as 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

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

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

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

Tabela 2.2: Postos rgX e rgY

2.6 Trabalhos Relacionados

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

Segundo Kendall Kim [64], 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.

A partir do trabalho de Danilo Pereira [65], pode-se simplificar os modelos de AT aplicados ao mercado financeiro em três metodologias centrais: modelos baseados em indicadores técnicos; modelos baseados em processos estocásticos; e modelos baseados em aprendizado de máquina.

2.6.1 Modelos Baseados em Indicadores Técnicos

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

próprias.

Diversos *traders*[5] e investidores utilizam este tipo de abordagem. Dentre eles podemos citar André Morais [40], de cujas contribuições servirão como base neste trabalho para um aperfeiçoamento via aprendizado de máquina.

2.6.2 Modelos Baseados em Processos Estocásticos

De acordo com Michael Godfrey *et al.* [66], a hipótese de que a flutuação de preços no mercado de ações poderia ser explicada por uma *Random Walk*[6] foi feita por Louis Bachelier [67]. 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, 68, 69], assim como estratégias [70]. Nota-se que até hoje utiliza-se *Random Walks* para testar a hipótese de eficiência de mercados [71].

Outra abordagem utilizada são os Modelos Ocultos de Markov (do inglês *Hidden Markov Model*) [72]. 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, Aishwary Jadhav *et al.* [73] utiliza um modelo HMM para previsão do preço de fechamento do dia seguinte para ações FAANG[7]. A partir da série histórica de preços OHLC[8], seu modelo atinge uma eficiência de 97%-99%,

---

[5]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 papél especulativo.

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

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

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

calculado a partir do erro percentual absoluto médio[9].

Uma outra aplicação de modelos HMM é dada por Luca De Angelis *et al.* [74], 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[10]. A ideia central está na modelagem de uma variância condicional, ou seja, que muda de acordo o instante da série [75]. Essa característica se faz muito útil em séries que possuem períodos de alta volatilidade se alternando com períodos de baixa volatilidade. Para um modelo genérico ARCH(q), seja ϵ_t o erro (resíduo) no instante t e α_0 um ruído branco, pode-se descrever a variância condicional de acordo com a Equação 2.7.

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

O modelo ARCH foi proposto por Robert Engle em 1982 para estimar a variância da inflação do Reino Unido [76]. A partir daí, várias derivações surgiram, como por exemplo: GARCH [11] por Tim Bollerslev [77] em 1986, EGARCH [12] por Daniel Nelson [78] em 1991, NGARCH [13] por Matthew Higgins *et al.* [79] em 1992, TGARCH [14] por Roger Rabemananjara *et al.* [80] em 1993, dentre outros. Alguns dos modelos da família ARCH podem ser encontrados nos trabalhos de Philip Franses *et al.* [81], de Juri Marcucci [82] e de Dima Alberg *et al.* [83].

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 [84], porém ainda em um estágio embrionário

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

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

[11]Generalised ARCH.

[12]Exponential Generalised ARCH.

[13]Non-linear Generalised ARCH.

[14]Threshold Generalised ARCH.

devido às dificuldades de processamento computacional e de acesso a dados na época. Por ser uma área de estudo extremamente dependente de ambas as questões, conforme elas foram evoluindo, mais trabalhos puderam ser realizados sobre o tema.

Isaac Nti *et al.* [85] 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[15] (*Artificial Neural Networks*) e SVM[16] (*Support Vector Machine*).

De forma semelhante, Dattatray Gandhmal *et al.* [86] verificou que a partir de uma análise detalhada de 50 trabalhos com o tema de predição do mercado financeiro, os algoritmos que mais costumam trazer resultados efetivos são ANN e técnicas baseadas em lógica *Fuzzy*[17]

É possível encontrar também modelos híbridos, com uma combinação de GARCH com ANN feita por Melike Bildirici *et al.* [87].

[15]Em português: Redes Neurais Artificiais.

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

[17]Em português: Difuso.

Capítulo 3

Metodologia e Implementação

3.1 Resumo

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

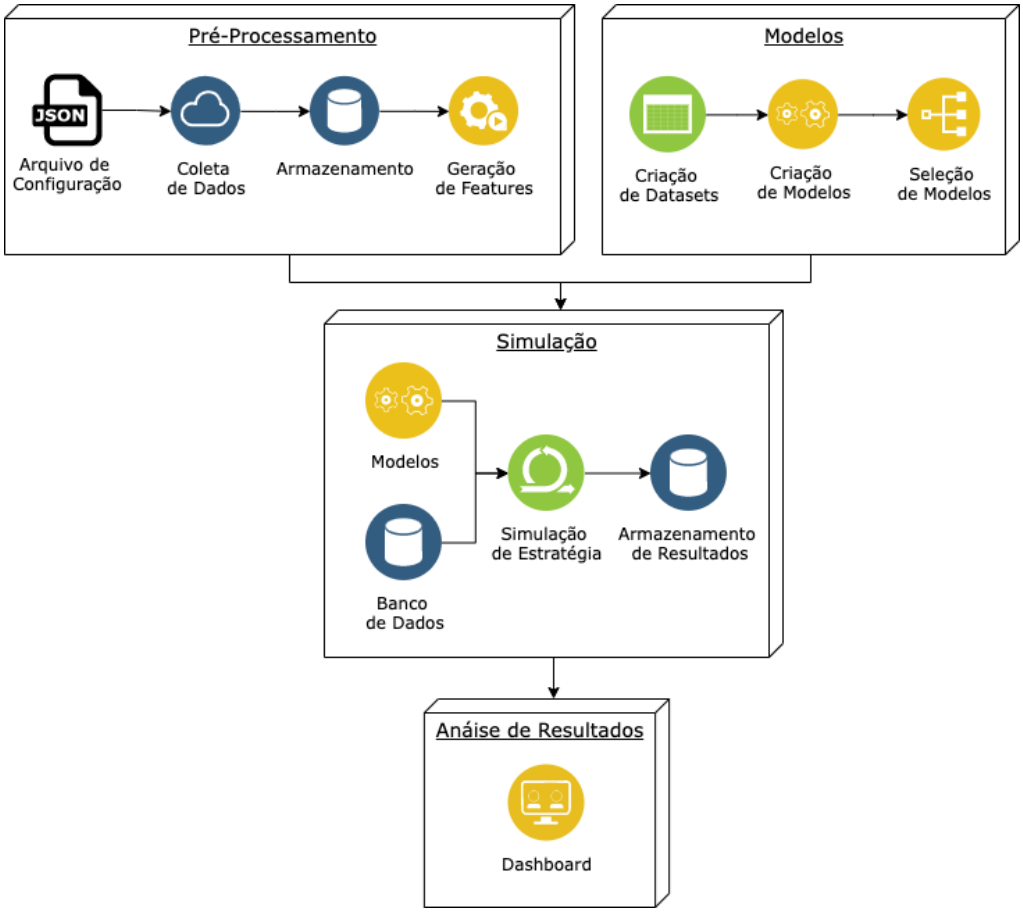

Figura 3.1: Estrutura do técnica do projeto

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

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

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, que ao final salva os resultados e as estatísticas no banco para posterior análise.

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

Em relação às tecnologias utilizadas, a aplicação foi desenvolvida em *Python* com o apoio das bibliotecas *yfinance*, *pandas*, *numpy*, *scikit-learn*, *multiprocessing*, *matplotlib* e *dash*. Foi estruturado um banco de dados PostgreSQL [88] para armazenamento dos *candlesticks* obtidos, das *features* geradas e das estratégias simuladas. Também foi incorporado o uso de *Docker* especificamente para a execução de estratégias sem a necessidade de configuração de ambiente.

A Figura 3.2 mostra a sequência lógica de refinamento dos parâmetros de simulação, abordada através das Seções 3.4.3, 3.4.4 e 3.4.6, respectivamente. Nota-se que o valor de saída de um parâmetro é utilizado durante o refinamento parâmetro do seguinte. Com exceção do Período Máximo de Dias por Operação, que utilizou o

intervalo de 2016 a 2018, os outros parâmetros foram refinados a partir de simulações no intervalo de janeiro de 2019 a março de 2020, denominado intervalo de refinamento de parâmetros de estratégia. Já o resultado da simulação final, disponível na Seção 4, utiliza o intervalo de abril de 2020 a dezembro de 2021, não havendo assim sobreposição.

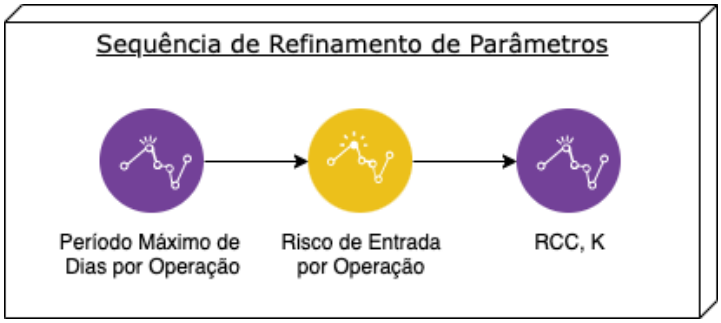

Figura 3.2: Sequência de Refinamento de Parâmetros

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

Ações Escolhidas (71)							
ABEV3	ALPA4	AMER3	B3SA3	BBAS3	BBDC3	BBDC4	BBSE3
BEEF3	BPAN4	BRAP4	BRFS3	BRKM5	BRML3	CCRO3	CIEL3
CMIG4	COGN3	CPFE3	CPLE6	CSAN3	CSNA3	CVCB3	CYRE3
DXCO3	ECOR3	EGIE3	ELET3	ELET6	EMBR3	ENBR3	ENEV3
ENGI11	EQTL3	EZTC3	FLRY3	GGBR4	GOAU4	GOLL4	HYPE3
ITSA4	ITUB4	JBSS3	JHSF3	LAME4	LCAM3	LREN3	MGLU3
MRFG3	MRVE3	MULT3	PETR3	PETR4	POSI3	PRIO3	QUAL3
RADL3	RENT3	SANB11	SBSP3	SULA11	TAEE11	TIMS3	TOTS3
UGPA3	USIM5	VALE3	VIIA3	VIVT3	WEGE3	YDUQ3	

Tabela 3.1: Ações Escolhidas

Os código fonte do projeto pode ser encontrado em seu repositório online no Github [89].

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 3.3 mostra sua estrutura.

Figura 3.3: Estrutura do Arquivo de Configuração

Nota-se que no topo são listados os parâmetros de uso geral, ou variáveis de escopo global, cujos valores precedem quaisquer outros listados na sequência, em caso de sobreposição. Em seguida abre-se o vetor de tipos de estratégias, onde o campo *name* representa o nome da classe selecionada, sendo este o elemento que conecta o usuário ao tipo de estratégia desejada. Ressalta-se que este trabalho compreende apenas um tipo de estratégia, embora o arquivo permita a leitura de qualquer nome. Na sequência, são configurados os parâmetros internos da estratégia. A Tabela 3.12 da Seção 3.4.7 descreve todos os parâmetros disponíveis.

Para se criar mais de um perfil de simulação, isto é, uma única execução do programa executar várias simulações via multiprocessamento, é necessário modificar o Arquivo conforme a Figura 3.4. Automaticamente, o código interpreta que existe mais de uma simulação a executar, com todos os parâmetros em comum exceto aqueles em formato de listas. Caso haja mais de um parâmetro no formato de lista, seus comprimentos precisam ser iguais. No caso da Figura 3.4, 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 3.4: Arquivo de Configuração para Execuções Múltiplas

3.2.2 Coleta de Dados

A Coleta de Dados ocorre através da biblioteca *open-source yfinance* [90], uma ferramenta não oficial que transmite dados obtidos através de APIs públicas da plataforma *Yahoo! Finance* [91], um subsistema da rede *Yahoo!*.

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

- Os valores de proventos que a biblioteca disponibiliza não são consistentes com

as declarações dos sites das próprias companhias, portanto não podem ser utilizados por este projeto. Testes internos confirmaram a presença de diversos proventos corretamente apresentados e ajustados pelos respectivos desdobramentos acumulados. O problema é que os mesmos estavam misturados com alguns *outliers* inexistentes na realidade, o suficiente para questionar o uso em escala (*i.e.*, para vários ativos sem verificação individual). O Apêndice A evidencia os problemas encontrados em mais detalhes.

- Até onde se pode verificar, os volumes de negociação disponibilizados coincidem em valores relativos com os volumes da plataforma *TradingView* [92], não tendo sido encontrada evidência do contrário. Em outras palavras, a variação percentual de volume entre dois pregões de um mesmo ativo é igual em ambas as plataformas.

- *Candlesticks* de janelas temporais inferiores à diária (*intraday*) são disponibilizados, porém o limite de busca de 730 dias inviabiliza seu uso.

Apenas os dados não existentes no banco são baixados via *yfinance*. Para isso, um *trigger*[1] é acoplado às tabelas de *candlesticks* e acionado sempre que operações de *insert*, *update*, *delete* e *truncate* são realizadas. Quando ativado, ele chama uma função responsável por atualizar a tabela de *status*, que registra o intervalo de tempo representado nas tabelas de *candlesticks* para cada *ticker* envolvido. Deve-se ressaltar que os devidos cuidados foram tomados para evitar buracos entre intervalos de tempo não adjacentes. Portanto, apenas uma consulta à tabela de *status* é executada para se verificar a necessidade de *download* de novos dados.

3.2.3 Armazenamento de Dados

O Armazenamento de Dados é realizado por um banco de dados *PostgreSQL*, criado com o objetivo de salvar: os resultados das simulações; as *features* de uso geral; e os *candlesticks* obtidos. As vantagens de um banco de dados em relação

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

a um arquivo CSV ou a uma planilha de Excel dispensam comentários. Contudo, quanto ao escopo deste trabalho, pode-se mencionar os seguintes pontos:

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

A figura 3.5 mostra o ERD[2] do banco de dados. Os *scripts* de criação e população inicial, bem como as *constraints* envolvidas podem ser encontrados em [93].

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

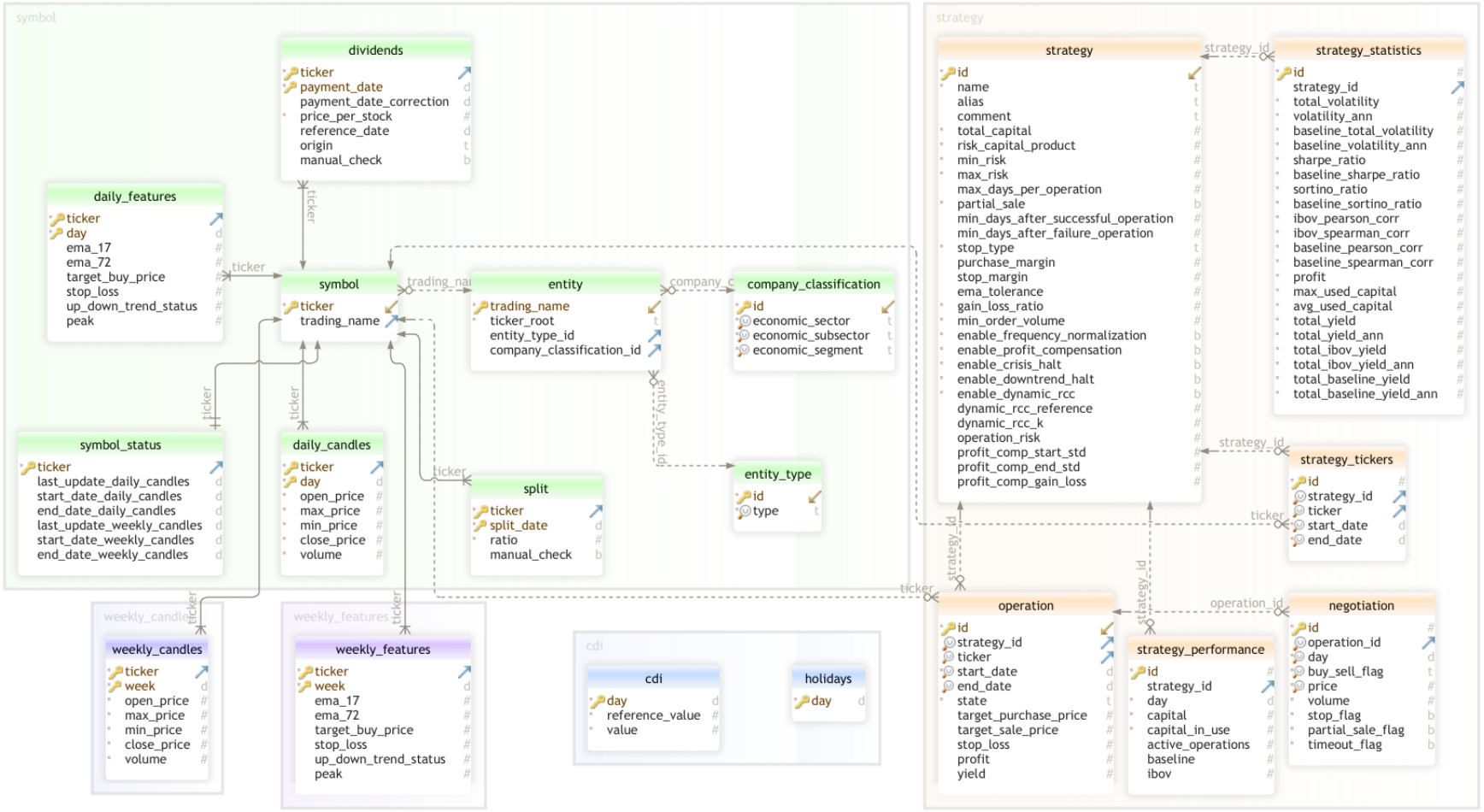

Figura 3.5: ERD do Banco de Dados

3.2.4 Geração de *Features* de Uso Geral

As *Features* de Uso Geral são características derivadas dos *candlesticks* que podem auxiliar qualquer decisão interna de uma estratégia. Mais especificamente, esta Seção aborda as *features* Risco Mínimo e Rísco Máximo, que possuem duas funções principais: auxiliar no processo de escolha do preços de compra, dos *stop loss* e dos preços alvo através do parâmetro Risco de Entrada por Operação; e auxiliar os modelos de ML a decidirem o momento apropriado de compra dos ativos, com base nos valores dos preços encontrados.

Devido à natureza genérica das *features*, que podem ser utilizadas por diversas estratégias, são calculadas antes do início das simulações e somente quando há necessidade, ou seja, quando os *candlesticks* são inseridos pela primeira vez no banco ou quando são atualizados. Ao final dos cálculos, são armazenadas nas tabelas de *features* para posteriores consultas durante as simulações.

O fato dos cálculos envolverem uma análise de dados do passado, é necessário garantir que as *features* geradas estejam sempre apontadas para um intervalo de tempo anterior ao dia no qual elas serão consumidas. Do contrário, um erro de não-causalidade poderia surgir, corrompendo a performance das simulações. Analogamente, durante a geração dos modelos (Seção 3.3.5) também é necessario se garantir que o período de treinamento é anterior ao período de teste, diretamente via a configuração das datas limites de cada período e indiretamente via remoção de datas cuja *features* possuem efeito memória, evitando assim que qualquer informação esteja presente simultaneamente dos dados de treinamento e de teste.

Antes de prosseguir, é necessário contextualizar o termo risco de entrada em uma operação, às vezes simplesmente chamado de risco. Este é caracterizado pela diferença percentual no qual o *stop loss* é colocado em relação ao do preço de compra. Por exemplo, se uma operação tem um preço de compra de R$10,00 e o *stop loss* é posicionado em R$9,00, diz-se o risco da operação é de 10%. A escolha do risco também determina o valor do preço alvo da operação, pois o mesmo é definido como 3 vezes a magnitude do risco escolhido, percentualmente e acima do preço de compra (ver Seção 3.4.1). Assim, no exemplo anterior, o preço alvo estaria situado

em R$13,00. Este assunto será novamente abordado na Seção 3.4.1, no entanto fez-se necessária uma antecipação pois esta terminologia é amplamente utilizada ao longo deste trabalho.

As *features* de Uso Geral utilizadas são:

- **Risco Mínimo**

  O Risco Mínimo é uma *feature* de suporte à escolha do risco de entrada em uma operação, não sendo assim consumido diretamente pelo modelo de ML, mas sim indiretamente. A fórmula é composta por uma parcela fixa somada a uma parcela variável, conforme mostrado pela Equação 3.1.

$$Risk_{min} = Risk_{min_f} + Risk_{min_v} \tag{3.1}$$

  Seja P_Δ a diferença entre o preço máximo e mínimo de um *candle* (Equação 3.2), pode-se definir $Risk_{min_f}$ como o valor mínimo de risco necessário para superar as oscilações diárias dos preços médios dos últimos 20 dias úteis (Equações 3.3 e 3.4). Nota-se que σ_{P_Δ} é o desvio padrão relativo aos últimos 20 dias úteis.

$$P_\Delta = P_{high} - P_{low} \tag{3.2}$$

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

$$Risk_{min_f} = \frac{\sigma_{P_\Delta}}{P_{mid}} \tag{3.4}$$

  A parcela variável $Risk_{min_v}$ está associada à tendência de queda de preço no curto prazo. Seu cálculo é realizado a partir da derivada de preços médios ajustada por um filtro digital IIR passa-baixas [94] (Equações 3.5, 3.6 e 3.7), onde α é o coeficiente de amortecimento. Observa-se que a derivada do preço médio foi normalizada para permitir que a tendência independa do valor absoluto do preço, tratando-se apenas de uma variação percentual. O sinal negativo na

Equação 3.7 indica que quanto maior a tendência de queda, maior precisa ser o risco associado.

$$\dot{P}_{mid(i)} = \frac{P_{mid(i)} - P_{mid(i-1)}}{\frac{1}{2}(P_{mid(i)} + P_{mid(i-1)})} \tag{3.5}$$

$$\dot{P}_{mid\_LPF(i)} = \alpha \dot{P}_{mid(i)} + (1-\alpha)\dot{P}_{mid\_LPF(i-1)}, \quad \text{onde} \quad 0 \leq \alpha \leq 1 \tag{3.6}$$

$$Risk_{min_v} = max(-\dot{P}_{mid\_LPF(i)}, 0) \tag{3.7}$$

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

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

As Figuras 3.6 e 3.7 e mostram os resultados do algoritmo para dois papéis de comportamentos distintos: MGLU3 representando um companhia com foco em alto crescimento, portanto mais volátil; e ABEV3 representando uma companhia já bem consolidada no mercado, portanto menos volátil.

- **Risco Máximo**

O Risco Máximo é uma *feature* de suporte à escolha do risco de entrada em uma operação, não sendo assim consumido diretamente pelo modelo de ML. Ressalta-se que o conceito de risco no escopo deste trabalho está relacionado à diferença de valor no qual o *stop loss* é colocado abaixo do preço de compra (Equação 3.17). A escolha do risco também implica no valor do preço alvo de uma operação, pois o mesmo é definido como 3 vezes a magnitude do risco escolhido, percentualmente acima do preço de compra (ver Seção 3.4.1). A ideia central está na análise estatística das subidas de preços entre os últimos picos identificados dentro do intervalo de 80 dias úteis.

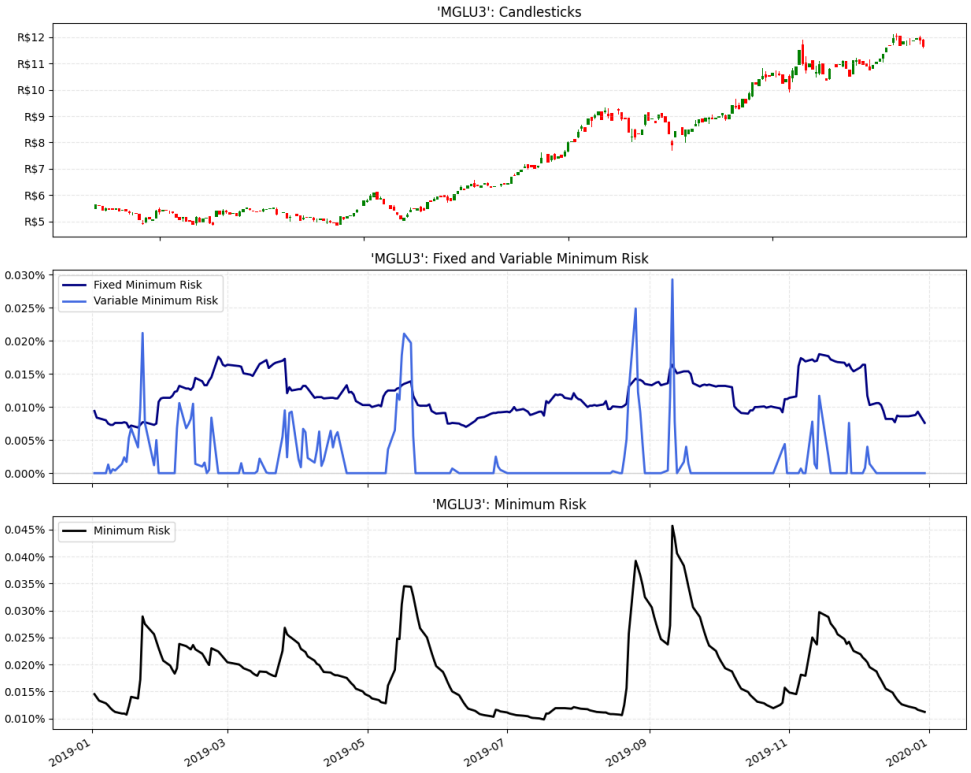

Figura 3.6: MGLU3 - Risco Mínimo (01/01/2019 a 31/12/2019)

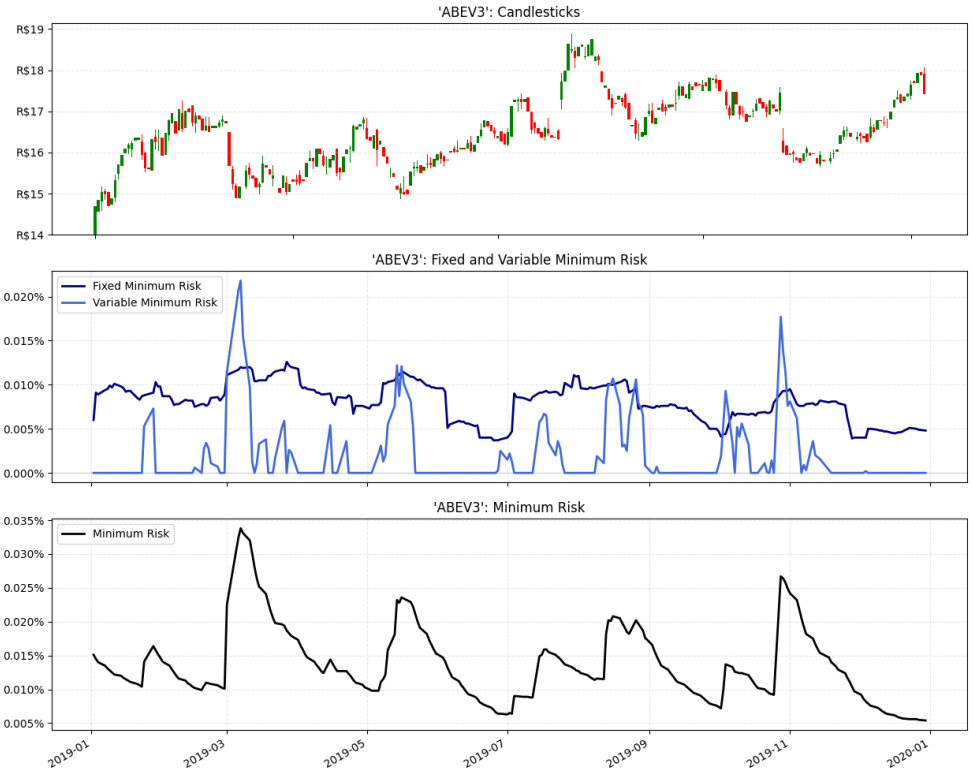

Figura 3.7: ABEV3 - Risco Mínimo (01/01/2019 a 31/12/2019)

Para se iniciar o cálculo, primeiro é necessário a criação de um um algoritmo de identificação de picos, conforme mostrado pela Figura 3.8. O método usa uma janela móvel de $W = 5$ *candles* que corre dia após dia até a data corrente e atribui 1 voto de máximo e 1 voto de mínimo aos preços de máximo e preços de mínimo encontrado na janela, respectivamente. Em todos os passos, o primeiro e o último *candle* da janela nunca recebem votos devido à falta de um *candle* adjacente. São elegíveis à picos apenas os *candles* que obtiveram um mínimo de $floor(W/2) = 2$ votos. Ao final, garante-se a alternância entre máximos e mínimos locais através da remoção de picos consecutivos de um mesmo tipo.

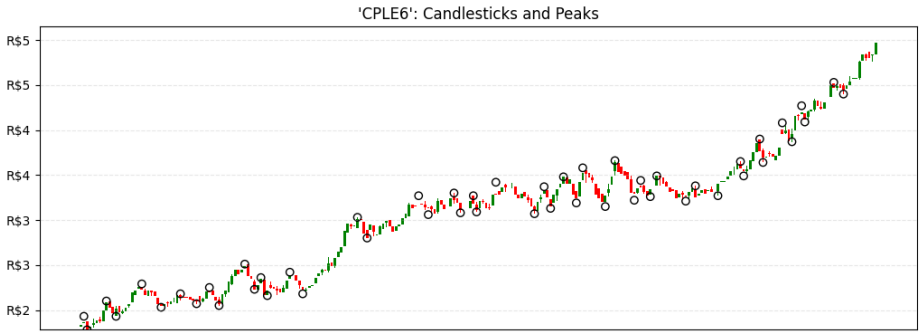

Figura 3.8: CPLE6 - Algoritmo de identificação de picos (01/01/2019 a 31/12/2019)

Depois da identificação de picos, extraem-se as n subidas de preços de cada mínimo para o máximo consecutivo no período designado, normalizados pelo pico de mínimo (Figura 3.9 e Equação 3.8). Por fim, o Risco máximo é obtido pelo cálculo da média $\overline{C_{(i)}}$ com um filtro digital IIR passa-baixas (Equações 3.9 e 3.10).

$$c_k = (P_{max(k)} - P_{min(k)})/P_{min(k)}, \quad \text{onde } 0 < k \leq n \tag{3.8}$$

$$\overline{C_{(i)}} = \frac{1}{n}\sum_{k=1}^{n} c_k \tag{3.9}$$

$$Risk_{max(i)} = \alpha\overline{C_{(i)}} + (1-\alpha)Risk_{max(i-1)} \tag{3.10}$$

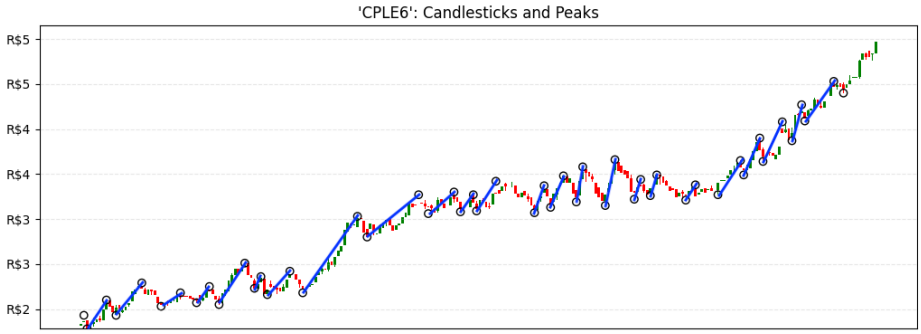

Figura 3.9: CPLE6 - Subidas de preços entre picos (01/01/2019 a 31/12/2019)

$$\overline{C_{LPF(i)}} = \alpha\overline{C_{(i)}} + (1-\alpha)\overline{C_{(i-1)}} \tag{3.11}$$

$$Risk_{max(i)} = \frac{\overline{C_{LPF(i)}} - 0.5\sigma_C}{G} \tag{3.12}$$

Foi utilizado $\alpha = 0.50$.

As Figuras 3.10 e 3.11 mostram o Risco Máximo para os ativos MGLU3 e ABEV3, respectivamente.

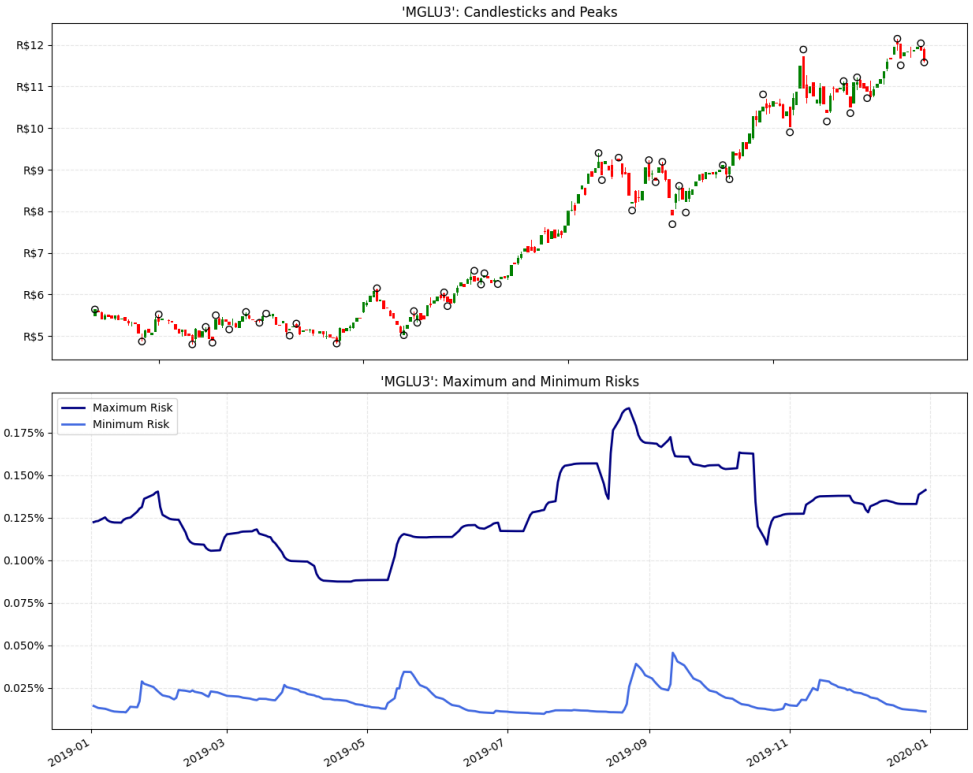

Figura 3.10: MGLU3 - Riscos Máximo e Mínimo (01/01/2019 a 31/12/2019)

Figura 3.11: ABEV3 - Riscos Máximo e Mínimo (01/01/2019 a 31/12/2019)

3.3 Modelos de Aprendizado Supervisionado

3.3.1 Resumo

Conforme mostrado na Seção 2.6.3, a literatura aponta para o uso de modelos dos tipos ANN, SVM e lógica *Fuzzy*. No entanto, escolheu-se *Random Forests* devido à sua baixa tendência de *overfitting* e à ausência de necessidade de escalamento das *features* consumidas. Apesar de ser um modelo com várias possibilidades de hiperparâmetros a serem ajustados, a prática mostra que poucos são aqueles que realmente trazem uma melhora de performance significativa, uma vez que uma RF que tenha atingido um número alto o suficiente de árvores.

A partir de *datasets* previamente populados, modelos são gerados para cada ação a cada intervalo de 3 meses de simulação (WFA). Um critério particular de performance foi criado para ranquear os melhores modelos, que são filtrados por uma varredura de alguns hiperparâmetros.

O período de treinamento e de teste de cada modelo foi separado utilizando WFA e pode ser verificado pela Tabela 3.2. Entende-se por validade o período de tempo durante a simulação no qual o algoritmo criado pode atuar. A Tabela também possui uma linha horizontal tracejada que separa os intervalos: de refinamento de parâmetros de estratégia; e de execução da simulação final.

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

Tabela 3.2: WFA - Intervalos de treinamento, teste e validade dos modelos

3.3.2 *Datasets* e *Feature Selection*

Os *datasets* são arquivos CSV criados para cada *ticker* através de uma varredura da série histórica. Registra-se dia após dia as *features* acumuladas e o resultado de uma operação hipotética iniciada no dia corrente. A Tabela 3.3 mostra a lista das *features* relevantes do arquivo, onde as linhas marcadas em negrito indicam as colunas utilizadas na entrada de dados dos modelos. A coluna Resultado da Operação indica a saída observada para o treinamento supervisionado.

O termo preço médio se refere ao definido pela Equação 3.3. Da mesma forma, a derivada do preço médio é indicada pela Equação 3.5. 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 do valor absoluto do preço de um ativo. 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.

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 3.3: Comparação de Resultados

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

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

3.3.3 Índice de Lucratividade

Durante a etapa de criação dos modelos, é necessário uma métrica para ranqueamento das performances de treinamento e de teste que acompanhe o contexto do problema. O impacto de um acerto por parte do modelo implica em um ganho de

3X para a carteira, onde X é o valor de entrada da operação. Assim como uma falha implica em uma perda de -X para a carteira.

O Índice de Lucratividade é um coeficiente entre 0 e 1 onde 0 significa o pior resultado possível, isto é, aquele no qual o modelo errou todas as operações no *dataset* de forma a trazer o maior prejuízo relativo, supondo que todos os aportes financeiros sejam fixos. Por outro lado, 1 significa o maior lucro que o modelo pode trazer se acertar todas as operações que o *dataset* permite e não errar nenhuma outra. Nota-se que o valor do Índice que representa o lucro zero não é necessariamente 0.5, mas sim algum valor intermediário que precisa ser calculado e varia para cada *dataset*. Portanto, essa métrica é útil para comparação de modelos que utilizem exatamente a mesma fonte de dados.

Seja A o número de operações de sucesso no *dataset* (classe 1) e B o número de operações de falha (classe 0), pode-se definir S_a como a soma dos riscos de todas as operações da classe 1 e S_b a soma dos riscos de todas as operações da classe 0. Como os aportes são fixos, as somas dos riscos é na verdade sinônimo do lucro ou do prejuízo acumulado. Assim, a Equação 3.13 representa uma função linear responsável por mapear o lucro L_m de um modelo no Índice de Lucratividde.

$$I_L = \frac{1}{(S_a - S_b)} L_m - \frac{S_b}{(S_a - S_b)} \tag{3.13}$$

A partir da matriz de confusão na Tabela 3.4, pode-se definir o lucro L_m através da Equação 3.14, onde VP é a contagem de verdadeiros positivos do modelos e FP é a contagem de falsos positivos. Cada VP representa uma tentativa bem sucedida de operação que já daria certo, portanto há um ganho de +3X, onde X é um valor genérico irrelevante. Cada FP, por outro lado, representa uma tentativa de sucesso em uma operação que na verdade falhou, contribuindo para uma perda de -X. Falsos negativos (FN) e verdadeiros negativos (VN) são desconsiderados do cálculo pois como não geram operações, uma vez que a aposta está na crença de que a suposta operação falharia, nenhum aporte é feito.

$$L_m = 3 \times VP - FP \tag{3.14}$$

Classe		Esperada	
		Positivo (1)	Negativo (0)
Real	Positivo (1)	VP (+3X)	FN (0X)
	Negativo (0)	FP (-X)	VN (0X)

Tabela 3.4: Matriz de Confusão

Por fim, o valor do Índice que representa o lucro zero pode ser encontrado a partir da Equação 3.13, basta impor a condição $L_m = 0$ (Equação 3.15).

$$I_{L_0} = -\frac{S_b}{(S_a - S_b)} \tag{3.15}$$

3.3.4 Balanceamento de Classes

Devido ao desbalanceamento dos *datasets* utilizados, que pode chegar até cerca de 90% de operações de falha contra 10% de operações de sucesso, faz-se imprescindível alguma técnica de balanceamento mencionada na Seção 2.3.3: *undersampling*, *oversampling* ou CSL. A natureza do problema também deve ser levada em consideração, pois um simples balanceamento via CSL daria a mesma importância à predição dos acertos e das falhas, no entanto o impacto de um acerto por parte do modelo gera um ganho de +3X para a carteira e uma falha gera uma perda de -X para os mesmos X aportados em uma operação genérica, conforme discutido anteriormente na Seção 3.3.3.

Levando em consideração as questões levantadas, optou-se por um balanceamento via CSL levando em considerações o número de amostras e o impacto das classes para a carteira (Tabela 3.5).

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

Tabela 3.5: Balanceamento via CSL

3.3.5 Geração de Modelos

Com o auxílio da biblioteca *Scikit-Learn* [95], a escolha do melhor modelo para um determinado, diversos outros são criados e excluídos a fim de se selecionar apenas o melhor

Os modelos *Random Forest* foram criados com o auxílio da biblioteca *Scikit-Learn* [95]. Para a escolha do modelo final, diversos outros são criados e excluídos a fim de se selecionar apenas o melhor. Todos modelos temporários compartilham os hiperparâmetros fixos indicados pela Tabela 3.6, porém diferem quanto aos hiperparâmetros variáveis da Tabela 3.7 ou quanto às sementes aleatórias utilizadas. A lógica de criação pode ser analisada da seguinte forma:

- Para cada um dos 71 *tickers* da carteira (Tabela 3.1) são gerados 12 modelos de acordo com a Tabela 3.2. Por trás de cada deles são avaliados 100 modelos temporários e organizados em dois níveis: os hiperparâmetros variáveis e sementes aleatórias.
- O nível de diferenciação dado pelos hiperparâmetros variáveis da Tabela 3.7 gera 10 pares de combinações entre *max_depth* e *max_features*: (3, 3), (3, 4), (4, 3), (4, 4), ..., (7, 4).
- Para cada par criado, são criados 10 modelos com sementes aleatórias diferentes.
- O modelo final escolhido possui os mesmos hiperparâmetros do grupo que possuir a maior razão de Índice de Lucratividade pelo desvio padrão do mesmo: I_L/σ_{I_L}.

A Figura 3.12 ilustra a lógica de criação dos modelos apresentada através de um diagrama.

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

Tabela 3.6: Hiperparâmetros fixos

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

Tabela 3.7: Hiperparâmetros variáveis

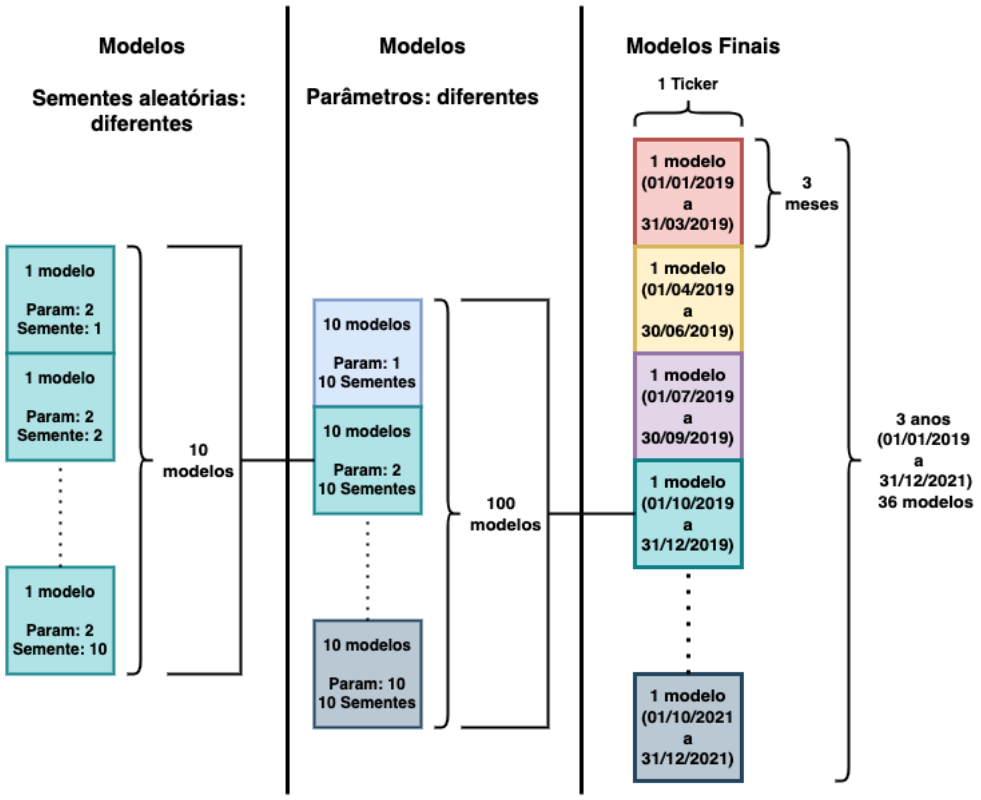

Figura 3.12: Diagrama de criação de modelos

3.3.6 Modelo *Baseline*

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

As Figuras 3.13 e 3.14 mostram o *baseline* calculado para os 71 *tickers* indicados na Tabela 3.1 no intervalo de refinamento dos parâmetros de simulação (01/01/2019 a 31/03/2020) e no intervalo das simulações finais (01/04/2020 a 31/12/2021), respectivamente. Adicionou-se o iBovespa (Seção 2.1.2) e o CDI[3] acumulado para fins de comparação. Nota-se a diferença de performance do *baseline* para o iBovespa, o que é razoável já que o *baseline* tem composição fixa e o iBovespa tem uma composição variável tanto na escolha dos ativos quanto em seus respectivos pesos. A Tabela 3.8 mostra os indicadores de performance para ambos os *baselines*.

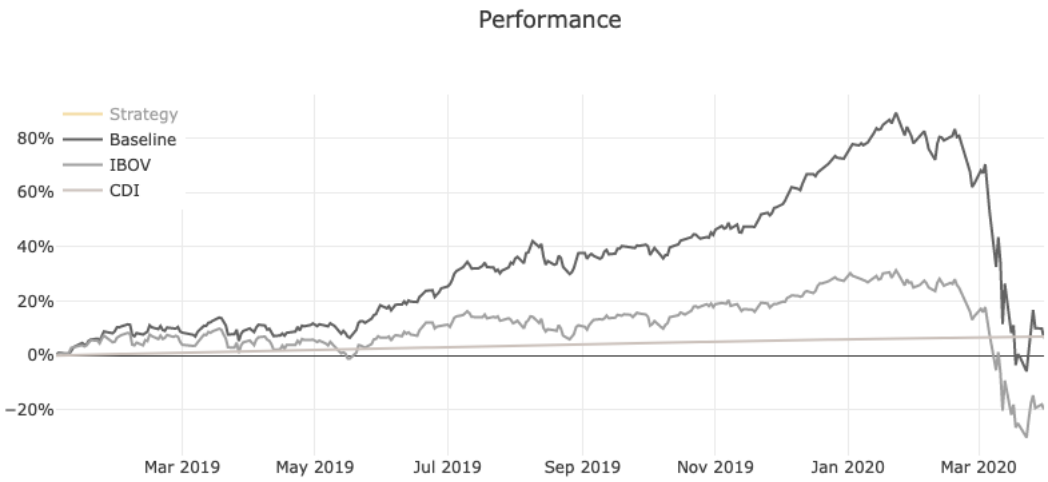

Figura 3.13: Rendimento do *baseline* para o intervalo de 01/01/2019 a 31/03/2020

[3]Certificado de Depósito Interbancário.

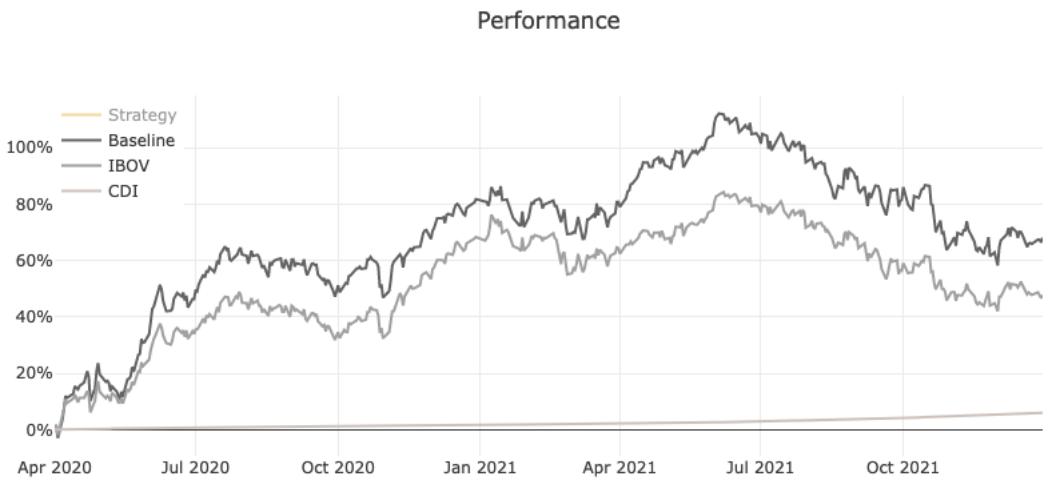

Figura 3.14: Rendimento do *baseline* para o intervalo de 01/04/2020 a 31/12/2021

Início	Fim	Rend. Final	Volatilidade	Sharpe	Sortino
01/01/2019	31/03/2020	6,39%	43,98%	0,20	0,18
01/04/2020	31/12/2021	68,07%	34,84%	1,15	1,71

Tabela 3.8: *Baseline* - Indicadores de Performance

3.4 Simulação de Estratégia

3.4.1 Estrutura

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

- Toda estratégia possui um **capital inicial**, que representa uma quantidade de capital pré-alocado para compra dos ativos financeiros. Essa quantia deve ser sempre respeitada ao longo da simulação de forma a não representar nunca um valor negativo.

- Toda estratégia deve possuir uma **carteira de ativos** (ou lista de ativos) com as respectivas datas iniciais e finais de validade, isto é, intervalos de tempo onde as operações podem ser realizadas. Embora sejam permitidos intervalos

diferentes, é convencionado a mesma data de início e de fim para todos os papéis.

- Define-se uma **operação** como o processo de compra única de um volume de ações de um ativo seguido pela venda de todo o volume comprado, independentemente do tempo, mesmo que esta ocorra em estágios. Nota-se que apenas a venda é cabível de ocorrer em estágios (*i.e.*, venda parcial).

- Toda operação possui um **preço alvo** e um ***stop loss***. O preço alvo é um valor acima do preço de compra e o *stop loss* é um valor abaixo do preço de compra. Quando o mercado atinge qualquer um dos dois valores, uma venda é disparada, encerrando a operação em vigor. No entanto, considera-se uma operação de sucesso aquela que encerrou por atingir o preço alvo e uma operação de falha aquela que encerrou por atingir o *stop loss*.

- Uma estratégia pode possuir no máximo **uma operação em vigência** para cada *ticker* em sua bolsa de ativos, portanto para que uma segunda compra ocorra no momento em que já existem papéis adquiridos, é necessários vendê-los primeiro.

- A **razão entre ganho e perda**, também denominada relação risco/ganho, é definida por André Moraes [40] e predetermina a relação entre o preço alvo e o *stop loss* em qualquer operação. Ela indica a razão entre a diferença do preço alvo P_{target} para o preço de compra P_{buy} sobre a a diferença do preço de compra para o *stop loss* P_{stop} (Equação 3.16). O valor recomendado é 3, portanto este é o valor que será utilizado em todo o escopo deste trabalho como uma constante.

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

  Utiliza-se o termo "risco de uma operação" ou simplesmente "risco" como sendo a diferença de valor no qual o *stop loss* é colocado abaixo do preço de compra (Equação 3.17). Por exemplo, se o preço de compra de uma operação é de R$10,00 e o seu risco é de 5%, então o *stop loss* se encontra em R$9,50 e o preço alvo em R$11,50 necessariamente.

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

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

3.4.2 Premissas

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

- O momento de decisão de **entrada em uma operação** por uma estratégia ocorre durante a abertura de mercado do dia corrente, mais precisamente no instante em que o preço de abertura é definido.
- No dia que houver a compra de um ativo, não pode haver a venda do mesmo. Em outras palavras, o **período mínimo de duração de uma operação é de 1 dia útil**.
- A **venda por *timeout*** ocorre quando o número máximo de dias de uma operações extrapola um valor definido (ver Seção 3.4.3)
- Devido a ausência de informações mais detalhadas que a janela de tempo diária, a seguinte ordem é priorizada durante a **venda de um ativo**:

  1. Venda por *stop loss*
  2. Venda parcial (caso habilitada)
  3. Venda por preço alvo
  4. Venda por *timeout*

- Caso algum **preço de venda seja pulado**, ou seja, a descontinuidade entre o preço de abertura do dia corrente e o preço de fechamento do dia anterior não englobe o valor de venda, utiliza-se o preço de abertura do dia corrente. A única exceção acontece para a venda por *timeout*, já que se trata de uma venda compulsória que sempre ocorre no preço de fechamento do dia designado.

3.4.3 Período Máximo de Dias por Operação

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

Nesta linha, um algoritmo auxiliar foi criado para varrer um período de dias passados e criar operações com diversos valores de risco, observando quais riscos levariam a operações de sucesso e quais levariam a operações de falha. Também analisou-se a distribuição de operações de sucesso de acordo com os valores de risco e o intervalo de dias corridos.

De início, foi fixado um intervalo máximo de 90 dias para cada operação hipotética que o algoritmo gerou. O valor é propositalmente excessivo, pois sua função é apenas não forçar *timeout* na maioria das operações. As Figuras 3.15 e 3.16 mostram dois histogramas dos dias das operações de sucesso que consideram o menor risco possível, isto é, o menor valor de risco que se pode utilizar a cada dia da série temporal de forma a tornar a operação um sucesso, caso ele exista. Se não existir, é considerado operação de falha, portanto está fora dos histogramas. A legenda indica faixas onde, no caso da linha tracejada em verde na Figura 3.15, 50% das contagens se encontram dentro dos 12 primeiros dias, e assim por diante.

Também foram analisados os histogramas de risco ótimo por operação, ou seja, o valor de risco que traz o melhor rendimento por operação considerando os dias corridos (Figuras 3.17 e 3.18).

A Figura 3.19 mostra a distribuição de todas as operações de sucesso possíveis no intervalo selecionado. Obseva-se que valores baixos de risco não costumam estar associados a longos períodos de operação e que a maior densidade de operações de sucesso se encontra em baixo risco e em baixa duração de operação.

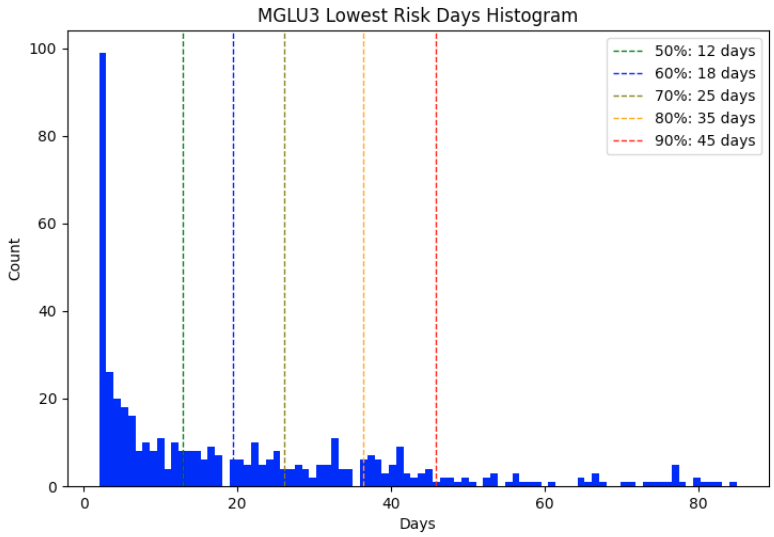

Figura 3.15: MGLU3 - Histograma de dias com risco mínimo em operações de sucesso (01/01/2016 a 31/12/2018)

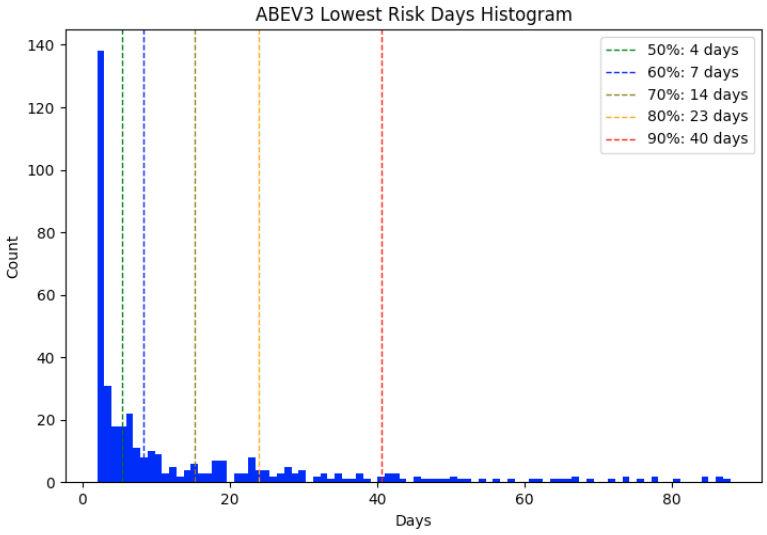

Figura 3.16: ABEV3 - Histograma de dias com risco mínimo em operações de sucesso (01/01/2016 a 31/12/2018)

A Tabela 3.9 resume o período de dias que engloba 90% das contagens dos histogramas conforme indicado nas Figuras 3.15, 3.16, 3.17 e 3.18.

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

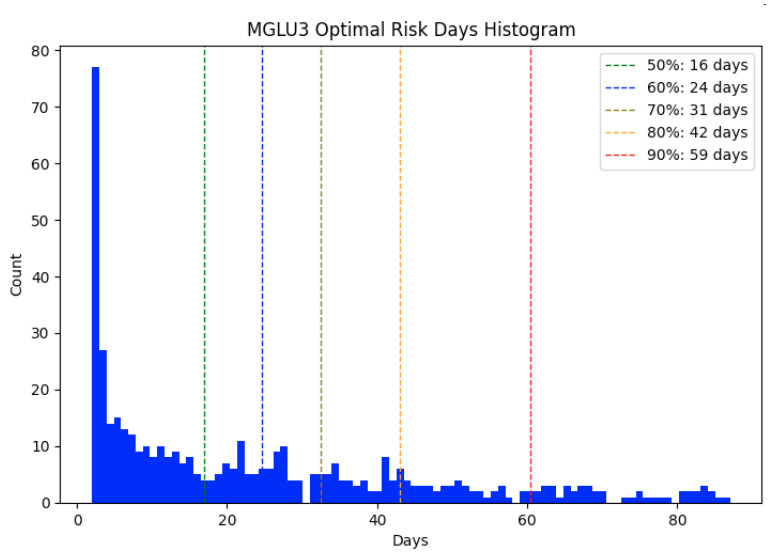

Figura 3.17: MGLU3 - Histograma de dias com risco ótimo em operações de sucesso (01/01/2016 a 31/12/2018)

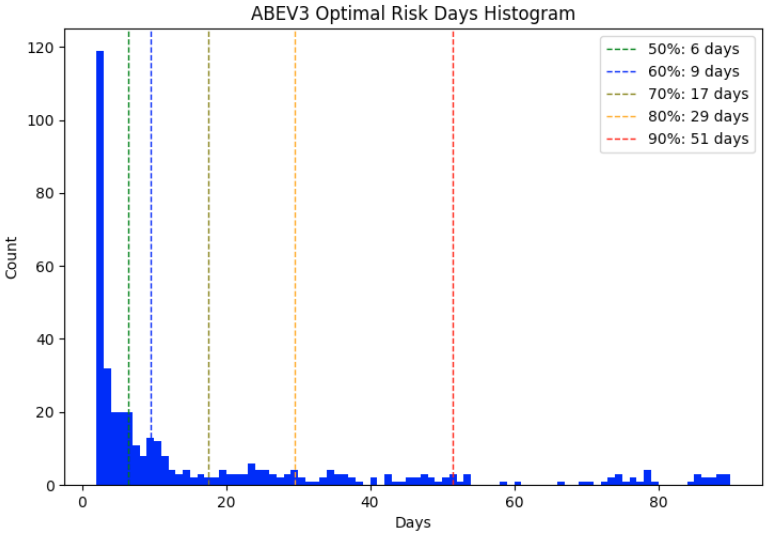

Figura 3.18: ABEV3 - Histograma de dias com risco ótimo em operações de sucesso (01/01/2016 a 31/12/2018)

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

Tabela 3.9: Período de dias que engloba 90% das contagens dos histogramas

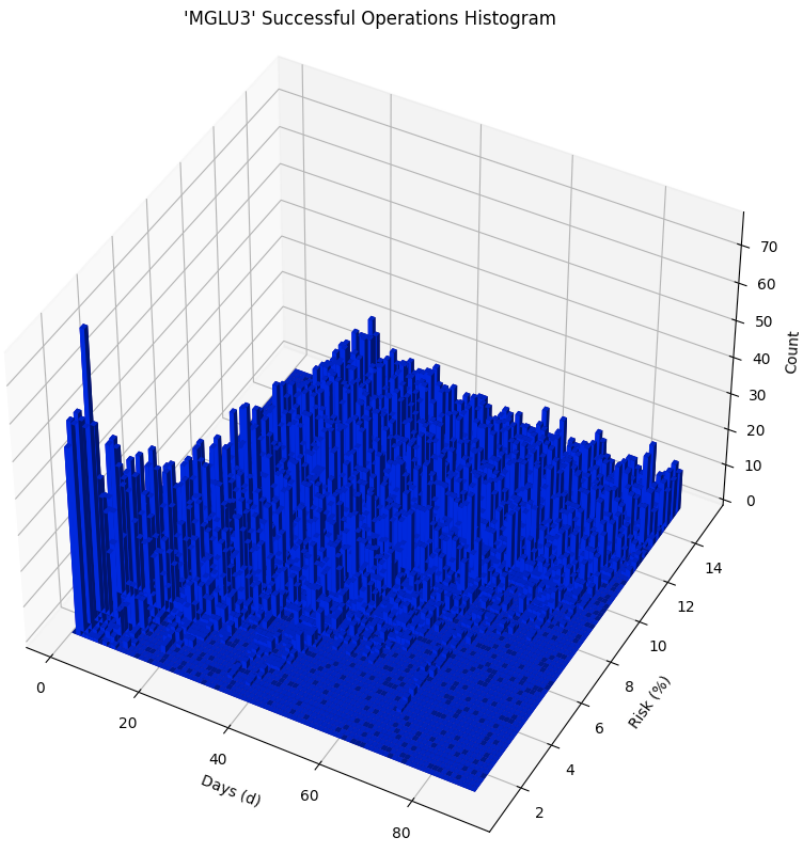

Figura 3.19: MGLU3 - Histograma de todas as operações de sucesso (01/01/2016 a 31/12/2018)

3.4.4 Risco de Entrada por Operação

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

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

Para a maioria dos casos, tem-se $Risk_{max} > Risk_{min}$. O valor do Risco de Entrada por Operação ($Risk_{operation}$) é definido pelo parâmetro $Risk_{coef}$ de acordo com a Equação 3.18, onde $0 \leq Risk_{coef} \leq 1$.

$$Risk_{operation} = Risk_{min} + Risk_{coef}(Risk_{max} - Risk_{min}) \tag{3.18}$$

A Figura 3.20 resume as simulações cujos valores de Risco de Entrada estão no intervalo fechado [0.10, 0.70]. Foram utilizados os 71 *tickers* da Tabela 3.1 no período de 01/01/2019 a 31/03/2020, além do Período Máximo de Dias por Operação de 45 dias (refinado na Seção 3.4.3). Utilizou-se um RCC de 0,10% apenas para evitar saturação de capital. Nota-se que o ponto de máximo para o índice de Sharpe e o índice de Sortino coincidiu para o Risco de Entrada de 0,43, que será o valor escolhido para as simulações seguintes. Apesar do ponto de máximo do rendimento final indicar o parâmetro de 0,25, este valor não será considerado pois o índice de Sharpe é uma alternativa mais robusta que engloba o anterior e considera também outros efeitos.

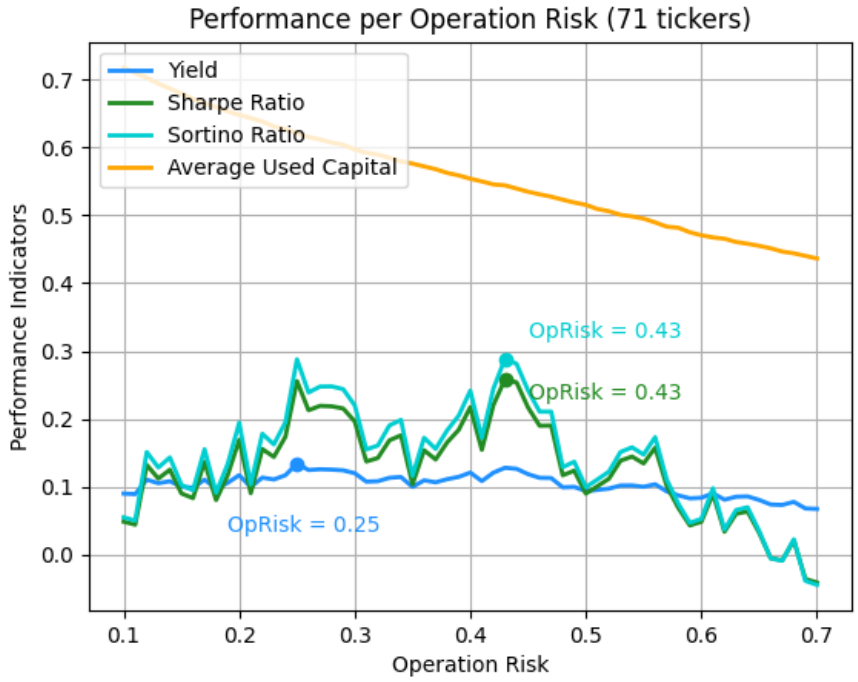

Figura 3.20: Indicadores de performance em função do risco de entrada (71 tickers, 01/01/2019 a 31/03/2020)

3.4.5 Gerenciamento de Risco

Segundo André Moraes [40], um bom Gerenciamento de Risco é essencial para a performance de uma estratégia. Afinal, não adianta obter uma alta taxa de acerto em operações de cujo lucro médio não compense as perdas acumuladas pelas operações que falham. Além disso, estar com o capital muito alocado em ativos de um único segmento é perigoso devido à exposição à fatores como falta de insumos industriais, mudanças na legislação, crises internas, instabilidade política, dentre outros.

Para mitigar as questões levantadas, algumas medidas foram tomas inspiradas no trabalho de André Moraes [40]. São elas:

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

O Coeficiente de Risco-Capital, definido pela Equação 3.19, 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 3.12) e vale para todos os ativos da carteira.

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

Durante uma simulação, a estratégia primeiro encontra o valor do risco desejado para entrar na operação, depois escolhe o capital a ser alocado. Dessa forma, a Equação 3.20 mostra de fato a aplicação do RCC. É evidente que quanto maior o risco envolvido, menor o capital a ser alocado e vice-versa.

$$Capital = \frac{RCC}{Risk} \tag{3.20}$$

O RCC influencia diretamente no uso médio de capital de uma estratégia. Por um lado, um uso de capital baixo significa um mau aproveitamento do capital, o que leva a uma performance ruim. Por outro lado, muito uso de capital implica

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

em pouco capital disponível para entrada em novas operações, o que também leva a uma performance ruim e instável, visto que a operação que iniciar imediatamente antes das demais leva quase todos o capital consigo.

A Figura 3.21 mostra o gráfico da relação entre os indicadores de performance em função do RCC para diversas simulações. Foi considerado o período de 01/01/2019 a 31/03/2020, os 71 *tickers* da Tabela 3.1, o Risco de Entrada por Operação de 0,43 (refinado na Seção 3.4.4) e o Período Máximo de Dias por Operação de 45 dias (refinado na Seção 3.4.3). Notam-se duas regiões de máximos locais no início e no fim do gráfico, onde os valores de RCC são 0,11% e 6,10%, respectivamente. Percebe-se também que a partir do RCC de 1,00%, o uso médio de capital por estratégia começa a saturar e o número percentual de operações totais começa a cair significativamente. Este efeito ocorre devido à falta de capital disponível para novas operações em momentos de saturação. Ou seja, quanto mais a direita no gráfico, menos operações são efetuadas e maior é o desvio padrão do capital de aporte nas operações restantes.

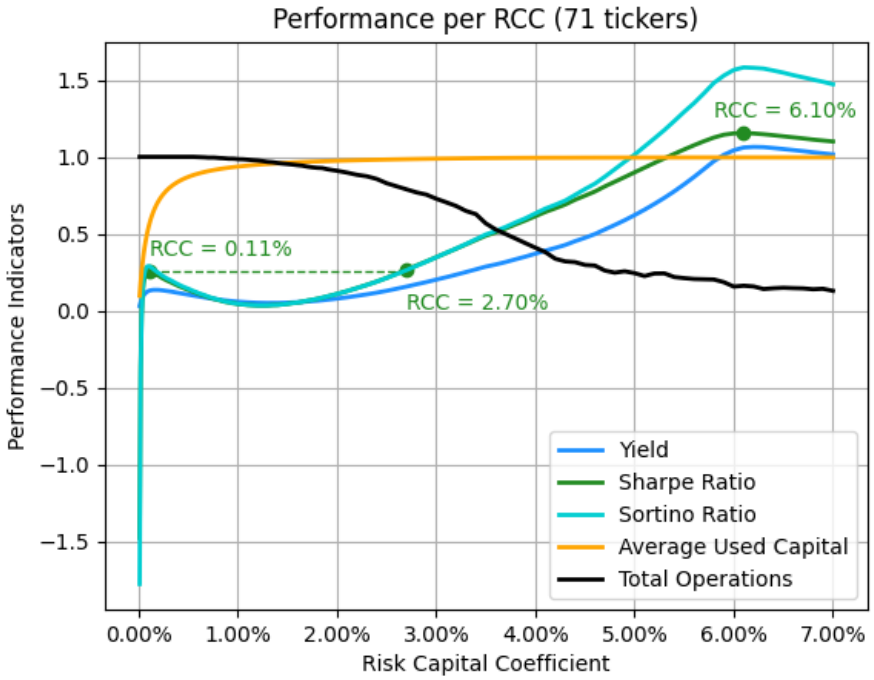

Figura 3.21: Indicadores de performance em função do RCC (71 tickers: 01/01/2019 a 31/03/2020)

A Tabela 3.10 mostra uma análise estatística do capital de entrada em operações para ambas as regiões de máximo local. Conforme esperado, o devio padrão cresce bastante, o que mostra que as operações estão com um aporte de capital elevadíssimo e assim que são finalizadas, o primeiro modelo da fila a indicar compra recebe quase todo o capital disponível. Por outro lado, tanto as operações que antes seriam finalizadas com sucesso quanto as que seriam finalizadas com falha são igualmente prejudicadas.

RCC	Uso Médio de Capital Geral	Média de Capital de Entrada por Operação	Desvio Padrão do Capital de Entrada por Operação
0,11%	56,85%	R$1188,89	R$449,74
6,10%	99,64%	R$4304,55	R$26584,57

Tabela 3.10: Análise estatística do capital de entrada em operações para as regiões de máximo local

Pode-se caracterizar o primeiro pico (RCC de 0,11%) como uma região de predominância mais uniforme dos modelos gerados, pois toda ordem de compra é acatada com capitais de entrada razoavelmente próximos entre si. O custo disso são os baixos valores dos indicadores de performance. Já o segundo pico (RCC de 6,10%) está situado na região de melhor performance, porém é onde a clareza operacional dos modelos pode ser questionada, pois não é fácil distinguir o quanto o efeito aparentemente aleatório da alavancagem de algumas operações em detrimento de outras entrou em simbiose com a operação dos modelos gerados.

Dentre as execuções que compõem a Figura 3.21, extraiu-se os gráficos de rendimento e de uso de capital para as simulações de RCC=0.11% e RCC=6,01% (Figuras 3.22, 3.23, 3.24 e 3.25, respectivamente). Assim, é possível verificar visualmente as questões levantadas.

Um problema geral e inerente à abordagem do RCC é a necessidade do conhecimento *a priori* de um valor razoável. Ou seja, sem algumas simulações prévias, não há como se ter indícios de um valor refinado. Uma proposta para se atenuar

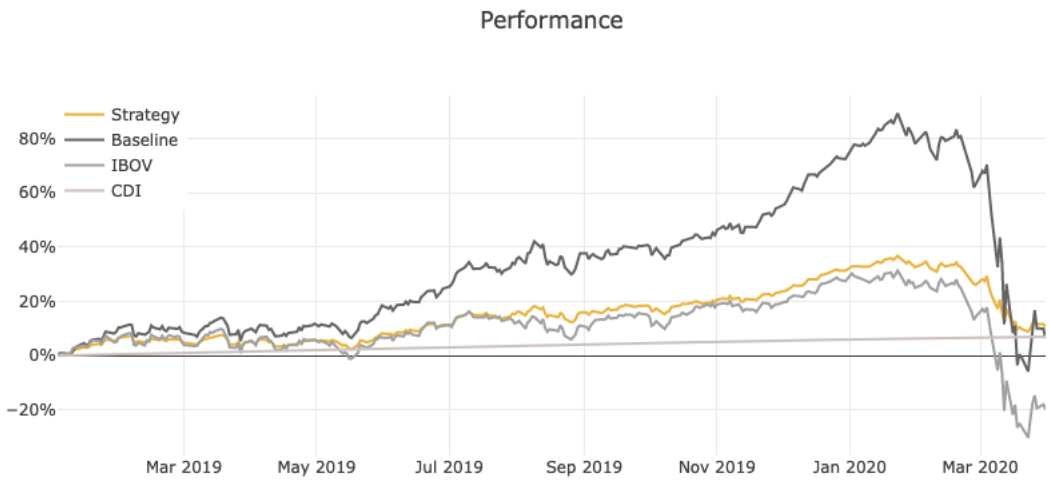

Figura 3.22: Rendimento (71 tickers, 01/01/2019 a 31/03/2020, RCC = 0,11%)

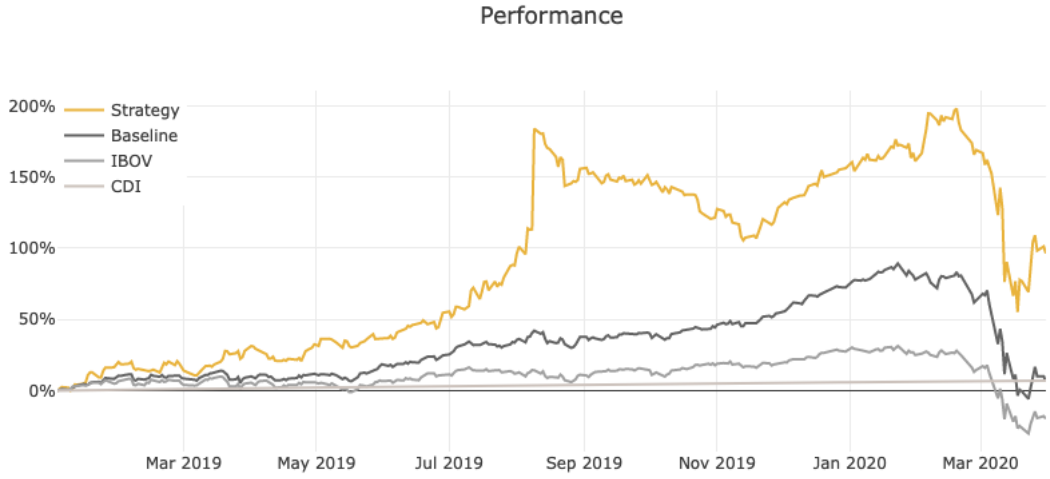

Figura 3.23: Rendimento (71 tickers, 01/01/2019 a 31/03/2020, RCC = 6,10%)

esse problema é através da criação de um controle proporcional que aumente o RCC geral em função do baixo aproveitamento do uso de capital ao longo dos dias e vice-versa. Essa alternativa também é chamada de RCC Dinâmico e abordada em mais detalhes na Seção 3.4.6.

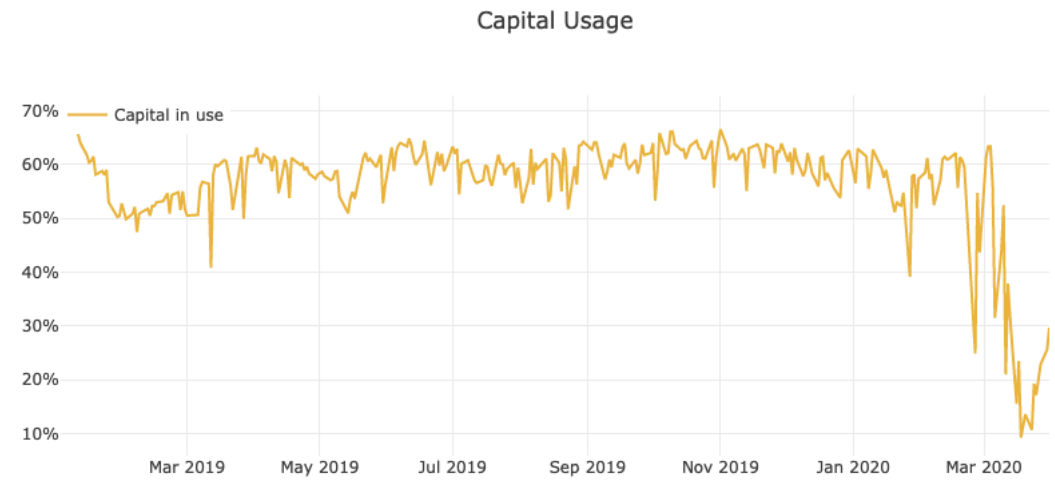

Figura 3.24: Uso de Capital (71 tickers, 01/01/2019 a 31/03/2020, RCC = 0,11%)

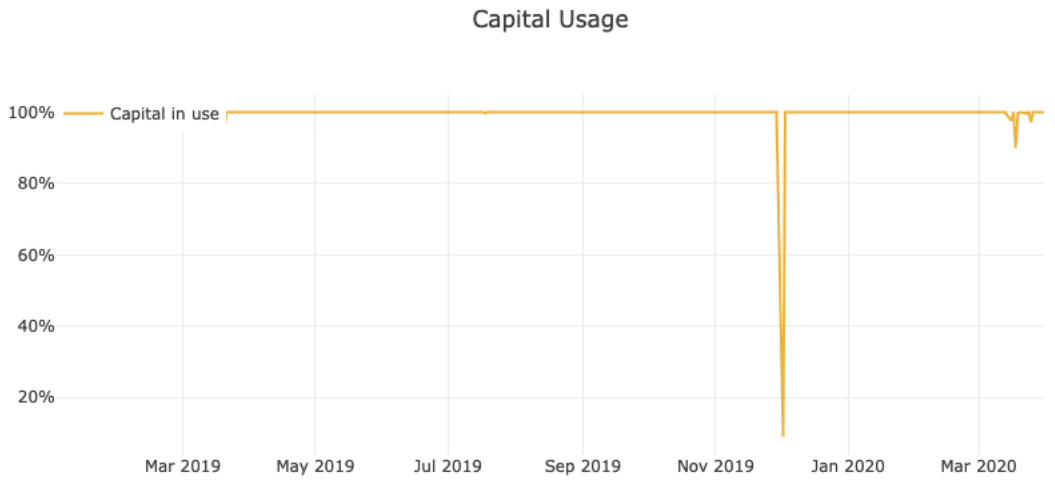

Figura 3.25: Uso de Capital (71 tickers, 01/01/2019 a 31/03/2020, RCC = 6,10%)

3.4.6 Controle Proporcional para Uso de Capital

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

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

As Equações 3.21 e 3.22 mostram o cálculo do erro e do RCC dinâmico (RCC_{din}) a partir do valor de referência para o uso médio de capital (C_{ref}), do uso médio de capital dos últimos 10 dias de simulação ($\overline{C_{10d}}$), do RCC fixo (RCC, definido pela Equação 3.19) e da constante de ganho proporcional (K).

$$e = C_{ref} - \overline{C_{10d}}, \quad \text{para } 0 \leq C_{ref} \leq 1, \text{e } 0 \leq \overline{C_{10d}} \leq 1 \tag{3.21}$$

$$RCC_{din} = RCC(1 + Ke) \tag{3.22}$$

A fim de simplificar a análise, o valor de C_{ref} foi configurado como 100%, deixando o processo de refinamento para os parâmetros RCC e K. Como ambos influenciam diretamente o uso de capital e todos os indicadores de performance, foram acoplados e analisados conjuntamente através das curvas de $RCC \times K$.

As Figuras 3.26, 3.27 e 3.28 apresentam indicadores de performance para valores de RCC e de K diferentes. Nessas simulações foram utilizados os 71 *tickers* indicados na Tabela 3.1 no intervalo de 01/01/2019 a 31/03/2020, junto com o Risco de Entrada por Operação de 0,43 (refinado na Seção 3.4.4) e o Período Máximo de Dias por Operação de 45 dias (refinado na Seção 3.4.3).

A primeira e mais notório observação que se pode fazer pelas Figuras 3.26 e 3.27 é que a região de melhor performance se encontra com baixo RCC e alto K. Também é possível verificar que o aumento graduau das curvas $RCC \times K$ encontrou um comportamento assintótico a partir de 1,40.

De forma semelhante ao comportamento observado na Seção 3.4.5, a Figura 3.28 mostra a tendência de saturação do uso médio de capital com o aumento do RCC e a suavização desse efeito que ocorre ao se diminuir a importância do RCC e aumentar a do K. Já a Figura 3.29 deixa claro que tanto aumentando o RCC quanto o K, em algum momento o número de operações totais cairá, porém é visível a presença de um ponto de máximo na região de $K = 500$ e $RCC = [0,08\%; 0,40\%]$ em cada curva.

Um segundo reflexo do efeito mencionado de suavização da saturação de capital através do aumento do K se faz presente na dimnuição do desvio padrão do capital de entrada por operação. Assim como foi tratado na Seção 3.4.5 pela Tabela 3.10, a Tabela 3.11 mostra os valores de média e desvio padrão do capital de entrada por operação, porém neste caso para as simulações de $RCC \times K = 1,80$.

RCC	K K	Uso Médio de Capital Geral	Média de Capital de Entrada por Operação	Desvio Padrão do Capital de Entrada por Operação
0,11%	-	56,85%	R$1188,89	R$449,74
6,10%	-	99,64%	R$4304,55	R$26584,57
0,000576%	312500	0.9721	R$2411,92	R$12306,90
0,002880%	62500	0.9725	R$2485,83	R$12107,00
0,014400%	12500	0.9767	R$2509,47	R$10650,77
0,072000%	2500	0.9798	R$2396,77	R$9554,60
0,360000%	500	0.9846	R$2332,00	R$7446,53
1,800000%	100	0.9896	R$2533,79	R$8617,72

Tabela 3.11: Análise estatística do capital de entrada em operações para $RCC \times K = 1,80$

Por fim, chama-se atenção para dois pares de valores de $RCC \times K$. O primeiro par é mostrado pela curva $RCC \times K = 1,80$ no ponto maior índice de Sharpe ($RCC = 0,00288\%, K = 62500$). Este curva foi escolhida, pois a partir de $RCC \times K = 1,40$, tanto o índice de Sharpe quanto o rendimento começam a apresentar um comportamento assintótico, não sendo necessário escolher uma curva de valor muito mais alto que este. Já o segundo par foi escolhido a partir do máximo local encontrado na curva $RCC \times K = 0,10$ da Figura 3.29 ($RCC = 0,1\%, K = 100$).

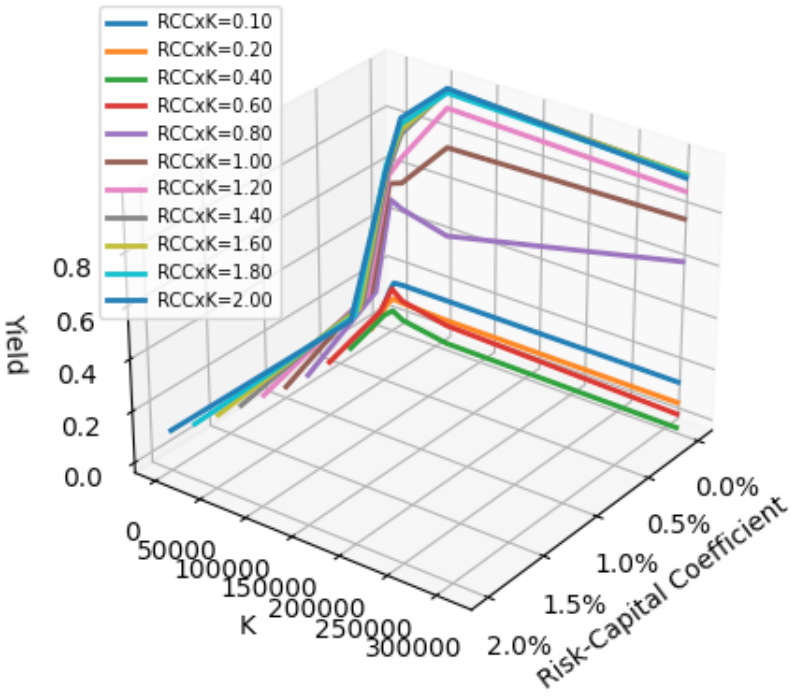

Figura 3.26: Rendimento final sob uso de RCC dinâmico

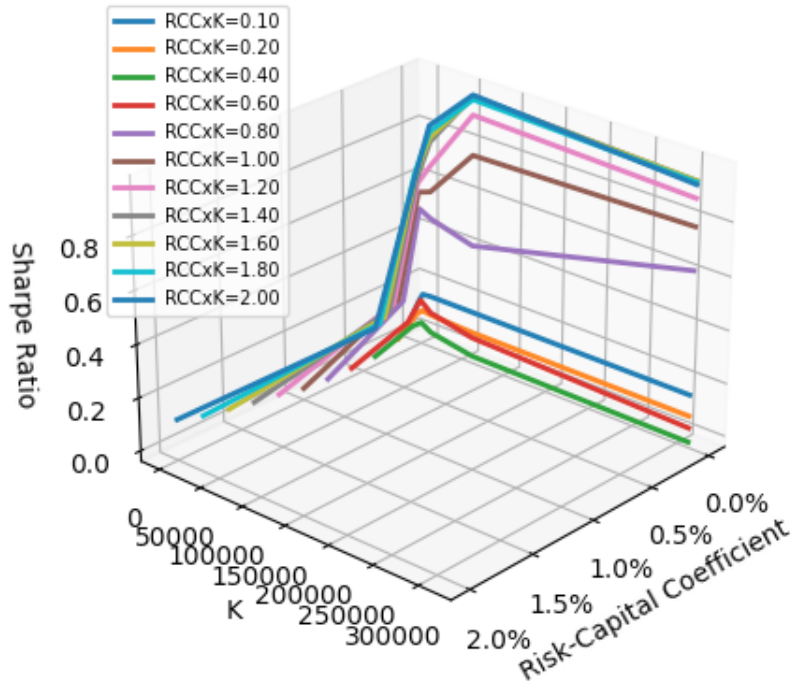

Figura 3.27: Índice de Sharpe sob uso de RCC dinâmico

Este, por sua vez, foi escolhido para representar a região de menor quantidade de operações perdidas.

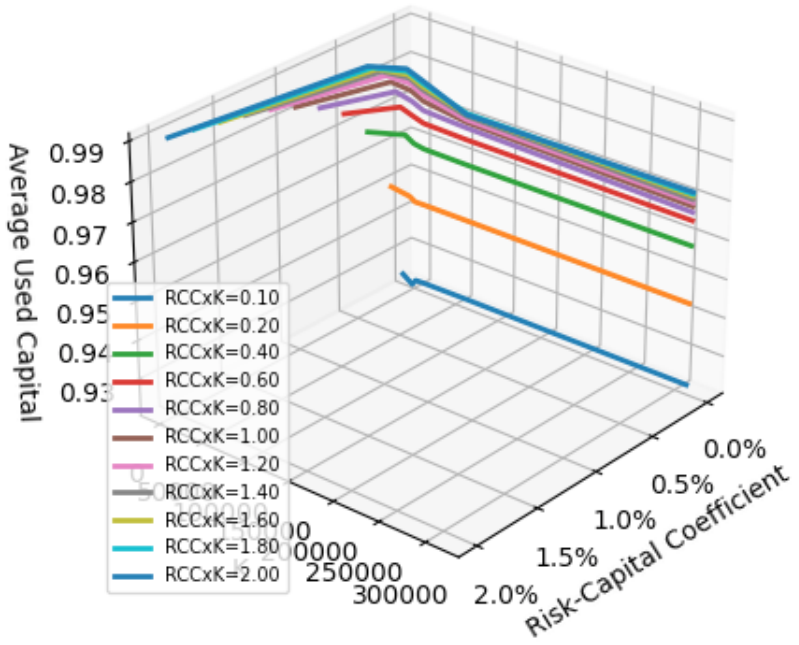

Figura 3.28: Uso médio de capital sob uso de RCC dinâmico

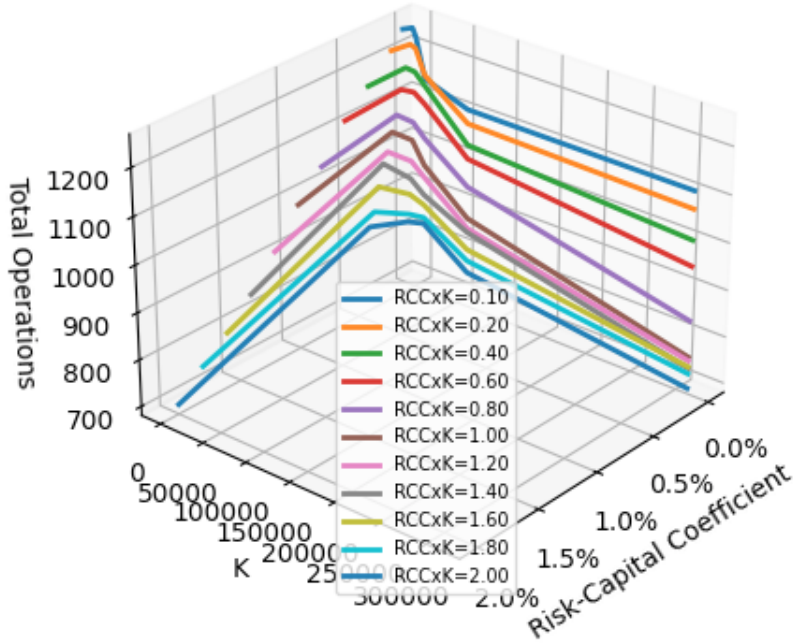

Figura 3.29: Total de operações sob uso de RCC dinâmico

3.4.7 Lista de Parâmetros de Configuração

A Tabela 3.12 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).

Lista de Parâmetros	
Nome do Parâmetro	Descrição
name	**(OBRIGATÓRIO)** Nome da estratégia a ser executada. Único valor válido: "ML". Tipo: *String*. Listável: Não.
alias	**(OBRIGATÓRIO)** Rótulo de Identificação. Tipo: *String*. *Default*: *String* vazia. Listável: Não.
stock_targets	**(OBRIGATÓRIO)** *Array* de ações a incluir na carteira. Formato indicado pela Figura 3.3.
comment	Comentário. Tipo: *String*. *Default*: *String* vazia. Listável: Não.
capital	Capital total da carteira em reais (R$). Tipo: *Float*. *Default*: 100000. Listável: Sim.
risk_capital_coefficient	Coeficiente de risco-capital (RCC) geral (Seção 3.4.5). Tipo: *Float*. *Default*: 0,001. Listável: Sim.
tickers_number	Número de ativos a escolher dentro de "stock_targets"em ordem de listagem. Tipo: *Int*. *Default*: 0 (todos). Listável: Sim.
min_order_volume	Volume mínimo por operação. Tipo: *Int*. *Default*: 1. Listável: Sim.
max_days_per_operation	Número máximo de dias por operação (Seção 3.4.3). 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	Risco mínimo por operação. Tipo: *Float*. *Default*: 0,003. Listável: Sim.
max_risk	Risco máximo por operação. Tipo: *Float*. *Default*: 0,15. Listável: Sim.

Continuação da Tabela 3.12	
Nome do Parâmetro	Descrição
operation_risk	Valor percentual de escolha do risco de entrada em operação (Seção 3.4.4). Tipo: *Float*. *Default*: 0,5. Listável: Sim.
enable_dynamic_rcc	Uso de Coeficiente de Risco-Capital dinâmico (Seção 3.4.6). Tipo: *Boolean*. *Default*: *False*. Listável: Sim.
dynamic_rcc_reference	Valor de referência de uso de capital médio no controle do RCC dinâmico (Seção 3.4.6). Tipo: *Float*. *Default*: 1,0. Listável: Sim.
dynamic_rcc_k	Valor do ganho proporcional K no controle do RCC dinâmico (Seção 3.4.6). Tipo: *Float*. *Default*: 10. Listável: Sim.
Fim da Tabela 3.12	

Tabela 3.12: Lista de parâmetros de simulação

3.4.8 *Dashboard*

Um *Dashboard* interativo é gerado por aplicação secundária a fim de auxiliar a análise dos resultados obtidos em cada simulação. O *framework Dash* [96] foi utilizado para criar uma interface web resumindo todas as informações pertinentes a uma simulação executada. As Figuras 3.30, 3.31, 3.32, 3.33 e 3.34 mostram em partes as seções de uma simulação genérica.

Figura 3.30: *Dashboard* - Performance

Parameters

Alias	2019-1 to 2020_1. All tickers. RCC K analysis.
Total Tickers	71
Start Date	01/01/2019
End Date	31/03/2020
Capital (R$)	100000
Risk-Capital Coefficient - RCC (%)	0
Gain-Loss Ratio	3
Minimum Order Volume	1
Minimum Operation Risk (%)	0.3
Maximum Operation Risk (%)	15
Enable Dynamic RCC	True
Dynamic RCC Reference (%)	100
Dynamic RCC K	62500
Operation Risk	0.43

Figura 3.31: *Dashboard* - Parâmetros de entrada

Results and Statistics

Strategy Total Yield (%)	**97.80**
Baseline Total Yield (%)	6.39
IBOVESPA Total Yield (%)	-19.77
CDI Total Yield (%)	7.02
Strategy Total Volatility (%)	**55.70**
Baseline Total Volatility (%)	43.98
Strategy Sharpe Ratio (-)	**1.24**
Baseline Sharpe Ratio (-)	0.2
Strategy Sortino Ratio (-)	**1.6**
Baseline Sortino Ratio (-)	0.18
Strategy-Baseline Spearman Correlation (-)	**0.87**
Strategy-IBOV Spearman Correlation (-)	0.78
Maximum Used Capital (%)	100.00
Average Used Capital (%)	**97.25**
Maximum Active Operations	71
Average Active Operations	63.73
Active Operations Standard Deviation	8.34
Profit (R$)	97797.46
Total Operations	783
---Successful Operations (hit 3:1 target)	**257 (32.8%)**
---Partial Sale Successfull Operations (hit 1:1 or 2:1 target)	0 (0.0%)
---Failed Operations	**455 (58.1%)**
---Timed Out Operations	68 (8.7%)
---Unfinished Operations	3 (0.4%)
Strategy Yield (% ann)	**74.10**
Baseline Yield (% ann)	5.16
IBOVESPA Yield (% ann)	-16.39
CDI Yield (% ann)	5.59
Strategy Volatility (%ann)	**50.22**
Baseline Volatility (%ann)	39.65

Figura 3.32: *Dashboard* - Resultados e estatísticas

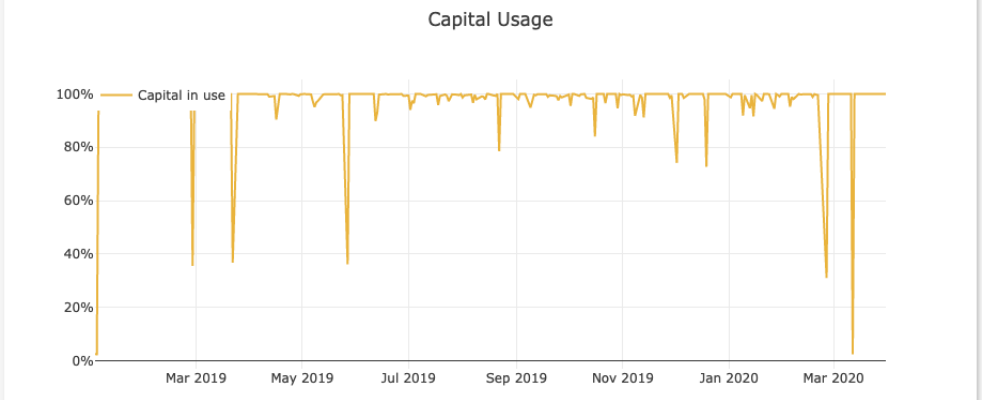

Figura 3.33: *Dashboard* - Gráfico de uso de capital

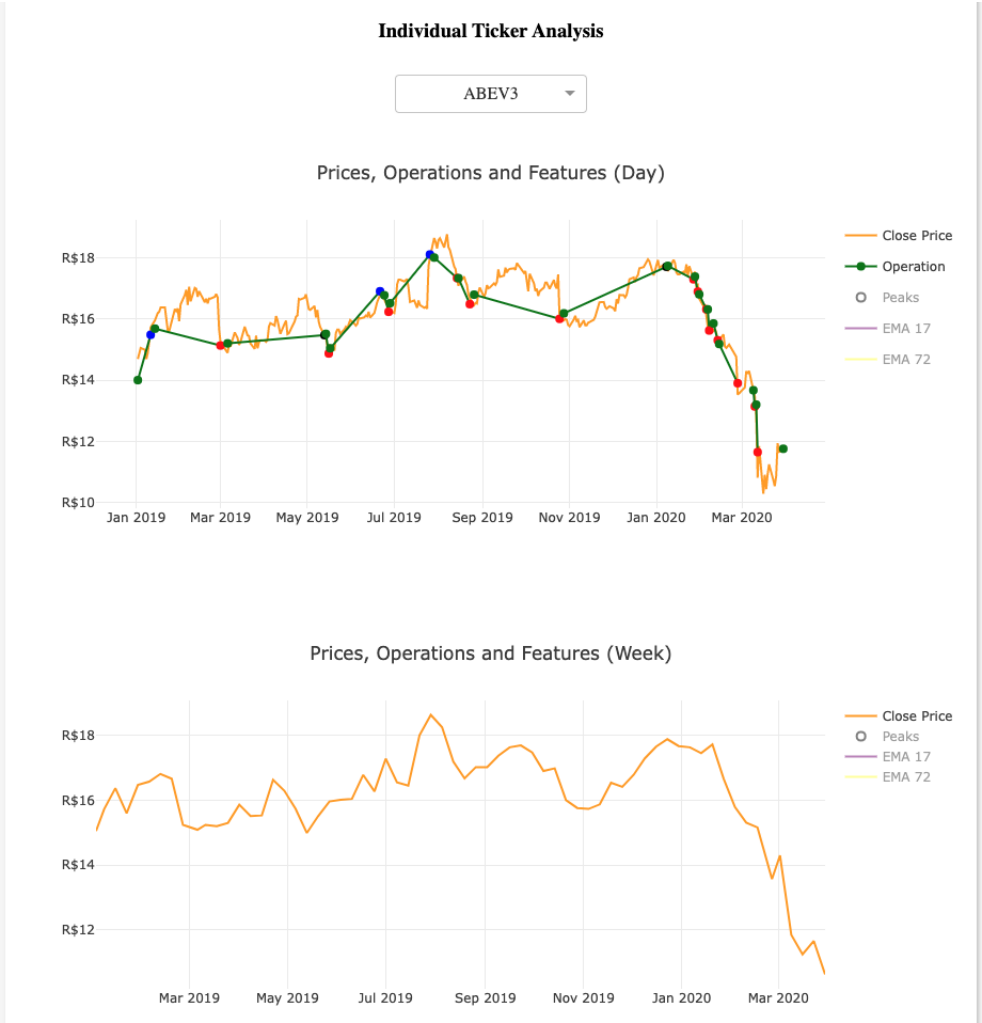

Figura 3.34: *Dashboard* - Gráficos de análise individual de ações

Capítulo 4

Resultados

A partir dos modelos criados na Seção 3.3 e dos parâmetros refinados na Seção 3.4, é possível prosseguir com as simulações finais, que compartilham das seguintes configurações:

- 71 *tickers* (Tabela 3.1)
- Período de simulação: 01/04/2020 a 31/12/2021
- Capital: R$ 100000,00
- Volume mínimo de ações por negociação: 1
- Risco de entrada por operação: 0,29
- Período máximo de dias por operação: 45

Observa-se que o período de simulação é posterior ao utilizado para refinamento dos parâmetros de simulação (01/01/2019 a 31/12/2020).

A Tabela 4.1 mostra 4 perfis de simulação diferentes, cada um motivado por uma questão diferente, são elas: (1) o máximo local na região de baixo RCC da Figura 3.21; (2) o máximo local da região de alto RCC, também da Figura 3.21; (3) a simulação de maior índice de Sharpe da Figura 3.27 e pertencente à curva $RCC \times K = 1,8$; e (4) a simulação de menor perda de operações na Figura 3.29, pertencente à curva $RCC \times K = 0,1$. Também são apresentados os respectivos resultados.

Parâmetro	Estratégia 1	Estratégia 2	Estratégia 3	Estratégia 4	*Baseline*
RCC	0,11%	6,10%	0,00288%	0,1%	-
K	-	-	62500	100	-
Rendimento Final	15,90%	-17,45%	-2,34%	33,25%	68,07%
Volatilidade	17,18%	53,77%	43,00%	31,81%	34,84%
Índice de Sharpe	0,45	-0,14	0,02	0,67	1,15
Índice de Sortino	0,67	-0,15	0,02	0,99	1,71
Cor. Spearman (*Baseline*)	0,96	0,39	0,18	0,90	-
Cor. Spearman (Ibovespa)	0,98	0,48	0,26	0,93	-
Uso Máximo de Capital	71,75%	100%	100%	99,98%	100%
Uso Médio de Capital	52,63%	99,36%	97,85%	92,29%	100%
Média de Cap. por Op.	R$1112,69	R$2038,49	R$2042,67	R$2145,08	-
Desvio P. de Cap. por Op.	R$434,26	R$11673,90	R$6601,82	R$2247,61	-
Operações Totais	1663	418	1171	1659	71
Operações de Sucesso	371 (22,3%)	92 (22,0%)	262 (22,4%)	370 (22,3%)	-
Operações de Falha	1060 (63,7%)	250 (59,8%)	753 (64,3%)	1057 (63,7%)	-
Operações de *Timeout*	164 (9,9%)	66 (15,8%)	104 (8,9%)	164 (9,9%)	-
Operações Incompletas	68 (4,1%)	10 (2,4%)	52 (4,4%)	68 (4,1%)	-

Tabela 4.1: Resultados finais

As Figuras 4.1, 4.2, 4.3 e 4.4 apresentam os rendimentos ao longo do tempo das 4 estratégias supracitadas.

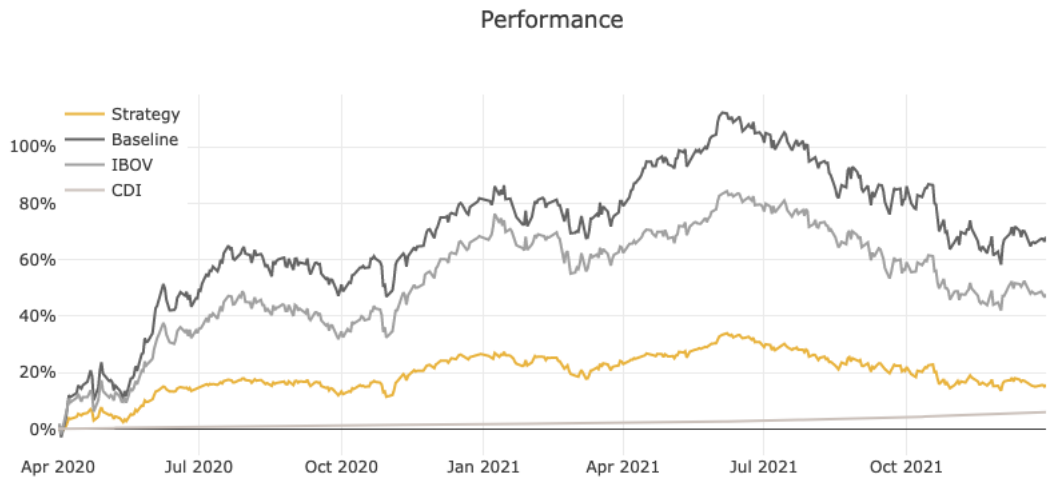

Figura 4.1: Rendimento da Estratégia 1

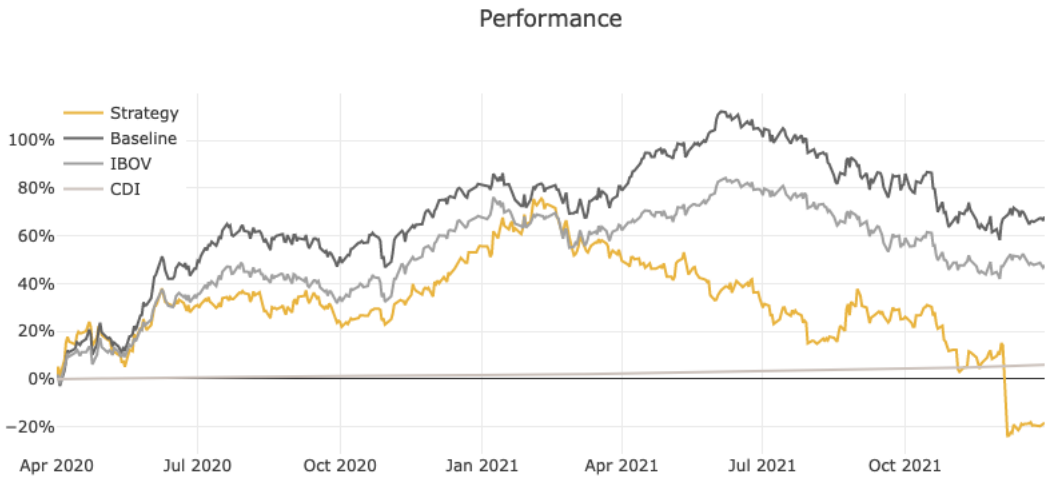

Figura 4.2: Rendimento da Estratégia 2

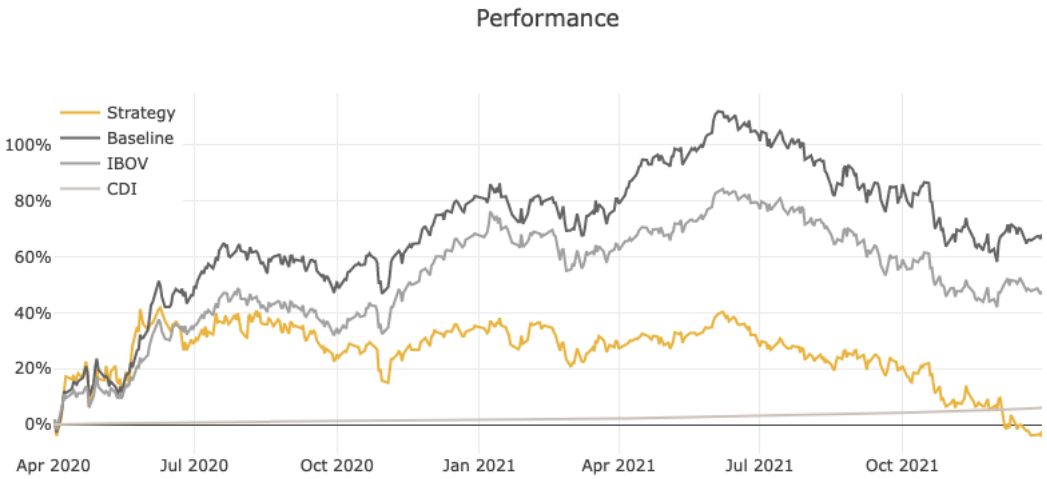

Figura 4.3: Rendimento da Estratégia 3

Começando a análise pelos pontos em comum entre as estratégias, percebe-se que todas obtiveram o rendimento final, o índice de Sharpe e o índice de Sortino significativamente menor que o *baseline*. Além disso, ficaram por baixo no gráfico de rendimento em quase todo o intervalo de simulação. A taxa de sucesso, de falha, de *timeout* e de incompletude das operações se mantiveram praticamente constantes. Apenas a Estratégia 2 que apresentou um ruído maior, pois esta foi a que mais forçou o uso de capital ao máximo, o que conforme visto na Seção 3.4.5, traz mais instabilidade.

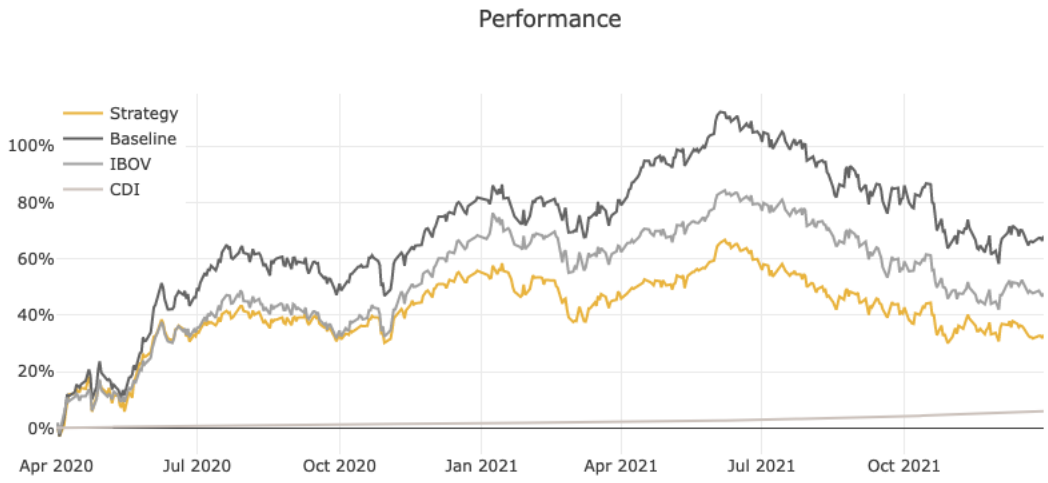

Figura 4.4: Rendimento da Estratégia 4

É possível ver que a ordem de progressão dos indicadores de performance sob um olhar geral segue o seguinte padrão ascendente: Estratégia 2, Estratégia 3, Estratégia 1 e Estratégia 4.

A Estratégia 2 mostra que a escolha do RCC do segundo pico da Figura 3.21 certamente não foi uma boa escolha, pois seus indicadores de performance são os piores dentre as outras três estratégias. Tal resultado não é uma surpresa porque já foi verificado que a disputa por capital em uma situação de alavancagem muito alta traz um efeito aleatório que ofusca a qualidade de atuação dos modelos, pois muitas operações deixam de existir para que outras monopolizem todo o capital da carteira.

A baixa performance da Estratégia 3 questiona o critério de escolha de um alto índice de Sharpe nas curvas apresentadas pela Figura 3.27. Nota-se aqui que há também bastante alavancagem de capital, visto que o uso médio de capital se encontra em 97,85%. Isto é significativamente menor que a Estratégia 2, pois nesta região assintótica da curva, um pequena diferença impacta bastante no desvio padrão de capital por operação, que quase cai pela metade entre ambas as estratégias.

Seguindo a ordem de progressão, a Estratégia 1 possui cerca de 50% de capital ocioso e isso não a impede de ser melhor que as estratégias anteriores, já que usaram quase todo o capital disponível. Tal afirmação contribui para o entendimento de

que o enfraquecimento da influência dos modelos causado pelo efeito aleatório de alavangem excessiva de capital não traz benefícios consistentes, o que não ficou evidente durante a análise da Figura 3.21. Para reforçar, vale a menção de que as duas melhores estratégias (1 e 4) são as que estão mais longe da saturação de capital. A Estratégia 1, dentre as quatro apresentadas, é a que tem menor volatilidade e isso contribui bastante para o índice de Sharpe de 0,45. Esta também é a única estratégia que apresenta o desvio padrão de capital por operação menor que a média do mesmo. Isso revela uma melhor distribuição de capital entre todos os ativos da carteira.

A análise sob a ótica das correlações de Spearman requer um pouco de cautela. Uma correlação muito próxima de 1 entre uma estratégia e sua referência pode indicar uma dependência muito alta, a ponto de atrapalhar na obtenção de ganhos acima da própria referência. Por outro lado, é importante analisar também as tendências de alta e de baixa do mercado, uma vez que não é desejável uma correlação alta com uma referência que só apresenta quedas, pelo contrário. Em suma, como regra geral, não se deseja uma correlação muito alta, pois é justamente nos espaços não correlacionados que uma estratégia tem abertura para se opor a uma tendência de mercado prejudicial. Todavia, muita oposição pode indicar ineficiência.

Portanto, verifica-se que a Estratégia 4, a mais performática, possui um índice de correlação um pouco menor que a Estratégia 1. É possível que essa diferença entre os índices represente ao menos parcialmente o ganho de inteligência entre as duas.

Em resumo, apesar da Estratégia 4 possuir os melhores indicadores, ainda se encontra distante do *baseline* proposto.

Capítulo 5

Considerações Finais

5.1 Conclusão

Neste trabalho, foi proposta a criação de uma estratégia de *swing trade* com a utilização de técnicas de aprendizado de máquina. Para isso, foi projetada uma estrutura de *software* que possibilitasse o manuseio desta estratégia de forma prática.

A partir dos resultados apresentados no Capítulo 4, foi possível observar que a melhor estratégia apresentada possui indicadores de performance ainda distantes de seu *baseline*. Mais especificamente, o índice de Sharpe foi de 0,67 enquanto o mesmo índice do *baseline* foi de 1,15.

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

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

5.2 Trabalhos Futuros

Este trabalho se baseou em premissas e regras que moldaram uma estrutura na qual os modelos pudessem operar. No entanto, alguns dos valores escolhidos podem ser revisitados a fim de serem aprimorados. Dentre eles, podemos citar a escolha do valor de 45 dias para o período máximo de dias por operação. Ao invés de um único valor geral, pode-se considerar uma análise estatística para cada ativo e encontrar um valor que seja mais adequado individualmente.

Deve-se mencionar também o parâmetro de razão entre ganho e perda, já que foi utilizado como uma constante de valor 3. Aqui também é cabível uma análise estatística para cada ativo a fim de se encontrar valores mais refinados.

Não houve consideração de proventos durante a simulação das estratégias pela dificuldade de se obter uma base de dados gratuita e confiável. Portanto, vale mencionar que a incorporação desta base pode enriquecer os estudos aqui realizados.

Referências Bibliográficas

[1] VELEZ, O. L., *Swing Trading*, v. 81. John Wiley & Sons, 2012.

[2] PYTHON, "The Python Tutorial", https://docs.python.org/3.10/tutorial/index.html, (Acessado em 06 de Outubro de 2022).

[3] 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).

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

[5] 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).

[6] 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).

[7] SIQUEIRA, A., "Robô de investimento: tudo o que você queria saber sobre essa tecnologia", https://blog.magnetis.com.br/robo-de-investimento/, (Acessado em 06 de Outubro de 2022).

[8] 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).

[9] CRAWFORD, M., KHOSHGOFTAAR, T. M., PRUSA, J. D., *et al.*, "Survey of review spam detection using machine learning techniques", *Journal of Big Data*, v. 2, n. 1, pp. 1–24, 2015.

[10] DULLAGHAN, C., ROZAKI, E., "Integration of machine learning techniques to evaluate dynamic customer segmentation analysis for mobile customers", *arXiv preprint arXiv:1702.02215*, , 2017.

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

[12] LADD, J. W., WRIGHT, R. M., "Obstáculos ao desenvolvimento do mercado brasileiro de capitais", *Revista de administração de empresas*, v. 5, pp. 71–104, 1965.

[13] 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).

[14] 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.

[15] 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).

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

[17] 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).

[18] 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).

[19] 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).

[20] 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).

[21] 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).

[22] 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).

[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] B3, "ETFs Listados", https://www.b3.com.br/pt_br/produtos-e-servicos/negociacao/renda-variavel/etf/renda-variavel/etfs-listados/, (Acessado em 10 de Outubro de 2022).

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

[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] KIRKPATRICK II, C. D., DAHLQUIST, J. A., *Technical analysis: the complete resource for financial market technicians*. FT press, 2010.

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

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

[39] INVESTIDOR, B. D., "Como Interpretar o Gráfico de Candlestick", https://www.bussoladoinvestidor.com.br/grafico-de-candlestick/, (Acessado em 5 de Abril de 2022).

[40] MORAES, A., *Se Afastando da Manada: Estratégias para vencer no Mercado de Ações*. Infomoney, 2016.

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

[42] MÜLLER, A. C., GUIDO, S., *Introduction to machine learning with Python: a guide for data scientists*. "O'Reilly Media, Inc.", 2016.

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

[44] 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.

[45] KHAN, A. I., AL-HABSI, S., "Machine learning in computer vision", *Procedia Computer Science*, v. 167, pp. 1444–1451, 2020.

[46] TRIPATHI, B. K., "On the complex domain deep machine learning for face recognition", *Applied Intelligence*, v. 47, n. 2, pp. 382–396, 2017.

[47] ZHOU, L., "Product advertising recommendation in e-commerce based on deep learning and distributed expression", *Electronic Commerce Research*, v. 20, n. 2, pp. 321–342, 2020.

[48] KUMAR, P., IQBAL, F., "Credit card fraud identification using machine learning approaches". In: *2019 1st International conference on innovations in information and communication technology (ICIICT)*, pp. 1–4, IEEE, 2019.

[49] RICHENS, J. G., LEE, C. M., JOHRI, S., "Improving the accuracy of medical diagnosis with causal machine learning", *Nature communications*, v. 11, n. 1, pp. 1–9, 2020.

[50] PROVOST, F., "Machine learning from imbalanced data sets 101". In: *Proceedings of the AAAI'2000 workshop on imbalanced data sets*, v. 68, pp. 1–3, AAAI Press, 2000.

[51] 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.

[52] STRANDS, "Unbalanced Datasets & What To Do About Them", https://blog.strands.com/unbalanced-datasets, (Acessado em 5 de Abril de 2022).

[53] PETERSON, L. E., "K-nearest neighbor", *Scholarpedia*, v. 4, n. 2, pp. 1883, 2009.

[54] WOLBERG, W. H., MANGASARIAN, O. L., "Multisurface method of pattern separation for medical diagnosis applied to breast cytology.", *Proceedings of the national academy of sciences*, v. 87, n. 23, pp. 9193–9196, 1990.

[55] BHATIA, N., OTHERS, "Survey of nearest neighbor techniques", *arXiv preprint arXiv:1007.0085*, , 2010.

[56] AHA, D. W., *Lazy learning*. Springer Science & Business Media, 2013.

[57] 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).

[58] LOBATO, E., UGEDO, A., ROUCO, L., *et al.*, "Decision trees applied to spanish power systems applications". In: *2006 International Conference on Probabilistic Methods Applied to Power Systems*, pp. 1–6, IEEE, 2006.

[59] PARDO, R., *The evaluation and optimization of trading strategies*. John Wiley & Sons, 2011.

[60] MULTICHARTS, "Walk Forward Optimization", https://www.multicharts.com/trading-software/index.php/Walk_Forward_Optimization, (Acessado em 28 de Junho de 2022).

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

[62] ROLLINGER, T. N., HOFFMAN, S. T., "Sortino: a 'sharper'ratio", *Chicago, Illinois: Red Rock Capital*, , 2013.

[63] SPEARMAN, C., "The proof and measurement of association between two things.", , 1961.

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

[65] PEREIRA, D. F. R., *Aprendizado de máquina e aprendizado profundo para apoio à decisão no mercado financeiro*. Dissertação de graduação, Escola Politécnica - Universidade Federal do Rio de Janeiro, https://monografias.poli.ufrj.br/repdownload.php?farquivo=monopoli10025708.pdf&fcodigo=3659, 2018.

[66] 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.

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

[68] 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.

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

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

[71] 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.

[72] 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.

[73] 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.

[74] 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.

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

[76] 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.

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

[78] NELSON, D. B., "Conditional heteroskedasticity in asset returns: A new approach", *Econometrica: Journal of the econometric society*, pp. 347–370, 1991.

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

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

[81] 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.

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

[83] 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.

[84] 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.

[85] 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.

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

[87] 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.

[88] POSTGRESQL, "About", https://www.postgresql.org/about/, (Acessado em 07 de Outubro de 2022).

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

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

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

[92] VIEW, T., "Trading View", https://www.tradingview.com/about/, (Acessado em 20 de Setembro de 2022).

[93] NORI, P., "Database Files in Project Github Page", https://github.com/Nori12/Projeto-Final/tree/master/database.

[94] HAYKIN, S., VAN VEEN, B., *Signals and systems*. John Wiley & Sons, 2007.

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

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

[97] LUIZA, M., “Relações com Investidores - Magazine Luiza”, https://ri.magazineluiza.com.br/, (Acessado em 20 de Setembro de 2022).

Apêndice A

Inconsistência de Proventos na Biblioteca *yfinance*

Apesar da praticidade de obtenção dos *candlesticks* diários que a biblioteca *yfinance* (Python) traz, seus valores de proventos (dividendos e juros sobre capital próprio) não são totalmente confiáveis. O estudo em questão mostra inconsistências tanto por duplicação quanto por inserção incorreta de proventos. Para isso, uma análise de caso foi realizada para a companhia Magazine Luiza (*ticker* MGLU3), onde foram comparados os dados obtidos do *yfinance* via *script* com o site de relações com investidores da mesma [97]. Além disso, utilizou-se a plataforma *TradingView* [92] para confirmação dos valores de preço de fechamento.

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

[1] Em inglês: *split*

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

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

Analisando a Tabela A.1 de cima para baixo, nota-se que a primeira irreularidade notável ocorre no dia 14/10/2020, onde a plataforma *TradingView* apresenta um preço de fechamento discrepante em relação ao site de RI da própria companhia e

do *yfinance*. Como se trata de um evento singular e não é o foco deste estudo, ele foi desconsiderado.

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

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

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

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