Universidade Federal de Uberlândia
Instituto de Matemática e Estatística

Bacharelado em Estatística

ENTRE O ACASO E A INTELIGÊNCIA:

Uma Comparação entre Investimentos na B3 Baseados em Aprendizado de Máquina e o Jogo da Roleta

Lucas Fernando Ribeiro Chaves

Uberlândia-MG
2025

Lucas Fernando Ribeiro Chaves

ENTRE O ACASO E A INTELIGÊNCIA:

Uma Comparação entre Investimentos na B3 Baseados em Aprendizado de Máquina e o Jogo da Roleta

Trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Dr. Pedro Franklin Cardoso Silva

Uberlândia-MG

2025

Universidade Federal de Uberlândia
Instituto de Matemática e Estatística

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, ____ de _____ de 20____

BANCA EXAMINADORA

_____

Dr. Pedro Franklin Cardoso Silva

_____

Dr. Lúcio Borges de Araújo

_____

Dr. Leandro Alves Pereira

Uberlândia-MG

2025

"Hardships often prepare ordinary people for an extraordinary destiny."

"Dificuldades frequentemente preparam pessoas comuns para destinos extraordinários."

Clive Staples Lewis

Agradecimentos

Primeiramente gostaria de agradecer a Deus, pelas inúmeras oportunidades de aprimoramento moral e intelectual que **Ele** tem frequentemente me ofertado nesta jornada humana em busca da evolução espiritual.

De forma muito especial, agradeço aos meus pais, **Edna** e **Nahor**, pelo apoio incondicional em todas as etapas da minha vida. Pela estabilidade financeira que sempre me proporcionaram e, sobretudo, pelo suporte afetivo e emocional, fundamentais para que eu pudesse vencer os desafios de cada nova jornada. Ao longo dos anos, a confiança, paciência e amor do meus pais foram os alicerces que sustentaram minhas conquistas: o bacharelado e o mestrado em Engenharia e, agora, com minha segunda graduação, o título de bacharel em Estatística. Esta realização é tanto deles (de vocês) quanto minha.

Agradeço a todos os professores e amigos que, ao longo de todo o meu percurso acadêmico, me proporcionaram um ambiente de intensa troca de ideias e constante incentivo aos meus objetivos — incluindo aqueles que tive a oportunidade de conhecer durante o período em que realizei intercâmbio internacional nos Estados Unidos, na University of Wisconsin.

Gostaria também de agradecer ao Prof. Dr. João Marcelo Vedovotto (FEMEC-UFU) por me apresentar ao extraordinário universo das avaliações estocásticas, o que despertou em mim um profundo interesse pela área da Estatística.

Por fim, agradeço ao Prof. Dr. Pedro Franklin Cardoso Silva (IME-UFU) pela confiança, orientação e apoio na realização deste trabalho. Sob sua supervisão, tive a oportunidade de cursar seis disciplinas de graduação, além desta, nas quais o professor demonstrou, de forma constante, dedicação, cordialidade e incentivo a mim e aos meus colegas. Sua postura sempre prestativa foi essencial para o êxito deste estudo e para o aprimoramento dos meus conhecimentos em inteligência artificial.

Resumo

O presente trabalho tem como objetivo investigar a utilização de algoritmos de aprendizado de máquina como suporte a estratégias de investimento de médio prazo na Bolsa de Valores do Brasil (B3), comparando seu desempenho percentual com o de sistemas de apostas do tipo roleta de cassino. Nesse contexto, três modelos de classificação supervisionada são testados, sendo eles: k-vizinhos mais próximos (k-NN), árvore de decisão e floresta aleatória. Assim, os sinais gerados por esses modelos são utilizados em operações de compra dos ativos-alvo, sem a execução de vendas, o que permite a avaliação do potencial de multiplicação patrimonial pela ótica da relação risco-retorno presente no mercado acionário. Em relação aos ativos avaliados, mais de 30 ações que compõem o índice Ibovespa são inicialmente examinadas; ao final desse processo de seleção, quatro delas nas escalas semanais e mensais de preços, são definidas como objeto de estudo, sendo elas: VALE3, ITUB4, PETR4 e SBSP3. Paralelamente à modelagem computacional associada ao mercado de ações, um sistema computacional baseado na lógica probabilística das apostas de roleta de cassino é implementado, a fim de estabelecer um referencial de comparação de natureza aleatória. Nesse sistema, a estratégia de jogo *Streets of Gold* é aplicada em busca da maximização dos lucros e redução dos efeitos aleatórios do referido jogo. Os resultados indicam que, em termos absolutos, nenhum modelo apresenta vantagem sob todos os ativos e horizontes; contudo, algumas combinações obtêm ganhos expressivamente maiores que os do sistema de aposta, sob as mesmas condições de tempo (rodadas) e capital disponível, garantindo além do elevado retorno financeiro uma menor exposição ao risco patrimonial. Dessa forma, o estudo contribui para a literatura acadêmica vigente ao demonstrar o potencial do aprendizado de máquina na formulação de estratégias financeiras, as quais oferecem subsídios práticos a investidores interessados em alternativas quantitativas frente a cenários de alta volatilidade. Adicionalmente, enfatiza o caráter meramente recreativo dos sistemas de apostas.

Palavras-chave: Aprendizado de Máquina, Bolsa de Valores, B3, k-NN, Árvore de decisão, Floresta aleatória, Roleta de cassino.

Abstract

This study aims to investigate the use of machine learning algorithms as support for medium-term investment strategies on the Brazilian Stock Exchange (B3), comparing their percentage performance with that of casino roulette betting systems. In this context, three supervised classification models are tested, namely: k-nearest neighbors (k-NN), decision tree, and random forest. Thus, the signals generated by these models are employed in buy-only transactions of the target assets—without any sell orders—allowing evaluation of their wealth multiplication potential from the perspective of the risk-return trade-off inherent in the equity market. Regarding the assets analyzed, over 30 stocks comprising the Ibovespa index are initially examined; at the end of this selection process, four of them—assessed on weekly and monthly price scales—are defined as the objects of study: VALE3, ITUB4, PETR4, and SBSP3. In parallel with the computational modeling related to the stock market, a computational system based on the probabilistic logic of casino roulette betting is implemented to establish a benchmark of random nature for comparative purposes. Within this system, the *Streets of Gold* betting strategy is applied in pursuit of profit maximization and mitigation of the randomness effects inherent to the game. The results indicate that, in absolute terms, no model demonstrates superiority across all assets and time horizons; however, certain combinations achieve significantly higher returns than the betting system under identical conditions of time (rounds) and available capital, delivering not only substantial financial gains but also lower exposure to wealth risk. Thus, this study contributes to the existing academic literature by demonstrating the potential of machine learning in formulating financial strategies that provide practical support to investors seeking quantitative alternatives in highly volatile market scenarios. Additionally, it underscores the purely recreational nature of betting systems.

Keywords: Machine Learning, Stock Market, B3, k-NN, Decision Tree, Random Forest, Roulette.

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

1.1 Contextualização do problema

Ao longo dos anos, o comportamento financeiro dos brasileiros tem mudado. Historicamente conservadores e afeitos à caderneta de poupança, cada vez mais brasileiros têm abandonado a segurança em favor de estratégias de maior risco, que incluem apostas e outros "jogos de azar". Conforme o estudo "Raio X do Investidor Brasileiro", realizado em novembro de 2023 com 5.814 participantes e conduzido pela Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais (ANBIMA) em colaboração com o Datafolha [14], constatou-se que uma parcela da população brasileira, incluindo as classes A a D, considera os "jogos de azar" uma opção financeira de investimento de capital.

De acordo com a pesquisa [14], 22% dos apostadores veem as apostas online (*bets*) como uma estratégia de investimento financeiro [1]. Isso ocorre com mais frequência entre os homens (25%) e nas classes D/E (25%), sendo também verificado, de forma expressiva nas classes A/B (22%) e C (21%). Esse comportamento está diretamente relacionado à caderneta de poupança, que perdeu grande parte da rentabilidade real diante da inflação [7], uma vez que a poupança oferece rendimento fixo de aproximadamente 0,5% ao mês [6]. Outro ponto a se destacar é a falta de conhecimento sobre opções de investimento mais lucrativas e menos arriscadas do que as apostas por parte dos investidores brasileiros, o que contribui para o referido fenômeno.

Diante desse cenário, os investimentos em renda variável (ações, derivativos, ETFs, fundos imobiliários) estão caindo cada vez mais no gosto de parte dos brasileiros, em contraponto ao perfil conservador dos que ainda preferem a poupança (renda fixa) e àqueles de perfil extremamente arrojado que apostam no caos dos jogos de azar. Informações da B3 [4] mostram que, apenas no ano de 2023, o total de investidores em renda variável no Brasil atingiu a cifra de 5,3 milhões, representando um aumento de 15% em comparação com o ano de 2022 e de 60% em relação ao ano de 2021. Tal aumento, o qual ocorre sobretudo entre os mais jovens (de 25 a 39 anos), está diretamente relacionado às baixas expectativas desse grupo em relação aos sistemas político-econômicos do país e à qualidade/quantidade de informação que recebem, especialmente ao fluxo de dados que circula nas redes sociais.

Nesse contexto, em que a busca por um crescimento financeiro é particularmente importante para quem deseja formar uma reserva de capital (patrimonial) de curto prazo, os investimentos

[1] Conforme [47], o investimento envolve avaliações cuidadosas do risco-retorno dos ativos nos quais se pretende investir, com base em análises de desempenho passado.

em ações aparecem como uma opção viável. Nesse tipo de investimento, é possível estimar, sob certos limites, os ganhos futuros, definindo margens de confiança e risco, graças às metodologias e técnicas fundamentalistas (como a análise de balanço) e às de perfil especulativo (como a análise gráfica). Sob tal perspectiva, os sistemas de aprendizado de máquina [35], também chamados de modelos de *machine learning*, têm-se mostrado fundamentais para compreender e prever o mercado financeiro, pois não apenas ajudam investidores a tomar decisões mais informadas e a reduzir riscos, como também ampliam sua capacidade analítica.

1.2 OBJETIVOS

O presente trabalho tem como objetivo central desenvolver uma estratégia de investimento de médio prazo (até três anos) na Bolsa de Valores do Brasil (B3), aplicando técnicas de aprendizado de máquina (ML). Nessa proposta, as decisões de compra de ações são tomadas exclusivamente a partir dos sinais gerados pelos modelos, não sendo realizadas operações de venda, sejam elas a descoberto ou não[2]. Desse modo, busca-se identificar, dentre os algoritmos clássicos, como k-NN e árvores de decisão, e métodos combinados, como os *ensembles* (floresta aleatória), aquele que apresente a melhor adequação preditiva à evolução temporal dos ativos negociados no país, visando à multiplicação patrimonial. Nesse sentido, os resultados obtidos serão comparados inicialmente com o desempenho de um sistema de referência baseado em investimento no formato livre e recorrente (sem apoio de técnicas gráficas ou computacionais) e em sequência a um sistema de apostas, a fim de analisar o equilíbrio entre risco e retorno dos investimentos em renda variável (ações), em contraposição aos aportes financeiros realizados em sistemas de apostas.

Como objetivo secundário, que se mostra essencial para o desenvolvimento e validação do objetivo principal, será criado um sistema computacional auxiliar baseado na aleatoriedade inerente aos jogos de fortuna[3]. Esse sistema será estruturado seguindo o modelo de "roleta de cassino" (apostas em números e cores), implementado e executado para simular as apostas realizadas online em sites de jogos de azar. As operações serão avaliadas com base na variação do capital após cada aposta, e sua performance será analisada individualmente e em comparação com o sistema principal de investimento em ações, baseado em aprendizado de máquina.

Por fim, ressalta-se que este estudo é relevante tanto para a academia quanto para o mercado financeiro. No campo acadêmico, contribui para a compreensão das dinâmicas de investimentos financeiros com aprendizado de máquina, possibilitando a comparação entre essa modalidade e as apostas. Já para o mercado financeiro, oferece *insights* para investidores e instituições ao associar estratégias de investimento tradicionais as de alto risco, auxiliando na tomada de decisão mais informada, especialmente em contextos de alta volatilidade econômica e inflação crescente, como ocorre frequentemente no cenário brasileiro.

[2]Operação a descoberta é aquela na qual um ativo é vendido sem que o vendedor seja detentor do mesmo. Em contrapartida na venda coberta o vendedor possuí o ativo comercializado

[3]Termo utilizado no meio jurídico.

1.3 Escopo do trabalho

O presente trabalho de conclusão de curso encontra-se dividido em cinco capítulos, um apêndice e um anexo, incluindo o atual (Capítulo 1), no qual o contexto social e financeiro dos brasileiros é apresentado. Adicionalmente, este capítulo descreve os objetivos e o escopo do trabalho. No Capítulo 2, apresenta-se a fundamentação teórica, na qual são descritas características técnicas do sistema de apostas do tipo roleta de cassino. Em seguida, é abordado o instrumento financeiro das ações, presente no mercado à vista da *Ibovespa*, correlacionando-o com a seção seguinte, a qual tratará das pormenoridades técnicas dos modelos de aprendizado estatístico aplicados a tais ativos. Ao final deste capítulo, as principais métricas de avaliação dos modelos de aprendizado são apresentadas e contextualizadas de acordo com o propósito do trabalho.

No Capítulo 3, a metodologia é apresentada e dividida nas seguintes seções: materiais e métodos utilizados na produção deste trabalho; análise e pré-processamento dos dados provenientes da Bolsa de Valores do Brasil (B3); modelagem preditiva, subdividida na parametrização dos algoritmos avaliados e seleção dos melhores atributos (*features*) de cada ativo-alvo; e, por fim, a estrutura e critérios do sistema de apostas. Nesta etapa, destaca-se que a estratégia de investimento em ações alvo deste trabalho é conduzida exclusivamente por meio dos sinais de compra produzidos pelos modelos de aprendizado, não sendo contempladas operações de venda em nenhuma circunstância.

Já no Capítulo 4, apresentam-se, inicialmente, os resultados dos investimentos livres e recorrentes, utilizados como referencial comparativo para o sistema seguinte. De modo que, na sequência, são discutidos os investimentos orientados por sinais, bem como as métricas que melhor discriminam e justificam tais resultados. Por último, são expostos os resultados das simulações associadas às apostas no sistema de roleta, os quais são confrontados com a evolução patrimonial proporcionada pelo sistema principal baseado em modelos estatísticos.

Por fim, o Capítulo 5 encerra o presente trabalho com as conclusões extraídas de todas as simulações realizadas e dos paralelos pertinentes entre os resultados observados, de modo a indicar a alternativa de aporte financeiro mais vantajosa dentre as opções avaliadas. Adicionalmente, são apresentadas sugestões para trabalhos futuros contemplando todas as áreas abordadas no presente estudo.

2. FUNDAMENTAÇÃO TEÓRICA

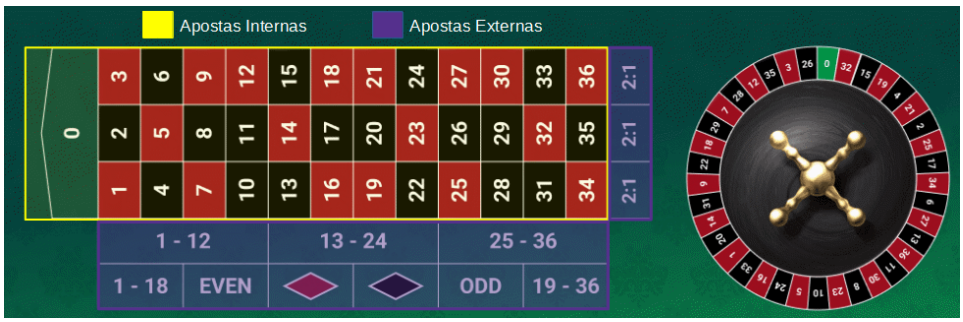
No presente capítulo são apresentados os conceitos fundamentais a respeito do sistema de apostas do tipo roleta de cassino, abordando suas variações, regras e possíveis estratégias de execução. Em relação ao sistema financeiro nacional, são apresentados os principais conceitos e agentes financeiros que o compõem, com ênfase no sistema do mercado de capitais. Adicionalmente, elencam-se os principais papéis comercializados na Bolsa de Valores do Brasil (B3). No que tange aos algoritmos de aprendizado de máquina, descreve-se matematicamente três modelos básicos, acompanhados de seus respectivos pseudocódigos e ilustrações que corroboram para a compreensão preditiva dos mesmos. Por fim, a última a seção é dedicada à avaliação de desempenho, na qual são apresentadas as métricas mais relevantes para a avaliação de modelos de aprendizado de máquina baseados em classificação, bem como a descrição técnica de como essas métricas podem ser interpretadas no contexto do presente trabalho.

Por fim, espera-se que este capítulo apresente os conceitos mínimos de cada um dos tópicos abordados, e estabeleça uma conexão lógica com os objetivos apresentados na seção 1.2, servindo como base conceitual para a compreensão e desenvolvimento do trabalho.

2.1 ROLETA DE CASSINO

O jogo da roleta é um dos jogos mais típicos e/ou tradicionais do ambiente dos cassinos, combinando uma aparente simplicidade com uma rebuscada estrutura probabilística. Sua origem é geralmente atribuída ao físico e matemático francês Blaise Pascal [36], **datando** no século XVII, embora não haja fontes irrefutáveis a respeito. Não obstante, é de conhecimento geral que seu nome é derivado do francês *roulette*, que significa "pequena roda", aludindo ao mecanismo circular que caracteriza o jogo, conforme pode ser visto na Figura 2.1.

Figura 2.1: Mesa de aposta e roleta europeia.



Fonte: Extraída de [33] - Alterada.

Acredita-se que a difusão da roleta ocorreu inicialmente nos cassinos da França espalhando-se posteriormente para a Europa (EU) e Estado Unidos da América (EUA), onde foi incorporada ao contexto cinematográfico e veio a ser um dos símbolos mais reconhecidos da cidade de Las Vegas-EUA. Esse jogo possui versões como a europeia, com 37 casas (0 a 36), e a americana, na qual se acrescenta a casa "00", totalizando 38 posições. Isso eleva a vantagem da casa de 2,70% para aproximadamente 5,26%.

A dinâmica do jogo é simples e envolve, fundamentalmente, o *croupier* e os apostadores como participantes. O *croupier* gira a roleta em uma direção e lança uma pequena bola na direção oposta. Quando a rotação da bola diminui, a mesma para em uma das casas numeradas e coloridas, definindo o resultado da rodada. A mesa de apostas, localizada ao lado da roleta (Figura 2.1), exibe os números e agrupamentos disponíveis, possibilitando que os jogadores decidam onde colocar suas fichas antes de cada giro da roleta, o que define o sistema/circuito de aposta. As opções de aposta podem ser classificadas em dois grupos principais: as internas (inside *bets*), que apresentam maior risco e consequentemente maior retorno, e as externas (outside *bets*), que oferecem maior probabilidade de acerto, mas, por sua vez, menor retorno.

As apostas, de acordo com [36], são sistematizadas nos tipos apresentados a seguir:

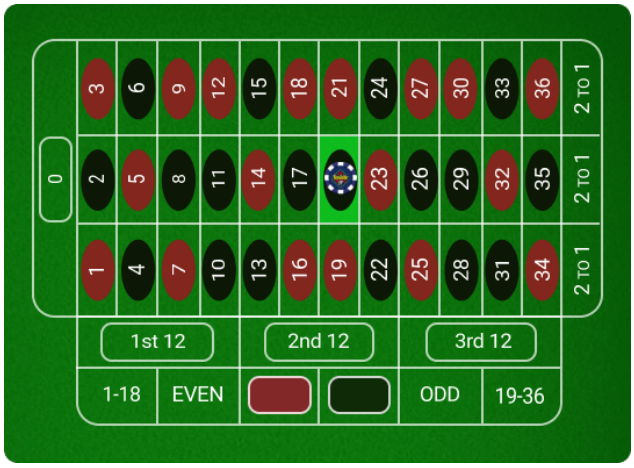
- aposta direta ou plena;
- aposta dividida (cavalo);
- aposta tripla (*street*);
- aposta de esquina (*corner*);
- aposta em cinco números;
- aposta de seis números;
- aposta em coluna;
- aposta em dúzia;
- aposta ímpar/par e;
- aposta em vermelho/preto.

2.1.1 TIPOS DE APOSTAS

A **aposta plena** (Figura 2.2) consiste em escolher um único número, podendo ser qualquer um entre 0 e 36 na roleta europeia. De modo que a ficha representando o valor apostado é posicionada no centro da casa escolhida (aposta interna), e, em caso de acerto, o pagamento é de 35:1. A probabilidade de acerto é de 2,70% na roleta europeia e 2,63% na americana (devido à presença do 00). Sendo essa a aposta mais arriscada e, ao mesmo tempo, a que oferece maior lucratividade. Já a **aposta dividida** ou *split* (Figura 2.3) permite cobrir dois números adjacentes na mesa, colocando a ficha sobre a linha que os separa. Um exemplo é apostar em "8

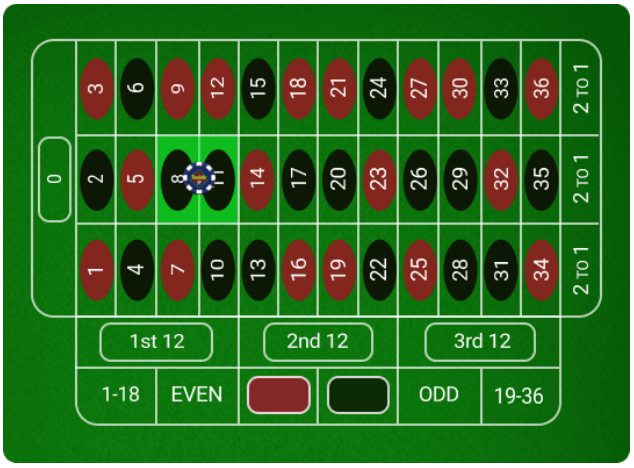
e 11". A probabilidade de acerto é de 5,40% na europeia e 5,26% na americana, com pagamento de 17:1.

Figura 2.2: Aposta direta ou plena.



Fonte: Extraída de [33]

Figura 2.3: Aposta dividida (cavalo).

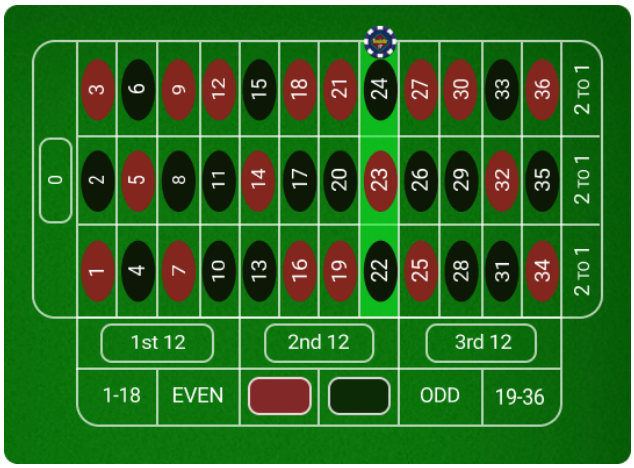


Fonte: Extraída de [33]

A **aposta tripla** ou em *street*, exemplificada pela Figura 2.4, cobre três números consecutivos de uma mesma linha horizontal. A ficha com o valor apostado é colocada na extremidade da linha e o pagamento é de 11:1, com probabilidade de acerto de 8,10% na roleta europeia.

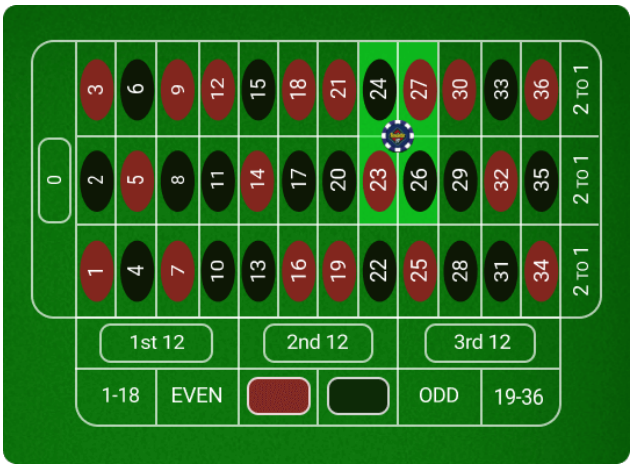
A **aposta de esquina** ou *corner* (Figura 2.5), assim como as duas últimas, é uma aposta dividida, a qual cobre quatro números que formam um bloco 2x2 na área de apostas, como por exemplo uma aposta feita nos números adjacentes "23, 24, 26 e 27". A ficha é posicionada na intersecção das quatro casas e o pagamento é de 8:1, com probabilidade de 10,80% na europeia.

Figura 2.4: Aposta tripla (*street*).



Fonte: Extraída de [33]

Figura 2.5: Aposta de esquina (*corner*).

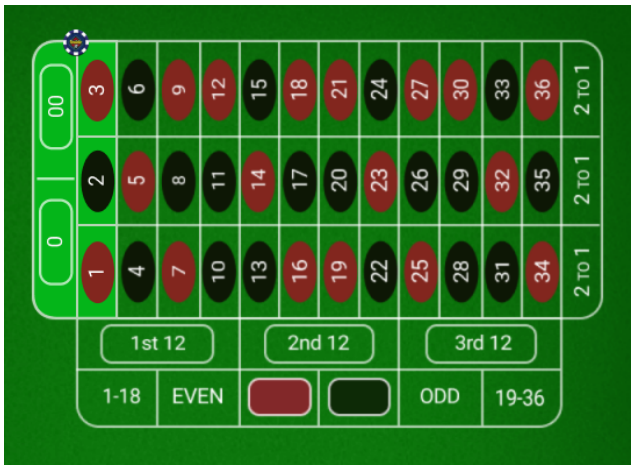


Fonte: Extraída de [33]

A **aposta em cinco números** (Figura 2.6) é exclusiva da roleta americana e cobre 0, 00, 1, 2 e 3 simultaneamente. Sendo o pagamento de 6:1 e a probabilidade de acerto de 13,16%, tornando-se a pior relação risco-retorno da roleta, com vantagem da casa de 7,89%.

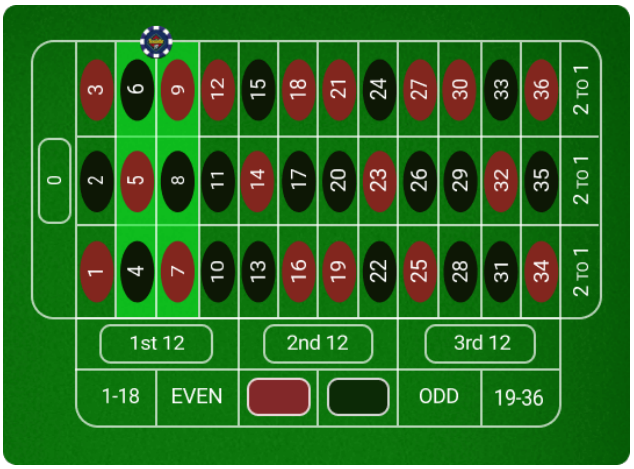
A **aposta de seis números** ou *six line* (Figura 2.7) cobre duas ruas adjacentes, totalizando seis números. A ficha é posicionada na intersecção da borda externa com a linha que separa as duas ruas. O pagamento é de 5:1 e a probabilidade de acerto é de 16,20% na roleta europeia.

Figura 2.6: Aposta em cinco números.



Fonte: Extraída de [33]

Figura 2.7: Aposta de seis números.

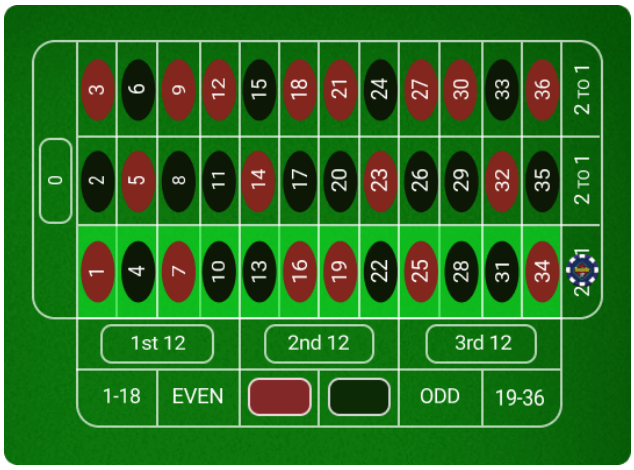


Fonte: Extraída de [33]

A **aposta em coluna** (Figura 2.8) cobre 12 números dispostos em uma mesma coluna vertical. A probabilidade é de 32,40% na roleta europeia, com pagamento de 2:1.

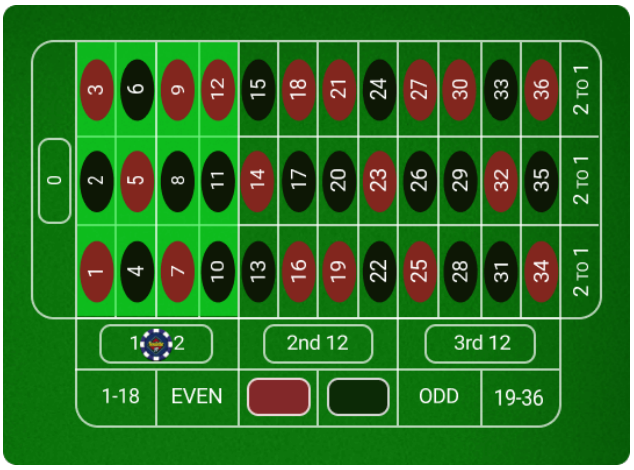
A **aposta em dúzia** (Figura 2.9) também cobre 12 números, mas organizados em três grupos sequenciais: 1–12, 13–24 e 25–36. A probabilidade de acerto é igual à da aposta em coluna, com pagamento de 2:1.

Figura 2.8: Aposta em coluna.



Fonte: Extraída de [33]

Figura 2.9: Aposta em dúzia.

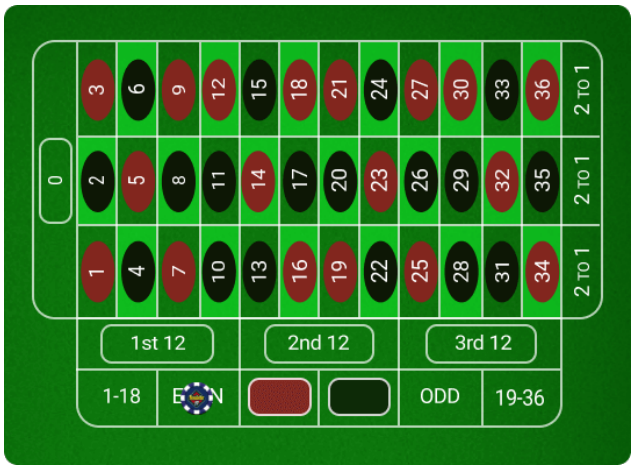


Fonte: Extraída de [33]

A **aposta ímpar/par** (Figura 2.10) cobre 18 números pares ou 18 ímpares, excluindo-se o(s) zero(s). A probabilidade de acerto é de 48,60% na roleta europeia e 47,40% na americana, com pagamento de 1:1.

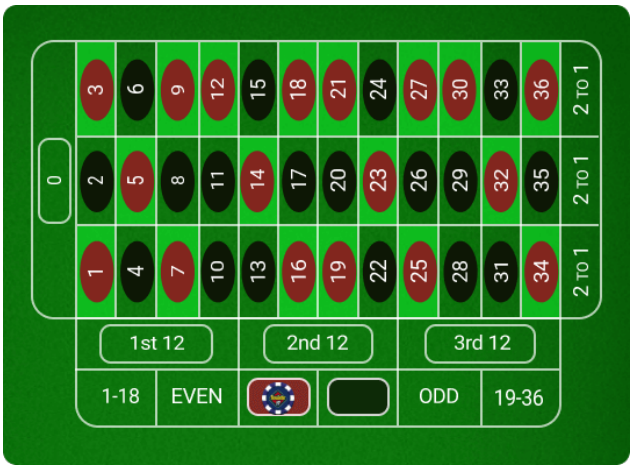
A **aposta em vermelho/preto** (Figura 2.11) segue a mesma lógica, cobrindo metade das casas de acordo com a cor. As probabilidades e o pagamento são idênticos aos da aposta ímpar/par.

Figura 2.10: Aposta ímpar/par.



Fonte: Extraída de [33]

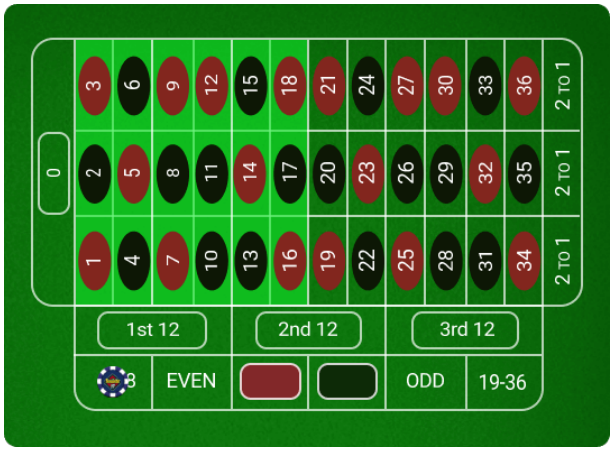
Figura 2.11: Aposta em vermelho/preto.



Fonte: Extraída de [33]

Por fim, a **aposta baixo/alto** (Figura 2.12) divide os números em dois grupos: 1 a 18 (baixo) e 19 a 36 (alto). Assim como as apostas ímpar/par e vermelho/preto, a probabilidade de acerto é de 48,60% na roleta europeia e 47,40% na americana, com pagamento de 1:1.

Figura 2.12: Aposta baixo/alto.



Fonte: Extraída de [33]

Como mencionado anteriormente, a probabilidade de sucesso de cada tipo de aposta varia proporcionalmente ao número de casas da roleta. Esse fato proporciona uma discrepância entre a probabilidade real e o pagamento oferecido, o que é amplamente conhecido como vantagem da casa, e é fator determinante para que o cassino obtenha lucro de forma consistente.

Nesse contexto, a roleta não é apenas um instrumento de entretenimento, mas também um objeto de estudo que abrange conceitos de probabilidade, estatística e até mesmo elementos psicossociais do comportamento humano. Com isso em mente, é possível considerar estratégias que visem maximizar os ganhos ou, pelo menos, estender o tempo que o jogador permanece na atividade.

2.1.2 Estratégias de apostas

Dentre as inúmeras táticas usadas na roleta, podemos identificar duas categorias principais: as progressivas e as não progressivas. As estratégias progressisvas se caracterizam pelo

aumento progressivo do valor das apostas, tanto após uma perda, buscando recuperar o que foi dispendido, quanto após uma vitória, visando potencializar os ganhos. Já as estratégias não progressivas, ou conservadoras, preservam o valor da aposta constante durante toda a partida, independentemente dos resultados.

Desta forma, podemos listá-las como é feito na tabelas, a seguir.

Tabela 2.1: Resultados mensal das compras realizadas pelos sinais dos modelos e pelas entradas aleatórias.

Tipo	Nome	Características
Progressiva	**D'Alembert**	Aumento ou redução linear (±1 unidade) após perda ou vitória.
	Fibonacci	Apostas seguem a sequência de Fibonacci; avança após perdas, recua após vitórias.
	Streets of Gold	Aposta em "ruas" (3 números), aumentando progressivamente após perdas.
Não progressiva	**Flat Betting**	Mantém sempre o mesmo valor de aposta.
	Colunas ou Dúzias	Aposta fixa em um dos três grupos de 12 números.
	Cor, Par/Ímpar, Alto/Baixo	Apostas constantes nos quase 50% de chance

Fonte: Elaborada pelo autor.

Das estratégias apresentadas na Tabela 2.1, duas se destacam pela praticidade e simplicidade de aplicação: a *flat betting* e a *streets of gold*, ambas associadas aos tipos de aposta tratados na subseção 2.1.1.

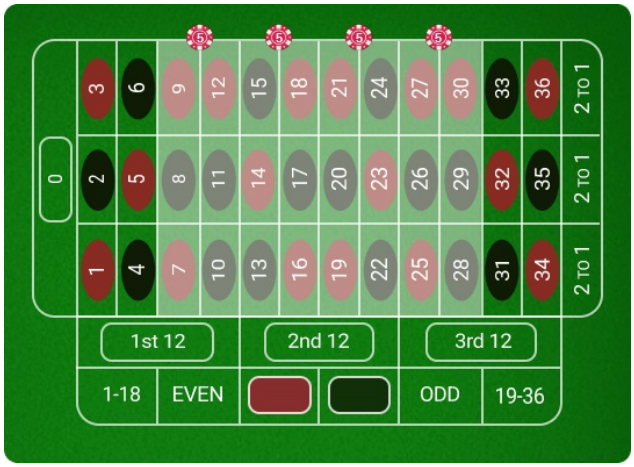
A *flat betting*, ao ser aplicada na roleta, baseia-se em uma lógica bastante simples: o jogador aposta invariavelmente o mesmo valor nas rodadas, independentemente do resultado obtido nas tentativas anteriores. Dito de outra forma, seja ganhando ou perdendo, a unidade apostada será fixa em todas as rodadas.

Tal uniformidade garante, na prática, maior previsibilidade ao jogador. Como aponta [40], a unidade de aposta (por exemplo, 5 u.d.f) deve ser fixada previamente e mantida em todos os períodos de aposta. Essa característica facilita a administração do capital, evitando grandes flutuações e diminuindo o risco de perdas acentuadas em curto prazo. Assim, esta é considerada uma estratégia conservadora, pois não se beneficia de possíveis sequências favoráveis nem possibilita um retorno rápido sobre as perdas acumuladas [40]. No geral, é mais recomendada para jogadores que buscam estender sua permanência no jogo e preservar o fundo/capital, uma vez que, o lucro e/ou o prejuízo é, de fato, pequeno em comparação aos sistemas progressivos.

Por sua vez, a estratégia *streets of gold* (SoG) é mais audaciosa. Ela consiste em apostas simultâneas em quatro sequências de seis números sucessivos, conhecidas como *double streets* ou *six line*. A combinação dos intervalos 9–12/15–18/21–24/27–30 é uma combinação clássica, conforme [41]; todavia, apostas em outras sequências de linhas também são permitidas, tais

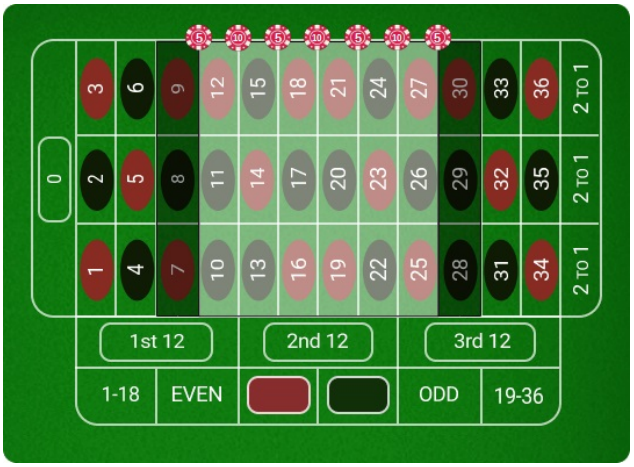
como: 3–6/9–12/15–18/21–24 ou 15–18/ 21–24/27–30/33–36, de modo que os intervalos totalizem 24 números do tabuleiro. Aqui, o jogador estabelece a unidade da aposta (eg.:R$) e a divide pelas quatro sequências, tal como representado na Figura 2.13.

Figura 2.13: Estratégia *streets of gold* no primeira passo ou em etapas pós-vitória.



Fonte: Extraída de [33] - Alterada.

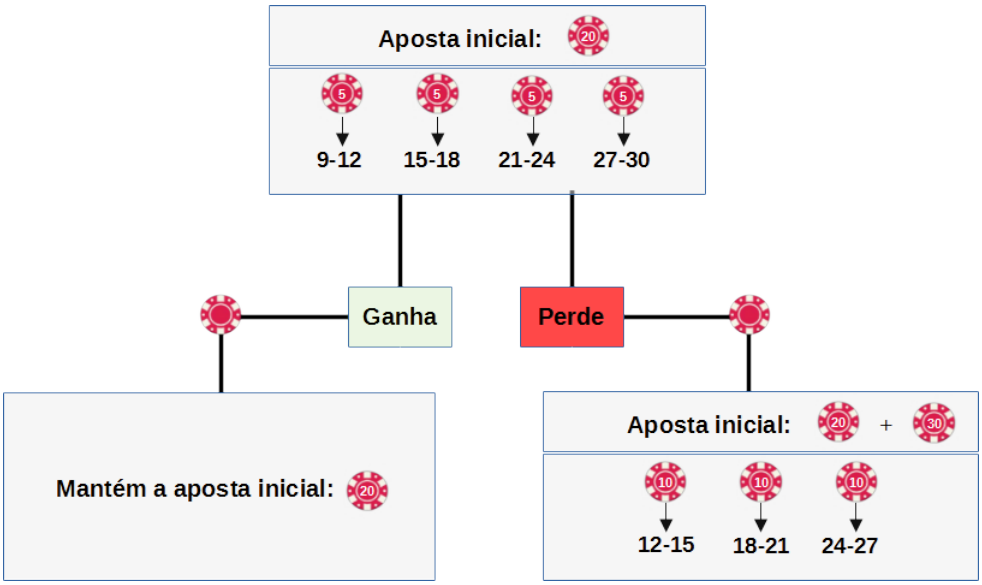
Figura 2.14: Estratégia *streets of gold* em etapa pós-derrota.



Fonte: Extraída de [33] - Alterada.

Em caso de perda, a SoG estabelece que o valor aportado em cada *street* inicial deve ser mantido, com o respectivo acréscimo do valor dobrado nas *double streets* adjacentes, Figura 2.14. Essa progressão amplia a cobertura e aumenta a chance de resgatar as perdas acumuladas. Após um ganho, o sistema volta à primeira configuração, tal como é representado na Figura 2.15.

Figura 2.15: Distribuição dos aportes na *Streets of Gold*.



Fonte: Extraída de [33] - Alterada.

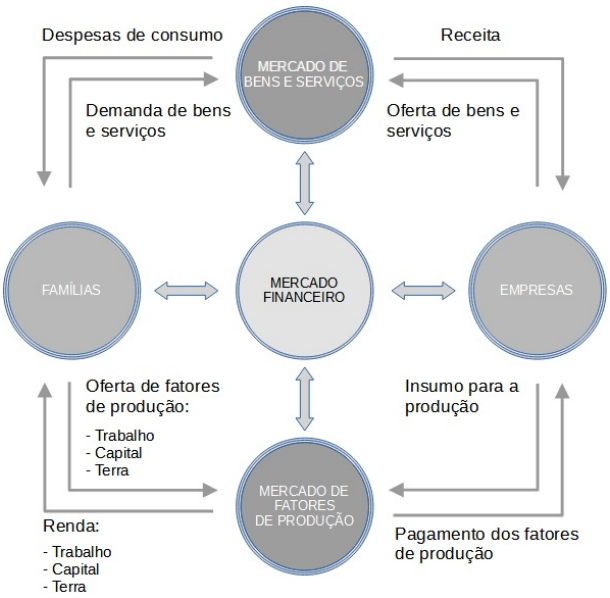
Esta estratégia (*streets of gold*) confere maior potencial de retorno, pois combina a diversificação dos aportes com um sistema progressivo de recuperação. Mas, exatamente por este motivo, exige um capital robusto capaz de suportar perdas acentuadas, associadas as duplicações exigidas. Assim, a *streets of gold* é considerada uma estratégia arrojada e arriscada. De

modo que, a mesma pode oferecer grandes possibilidades de ganhos no curto prazo, mas também expõe os jogadores a maiores riscos em contrapartida a outras abordagens mais cautelosas.

2.2 O SISTEMA FINANCEIRO E O MERCADO DE CAPITAIS

O sistema financeiro de um país pode e deve ser compreendido como um conjunto de instituições, mercados, produtos, agentes econômicos e regras que têm como função principal intermediar a transferência de recursos, de forma mutuamente acordada, entre os agentes que têm poupança (poupadores ou investidores) aos agentes que precisam de capital para investir ou consumir (tomadores), tal como dissertado em [12] e representado na Figura 2.17. De modo que, este sistema é considerado a espinha dorsal da conhecida economia moderna; um ecossistema complexo e composto por macrointerações, particionado em conglomerados (mercado de bens e serviços, famílias, empresas e mercado de fatores de produção) dinamicamente ativos por meio de microinterações, que asseguram a eficiente circulação de dinheiro, estimulando o crescimento econômico e a criação de empregos e a inovação.

Figura 2.16: Representação esquemática do fluxo circular da economia entre famílias, empresas, mercado de bens e serviços e mercado de fatores de produção - fluxo financeiro.



Fonte: Extraída de [12] - Alterada.

Figura 2.17: Ciclo financeiro ou mercado financeiro.

Fonte: Extraída de [12].

O sistema financeiro, neste contexto, representado na Figura 2.16, é constituído pela atuação dos intermediários financeiros e das diferentes divisões que integram o que conhecemos como mercado financeiro (Figura 2.17). Os bancos comerciais, nesse processo, têm papel essencial, captando recursos via depósitos em contas correntes, contas-poupança e certificados de depósito bancário (CDBs), e, em sequência, oferecendo esses recursos sob a forma de crédito a indivíduos e as empresas, formando o que se denomina de mercado de crédito. Paralelamente ao sistema mencionado, seguradoras, fundos de pensão, corretoras de valores e outras instituições não bancárias formam o conhecido mercado de capitais, segmento responsável pela captação de recursos

de longo prazo e pela intermediação de investimentos no âmbito corporativo e individual. Por sua vez, o Banco Central (BC), integrante do mercado monetário, possui papel fundamental na regulação e supervisão das instituições mencionadas acima, bem como na implementação da política monetária nacional, a qual impacta diretamente sobre o consumo do cidadão comum e na dinâmica econômica nacional. Adicionalmente, o BC regula e age (em regra, por intermédio das instituições financeiras) nas relações de câmbio, formando o que chamamos de mercado de câmbio.

De maneira particularizada, podemos definir cada um destes intermediários/membros do mercado financeiro, com o apresentado a seguir.

Mercado monetário:

O mercado monetário constitui a parte do sistema financeiro voltada exclusivamente à concessão e à captação de recursos de curto prazo. Sendo, essas operações realizadas predominantemente entre instituições bancárias e o Banco Central. A função primordial do mercado monetário consiste em garantir a liquidez imediata do sistema, por meio do ajuste diário das posições de caixa (moeda) das instituições bancárias, evitando, dessa forma, que uma entidade finalize o dia com um excesso ou *déficit* significativo de recursos.

Esse mercado constitui o principal meio para a execução da política monetária do Banco Central, que atua nesse contexto por meio de dois mecanismos interligados. O primeiro mecanismo refere-se às operações de mercado aberto [8], nas quais o BC vende títulos com o intuito de remover o excesso de moeda da economia ou, ao contrário, os adquire, injetando liquidez conforme o cenário macroeconômico. O segundo, e mais relevante mecanismo, refere-se ao controle da taxa básica de juros, conhecida como *selic*[1]. Essa taxa funciona como o custo do capital nas transações interbancárias de curto prazo, sendo o principal mecanismo de regulação da atividade econômica no país.

Assim, se o Banco Central busca reduzir a liquidez e combater a inflação, ele eleva a taxa de juros, onerando o crédito, desestimulando sua concessão e incentivando as aplicações financeiras, o que, consequentemente, retira moeda de circulação. Por outro lado, quando a meta é impulsionar a economia, a taxa de juros é reduzida, diminuindo o custo do crédito, fomenta investimentos e o consumo, e, dessa forma, são injetados mais recursos financeiros na economia. Dessa forma, o mercado monetário configura-se como o principal canal por meio do qual a autoridade monetária exerce influência sobre a atividade econômica, o crédito e a inflação.

Mercado de câmbio:

O mercado cambial é o principal meio pelo qual as transações econômicas entre os países são viabilizadas, constituindo uma teia internacional de operações de conversão de moedas. No Brasil, o controle e fiscalização desse mercado são feitos de maneira centralizada pelo Banco Central, que é o principal agente responsável pela supervisão e pela ativa implementação da

[1]Sistema Especial de Liquidação e Custódia (Selic), definida pelo Comitê de Política Monetária (Copom) do Banco Central e pode ser consultada em [39].

política cambial nacional. De modo que, todos os operadores do comércio exterior - empresas exportadoras e importadoras, pessoas não jurídicas e investidores institucionais - dependem desse mecanismo para atuar, isto é, converter a moeda nacional em outras moedas e vice-versa.

Esse mercado (cambial) no Brasil possui uma estrutura em camadas, na qual o Banco Central ocupa posição orquestradora no nível mais alto da organização. Sendo, este, o responsável por determinar as diretrizes, fiscalizar a conformidade com as normas e intervir quando necessário, sempre tendo em vista garantir a previsibilidade e evitar variações excessivas do atual câmbio flutuante. Na sequência, os bancos autorizados operam entre si e com o público por meio de um ambiente denominado mercado interbancário, constituindo-se como principal intermediário na compra e venda de moeda estrangeira. Na base dessa pirâmide encontram-se os usuários finais, que abrangem as empresas que atuam no comércio exterior, os investidores em ativos internacionais, os turistas e demais usuários que realizam operações de envio/recebimento de pagamentos ao/e do exterior.

MERCADO DE CRÉDITO:

De acordo com [12], o setor de crédito constitui-se como uma das partes do mercado financeiro, na qual as instituições financeiras mobilizam fundos de fontes diversas, mais especificamente, de agentes superavitários, concedendo, por meio destes, empréstimos aos entes adjacentes do sistema financeiro (famílias, empresas e mercado), tal como apresentado nas Figuras 2.16, recebendo por essa arbitragem uma compensação financeira proveniente da diferença entre o custo de captação e as taxas aplicadas aos tomadores. Sendo a diferença entre valor emprestado e arrecadado conhecida *spread*. Dessa forma, as instituições financeiras nesse setor desempenham como função principal a intermediação financeira em si e público.

Figura 2.18: Fluxo de intermediação de crédito com captação de recursos (poupador), intermediação financeira (com taxa de *spread* x%) e empréstimo ao tomador.

Fonte: Extraída de [12].

Assim, e como descrito pela figura anterior (Figura 2.18), o mercado de crédito tem como função essencial a conversão da poupança em investimento e consumo. Neste contexto, e de acordo com [12], as instituições financeiras executam duas funções críticas: concentram os riscos e protegem os credores contra perdas, ao passo que aperfeiçoam as avaliações de crédito; além disso, atuam como o elo entre os agentes econômicos, que têm expectativas distintas sobre os prazos e quantias. Sem esse sistema, ou em caso de falha do mesmo, grande parte da demanda por capital não seria atendida/direcionada aos investimentos ou empréstimos solicitados, impedindo a circulação de recursos na economia, levando a uma interrupção desastrosa da atividade econômica.

Todavia, em algumas situações, o mercado de crédito, como posto, não supre de maneira eficiente as necessidades dos entes do sistema financeiro, seja pelos prazos, taxa e/ou pelos montantes de capital envolvidos. Neste contexto, dois caminhos podem ser seguidos pelos tomadores de crédito: a solicitação, via BNDES (Banco Nacional de Desenvolvimento Econômico e Social) de linhas de crédito destinadas a atividades de interesse nacional, ou a captação de recursos via mercado de capitais, sendo este último apresentado no tópico a seguir.

Mercado de Capitais:

O mercado de capitais é considerado uma das principais e mais dinâmicas áreas do sistema financeiro [34], encabeçando o financiamento de médio e longo prazo, em contraste ao crédito bancário associado ao mercado monetário, que se destina a operações de curto prazo [32]. Esse, por sua vez, tem por função o direcionamento da poupança dos investidores às empresas (Figura 2.19), permitindo que estas realizem a captação de recursos de forma direta, sem a necessidade da intermediação bancária convencional. Aqui, a figura do intermediador financeiro é substituída pela do prestador de serviços financeiros que não opera com os valores custodiados.

A eliminação desse intermediário financeiro altera, de forma significativa, a relação risco-retorno. Visto que, enquanto nas operações bancárias o custo para o investidor se traduz, em geral, em uma margem percentual acrescida pela instituição financeira (Figura 2.18), no mercado de capitais o encargo assume a forma de um valor fixo e reduzido por operação. Assim, passa a existir um potencial para a maior remuneração do investimento em comparação ao crédito bancário. Todavia, esse ganho vem acompanhado de um risco adicional, já que o investidor passa a assumir integralmente a incerteza financeira que seria de responsabilidade dos bancos.

Figura 2.19: Fluxo de recursos no mercado de capitais, na qual o intermediador financeiro dá conhecimento entre as partes.

Fonte: Extraída de [12].

Neste contexto, esse mercado pode ser dividido em dois grandes segmentos. Sendo o primeiro o mercado de ações, o qual também é chamado de mercado de capitais próprios e corresponde ao cenário no qual as empresas escolhem abrir seu capital, emitindo ações que representam frações de sua propriedade [13] na bolsa de valores. Essas ações são listadas em proporções de importância, tal como apresentado na Tabela 2.2[2].

No cenário apresentado, o investidor adquire uma ação, tornando-se coproprietário da empresa. Isso significa que ele compra o direito de receber parte dos lucros na forma de dividendos

[2]A lista completa das ações que compõem o índice Ibovespa é apresentada no Apendice A pelas Tabelas A.1, A.2 e A.3

e, adicionalmente, lucra com a valorização das cotas que possui (acúmulo patrimonial). As transações ocorrem em bolsas de valores, como na B3, no Brasil, que disponibiliza uma infraestrutura que garante maior liquidez e transparência nas negociações [17].

Tabela 2.2: As oito empresas com maior participação percentual no índice ibovespa (ibov) - carteira teórica.

Ativo/Ticker	Descrição	Setor	Participação [%]
VALE3	Vale S.A.	Materiais Básicos	10,810
ITUB4	Itaú Unibanco Holding S.A.	Financeiro	8,230
PETR4	Petróleo Brasileiro S.A. – Petrobras	Petróleo, Gás e	6,496
PETR3	Petróleo Brasileiro S.A. – Petrobras	Biocombustíveis	4,546
BBDC4	Banco Bradesco S.A.	Financeiro	3.907
SBSP3	Companhia de Saneamento Básico do Estado de São Paulo – Sabesp	Utilidade Pública	3,692
B3SA3	B3 S.A. – Brasil, Bolsa, Balcão	Financeiro	3,508
ELET3	Centrais Elétricas Brasileiras S.A. – Eletrobras	Utilidade Pública	3,397

Fonte: Extraída de [5] - Alterada.

Já o mercado de dívida, ou de capitais de terceiros, é o segundo segmento mencionado, no qual são emitidos títulos, como debêntures e certificados de recebíveis imobiliários (CRIs) por empresas e governos. Nessas situações, o investidor assume a posição de credor, recebendo juros regulares e, ao término do prazo, o valor principal do investimento. No entanto, diferentemente do mercado acionário, em nenhum momento o investidor/credor torna-se parte do quadro acionário da empresa/empreendimento alvo das aplicações realizadas.

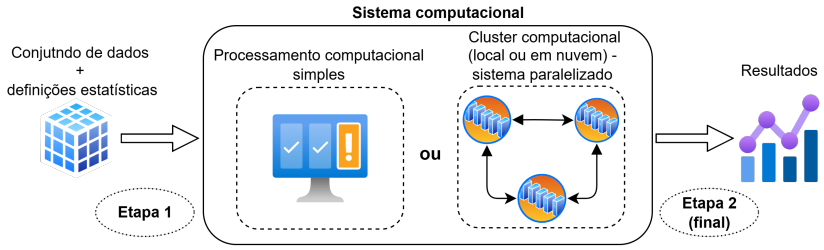
2.3 Algoritmos de aprendizado máquina

A área da inteligência artificial (IA) foi, por muito tempo, considerada o estado da arte da computação e definida, conforme destacado em [30], "como o ramo da ciência da computação dedicado à automação do comportamento inteligente". Com os avanços tecnológicos e a crescente demanda por soluções mais eficientes em termos de custo e tempo de processamento, esse paradigma estritamente computacional e mecanicista — centrado apenas na automação de sistemas que imitavam a lógica humana — passou a ser modificado e ampliado pela noção de sistemas adaptativos associados à matemática e probabilidade. Assim, a essência da IA tem sido reinterpretada como um conjunto de métodos estatísticos voltados à resolução de problemas complexos [31], especialmente em contextos em que há a exigência do processamento de dados em tempo real por sistemas computacionais robustos. Esse movimento culminou no desenvolvimento do que hoje chamamos de modelos de aprendizado de máquina (*machine learning* - ml) e suas variantes especializadas (redes neurais, *deep learning* e os modelos de IAs generativas).

Nesse cenário, em que a inteligência artificial passa a ser compreendida a partir de métodos estatísticos e de aprendizado de máquina, os sistemas computacionais assumem papel essencial ao possibilitar o processamento de grandes volumes de dados. Os quais, em geral, demandam sistemas distribuídos (com processamento paralelo ou multi-core) em nuvem ou localmente (*on-premise*), similares àqueles oferecidos por empresas como Databricks, SnowFlake, entre outras,

representados genericamente pela Figura 2.21 a seguir.

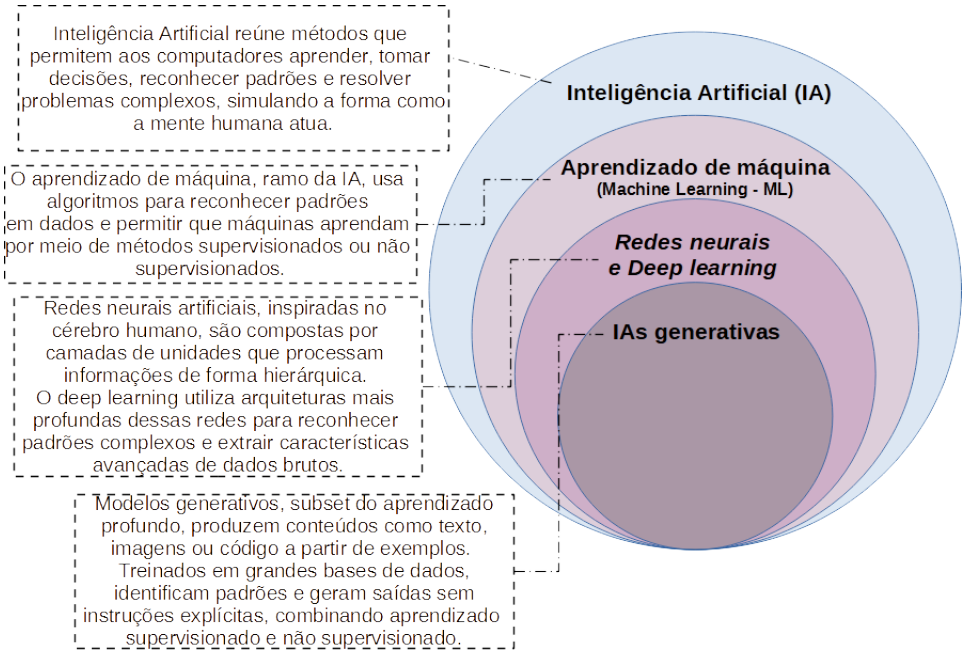
Figura 2.20: Sistema de processamento de dados e de simulações usando IA.



A Figura 2.21, ao apresentar alguns dos diferentes tipos de sistemas computacionais disponíveis, também evidencia a motivação de sua existência, fundamentada no conjunto de dados e nas técnicas estatísticas aplicadas a esses sistemas. Assim, cada algoritmo de aprendizado de máquina, com suas particularidades procedimentais, mostra-se mais ou menos adequado a determinados conjuntos de dados e, consequentemente, exige recursos computacionais específicos, seja em processamento simples ou em paralelo. Dessa forma, a ideia apresentada por [30] é invertida, uma vez que a preocupação científica deixa de ter na automação o seu eixo de estudo e passa a concentrar-se na compreensão e no aprimoramento do "pensamento inteligente" dos modelos estatísticos.

Neste contexto, compreender a sequência das camadas e conhecer os principais algoritmos de aprendizado de máquina torna-se imprescindível para o bom tratamento dos dados e obtenção de resultados significativos.

Figura 2.21: Infográfico em camadas descrevendo os sistemas inteligentes existentes.



Fonte: Extraída de [28] - Alterada.

2.3.1 Classes dos algoritmos de aprendizado

Os algoritmos de aprendizado de máquina podem ser classificados com base na natureza dos dados disponíveis e pelo tipo de resposta gerada durante o processo de treinamento [46]. De modo que, tal classificação é fundamental para identificar o método mais adequado para a resolução do problema avaliado. Segundo [1], os algoritmos podem ser classificados em três classes: os de aprendizado supervisionado, não supervisionado e semi-supervisionado, sendo este último o menos relevante para os objetivos do presente estudo." Neste contexto, o que irá diferenciar cada uma das classes serão as técnicas empregadas para ajustar cada um dos modelos às informações disponíveis. Podendo, tais modelos, ser utilizados para tarefas de classificação e regressão, dependendo dos objetivos a serem alcançados. Por fim, como este trabalho tem como foco a identificação de padrões categóricos, a ênfase recairá sobre os algoritmos de classificação, não abrangendo os modelos regressivos, os quais extrapolam o escopo da pesquisa.

Algoritmos supervisionados

Os algoritmos de aprendizado de máquina supervisionados, em particular os de classificação, têm seu funcionamento estruturado na simples ideia de aprendizado a partir de exemplos, ou seja, os rotulados dos dados são previamente apresentados ao modelo, como descreve [1, 44]. Neste conjunto de dados, cada entrada possui um rótulo que indica a resposta correta. Desse modo, o algoritmo deve aprender as conexões entre os dados de entrada (ou atributos/variáveis independentes) e as respostas (ou resultados) de saída (ou variável dependente), tornando-se capaz de classificar dados não rotulados.

Ainda de acordo com [1, 44], podemos generalizar o mecanismo ou processo de treinamento dos algoritmos supervisionados em três etapas distintas. Na primeira etapa, o modelo é treinado utilizando a porção dos dados destinada a esse fim, associando para cada uma das características (x_i) do conjunto de dados uma resposta y_i. Esse processo de equivalência $x_i \rightarrow y_i$ permite ao algoritmo aprender as relações entre as variáveis. Na etapa seguinte, temos a calibração dos parâmetros do modelo; aqui, os pesos são ajustados de modo a reduzir os erros associados às predições.

Na última, o modelo é validado utilizando a validação cruzada, processo este que consiste na divisão do conjunto de dados em várias (k) partições, chamadas "*folds*". Em cada iteração da simulação computacional, uma dessas partições (*fold*) é utilizada para o treinamento do modelo, enquanto as demais servem para a validação, assegurando que, ao longo do ciclo, todas as observações do conjunto de dados sejam utilizadas. Esta prática reduz a chance de *overfitting*, fornecendo uma estimativa mais precisa do desempenho (apresentado na seção 2.4), e garantindo, dessa forma, que o modelo treinado seja representativo e capaz de generalizar as características latentes do conjunto de dados em estudo, ao invés de ser meramente o resultado (o produto) da memorização dos dados deste. A dê-se salientar que o processo de validação cruzada é um grande aliado do processo de busca recursiva dos melhores parâmetros dos modelos, o chamado *autoML*.

Por fim, é importante que os principais algoritmos representantes dessa classe de aprendizado de máquina sejam pontuados, tal como é feito a seguir.

Algoritmos de classificação supervisionada:

- Naive Bayes;
- k-vizinhos mais próximos – knn;
- Árvore de decisão;
- Floresta aleatório;
- Máquina de vetores de suporte ou SVMs;
- Regressão logística.

Algoritmos não–supervisionados

Ao contrário dos algoritmos supervisionados, os não supervisionados não têm marcadores explícitos de resposta que corroborem para o processo de treinamento. Pelo contrário, esses algoritmos são responsáveis pela autodetecção de estruturas de similaridade geralmente ocultas à percepção de um observador humano, as quais são utilizadas no processo de classificação. Ou seja, tais algoritmos buscam inferir relações entre as variáveis de entrada x_i, de modo que seja possível segmentar o conjunto de dados analisado em partições não etiquetadas. Por esse motivo, segundo [45, 49], tais algoritmos são de grande utilidade na análise de grandes volumes de dados não rotulados, dos quais se deseja compreender a organização de proximidade entre os elementos e efetivamente criar fronteiras de classificação.

Todavia, o processo de treinamento dessa classe de algoritmos não deve ser realizado sem que seja feita a curadoria adequada do conjunto de dados. Isto porque, a inserção de variáveis (dados) altamente correlacionadas ou em escalas diferentes pode produzir distorções na medida de similaridade do processo, gerando agrupamentos artificiais ou sobreposição de padrões que não descrevem a estrutura real dos conjunto de dados. Dessa forma, a presente etapa é fundamental [49], permitindo que os padrões mais importantes possam ser identificados e, por consequência, o processo de classificação seja satisfatoriamente realizado.

Assim, com o conjunto de dados filtrado, o algoritmo escolhido para o processo de treinamento realiza o aprendizado dos dados, buscando padrões ocultos e relações subjacentes presentes nas características do conjunto. Tal processamento pode ser realizado por meio da redução de dimensionalidade, a qual visa simplificar o subespaço das variáveis, determinando aquelas mais significativas, pelo processo de agrupamento baseado em distâncias ou até mesmo pela identificação de pontos que possam distorcer o padrão geral do conjunto de dados (anomalias), tal como apresentado por [45, 50].

Por fim, e mais uma vez, ao contrário do realizado no aprendizado supervisionado, o não supervisionado não possui, por padrão, uma medida que possibilite aferir o acerto ou erro das classificações realizadas. De modo que, a avaliação para este conjunto de algoritmos fica

restrita a métricas específicas (de inércia e variância explicada) as quais não serão tratadas neste trabalho.

Algoritmos de classificação não–supervisionada:

- Mistura gaussiana;
- Clusterização k-*means*;
- Clusterização hierarquizada;
- Máquina de vetores de suporte ou SVMs.

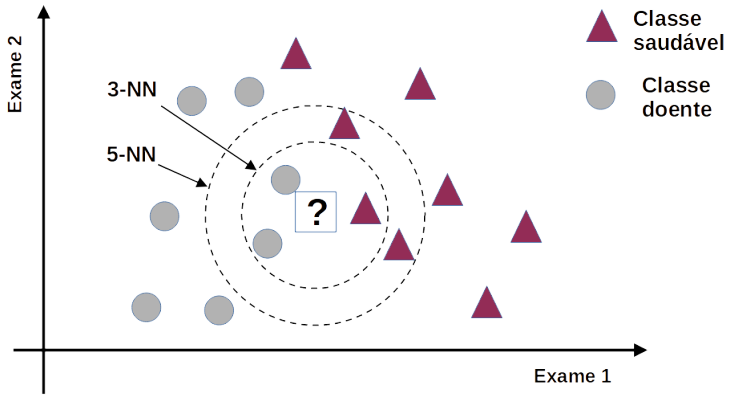
2.3.2 K-VIZINHOS MAIS PRÓXIMOS – KNN

Como apresentado anteriormente, o k-NN, ou k-*nearest neighbors* (em português k-vizinhos mais próximos), é um algoritmo pertencente à família dos algoritmos supervisionados. Isso, pois cada característica x_i está associada a uma resposta (rotulada) y_i correspondente $(\mathbf{x}_i, y_i)$, tanto na etapa de treinamento quanto para a de predição.

Entretanto, mesmo sendo habitual a utilização do termo "treinamento" para nomear esse processo, esta palavra não descreve corretamente o funcionamento do k-NN. Isso porque, ao contrário de outros métodos supervisionados, aqui não temos a construção de um modelo que justifique tal denominação. O que efetivamente ocorre é um ajuste na exploração do hiperparâmetro k e na associação entre $x_i \rightarrow y_i$ no durante a classificação preditiva [10].

De acordo com [18, 29], a predição (ou classificação) obtida pelo k-NN é dada pela classe majoritária entre os k vizinhos mais próximos de cada ponto, dentro do conjunto de dados de treino e teste. Desse modo, a classificação é obtida ao se comparar todos os k dados mais próximos ao ponto avaliado, tal como apresentado na Figura 2.22.

Figura 2.22: Exemplificação do algoritmo k-NN para k igual a 3 e 5.



Fonte: Extraída de [15] - Alterada.

A Figura 2.22 ilustra um cenário bidimensional com duas classes: triângulos representando os indivíduos *saudáveis* e círculos os indivíduos *doentes*. O marcador "?" (interrogação) corresponde ao paciente cuja classe se deseja determinar (classificar). Para k igual a 3, os três vizinhos mais próximos do ponto pertencem majoritariamente à classe *doente*; já para k igual a

5, a maioria passa a ser *saudável*. O exemplo evidencia que a escolha do parâmetro k influencia diretamente a decisão do modelo, de modo que valores pequenos tornam a fronteira mais sensível a ruídos, enquanto valores maiores a tendem a suavizá-la. Por esse motivo, a seleção de k costuma ser realizada por meio de validação cruzada, como discutido no tópico dos algoritmos supervisionados, podendo-se ainda ponderar os vizinhos pela distância. Vale destacar que que a determinação da classe é feita por votação, com ou sem peso, assim, quanto mais indivíduos de uma classe estiverem próximos da amostra a ser classificada, maior é a chance desta assumir a mesma classificação desses elementos.

Por fim, a noção de "vizinhança" no k-NN depende da métrica de distância $d(\cdot,\cdot)$ adotada para comparar exemplos. Uma forma geral é a distância de *Minkowski* de ordem p:

$$d_p(\mathbf{x},\mathbf{z}) = \Big(\sum_{j=1}^{m} |x_j - z_j|^{\,p}\Big)^{1/p}, \tag{2.1}$$

cujos casos particulares incluem a distância *euclidiana* ($p = 2$) e a *Manhattan* ($p = 1$); no limite $p \to \infty$, obtém-se a *Chebyshev*. Quando os atributos estão em escalas distintas, recomenda-se a padronização ou o uso de pesos para cada atributo, resultando em

$$d_{p,w}(\mathbf{x},\mathbf{z}) = \Big(\sum_{j=1}^{m} w_j\,|x_j - z_j|^{\,p}\Big)^{1/p}. \tag{2.2}$$

A métrica escolhida, portanto, define quem são os k vizinhos "mais próximos" e, em conjunto com a votação (simples ou ponderada pela distância), determina a classe atribuída ao novo ponto.

PSEUDOCÓDIGO PARA O *K*-NN

Tendo como referência o apresentado até aqui, a Figura 2.25 exemplifica a sequência lógica de construção e execução do k-NN.

Figura 2.23: Pseudocódigo do k-NN para classificação.

Classificação por k-Vizinhos Mais Próximos (k-NN)

Entrada:

X_{train} - matriz de atributos de treino
Y_{train} - rótulos/classes de treino
X_{entrada} - amostra a classificar
k - número de vizinhos
ζ - métrica de distância (Euclidiana, Manhattan, Minkowski, ...)
Saída: $\hat{Y}_{pred}$ - classe prevista para X_{entrada}

para cada $i \leftarrow 1$ ***até*** $|X_{train}|$ **faça**
  $d_i \leftarrow \text{Distância}_{\zeta}\big(X_{\text{entrada}}, X_{train}^{(i)}\big)$

Ordene os pares $\{(d_i, Y_{train}^{(i)})\}_{i=1}^{|X_{train}|}$ em ordem crescente de d_i $\mathcal{V}_k \leftarrow$ os k primeiros pares do conjunto ordenado

Conte as frequências de cada classe entre os rótulos em $\mathcal{V}_k$ $\hat{Y}_{pred} \leftarrow \arg\max_c \text{freq}(c \mid \mathcal{V}_k)$

retorna $\hat{Y}_{pred}$

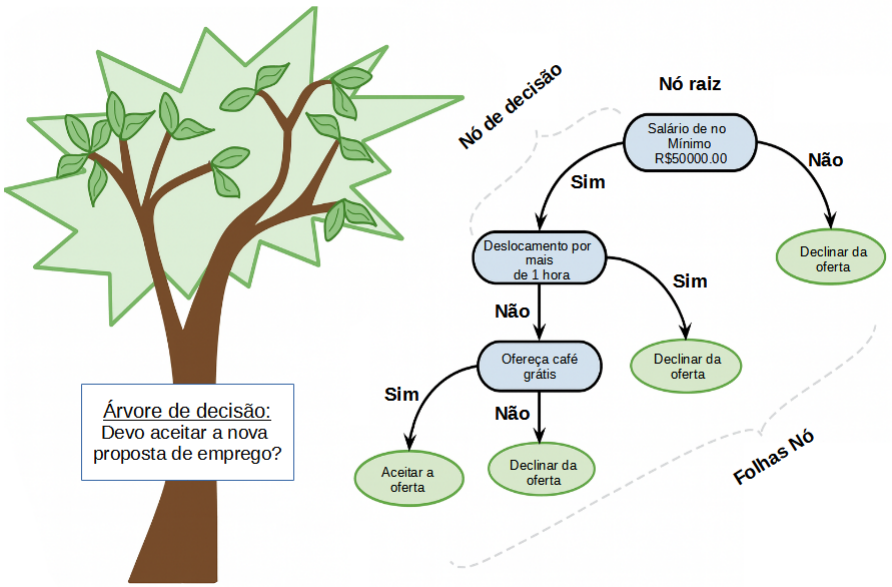
Fonte: Extraída de [16] - Alterada.

Em formato esquemático, o pseudocódigo mostra como funciona o algoritmo do k-NN por meio da definição do conjunto de treinamento, que contém as entradas (X_{train}, Y_{train}, $X_{entrada}$, k, ζ) e a saída esperada ($\hat{Y}_{pred}$). Nesta etapa (inicial), deve-se especificar também a métrica de distância a ser utilizada; podendo, essa métrica, ser definida como as distâncias *Euclidiana*, de *Manhattan* ou de *Minkowski*, bem como o número de vizinhos que serão considerados (k). O passo seguinte consiste no cálculo das distâncias entre a nova amostra e o restante dos pontos da base de dados, sendo, a partir desse valor, possível ordenar o conjunto de forma crescente em relação a proximidade deste ao dado considerado, elegendo assim os k-vizinhos mais próximos. Assim, a classe atribuída à nova observação será aquela a qual houver predominância de vizinhos, isto é, a mais recorrente ao redor do ponto avaliado. Desse modo, o algoritmo realiza previsões de forma prática e direta, sem a necessidade de construído um modelo paramétrico, ao contrário dos métodos que apresentados nas seções 2.3.3 e 2.3.4. Em essência, o método baseia-se exatamente na ideia de realizar comparações de similaridades entre os dados para a tomada de decisão.

2.3.3 ÁRVORES DE DECISÃO

A árvore de decisão, segundo algoritmo supervisionado apresentado neste trabalho, assim como o k-NN, se sobressai pela transparência no processo de escolha e tomada de decisão. Sendo, um dos métodos mais tradicionais e utilizados na área de aprendizado de máquina, uma vez que, combina facilidade de interpretação com resultados confiáveis [2, 3, 27]. A utilização deste método também é muito comum em campos como engenharia, marketing e finanças, principalmente por sua estrutura fundamentada em regras do tipo "se-então", que estabelecem visual os caminhos até a resposta predita. Conceitualmente, o modelo é estruturado como uma árvore invertida, com o nó raiz no topo e, a partir dele, os ramos que levam às decisões intermediárias e finais (Figura 2.24).

Figura 2.24: Representação genérica de uma árvore de decisão.



Fonte: Extraída de [48] - Alterada.

Na Figura 2.24, acima, observa-se a analogia mencionada entre a estrutura do algoritmo e a forma de uma árvore real. Sendo, os nós internos os atributos ou características do conjunto de dados, enquanto os ramos representam as regras ou decisões provenientes desses atributos. Já os nós folha representam os resultados ou as classes previstas, formando, em conjunto, a "copa" da árvore. Por esta simplicidade cognitiva, o modelo se sobressai em relação a outros mais rebuscados.

Em termos de funcionamento, a árvore de decisão (AD) segue um processo progressivo de particionamento do espaço de características dos dados de treinamento, buscando organizar as instâncias em subconjuntos mais homogêneos. Para ilustrar, pode-se compará-la à tarefa de organizar brinquedos em caixas segundo a cor: brinquedos de cor uniforme seriam facilmente classificados, enquanto outros, multicoloridos mas com predominância de um tom, tornariam a caixa em que fossem colocados menos homogênea. Todavia, ainda assim, seriam designados a referida caixa em função da cor predominante de sua fabricação. De modo análogo, o algoritmo AD inicia o processo de classificação, procurando as divisões mais adequadas. Sendo o objetivo central, conforme [27], a formação de subconjuntos de alta similaridade (pureza), isto é, compostos majoritariamente por instância de uma única classe.

A divisão e alocação das informações/dados mais ou menos homogêneos é realizada não por inspeção visual ou por contagem simples; mas utilizando métricas estatísticas, como a da entropia da informação ou do índice de Gini, conforme mencionado por [27].

Índice de Gini (IG):

O índice de Gini é definido por:

$$Gini(t) = 1 - \sum_{i=1}^{C} p_i^2 \tag{2.3}$$

- t representa o nó avaliado;
- **C** o número de classes;
- $\mathbf{p}_i$ a proporção de instâncias da classe.

Sendo, um IG de 0 a pureza total (conjunto totalmente homogêneo) e 1 totalmente impuro (heterogêneo). Assim, a AD busca a rigor minimizar o IG a cada nova iteração.

Entropia:

A entropia é definida por:

$$Entropia(t) = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{2.4}$$

- t representa o nó avaliado;

- **C** o número de classes;
- $\mathbf{p}_i$ representa a proporção de instâncias da classe i no nó t.

Aqui, assim como no IG, 0 representa pureza total (conjunto totalmente homogêneo) e 1 totalmente impuro (máxima impureza - heterogêneo).

Neste contexto, quanto menor o IG e/o a entropia mais homogêneo é o conjunto após a divisão e melhor é considerado o atributo escolhido. Assim, uma vez escolhido o atributo, o nó é particionado em ramos, cada um correspondente a um possível valor deste atributo (limite de decisão) [27]. Desse modo, o processo continua recursivamente até que seja atingido o limite máximo de agrupamentos (profundidade máxima), ou quando todas as instâncias de um nó pertencem à mesma classe, ou ainda quando uma nova divisão não resulta em ganho de pureza.

Por fim, ao se atingir qualquer dos critérios de parada mencionados anteriormente, cada nó folha é associado a uma classe. Em geral, a predição corresponde à classe de maior predominância entre as instâncias do nó, permitindo que a árvore realize previsões sem ambiguidades para novas observações. Nesse ponto, vale ressaltar que, em situações extremas com árvores de grande profundidade, sua capacidade de generalização tende a diminuir, tornando-se excessivamente ajustadas aos dados de treinamento. Esse efeito, conhecido com sobreajuste (ou *overfitting*), pode ser reduzido controlando a profundidade da árvore, o que é feito por meio de técnicas de poda, ou ainda utilizando abordagens de crescimento progressivo disponibilizadas por ferramentas de *AutoML*, recurso utilizado neste trabalho.

PSEUDOCÓDIGO PARA A ÁRVORE DE DECISÃO

Pelo descrito anteriormente, o pseudocódigo a seguir exemplifica a sequência lógica de treinamento da árvore de decisão classificatória.

Figura 2.25: Pseudocódigo para a árvore de decisão.

Árvore de Decisão

Entrada: S - dados de treinamento
$\mathcal{A}$ - atributos candidatos
k - profundidade máxima
ζ - critério de impureza (Gini ou Entropia)
Saída: $\mathcal{T}$ - árvore treinada

se *todas as instâncias em* S *têm a mesma classe* ***ou*** $|S|$ *pequeno* ***ou*** *profundidade* $= k$ **então**
  crie nó folha com classe majoritária; **retorna** nó

para cada *atributo* $a \in \mathcal{A}$ **faça**
  calcule divisões possíveis θ de a para cada θ, calcule ganho $= \zeta(S) - [p_L\zeta(S_L) + p_R\zeta(S_R)]$
  escolha (a, θ) com maior ganho

se *nenhum ganho positivo* **então**
  crie nó folha com classe majoritária; **retorna** nó

Crie nó interno com (a, θ); nó.esq $\leftarrow$ Treinar(S_L) nó.dir $\leftarrow$ Treinar(S_R)

retorna $\mathcal{T}$ `// Árvore treinada`

Fonte: Extraída de [16] - Alterada.

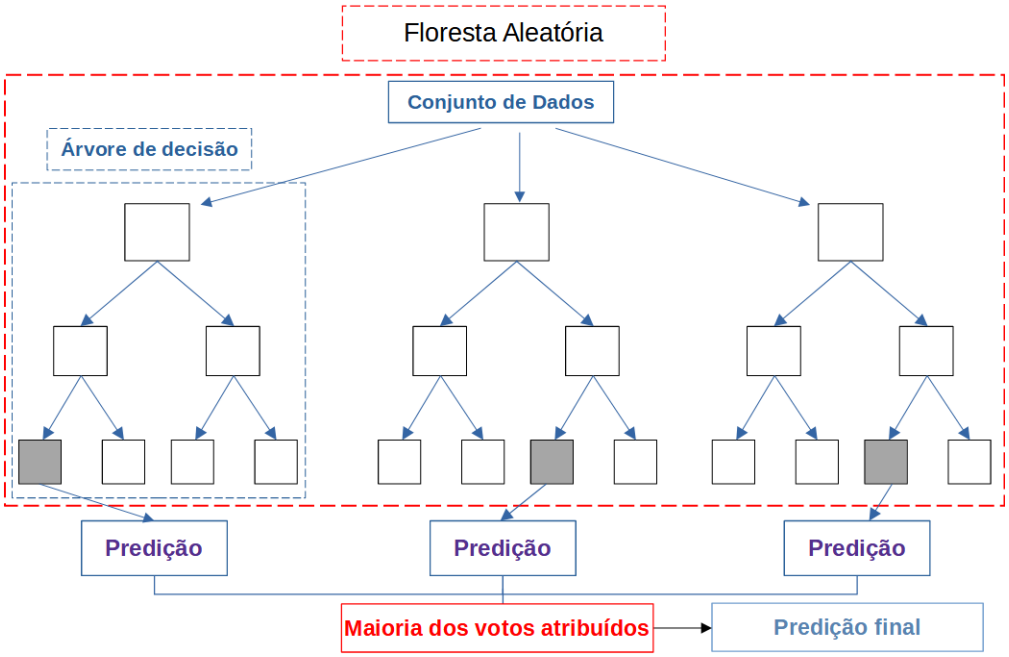
O pseudocódigo da árvore de decisão, tem como entrada: o conjunto de treinamento S, o conjunto de atributos candidatos $\mathcal{A}$, a profundidade máxima k e o critério de impureza ζ,

sendo a saída da modelagem a árvore $\mathcal{T}$ treinada. Assim, o processo inicia com a verificação dos critérios de parada: se todas as instâncias em S pertencem à mesma classe, se o número de elementos em $|S|$ é pequeno e se a profundidade atual é igual a k, então cria-se um nó folha com a classe majoritária e este é retornado. Caso contrário, cada atributo $a \in \mathcal{A}$ é testado, sendo calculadas as suas divisões possíveis θ. Para cada divisão, computa-se o ganho de informação ganho(a, θ) , onde SL e SR representam os subconjuntos gerados pela divisão e pL, pR são as proporções relativas. Assim, a divisão que maximiza o ganho, $(a,\ \theta)$, será escolhida. Se nenhum ganho positivo for obtido, um nó folha com a classe majoritária é retornado. Caso contrário, será criado um nó interno definido por $(a,\ \theta)$ e o algoritmo será chamado recursivamente nos subconjuntos S_L e S_R, definindo, respectivamente, os ramos esquerdo e direito. O processo se repete até que novos critérios de parada sejam atingidos, e ao final retorna-se a árvore $\mathcal{T}$, que constitui o modelo treinado.

2.3.4 FLORESTA ALEATÓRIA

A floresta aleatória, conhecida em inglês como *random forest* (RF), é mais um dos algoritmos supervisionados da família dos modelos de aprendizado de máquina clássicos. Assim como a árvore de decisão, ela é treinada com base em rótulos e fundamenta sua estrutura de aprendizado em regras do tipo "se-então"; todavia, torna-se mais robusta que sua antecessora pela pela capacidade de construir múltiplas sessões/instâncias de treinamento simultaneamente em formato de árvore. Tal como apresentado na Figura 2.26, a seguir.

Figura 2.26: Representação genérica da floresta aleatória.



Fonte: Extraída de [26] - Alterada.

Na Figura 2.26, é possível visualizar o processo de construção das diversas árvores de decisão que compõem o algoritmo da floresta aleatória. Nele, cada instância da floresta faz uma predição, a qual integra o conhecimento coletivo do modelo e, assim, potencializa o mecanismo

de votação responsável por gerar a classifica final. Neste sentido, e de acordo com [42] e [20], as múltiplas árvores contribuem para o aumento da precisão, reduzindo os riscos de sobreajuste (*overfitting*).

Assim, a floresta aleatória baseia-se na ideia de que a combinação de diferentes modelos podem levar a previsões mais robustas do que aquelas fornecidas por uma única árvore de decisão [21]. Para isso, cada árvore é modelada para apresentar grandes diferenças em relação aos seus pares, o que é essencial para o bom funcionamento do modelo [43]. Essa variabilidade é proporcionada, em grande parte, através da amostragem *bootstrap* (*bootstrap aggregating*), também chamada de *bagging* (Figura 2.27), e da escolha aleatória de características.

Figura 2.27: Representação do processo de *bootstrap aggregation* (*bagging*).

Dados Iniciais			
Col_1	Col_2	Col_3	Resp
x_{11}	x_{12}	x_{13}	y_1
x_{21}	x_{22}	x_{23}	y_2
x_{31}	x_{32}	x_{33}	y_3
x_{41}	x_{42}	x_{43}	y_4
x_{51}	x_{52}	x_{53}	y_5
x_{61}	x_{62}	x_{63}	y_6

Bootstrap 1			
Col_1	Col_2	Col_3	Resp
x_{31}	x_{32}	x_{33}	y_3
x_{21}	x_{22}	x_{23}	y_2
x_{11}	x_{12}	x_{13}	y_1
x_{61}	x_{62}	x_{63}	y_6
x_{41}	x_{42}	x_{43}	y_4
x_{51}	x_{52}	x_{53}	y_5

Bootstrap 2			
Col_1	Col_2	Col_3	Resp
x_{21}	x_{22}	x_{23}	y_2
x_{31}	x_{32}	x_{33}	y_3
x_{11}	x_{12}	x_{13}	y_1
-	-	-	-
x_{61}	x_{62}	x_{63}	y_6
x_{41}	x_{42}	x_{43}	y_4

Fonte: Extraída de [24] - Alterada.

No processo de *bagging* apresentado na Figura 2.27, múltiplos subconjuntos de dados provenientes do conjunto original são formados a partir do processo de amostragem, o qual ocorre com reposição. Dessa forma, um mesmo registro pode ser selecionado várias vezes ou, até mesmo, não aparecer em determinado subconjunto. Assim, cada árvore da floresta aleatória é treinada em um desses diferentes subconjuntos de dados, fato que promove a redução da variância e o aumento da capacidade de generalização do modelo [23].

Figura 2.28: Representação do processo de seleção da características (*features*).

Dados Iniciais			
Col_1	Col_2	Col_3	Resp
x_{11}	x_{12}	x_{13}	y_1
x_{21}	x_{22}	x_{23}	y_2
x_{31}	x_{32}	x_{33}	y_3
x_{41}	x_{42}	x_{43}	y_4
x_{51}	x_{52}	x_{53}	y_5
x_{61}	x_{62}	x_{63}	y_6

Seleção 1		
Col_2	Col_3	Resp
x_{32}	x_{33}	y_3
x_{22}	x_{23}	y_2
x_{12}	x_{13}	y_1
x_{62}	x_{63}	y_6
x_{42}	x_{43}	y_4
x_{52}	x_{53}	y_5

Seleção 2		
Col_1	Col_2	Resp
x_{21}	x_{22}	y_2
x_{31}	x_{32}	y_3
x_{11}	x_{12}	y_1
x_{51}	x_{52}	y_5
x_{61}	x_{62}	y_6
x_{41}	x_{42}	y_4

Fonte: Extraída de [24] - Alterada.

Adicionalmente ao *bagging*, na criação de cada árvore, o algoritmo da floresta pode, de maneira aleatória, considerar apenas uma fração das variáveis disponíveis (colunas), conforme

apresentado na Figura 2.28, para definir os pontos de divisão de cada nó. Essa estratégia reduz a correlação entre as árvores e garante diversidade no processo de treinamento, já que cada uma utiliza diferentes combinações de atributos. Assim, por se basear na pluralidade das árvores e dos conjuntos de treinamento, a floresta aleatória dispensa tanto o processo de poda quanto o controle de crescimento aplicados naturalmente na modelagem propriamente dita das árvores de decisão. Como resultado, muitas árvores crescem até o limite máximo, e o eventual sobreajuste gerado isoladamente é compensado pelo desempenho agregado do conjunto.

Por fim, o processo de agregação das previsões (Figura 2.26) permite consolidar os resultados individuais de cada árvore em uma única saída. Assim, nos problemas de classificação, adota-se o critério de votação majoritária: a classe mais votada entre todas as árvores passa a ser a previsão final.

PSEUDOCÓDIGO PARA A FLORESTA ALEATÓRIA

Tal como feito para os algoritmos anteriores, o pseudocódigo a seguir (Figura 2.29) representa a sequência lógica de treinamento do algoritmo da floresta aleatória para o sistema de classificação.

Figura 2.29: Pseudocódigo para a floresta aleatória.

Floresta Aleatória (Classificação)

Entrada: S - dados de treinamento
$\mathcal{A}$ - conjunto de atributos
B - número de árvores
$mtry$ - nº de atributos sorteados em cada nó ($1 \leq mtry \leq |\mathcal{A}|$)
$k_{\max}$ - profundidade máxima (opcional)
ζ - critério de impureza (Gini ou Entropia)
Saída: $\mathcal{F} = \{\mathcal{T}_1, \ldots, \mathcal{T}_B\}$ - floresta treinada

para $b \leftarrow 1$ **to** B **faça**
  $S_b \leftarrow$ Bootstrap(S)
  $\mathcal{T}_b \leftarrow$ Treinar_Arvore($S_b, \mathcal{A}, mtry, k_{\max}, \zeta$) **adicione** $\mathcal{T}_b$ em $\mathcal{F}$
retorna $\mathcal{F}$

Função `Treinar_Arvore`(S, $\mathcal{A}$, $mtry$, $k_{\max}$, ζ, $prof$=0):
  se *todas as instâncias em* S *têm a mesma classe* ***ou*** $|S|$ *pequeno* ***ou*** $prof = k_{\max}$ **então**
    retorna nó folha com classe majoritária em S
  $\mathcal{A}' \leftarrow$ Amostra_SemReposição($\mathcal{A}, mtry$)
  para cada $a \in \mathcal{A}'$ **faça**
    calcule divisões possíveis θ de a **para cada** θ **faça**
      ganho$(a, \theta) \leftarrow \zeta(S) - [p_L\zeta(S_L) + p_R\zeta(S_R)]$
  escolha (a, θ) com maior *ganho* **se** *ganho* $(a, \theta) \leq 0$ **então**
    retorna nó folha com classe majoritária em S
  crie nó interno com (a, θ) nó.esq ← `Treinar_Arvore`(S_L, $\mathcal{A}$, $mtry$, $k_{\max}$, ζ, $prof$+1) nó.dir
   ← `Treinar_Arvore`(S_R, $\mathcal{A}$, $mtry$, $k_{\max}$, ζ, $prof$+1) **retorna** nó

Fonte: Elaborada pelo autor.

O pseudocódigo da floresta aleatória de classificação tem início pela definição das entradas, sendo estas: o conjunto de treinamento S, o conjunto de atributos $\mathcal{A}$, o número de árvores B, o número de atributos sorteados e cada nó, $mtry$ ($1 \leq mtry \leq |\mathcal{A}|$), a profundidade máxima $k_{\max}$ (opcional) e o critério de impureza ζ (Gini ou Entropia) a ser utilizado. Já a saída, trata-se propriamente do resultado do treinamento, o modelo da floresta aleatória $\mathcal{F} = \{\mathcal{T}_1, \ldots, \mathcal{T}_B\}$.

Contudo, o treinamento propriamente dito tem início apenas com o processo de *bagging*. Nessa etapa, para cada árvore $b = 1, \ldots, B$, é gerado um conjunto S_b por amostragem *bootstrap* a partir de S (isto é, com reposição), sobre o qual se treina a respectiva árvore $\mathcal{T}b$. A rotina de crescimento de cada árvore segue critérios de parada semelhantes aos adotados no algoritmo da árvore de decisão. Assim, quando todas as instâncias de S pertencem à mesma classe, quando o número de observações $|S|$ é muito reduzido ou quando a profundidade atual atinge o limite $k$max, o processo retorna um nó folha atribuído à classe majoritária. Caso contrário, em cada nó, seleciona-se aleatoriamente um subconjunto de atributos $\mathcal{A}' \subset \mathcal{A}$ de tamanho $mtry$ e, para cada atributo $a \in \mathcal{A}'$, são calculadas divisões candidatas θ. Em seguida, determina-se o ganho de informação correspondente, em que S_L e S_R representam os subconjuntos formados pela divisão e p_L, p_R suas proporções relativas. A divisão ótima (a, θ) é então escolhida e, caso o ganho obtido não seja positivo, o nó é finalizado como folha com a classe predominante. Do contrário, cria-se um nó interno parametrizado por (a, θ), e o procedimento é repetido recursivamente sobre S_L^* e S_R^*, definindo os ramos esquerdo e direito até que novos critérios de parada sejam satisfeitos.

Ao término do laço em b, a coleção de árvores forma a floresta $\mathcal{F}$, a qual, em geral, dispensa o uso de poda explícita, uma vez que a redução do sobreajuste decorre tanto do processo de *bagging* quanto da aleatoriedade introduzida na seleção de variáveis definida por $mtry$.

2.4 AVALIAÇÃO DE DESEMPENHO

De acordo com [15], a análise de algoritmos de aprendizado de máquina é, em geral, feita verificando-se o quão bem o modelo consegue classificar novas instâncias que não foram utilizadas em seu treinamento. Por isso, entender as principais métricas de avaliação em modelos de classificação é essencial para decidir pela aceitação ou rejeição de uma determinada abordagem. Também é importante ter clareza quanto às necessidades do sistema em estudo, já que a escolha do modelo depende diretamente desse decisão.

Em contrapartida, em sistemas em que os resultados são totalmente determinados pelo acaso (seção 2.1), as métricas de aprendizado de máquina deixam de fazer sentido. Isso porque, nesse contexto, não existe espaço para previsões baseadas em fundamentos econômicos ou séries históricas de dados. Nesses casos, o desempenho é medido apenas pela relação entre acertos e erros, refletindo ganhos e perdas de forma direta. Desse modo, essas avaliações não serão descritas aqui, visto a simplicidade técnica e intuitiva de obtê-las.

2.4.1 MÉTRICAS DE AVALIAÇÃO DE MODELOS

Para a avaliação de modelos de aprendizado de máquinas do tipo classificatório, diferentes métricas podem ser utilizadas, tais como: precisão, sensibilidade, especificidade e acurácia. Essas descrevem de forma quantitativa os acertos e erros das classes positivas e negativas em relação ao objeto avaliado. Tal como apresentado na Tabela 2.3 (matriz de confusão), a seguir.

Tabela 2.3: Matriz de confusão.

Valor Predito	**Valor Verdadeiro**	
	y=0	y=1
y=0	VN (verdadeiro negativo)	FN (falso negativo)
y=1	FP (falso positivo)	VP (verdadeiro positivo)

Fonte: Extraída de [22].

Na matriz de confusão, Tabela 2.3, a primeira sigla (ou descrição) das entradas da tabela refere-se aos resultados das predições do modelo de ML, os quais indicam a ocorrência de acerto ou erro (Verdadeiro - V ou Falso - F) em relação ao conjunto de dados de teste. O segundo termo (Positivo - P ou Negativo - N), por sua vez, aponta para a presença ou ausência da característica analisada, como, por exemplo, a distinção entre indivíduos doentes (1) e não-doentes (0) em um conjunto de dados qualquer de uma pesquisa de doenças infecciosas. Dessa forma, pode-se resumir a matriz de confusão como o exemplificado por [15] e apresentado a seguir.

- **VP (Verdadeiros Positivos):** indica a quantidade de elementos da classe positiva (P) que foram corretamente (V) identificados pelo modelo;
- **VN (Verdadeiros Negativos):** representa o total de elementos da classe negativa (N) que também foram classificados de forma correta (V);
- **FP (Falsos Positivos):** corresponde aos casos em que a classe real é nergativa (N), mas o modelo erra (F) ao classificá-los como positivos;
- **FN (Falsos Negativos):** refere-se aos exemplos que pertenciam à classe positiva (P), mas que o modelo previu, de maneira equivocada (F), como negativos.

Neste contexto, o presente trabalho, apresentará a análise e contextualização dessas métricas sob a perspectiva de dois cenários de investimento distintos, os quais são apresentados a seguir. Sendo, apenas o segundo cenário o objeto de estudo e análise dos capítulos 3 e 4.

- **Cenário 1 - Investidor de curto prazo (*day* ou *swing trade*):** Procura detectar, da maneira mais precisa possível, os melhores pontos de entrada e saída do mercado, sempre comprando em pontos de menor preço ou em tendência de alta, e vendendo em pontos de preço mais elevado ou em tendência de queda, ou vice-versa;
- **Cenário 2 - Investidor patrimonial (*buy* & *hold* - bh):** Faz compras periódicas de ações, priorizando compras de ações em momentos de preço reduzido e, assim acumulando patrimônio a longo prazo, uma vez que, nesse modo, o investidor não realiza vendas.

Acurácia

A acurácia representa a proporção de acertos do modelo, considerando tanto os casos positivos quanto negativos, tal como apresentado na Equação 2.5 a seguir.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.5}$$

No cenário de **day/swing trade**, a acurácia explicita a capacidade global do modelo de prever corretamente os momentos de compra e venda. Entretanto, uma alta acurácia pode ser enganosa caso o modelo tenha dificuldade em prever corretamente momentos críticos de entrada ou saída. Já no caso do ***buy & hold***, a acurácia mostra o quão consistente o modelo é em indicar preços realmente baixos ao longo do tempo, ainda que, nesse contexto, métricas mais específicas assumam maior relevância (como por exemplo a sensibilidade).

PRECISÃO

A precisão (Equação 2.6) mede a proporção de previsões positivas que, de fato, eram positivas.

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{2.6}$$

Para o **day/swing trade**, a precisão representa um ponto relevante de avaliação, uma vez que confere credibilidade ao tipo de sinal emitido pelo modelo (compra ou venda). De modo que, neste tipo de investimento (de alta frequência), uma baixa precisão tenha impacto considerável, provocando grandes prejuízos em um curto espaço de tempo. Já para o investidor **buy hold** ou de longo e médio prazo, a precisão assegura que, quando o modelo sinalizar a condição de preço baixo, essa indicação se confirme como um ponto de mínimo relevante no horizonte temporal analisado.

SENSIBILIDADE

A sensibilidade, representada matematicamente pela Equação 2.7, avalia a capacidade do modelo em identificar corretamente os casos positivos.

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \tag{2.7}$$

No ***day/swing trade***, uma alta sensibilidade indica a capacidade de capturar a maioria das oportunidades lucrativas, evitando que boas entradas ou saídas passem despercebidas. Vale ressaltar que, nessa modalidade, podem ser exploradas tanto a ponta comprada quanto a vendida do processo de investimento, ou seja, a faixa de operação na qual o lucro será realizado. Nesse contexto, altos valores de sensibilidade nem sempre correspondem a ganhos acentuados, já que a métrica pode quantificar pontos "ótimos" de compra e venda (ou vice-versa) muito próximos entre si, o que reduz o lucro líquido e aumenta a exposição às variações do mercado de financeiro. Já no ***buy & hold***, como mencionado na descrição da acurácia, a sensibilidade torna-se ainda mais central. Cada oportunidade perdida de compra (considerando apenas a ponta comprada) a preços reduzidos compromete a capacidade de acúmulo de capital no longo

prazo. Assim, para esse tipo de investidor, a sensibilidade reflete diretamente o que é conhecido no mercado financeiro como custo de oportunidade, e o que de fato buscamos reduzir no presente trabalho.

ESPECIFICIDADE

A especificidade mede a capacidade do modelo de reconhecer corretamente os casos negativos, sendo representada, a seguir, pela Equação 2.8.

$$\text{Especificidade} = \frac{VN}{VN + FP} \tag{2.8}$$

Para o **investidor de curto prazo**, uma alta especificidade reduz a quantidade de sinais falsos de operação, protegendo-o contra entradas desnecessárias no mercado. Isso reduz a exposição do investidor às oscilações do mercado ao passo que ele não se posiciona (opera) de forma enganosa. Assim, o mesmo consegue realizar lucros maiores com menos operações efetivadas. No ***buy & hold***, a especificidade evita que preços elevados sejam erroneamente classificados como baixos, ajudando a preservar a disciplina de compra apenas em condições vantajosas. Deste modo, a atribuição da especificidade por mais óbvia que possa parecer, é de extrema importância no mercado financeiro, sobretudo sob a perspectiva de *trades* orientados por algoritmos e modelos de ML. Isso porque as oscilações do mercado financeiro, conhecidas como volatilidade, costumam gerar pertubações as quais simulam condições favoráveis aos algoritmos quando estes não estão devidamente calibrados em relação a essa métrica.

Por fim, é possível verificar que cada tipo de métrica de avaliação auxilia em maior ou menor intensidade a tomada de decisão. Assim, escolher a métrica que melhor orienta os propósitos dos projetos é essencial na obtenção de resultados positivos.

2.5 SELEÇÃO DE VARIÁVEIS

A seleção de variáveis, ou do inglês *feature selection*, é um dos processos de maior importância no estudo e construção de modelos de aprendizado de máquina [19], processo este que ainda em expansão. Esse sistema concentra-se na seleção das variáveis preditoras que impactam de forma decisiva os modelos, nos seguintes aspectos:

- redução do *overfitting*;
- aumento da interpretabilidade e compreensão do modelo;
- redução dos custos computacionais;
- melhora da generalização da predições, e;
- melhora da precisão do modelo;

Em muitos casos, o processo de seleção de variáveis é realizado utilizando métodos de redução de dimensionalidade, tais como a análise discriminante linear (LDA - *Linear Discrimi-*

nant Analysis) e a análise de componentes principais (PCA - *Principal Component Analysis*) [37]. Contudo, estas técnicas, ao transformarem o conjunto original de dados em um novo espaço de eixos (componentes) com um número menos preditores, e portanto, simplificando o conjunto de dados, acabam por comprometer a interpretabilidade do conjunto analisado. Fato o qual, é indesejado e contrário aos pontos listados anteriormente.

Neste contexto, técnicas e procedimentos que mantêm tanto a consistência quanto a interpretabilidade dos dados são preferíveis. De tal sorte, que os próprios modelos de aprendizado de máquina podem ser considerados para a referida tarefa, em especial os algoritmos de árvore de decisão e floresta aleatória apresentados na seção 2.3. Isso ocorre porque, de acordo com [25], a seleção de variáveis ocorre inerentemente durante a construção da árvore, na qual o algoritmo escolhe a variável que melhor divide os dados em cada nó, maximizando a pureza das classes resultantes ou reduzindo a variância. Do mesmo modo, temos a floresta aleatória, a qual herda as vantagens das árvores com o acréscimo das técnicas de *bagging* e permutação de *features*, as quais criam subconjuntos de dados que nada mais são do que um processo intrínseco de seleção de variáveis em tempo de execução do próprio algoritmo. Ademais, outras técnicas como seleção progressiva também podem auxiliar nesse processo.

3. METODOLOGIA

No presente capítulo, são apresentadas as principais etapas para a construção do trabalho e obtenção dos resultados (capítulo 4). O objetivo é descrever em detalhes a criação da estratégia de investimento fundamentada em aprendizado de máquina, bem como o desenvolvimento do sistema de simulação de apostas (roleta de cassino), a fim de assegurar a transparência e a possibilidade de reprodução do estudo.

Na primeira seção, materiais e métodos, o conjunto de dados da B3 é apresentado, detalhando-se as escalas, o período de análise e os softwares utilizados no processamento desses dados. A seção seguinte, de pré-processamento, aborda o tratamento e padronização dos dados, bem como a geração de novos atributos e a definição da variável-alvo (de interesse) avaliada no presente trabalho. Por fim, mas ainda na seção de pré-processamento, uma análise descritiva preliminar dos dados é apresentada.

Na etapa de modelagem preditiva, é apresentada a justificativa para as escolhas dos algoritmos de classificação utilizados no trabalho, assim como são detalhados os processos de treinamento dos modelos e o sistema de validação cruzada, além do critério utilizado para a seleção das variáveis da modelagem final. Posteriormente, apresenta-se a simulação do sistema de apostas, com a estruturação do algoritmo da roleta de cassino, os tipos de estratégias consideradas e os critérios para contabilização dos resultados. Por conseguinte, na seção sobre as métricas de avaliação, são abordadas as definições dos indicadores estatísticos aplicados aos modelos, bem como a justificativa para a escolha destes, estabelecendo ainda os parâmetros de comparação entre a estratégia de investimento quantitativa e o sistema de apostas.

Por fim, ressalta-se que todas as etapas foram implementadas na linguagem de programação *R* e os códigos estão disponíveis no repositório público *git* do autor do projeto [11]. Além disso, importante pontuar que a ordem de apresentação adotada neste capítulo prioriza a clareza didática, podendo diferir parcialmente da lógica de execução dos *scripts* computacionais disponibilizados.

3.1 MATERIAIS E MÉTODOS

Para a execução do presente trabalho, dois sistemas de extração de dados foram testados durante a fase de obtenção dos valores (de negociação) das ações no mercado financeiro brasileiro. O primeiro utilizou a conexão com a plataforma de negociação de ativos Metatrader 5 (MT5). Nesse sistema, a comunicação era processada utilizando a linguagem de programação

Python, que, ao contrário da linguagem *R*, possui integração nativa com o MT5. Assim, o fluxo de solicitação dos preços das ações tinha início com a chamada do código R ao código em *Python*; de modo que este enviava a requisição ao MT5, o qual, por sua vez, fazia a ponte com a corretora de valores retornando as informações necessárias. Essa estrutura possibilitou a coleta de preços de abertura (*open*), fechamento (*close*), máxima (*high*), mínima (*low*) e volume (*volume*) de mais de 80 ativos da B3, nas escalas diária, semanal, mensal e anual, no período de 2002 a 2024, totalizando 23 anos de dados. Todavia, esse caminho apresentou inconvenientes relevantes e, por essa razão, foi desconsiderado. Entre os principais, destacam-se a obrigatoriedade de criação de conta em corretora; o uso de um código intermediário que funcionava apenas como ponte entre R e MT5; e, por fim, a limitação do próprio MT5, que opera de forma adequada apenas em sistema operacional Windows, o que restringe a reprodução deste estudo em diferentes ambientes (distribuições baseadas em *Unix*).

Neste contexto, o segundo sistema de extração de dados foi escolhido e mantido como fonte de informações do projeto. Uma vez que, esse sistema utiliza exclusivamente a linguagem *R* e bibliotecas disponíveis no CRAN, entre elas a *yfR*, que se conecta ao *Yahoo Finance* e provê os dados necessários dos ativos da B3 sem conexões intermediárias. Assim, por não depender de softwares auxiliares, este sistema conferiu maior versatilidade ao projeto, uma vez que dispensa, além dos softwares auxiliares, também a mediação de uma corretora de valores. Possibilitando, dessa forma que o mesmo seja utilizado em múltiplos sistemas operacionais, tais como: Windows, Linux e MacOS. Entretanto, apesar das vantagens enumeradas, o método de extração de dados em questão apresenta limitações de volume de dados, já que nem todos os ativos dispõem de dados completos nas escalas e nos períodos solicitados. Como consequência, dos mais de 80 ativos que compõem o índice Ibovespa, apenas 30 puderam ser extraídos no momento da execução do presente trabalho, sendo estes listados na Tabela 3.1, apresentada a seguir.

Tabela 3.1: Conjunto de (30) ativos disponíveis para a consulta e extração via *api* do *yfinance* em R.

Tickers – Ativos				
ABEV3	CMIG4	EMBR3	PETR3	VIVT3
BBAS3	CPFE3	GGBR4	PETR4	WEGE3
BBDC3	CPLE6	GOAU4	RADL3	
BHIA3	CSNA3	ITSA4	UGPA3	
BRAP4	CYRE3	ITUB4	SBSP3	
BRFS3	EGIE3	LREN3	USIM5	
BRKM5	ELET3	PCAR3	VALE3	

Fonte: Elaborada pelo autor.

Na tabela apresentada acima, Tabela 3.1, as siglas correspondem aos *tickers* dos ativos extraídos do *Yahoo Finance* via *api*, descritos no Apêndice A. Esses *tickers* representam os ativos que puderam ser requisitados por meio da *api* mencionada e que como base nas datas de solicitação (início e fim), bem como nas escalas (diária, semanal, mensal e anual), retornavam mais de 75% das informações solicitadas. Tal porcentagem, mínima, de retorno é um

setup/requerimento da própria *api* para o envio das informações ao requerente.

3.2 Pré-processamento dos Dados

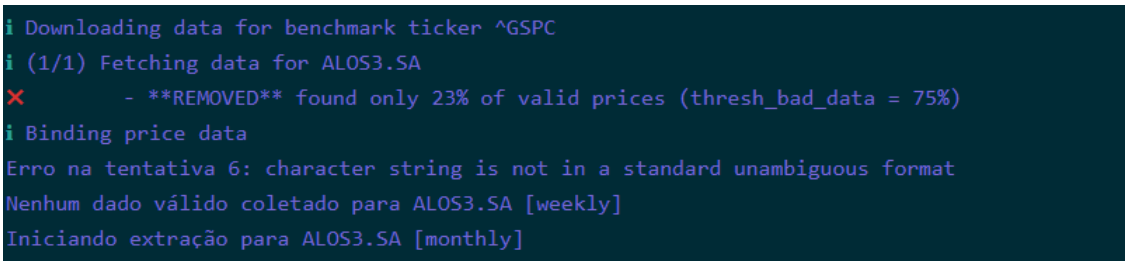
O tratamento dos dados, no que tange ao sistema financeiro simulado, foi realizado sob a perspectiva dos tópicos apresentados abaixo.

- Tratamento de dados ausentes ou irregulares;
- Filtragem e análise descritiva dos dados;
- Criação de variáveis derivadas (médias móveis, volatilidade, retornos etc.);
- Definição do grupo de classes.

Tratamento de dados ausentes ou irregulares

O tratamento de dados ausentes foi avaliado sob duas perspectivas. A primeira relacionada à própria estrutura **"NAN"** (*not a number*) que caracteriza a ausência de dados. Neste caso, o retorno da *api* de consulta no *Yahoo Finance* já realiza o tratamento em relação a ausência desses, uma vez que esta, performa, classificando como inapropriado o envio das bases que não satisfazem o limite mínimo de 75% de informações disponíveis em relação ao período e escala solicitada, tal como apresentado a seguir (Figura 3.1).

Figura 3.1: Limite mínimo de envio de dados.



Fonte: Elaborada pelo autor.

No exemplo da Figura 3.1, a plataforma de serviços Allos (**ALOS3.SA**) teve para a escala semanal (*weekly*) disponível apenas 23% dos dados solicitados, desta forma, a própria *api* classificou como inapropriado o envio desses dados ao usuário.

Em relação a segunda perspectiva, o tratamento de dados está associado a variação ou magnitude destoante de alguns dos valores (preços ou volumes) presentes no conjunto de dados, os quais no mercado financeiro frequentemente retratam o procedimento de *split*, *inplit* ou de ajuste de preços. No caso da **UGPA3.SA** (grupo Ultrapar), apresentado na Figura 3.2, o processo realizado foi de *split* de ações, no qual a empresa aumenta a quantidade de papéis em circulação, reduzindo proporcionalmente o preço unitário destes, sem no entanto alterar o valor total aplicado pelos acionistas/investidores ou o valor de mercado da companhia, tal como exemplificado por [38].

Figura 3.2: Dados anuais da UGPA3.SA antes e depois do *split* de ações.

UGPA3_yearly.csv

Data: UGPA3_yearly.csv as Plain Text File

Date DATE	Open DOUBLE	High DOUBLE	Low DOUBLE	Close DOUBLE
2002-12-11	3780717.50	4347826.00	3308128.25	3780717.50
2003-01-01	6162571.50	6162571.50	3780717.50	6162571.50
2004-01-01	6068052.00	6427221.50	4725898.00	6068052.00
2005-01-03	6068052.00	6068052.00	6068052.00	6068052.00
2006-01-02	6068052.00	6068052.00	6068052.00	6068052.00
2007-01-02	3.46	6068052.00	2.90	3.46
2008-01-02	3.37	3.48	3.37	3.37
2009-01-02	3.37	3.37	3.37	3.37
2010-01-04	2.88	3.37	2.88	2.88

Fonte: Elaborada pelo autor.

Desta forma, os dados pré-*split* da Ultrapar (anteriores a 2007) foram removidos do conjunto de análise, para isso a regra apresentada na Equação 3.1, estruturada via inspeção, foi aplicada.

$$\frac{valor_t}{valor_{t-1}} < \frac{1}{10} \quad ou \quad > 10; \tag{3.1}$$

Assim, ao se aplicar a Equação 3.1 sobre os valores de abertura, máxima, mínima e fechamento, da empresa Ultrapar o conjunto de dados referente ao período de 2002 a 2006 foi removido, preservando apenas aqueles que representam a evolução regular do ativo após o processo de *split*. Essa etapa, apesar de simples, é fundamental tanto para a análise descritiva quanto para o treinamento dos modelos de aprendizado de máquina, uma vez que impede que valores considerados *outliers* distorçam a distribuição dos dados, favorecendo ou prejudicando a seleção/escolha do ativo em questão. Dessa forma, o mesmo procedimento foi estendido a toda a base de dados, tendo como referência a escala anual, escolhida pelo autor do presente trabalho como critério de filtragem na etapa de tratamento e seleção dos ativos.

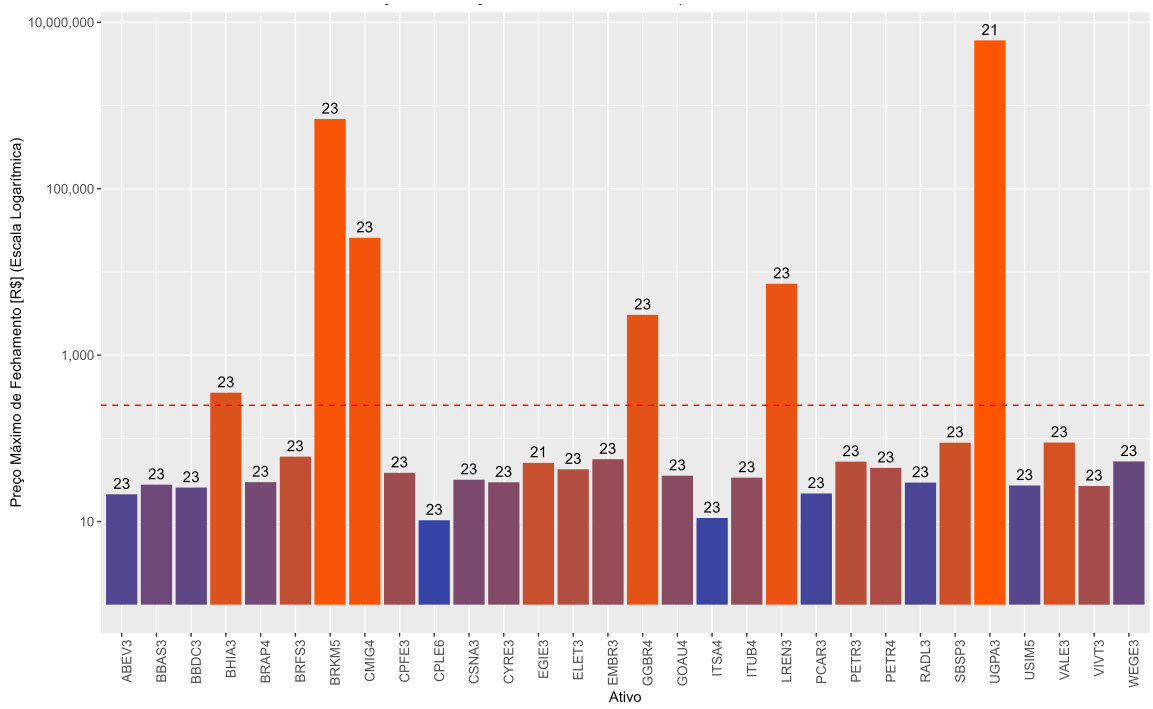
Por fim, outras fontes de perturbação nos dados, que eventualmente surgiram, também foram corrigidas por meio da Equação 3.1. Isso porque, embora inicialmente elaborada para um propósito específico, essa equação mostrou-se eficaz em ajustar distorções numéricas diversas, uma vez que a mesma realiza a avaliação do limite de corte estabelecido entre os valores tabulares para a exclusão ou não dos dados que venham a apresentar distorções.

FILTRAGEM

O processo de filtragem, e, portanto, de pré-seleção dos ativos, deu-se utilizando a escala anual do conjunto de dados obtidos do *yfinance*, assim como mencionado na subseção acima. Tal escala foi utilizada, pois permitiu que longos períodos de dados fossem avaliados de forma concisa. Neste contexto, como medida de corte e seleção (filtragem), três critérios foram estabelecidos, sendo os dois primeiros: a cobertura temporal de 23 anos consecutivos e completos (2002 a 2024) em relação ao conjunto de dados avaliados, e o limite de preço máximo de fechamento

no período mencionado, fixado arbitrariamente em R$ 250,00, um quarto do milhar. Deste modo, os dois primeiros critérios, apresentados, foram definidos buscando assegurar que todos os papéis possuíssem séries históricas equivalentes e, ao mesmo tempo, permanecessem acessíveis aos investidores comuns, simulando um cenário em que se dispõe de recursos limitados, mas com um histórico completo para análise.

Figura 3.3: Preço máximo de fechamento (escala logarítmica [R$]) e frequência anual de preços com registros diferentes de zero por ativo entre os anos de 2002 e 2024. Elaborada pelo autor.



Fonte: Elaborada pelo autor.

A aplicação dos dois filtros mencionados anteriormente pode ser observada na Figura 3.3. Nela, para cada um dos 30 ativos avaliados, é apresentado o preço anual máximo em escala logarítmica, acompanhado da respectiva frequência de anos disponíveis. A linha tracejada em vermelho indica o limite de preço de R$ 250,00, adotado como ponto de corte superior para exclusão de papéis historicamente caros. Pelo referido processo de filtragem, foram mantidos apenas os ativos que, além de respeitar esse limite de preço, possuíam séries históricas completas em todos os anos do intervalo estudado, ou seja, 23 anos de dados completos. Essa seleção, resultante da exclusão de **BHIA3**, **BRKM5**, **CMIG4**, **GGBR4**, **LREN3**, **UGPA3** e **EGIE3**, garantiu um conjunto de ativos mais homogêneos, acessíveis e historicamente consistentes, adequado para a aplicação do critério seguinte, o de definição final dos ativos-alvo.

Assim, de posse do conjunto final de ativos candidatos, os 24 restantes foram avaliados sob a perspectiva quantitativa de participação no índice nacional da bolsa de valores, apresentado na seção 2.2, Tabela 2.2. Nessa tabela constam os 8 ativos de maior participação na composição do índice Ibovespa. Dentre esses, foi definido pelo autor que 4 seriam selecionados, sendo considerados aqueles com maior representatividade percentual e pertencentes a setores distintos. Dessa forma, chegou-se à seguinte composição, apresentada a seguir:

Tabela 3.2: Ativos selecionados para as simulações computacionais.

Ativo/Ticker	Descrição	Setor	Participação [%]
VALE3	Vale S.A.	Materiais Básicos	10,810
ITUB4	Itaú Unibanco Holding S.A.	Financeiro	8,230
PETR4	Petróleo Brasileiro S.A. – Petrobras	Petróleo, Gás e Biocombustíveis	6,496
SBSP3	Companhia de Saneamento Básico do Estado de São Paulo – Sabesp	Utilidade Pública	3,692

Fonte: Extraída de [5] - Alterada.

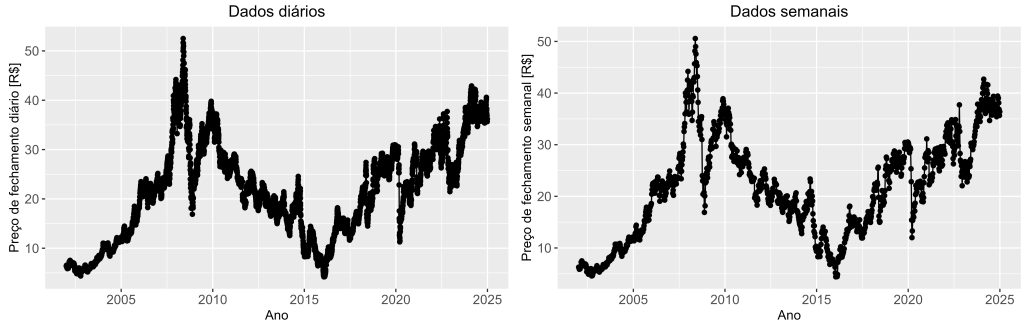
Neste contexto, tendo estabelecido os ativos finais de avaliação, Tabela 3.2, é importante salientar que embora tenham sido escolhidos 4 dos 8 primeiros papéis que compõem o Ibovespa, tal escolha não reduz a importância das duas primeiras etapas/critérios de seleção. Pelo contrário, tais critérios permanecem indispensáveis, uma vez que possuem a capacidade de modificar a lista em análise dos ativos de maior participação no índice, isto em função da disponibilidade destes, e por consequência o resultado final da seleção. Deste modo, a escolha final reflete não apenas o peso percentual de cada ativo no índice ibovespa, mas também o limite de preço por ativo, a consistência e a disponibilidade dos dados.

Análise descritiva dos dados

Na análisa descritiva dos dados, as séries históricas de Petrobras (**PETR4.SA**) são utilizadas como exemplo descritivo, Figura 3.4. Tal avaliação é de suma importância para a compreensão de dois fenômenos importantes. O primeiro associados a frequência temporal dos dados, os quais são dispostos nas escalas diária, semanal, mensal e anual. E o segundo, que corresponde a razão pelo qual as representações mais significativas de movimentação da série são dadas pelos retornos e não pelo nível real dos preços.

Assim, analisando de forma generalista a escala diária, da Figura 3.4, verifica-se que o conjunto de preços apresenta um alto nível de detalhamento, com oscilações rápidas causadas tanto pelos movimentos de mercado de curto prazo (volume de negociação) quanto por fatores sazonais. Esse tipo de representação apresenta uma periodicidade caracterizada por saltos bruscos seguidos de correções, o que evidência o que conhecemos por "heterocedasticidade" da série.

Figura 3.4: Série histórico dos preços de fechamento [R$] de petrobras (PETR4) nas escalas diária e semanal no período de 2002 à 2025.



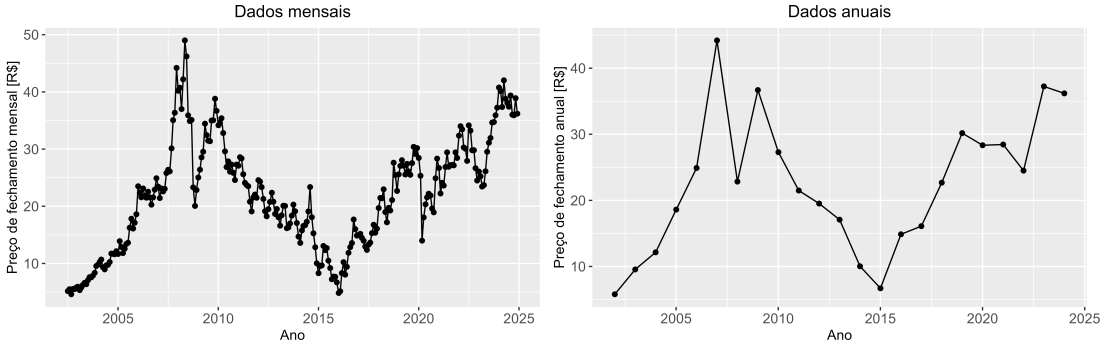
Fonte: Elaborada pelo autor.

Já na escala semanal, como pode ser observado pela figura anterior, parte dos ruídos contidos

nos dados diários são suavizados, o que possibilita uma visualização mais clara dos movimentos de médio prazo. Essa agregação destaca oscilações que se estruturam em ondas, tornando mais fácil a detecção de tendências mais sólidas (também conhecidas como macrotendências) e ciclos de valorização e/ou desvalorização. Assim, em séries semanais é normal que os padrões de volatilidade se tornem mais claros, uma vez que os choques pontuais perdem força em relação à consolidação dos preços ao longo dos pregões.

Da mesma forma, nas escalas mensais e anuais (Figura 3.5) as oscilações de preços ficam ainda mais filtradas. Destacando-se apenas as variações que têm relevância estrutural para composição da série histórica. Assim, ao consolidar as observações de aproximadamente 20 a 23 pregões para a escala mensal e de 250 pregões em média para a anual, os ruídos de curto prazo e eventos puramente especulativos de poucos dias praticamente desaparecem, o que possibilita identificar com clareza as tendência e mudanças de regime do ativo avaliado, bem como as amplas zonas de suporte e resistência que impedem que o preço do ativo se distancie da média histórica acumulada.

Figura 3.5: Série histórico dos preços de fechamento [R$] de Petrobras (PETR4) nas escalas mensal e anual no período de 2002 à 2025.



Fonte: Elaborada pelo autor.

Neste contexto, o processo de filtragem via escala/frequência dos dados (diária, semanal, mensal e anual) nos permite compreender a dinâmica dos preços de ativos financeiros em janelas distintas. De modo que, cada escala fornece um tipo de informação específica, sendo a escala:

- **diária:** o comportamento granular e heterocedástico típico do mercado financeiro no curtíssimo prazo;
- **semanal:** evidencia tendências intermediárias e suaviza a volatilidade diária, permitindo identificar padrões consistentes em horizontes de curto a médio prazo;
- **mensal:** destaca movimentos de médio prazo, refletindo ajustes sazonais e oscilações relevantes sem o ruído de curtíssimo prazo;
- **anual:** evidencia movimentos estruturais de longo prazo, associados a ciclos econômicos mais amplos e à consolidação de tendências.

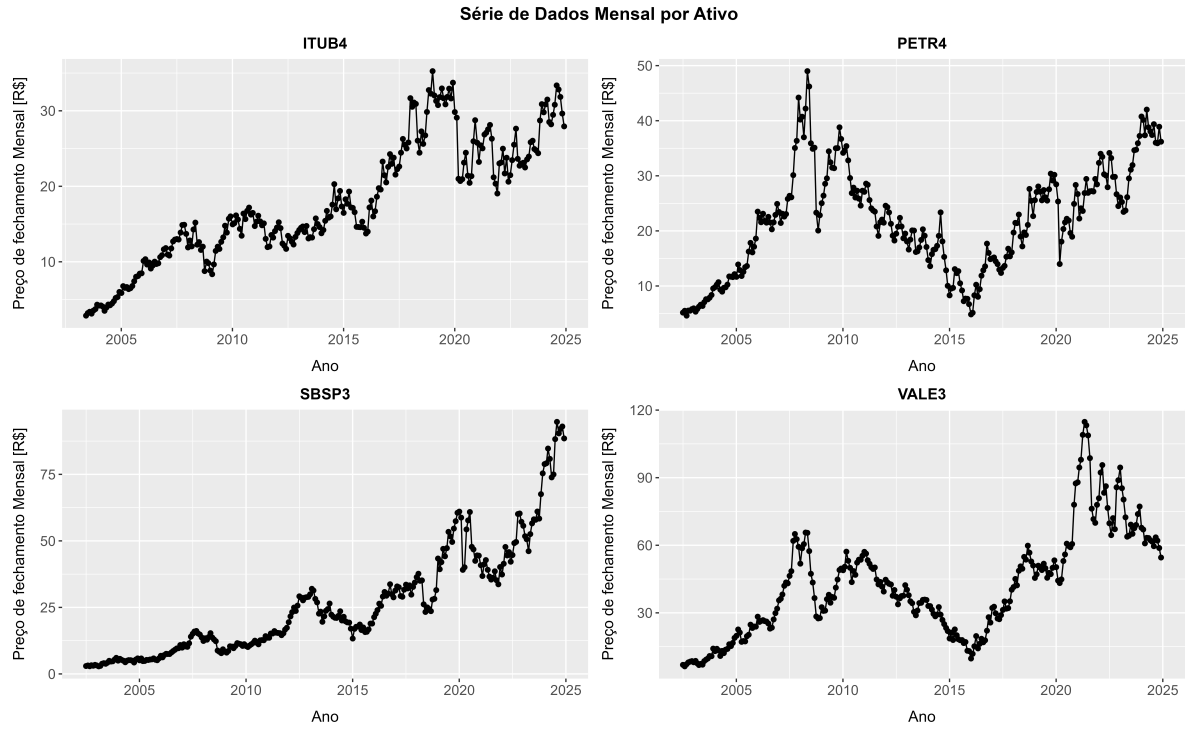
Já, o segundo fenômeno, mencionado aqui, apoia-se no fato de que apenas observar os níveis de preços não ajuda a captar a dinâmica do mercado financeiro, uma vez que esses valores são,

muitas vezes, representações especulativas associadas ao volume temporário das negociações de um ativo, o que distorce a percepção de suas variações e, por esse motivo, leva à perda da correlação natural entre períodos ($preço_t/preço_{t-1}$). Desse modo, os retornos mostram-se muito mais expressivos para a movimentação da série, pois restabelecem de forma forçada a referida correlação, tornando-se essenciais tanto para a comparabilidade entre ativos quanto para a geração de sinais de decisão dentro da própria série temporal avaliada. Tal correlação, garante uma estacionariedade aproximada dos valores e facilita a modelagem dos algoritmos de aprendizado de máquina.

Desta forma, e pela análise desenvolvida até aqui, é razoável definir que as escalas de interesse para as séries temporais dos ativos da Tabela 3.2 sejam a semanal e mensal. Isto, pelas característica suavizadas ou menos ruidosa de seus movimentos de curto e médio prazo e pela captação dos ajustes sazonais de médio-longo prazo, fatores que atendem aos requerimentos da modelagem em aprendizado de máquina.

Assim, a Figura 3.6, apresentada a seguir na escala mensal, pode ser tomada como referência para a apresentação dos perfis temporais dos ativos **ITUB4**, **PETR4**, **SBSP3** e **VALE3**, sem prejuízo à frequência semanal, a qual, assim como a primeira é um dos focos deste trabalho.

Figura 3.6: Série histórico dos preços de fechamento [R$] de ITUB4, PETR4, SBSP3 e VALE3 na escala mensal no período de 2002 à 2025.



Fonte: Elaborada pelo autor.

Na série histórica de **ITUB4**, é possível visualizar um perfil crescente com pontos de descontinuidade (variações moderadas), perfil típico de um ativo bancário consolidado. Contudo, nota-se quedas acentuadas de caráter pontual no entre período de 2007-2008, o qual é caracterizado pela crise do setor financeiro estadunidense que se alastrou para as mais diversas economias do mundo, e também no período inicial de 2020, materializando a crise de saúde

mundial da COVID-19.

Em contra partida, **PETR4** apresenta um perfil bem mais volátil, marcado por longos períodos de perda de valor com posteriores recuperações. Sendo o período de 2010 a 2015 o mais marcante no gráfico avaliado, e o qual retrata crises políticas internas somadas a queda do preço do petróleo no mercado internacional. Já, Sabesp (**SBSP3**), por outro lado, apresentou um perfil de crescimento conservador de 2005 a 2015 seguido por um *boom* associado a microcenários de volatilidade até o presente momento, delineando, desta forma, uma forte tendência de alta no contexto geral. Tendo, recentemente sido impulsionada pela atual tendência de privatização da empresa por parte do governo estadual.

Por fim, a VALE3, assim como a Petrobras, pela forte exposição ao mercado internacional de *commodities*, apresenta um perfil volátil que faz com que os preços de suas ações oscilem em torno de um valor médio de R$ 60,00 por ação. De modo que, a empresa, apesar de possuir um nível de diversificação de produtos superior à PETR4 ainda apresenta volatilidade similar, visto que ainda está muito ligada ao seu principal consumidor, a China, fato que representa um risco de mercado relevante e o qual é refletido na evolução histórica dos preços da mesma.

Criação de variáveis derivadas

Em relação a definição de variáveis derivadas, o presente trabalho utilizou um sistema de engenharia de atributos que criou uma sequência de variáveis secundárias buscando descrever ao máximo o comportamento das séries financeiras analisadas. Tal sistema foi implementado por meio da classe *FeatureEngineer* presente em [11], a qual gera automaticamente diferentes indicadores a partir dos dados de fechamento e de volume, enriquecendo a base de dados em diferentes horizontes/janelas (1, 3, 5, 20 e 30 períodos). Tais variáveis podem ser visualizadas na Tabela 3.3, a qual também apresenta as variáveis principais (sublinhas): *Date* (data), *Open* (abertura), *High* (máxima), *Low* (mínima), *Close* (fechamento), *Volume* (volume).

Tabela 3.3: Conjunto de *features* avaliadas para cada ativos.

Feature	Descrição
Date	Data associada ao preço avaliado do ativo objeto.
Open	Preço de abertura do ativo objeto no pregão.
High	Preço máximo do ativo objeto no pregão.
Low	Preço mínimo do ativo objeto no pregão.
Close	Preço de fechamento do ativo objeto no pregão.
Volume	Quantidade de negociações (volume de comercialização) do ativo objeto.
Return_{num}c	Retorno percentual acumulado em janelas de {num} pregões
Volatility_{num}c	Volatilidade histórica do ativo medida em janelas de {num} pregões
Avg_volume_{num}c	Volume médio negociado do ativo em janelas de {num} pregões
EMA_{num}	Média móvel exponencial do preço de fechamento calculada sobre {num} pregões
VolumeChange_{num}	Variação percentual do volume negociado em relação à média dos últimos {num} pregões
Return1c_EMA{num}	Retorno do ativo em 1 pregão comparado com a EMA de {num} períodos
Return1c_SMA{num}	Retorno do ativo em 1 pregão comparado com a média móvel simples (SMA) de {num} períodos

Fonte: Elaborada pelo autor.

Neste contexto, a criação das variáveis derivadas teve como objetivo ampliar a capacidade explicativa e preditiva dos modelos de aprendizagem, indo além do simples uso dos preços brutos

de fechamento. Visto que, em séries financeiras, movimentos de curto prazo são influenciados por fatores como volatilidade, tendência, liquidez e intensidade de negociação. Por isso, cada grupo de variáveis foi pensado com vista a capturar um desses aspectos considerados, inicialmente, centrais.

De forma mais clara, podemos sumarizar e abstraír a ideia de cada uma dessas variáveis pelas descrições dos seus respectivos agrupamentos, apresentados a seguir.

- **Retornos:** variações percentuais acumuladas em diferentes janelas de tempo, permitindo capturar movimentos de curto e médio prazo;
- **Volatilidade:** medida do risco recente do ativo, calculada como a dispersão dos retornos em janelas móveis;
- **Médias móveis (SMA e EMA):** indicadores de tendência obtidos a partir do preço de fechamento, no qual a EMA (*exponential moving average*) atribui maior peso às observações mais recentes em relação a SMA (*simple moving average*);
- **Momentum (ROC):** taxa de variação contínua dos preços, útil para identificar a aceleração ou perda de força do movimento/tendência (de alta ou de baixa);
- **Retornos suavizados:** retorno de um período ajustado por médias móveis, reduzindo ruídos de curto prazo;
- **Variação de volume:** comparação percentual entre volumes negociados em diferentes janelas, aproximando o efeito da liquidez.

DEFINIÇÃO DO GRUPO DE CLASSES

O grupo de classes no presente trabalho foi definido tal como apresentado na subseção 2.4.1 (*Métricas de Avaliação de Modelos*), mais especificamente no que tange o *cenário 1* de investimento. Neste, o investimento é feito objetivando o acúmulo patrimonial por meio de múltiplas compras. Assim, a classe definida como variável alvo (dependente) foi a de *compra simples*, a qual é construída sobre a variável derivada do retorno de janela de 1 período, de modo que a regra de decisão segue a seguinte lógica:

- **Compra**: ocorre quando o retorno esperado r_t é maior ou igual ao limite de referência (LMR) previamente definido.
  Nesse caso, quando o sinal de compra é emitido entende-se que há grandes chances de valorização do ativo, considerando-se a relação presente-passado associada ao retorno do preço de fechamento, o que justifica a realização da operação;
- **Sem compra**: ocorre quando o retorno esperado r_t é menor que o limite de referência definido.
  Nessa situação, a expectativa de ganho é considerada insuficiente, e a operação não é realizada, tendo em vista a mesma relação presente-passado avaliada no mecanismo de compra.

Assim, a estrutura descrita acima garante que a decisão de *compra* esteja fundamentada não apenas no nível bruto do preço de fechamento dos ativos, mas também na relação direta entre o retorno dos períodos anteriores da série histórica. Uma vez que, o retorno é a interação atrasada (ou adiantada a depender do contexto) entre os preços do ativo avaliado.

Neste ponto, portanto, é importante elucidar o que foi definido como limite de referência (LMR) na descrição da classe objetivo (de decisão), e que é utilizado como ponto de operação no presente trabalho. O LMR trata-se de um valor estatístico extraído diretamente da série histórica dos retornos de cada ativo, o qual é obtido por meio da avaliação descritiva das medidas de tendência central, ou seja, para cada ponto de retorno dos ativos-alvo em sua respectiva escala (semanal ou mensal) as medidas de tendência central são calculadas, tal como apresentado na Tabela 3.4, e utilizadas como referência comparativa para a decisão de compra (ou não) em relação ao retorno testado do período seguinte.

Tabela 3.4: Avaliação descritiva (tendência central) dos dados relativos ao retorno de 1 período para os preços de fechamento dos ativos-alvo.

Ativo	Escala	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
PETR4	**Semanal**	0,0000	0,0139	0,0296	0,0397	0,0545	0,4825
VALE3		0,0000	0,0131	0,0298	0,0377	0,0517	0,5045
SBSP3		0,0000	0,0115	0,0277	0,0357	0,0495	0,3301
ITUB4		0,0000	0,0099	0,0248	0,0327	0,0446	0,3314
PETR4	**Mensal**	0,0003	0,0362	0,0707	0,0867	0,1072	0,6167
VALE3		0,0007	0,0314	0,0665	0,0796	0,1120	0,3140
SBSP3		0,0000	0,0361	0,0633	0,0774	0,1046	0,3768
ITUB4		0,0000	0,0238	0,0507	0,0637	0,0892	0,2784

Fonte: Elaborada pelo autor.

A Tabela 3.4, tal como mencionado acima, apresenta os resultados dos cálculos das medidas de tendência central para os ativos selecionados. Esses valores nos permitem observar como cada estatística gera um limite de referência distinto tanto em relação ao ativo como em relação a escala avaliada. Por exemplo, Petrobras tem um limite de referência de compra de 2,98% (0,0298) na escala semanal considerando a mediana como referência estatística para efetivar um *compra*, ou seja, caso o retorno entre o preço atual e o anterior atinja esse limite, então a compra deve ocorrer. Sob essa perspectiva, valores menos conservadores, como o mínimo ou o primeiro quartil, têm a tendência em produzir um número maior de sinais de compra, já que o critério de igualdade ou superação é mais facilmente atendido. Por outro lado, medidas mais conservadoras, como a média, a mediana (apresentada no exemplo acima) e o terceiro quartil, tendem a elevar o nível de corte, reduzindo a quantidade de operações, e assim, concentrando as compras em momentos que teoricamente sejam de maior expectativa de valorização do ativo avaliado.

Neste contexto, a escolha da medida de tendência central (MTC) a ser utilizada é de fato

fator determinante, tanto para o treinamento quanto para as predições dos modelos de aprendizado de máquina. De modo que, a Tabela 3.5 elaborada como parâmetro de decisão, e obtida a partir das Tabelas B.1 e B.2, é apresentada a seguir.

Tabela 3.5: Média dos valores de tendência central referentes ao retorno dos preços de fechamento classificados percetualmente em *Compra* e *Sem Compra* para os ativos-alvo (**PETR4**, **VALE3**, **SBSP3**, **ITUB4**).

Medida de Tendência	**Grupo**	**Escala**	
		Semanal [%]	**Mensal [%]**
Mínimo	**Compra**	47,54	45,81
	Sem Compra	52,46	54,19
1º Quartil	**Compra**	36,81	33,69
	Sem Compra	63,19	66,31
Mediana	**Compra**	24,33	21,89
	Sem Compra	75,67	78,11
Média	**Compra**	19,00	17,77
	Sem Compra	81,00	82,23
3º Quartil	**Compra**	12,65	10,95
	Sem Compra	87,35	89,05
Máximo	**Compra**	0,02	0,66
	Sem Compra	99,98	99,34

Fonte: Elaborada pelo autor.

Na Tabela sumarizada 3.5, os valores apresentados foram agregados pelas MTCs ao invés de manter a separação por ativo e escala, como nas tabelas anteriormente mencionadas, evidenciando, assim, as médias percentuais gerais para cada grupo associado ("*compra*" e "*sem compra*") às medidas de tendência. Fato que permite consolidar os efeitos médios de cada estatística e produzir como informação adicional o balanceamento da variável-alvo por medida.

Neste sentido, e buscando um conjundo de variáveis respostas com perfil naturalmente balanceado e que permita a emissão de sinais de compra com maior frequência, a MTC mais adequada aos propósitos do presente trabalho são os valores associados ao mínimo da Tabela 3.4, os quais geram um conjunto de dados balanceados na proporção semanal média de 47,54% de dados "*compra*" contra 52,46% "*sem compra*", e de 45,81% "*compra*" contra 54,19% "*sem compra*" na escala mensal.

Assim, no presente trabalho utilizou-se como limite de referência para os sinais de compra a coluna de mínimo da Tabela 3.4 para ambas as escalas testadas.

3.3 Modelagem Preditiva

A modelagem preditiva do presentes trabalho, deu-se de maneira compartimentalizada e seguindo, desta forma, as etapas listas a seguir.

- Algoritmos utilizados;

  - Justificativa geral;
  - Hiperparâmetrização e validação cruzada;
  - Quantidade de modelos testados.

- Escolha dos melhores atributos

Como característica de criação dos modelos de aprendizado de máquina, o conjunto de dados foi particionado para todos os casos utilizando a proporção 80%-20%. Na qual 80% dos dados são direcionados ao treinamento e 20% associados a etapa de teste dos modelos. De modo que, em cada ano considerado, os doze meses completos foram utilizados como balizadores da modelagem, preservando as variações sazonais ao longo do tempo. Assim, os anos de 2002 a 2017 compuseram a etapa de treinamento, enquanto o período de 2018 a 2021 foi reservado para os testes, sendo os anos de 2022 a 2024 direcionados às operações avaliativas dos modelos finais. Aqui, ainda é importante ressaltar que, apesar de todos os doze meses de cada ano terem sido empregados para o treino e teste (como já mencionado), dezembro não foi incluído nas operações de compra sinalizadas pelos modelos na etapa avaliativa. Isso porque, embora seja importante para o aprendizado numérico dos modelos, o referido mês apresente comportamento de mercado pouco lucrativo, com baixa liquidez e alta volatilidade, o que poderia prejudicar uma avaliação prática de desempenho das estratégias.

Algoritmos utilizados

Os algoritmos utilizados nas simulações computacionais do referido trabalho, são os mesmos daqueles apresentados no capítulo 2. Sendo estes: o k-NN, árvore de decisão e a floresta aleatória, todos considerados algoritmos de aprendizado de máquina clássicos de característica supervisionada.

Deste modo, ***k*-NN** foi utilizado no presente trabalho pela simplicidade técnica e conceitual, uma vez que este algoritmo tem como pressuposto a classificação de novas entidades por meio da comparação da distância entre os elementos vizinhos e o objeto em avaliação. Adicionalmente, este algoritmo se destaca por seu elevado custo-benefício, visto que, combina baixo esforço de implementação e interpretação a resultados consistentes em diversos cenários. Sob tal perspectiva, o referido modelo foi ajustado (hiperparametrizado) utilizando o conceito de *pipeline* em grade do sistema *tidymodel* da linguagem de programação R, construído com base em três fatores principais.

- **Número de vizinhos** (k): tal parâmetro controla a quantidade de elementos vizinhos considerados durante a etapa de classificação, influenciando diretamente o viés e a variância do modelo. Assim, o mesmo foi posto a variar de 1 até 10, em passos unitários.

- **Função de ponderação das distâncias** (`weight_func`): neste hiperparâmetro temos o controle de como cada vizinho é ajustado em relação a sua distância ao ponto-alvo, dando via modificação da função de ponderação maior relevância aos vizinhos mais próximos e

reduzir o impacto dos mais distantes. De modo que, aqui foram testados três formas distintas de atribuir pesos aos vizinhos, sendo as formas `rectangular`, `triangular` e `gaussian`. Sendo a primeira responsável por atribui pesos iguais a todos os vizinhos, ao passo que as demais privilegiam observações mais próximas do ponto a ser classificado.

- **Potência da distância (`dist_power`)**: este hiperparâmetro, em suma, representa o coeficiente p da equação apresentada na sub-seção 2.3.2, o qual determina o tipo de equação a ser usado no cálculo da distância entre as entidades avaliadas, distância Manhattan (L_1) e Euclidiana (L_2). Podendo, portanto, p ser definido como 1 ou 2.

Assim, a combinação desses parâmetros resultou em uma busca combinada (*grid search*) de $10 \times 3 \times 2 = 60$ configurações distintas do modelo. Em seguida, cada configuração foi avaliada em diferentes esquemas de validação cruzada ($v = 3$, $v = 5$, $v = 7$ e $v = 10$), totalizando 240 testes (60×4). Por fim, a seleção do melhor conjunto de hiperparâmetros foi realizada de acordo com a métrica da sensibilidade, a qual, de acordo com o apresentado na capítulo 2, prioriza a correta identificação do momento de *"compra"* mediante ao custo de oportunidade.

Já o algoritmo de **árvore de decisão** foi aplicado no presente trabalho pela sua capacidade de predição associada ao seu excelente processo de visualização hierárquica, ambos aliados a um eficiente sistema de (re)amostragem para a construção do processo de classificação. De modo que, essas características geram uma garantia de maior generalidade e controle sobre o modelo (com a definição da profundidade da árvore), além de favorecerem a interpretabilidade das decisões. Tal como no k-NN, este modelo foi calibrado utilizando o pacote *tidymodels*, em uma estrutura de *grid search*, sendo avaliados os principais hiperparâmetros, os quais são apresentados a seguir.

- **Custo de complexidade (`cost_complexity`)**: este hiperparâmetro controla a penalização associada à adição de novos ramos na árvore. De modo que, um custo de complexidade mais baixo permite estruturas mais profundas e complexas, enquanto valores maiores de custo favorecem as árvores mais simples e generalistas. Assim, pela capacidade em promover ou reduzir o processo de crescimento da árvore este hiperparâmetro está diretamente relacionado ao processo de *poda*, pois atua reduzindo divisões que não trazem ganho substancial de desempenho ao modelo, evitando o processo de *overfitting*.

- **Profundidade máxima (`tree_depth`)**: este hiperparâmetro controla desde o início do processo o número de divisões sucessivas da árvore. Dessa forma, regula a capacidade do modelo em capturar relações não lineares complexas, ao mesmo tempo em que previne o sobreajuste (*overfitting*). Embora não seja exatatamente um processo de poda como regularmente conhecemos, este exerce papel semelhante no controle da complexidade do modelo, estabelecendo um limite estrutural da árvore, característica frequentemente ajustada em abordagens de *AutoML*.

- **Número mínimo de observações por nó terminal (`min_n`)**: assim com pode ser inferido pela nomenclatura do tópico, este hieprparâmetro define o tamanho mínimo para

que um nó seja considerado válido. De modo que, valores mais altos restringem a criação de divisões muito específicas, promovendo maior estabilidade do modelo e por consequência reduzindo o *overfitting*.

Utilizando o conjunto de hiperparâmetros apresentados acima, o *grid search* (grade de busca) construído no presente experimento gerou um conjunto de avaliação que totalizou 216 combinações (6×6×6). Sendo cada configuração avaliada sob esquema de validação cruzada com diferentes números de *folds* ($v = 3$, $v = 5$, $v = 7$ e $v = 10$), somando, deste modo, 864 modelos testados. Ao fim, a métrica de sensibilidade foi priorizada para a escolha do melhores conjunto de hiperparâmetros.

Por fim, o modelo de **floresta aleatória** também foi utilizado no presente trabalho por se tratar de um algoritmo *ensemble* que se vale dos principais pontos positivos do modelo de árvores de decisão para a criação de múltiplas instâncias de avaliação, ponto este que introduz aleatoriedade tanto na seleção dos preditores (`mtry`) quanto na amostragem dos dados (*bagging*). De modo que, tal fato força cada árvore a seguir caminhos diferentes, mesmo quando treinadas sobre dados semelhantes, aumentando a diversidade do *ensemble*. Deste forma, esse procedimento reduz de maneira significativa a variância do modelo final e aumenta a robustez preditiva, característica fundamental em cenários de elevada volatilidade. Assim, o presente modelo foi calibrado por meio de cinco valores igualmente espaçados para cada um dos três hiperparâmetros avaliados, como apresentado a seguir.

- **Variáveis candidatas (`mtry`)**: corresponde à quantidade de variáveis preditoras escolhidas aleatoriamente em cada separação da árvore, promovendo diversidade entre as árvores e diminuindo a correlação entre elas. Assim, buscando o melhor valor para o contexto do trabalho associado a esta variável, o limites da mesma foram fixados em 1 (limite inferior) e p (limite superior), com p sendo a quantidade total de preditores do conjunto de dados;

- **Quantidade mínima de observações (`min_n`)**: determina o número mínimo de instâncias/observações necessárias para que um nó possa ser dividido, controlando, desta forma, a granularidade da árvore e evitando que divisões excessivamente específicas sejam geradas. Desta forma, a otimização dessa variável se deu aplicando o intervalo de 2 instânceas a 30.

- **Número de árvores do ensemble (`trees`)**: este hiperparâmetro estabelece a quantidade de árvores independentes que vão compor a floresta aleatória, de modo que, valores mais altos tendem a reduzir a variância do modelo estabilizando as estimativas, embora impliquem maior custo computacional pela criação de novas instânceas. No processo de otimização do referido hiperparâmetro foram considerados cinco valores igualmente espaçados entre 100 e 1000.

Na floresta aleatório assim como nos algoritmos anteriormente mencionados, também foi realizado o processo de validação cruzada. O que resultou na experimentação de 500 modelos, contabilizados da seguinte forma: 5 (`mtry`) × 5 (`min_n`) × 5 (`trees`) = 125 × 4 (cv) = 500.

Finalmente, e após todo o contexto da modelagem preditiva aqui apresentado, a Tabela 3.6 apresenta de forma resumida a quantidade de experimentos/modelos gerados.

Tabela 3.6: Quantidade de modelos gerados sem/com validação cruzada.

Modelos	Sem validação cruzada	Com validação cruzada
KNN	60	240
Árvore de Decisão	216	864
Floresta Aleatória	125	500
Toal	**401 x 4 ativos = 1.604**	**1.604 x 4 ativos = 6.416**

Fonte: Elaborada pelo autor.

ESCOLHA DOS MELHORES ATRIBUTOS

No que se refere à escolha dos atributos, dois procedimentos foram conduzidos em etapas distintas. Na primeira etapa, adotou-se a estratégia de seleção pautada na relevância e na contribuição média efetiva de cada variável (derivada) ao processo de classificação dos ativos-alvo. Esse procedimento foi conduzido utilizando um algortimo de aprendizado de máquina baseado em árvores sem o processo de validação cruzada, uma vez que o objetivo, nesta etapa, era apenas o de *rankear* as variáveis testadas, buscando identificar aquelas de maior preponderância em suas respectivas escalas.

Mais especificamente, o modelo de floresta aleatória foi utilizado para a obtenção das importâncias relativas, tanto nesta quanto na etapa seguinte, executado separadamente por ativo-alvo e por escala, sempre em dados de treino e com controle de reprodutibilidade. Aqui, as importâncias resultantes das variáveis preditoras foram consolidadas por agregação média dos seus valores percentuais de contribuição, de modo a produzir um ranqueamento global de atributos para os ativos **PETR4**, **VALE3**, **SBSP3** e **ITUB4**. Esse processo ordenou na escala númerica de 100-0 % a importância das variáveis preditoras para as frequências semanal e mensal, de modo que, um corte arbitrário de 9% era necessário e foi estabelecido pelo autor. Assim, apenas variáveis com importância igual ou superior a 9% seriam consideradas para a etapa seguinte.

Na segunda etapa, buscando reduzir a dimensionaldiade do conjunto de dados e obter maior generalização, reduzindo as chances de *overfitting*, adotou-se o processo de seleção progressiva (*forward selection* - FS) de variáveis, conforme apresentado por [9]. Esse método teve por princípio a adição, de forma sequencial, de variáveis preditoras ao modelo de *ML* utilizado. De modo que, a cada nova inclusão, das variáveis pré-selecionadas o modelo era reestimado e avaliava-se a métrica de desempenho que melhor descrevia o processo. Caso a métrica apresentesse melhora, a variável era mantida; caso contrário, descartada.

Por fim, após a aplicação das duas etapas de seleção de variáveis mencionadas, as quais foram o ranqueamento inicial por importância agregada com corte e a seleção progressiva orientada à decisão, obteve-se uma lista final de preditores (apresentada no capítulo 4). Tal lista, foi considerada como a que melhor descrevia as características médias dos diferentes ativos e escalas, assegurando maior homogeneidade na base de dados dos ativos-alvo e permitindo que comparações futuras de resultados patrimoniais e de investimento sejam realizadas em cenários equivalentes.

3.4 Simulação do Sistema de Apostas

As simulações associadas ao sistema de aposta, descrito na seção 2.1, foram implementados no presente trabalho utilizando, assim como para a modelagem de aprendizado de máquina, a linguagem de programação *R*. Tal implementação deu-se tendo como referência a roleta europeia tradicional composta por 37 números (0 a 36), sendo o código estruturado com o objetivo de emular da forma mais fiel possível o funcionamento do jogo e, ao mesmo tempo, oferecer flexibilidade para a avaliação de diferentes tipos de apostas. Assim, o mesmo foi compartimentalizado como o apresentado a seguir:

- Estrutura do algoritmo da roleta
- Parâmetros de entrada e critérios de contabilização dos resultados

Estrutura do algoritmo da roleta

O algoritmo de simulação da roleta de cassino foi estruturado de acordo com os componente principais do referido jogo, sendo eles o conjunto de números, a classificação em cores (vermelho e preto) e a disposição em dúzias e colunas; de modo que tais componentes permitiram o desenvolvimento de duas funções essenciais ao processo, as quais são apresentadas a seguir:

- **Função de sorteio (`girar_roleta`)**: esta função é a responsável por selecionar de forma aleatória um número entre 0 e 36, representando o resultado de cada simulação (rodada da roleta). De sorte que, a função `sample()` é responsável pela característica aleatória e equiprovável dos resultados.
- **Função principal de apostas (`apostar`)**: corresponde a estrutura a lógica dos diferentes tipos de aposta, abrangendo opções clássicas como número único, paridade (par/impar), cor (vermelho/preto), dúzias e colunas. De modo que, para cada caso, são definidos os critérios de vitória e o valor de pagamento correspondente (e.g., 35:1 para número exato, 1:1 para paridade ou cor, 2:1 para dúzia e coluna), os quais foram descritos na sub-seção 2.1.1.

O sistema de simulação da roleta ainda tem como funcionalidade adicional o controle de saldo, o qual valida a existência ou não de recursos para a nova rodada de aposta. Com isso, caso seja necessário, a estrutura permite avaliar tanto o desempenho probabilístico de cada tipo de aposta quanto a dinâmica de variação do capital sob diferentes estratégias.

Parâmetros de entrada e critérios de contabilização dos resultados

A definição dos parâmetros iniciais da roleta desempenha um papel fundamental no processo de simulação deste trabalho. Assim, para se ter uma comparação entre o sistema de investimento e o sistema de apostas, o capital inicial da roleta foi definido com base no valor do capital de investimento calculado para o modelo que apresentou o melhor desempenho, mostrado na

seção 4.1.4. Como critério de parada, foi considerada tanto a redução total do capital até zero quanto o limite de rodadas (apostas) igual ao número de operações realizadas pelo sistema de investimento mencionado anteriormente. Já, o valor do aporte inicial da primeira rodada na roleta foi definido para que representasse o montante mínimo necessário que em uma simulação ideal com ganhos consecutivos utilizando a estratégia *Streets of Gold*, tenha havido condições de igualar ou superar o resultado do modelo de investimento em ações.

Por fim, o critério de contabilização foi criado para registrar, nada mais do que, o aspecto financeiro do jogo da roleta. De modo que, a cada rodada o algoritmo registra:

- Saldo inicial e final;
- Resultado do sorteio;
- Status da aposta;
- Histórico agregado.

Assim, o sistema de contabilização descrito aqui desempenha papel de extrema importância no processo de simulação deste trabalho, haja visto que ele possibilita o registro dos resultados das apostas (simuladas) na roleta. Sendo esse armazenamento essencial para que, em etapa posterior, sejam realizadas as comparações entre o desempenho das estratégias de apostas e os resultados obtidos por meio dos sistemas baseados em modelos de aprendizado de máquina aplicados ao mercado financeiro.

4. RESULTADOS

No presente e último capítulo, de caráter técnico, são apresentados os principais resultados do referido trabalho, obtidos a partir de investimentos puramente recorrentes (livres) e daqueles orientados por simulações computacionais. Sendo, os procedimentos computacionais realizados com modelos de aprendizado de máquina, os quais totalizaram mais de seis mil simulações numéricas, e com o sistema do tipo fortuna, o qual simulou decisões puramente estocásticas, análogas às apostas em um cassino. Neste contexto, o objetivo aqui é sintetizar os achados mais relevantes, destacando o desempenho financeiro dos modelos avaliados, contrastando-os tanto com os resultados de investimentos não orientados por modelos quanto com aqueles provenientes das apostas realizadas na roleta de cassino.

Deste modo, os resultados são detalhados de forma progressiva, passando pela seleção das variáveis preditoras utilizadas pelos modelos de aprendizado de máquina, pela evolução patrimonial das aplicações via entradas aleatórias e dos investimentos utilizando modelos estatísticos, seguido pelas métricas de avaliação dos modelos de melhor desempenho, chegando até os resultados das apostas.

Assim, o capítulo está organizado em duas seções principais: investimento em ações do Ibovespa e apostas na roleta de cassino, possibilitando não apenas a exposição dos resultados de cada uma das abordagens, mas também uma comparação crítica entre ambas, evidenciando as implicações práticas dessa análise para a tomada de decisão em contextos de risco e incerteza.

4.1 INVESTIMENTO EM AÇÕES IBOVESPA

No que se refere aos investimentos em ações, o presente trabalho tratou o processo sob a perspectiva quantitativa e pelo prisma de investimentos recorrentes não orientados, ou seja, todas as entradas foram realizadas por sinais de compra emitidos por modelos de aprendizado de máquina ou efetivadas sem nenhum critério matemático, seguindo apenas o fechamento dos preços semanais e mensais.

Assim, nessa seção serão apresentados:

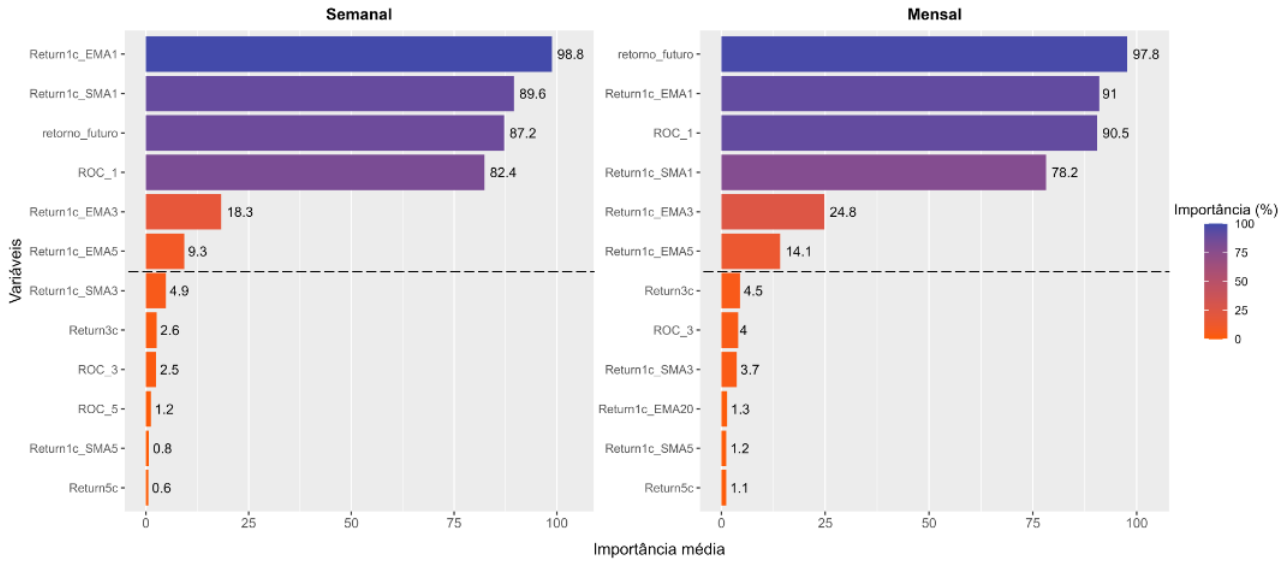
- as variáveis preditoras selecionadas para a modelagem preditiva;
- os resultados operacionais de entradas aleatórias e utilizando modelos de *ML*;
- as métricas de avaliação dos modelos de melhor desempenho;
- características do modelo escolhido como "campeão".

Neste contexto, vale destacar que os resultados operacionais serão apresentados anteriormente as métricas de avaliação, e até mesmo aos parâmetros de modelagem, pois estes balizaram a escolha tanto do ativo de melhor desempenho como do modelo que propiciou tal resultado.

4.1.1 VARIÁVEIS PREDITORAS

Como mencionado no capítulo anterior (3), as variáveis preditoras foram escolhidas após duas etapas de seleção. Na primeira etapa, **PETR4**, **VALE3**, **SBSP3** e **ITUB4** nas escalas semanal e mensal, definidas na subseção 3.2, tiveram seus respectivos conjuntos de dados processados pelo algoritmo da floresta aleatória. Produzindo como resultado o *rank* da Figura 4.1, no qual observa-se que determinadas variáveis assumem maior peso relativo na explicação dos sinais de compra, destacando-se como preditores centrais no processo de classificação.

Figura 4.1: Atributos (*features*) com maior importância média no processamento das predições dos pontos de entrada nos ativos-alvo, considerando as escalas semanal e mensal.



Fonte: Elaborada pelo autor.

Como resultado parcial, a ordenação da figura anterior (Figura 4.1) apresenta o ranqueamento médio das importâncias percentuais dos ativos-alvo. Sob tal perspectiva, nota-se que atributos como retornos defasados e médias móveis apresentaram maior peso relativo, assumindo papel central na explicação dos sinais de compra. Assim, e a partir desses resultados, foram consideradas como relevantes as variáveis que obtiveram importância média igual ou superior a 9%, resultando na seguinte seleção de variáveis preditoras:

- Return1c_EMA1;
- Return1c_SMA1;
- Return1c_EMA3;
- Return1c_EMA5;
- retorno_futuro;
- ROC_1.

Em etapa seguinte, buscando-se refinar o conjunto das variáveis preditoras, e consequentemente de reduzir as chances de *overfitting*, adotou-se o processo de seleção progressiva (*forward*

selection - FS) de variáveis, conforme apresentado por [9] e na seção metodológica. Resultando na escolha das variáveis preditoras Return1c_EMA3 e Return1c_EMA5. Tais variáveis, juntamente com as definidas previamente pelo autor e apresentadas na Tabela 3.3 (*Open*, *High*, *Low*, *Close* e *Volume*), compuseram, assim, o conjunto final de preditores utilizados em todas as simulações dos modelos de aprendizado de máquina realizadas neste trabalho.

Assim, podemos elencar o conjunto resultante de variáveis preditoras obtidas.

- **Definidas pelo autor:**
  - *Open*;
  - *High*;
  - *Low*;
  - *Close*;
  - *Volume*;
- **Obtidas pelo processo de seleção de variáveis:**
  - Return1c_EMA3;
  - Return1c_EMA5.

Por fim, o resultado da seleção das variáveis preditoras orientou a modelagem preditiva (seção 3.3) do presente trabalho, uma vez que, tornou o processo de modelagem generalista e com menor custo de processamento computacional.

4.1.2 RESULTADOS OPERACIONAIS

Antes de avaliar o desempenho operacional dos investimentos com modelos de aprendizado de máquina, e mesmo antes de compará-los aos resultados obtidos pelo sistema de apostas, é necessário estabelecer uma referência comparativa para essa modalidade de aporte financeiro, a qual permita verificar se o emprego de técnicas computacionais sofisticadas, como as de apredizado de máquina, são justificadas. Para isso, o sistema de investimento de característica livre e recorrente (aleatório) em ativos da bolsa de valores é utilizado. Nessa modalidade, os valores são investidos sem o direcionamento de estratégias econométricas, fundamentalistas ou de *softwares* computacionais, seguindo apenas a premissa de aportes recorrentes no fechamento das escalas escolhidas (semanais e mensais). Tal como apresentado na Tabela 4.1, a seguir.

Tabela 4.1: Resultados semanais e mensais das compras realizadas de forma aleatória.

Ativo	Escala	Invest. [R$]	Valor da carteira [R$]	Retorno [%]	Nº de compradas
ITUB4	**Semanal**	3.782,25	4.267,64	12,83	144
PETR4		4.732,93	5.601,60	18,35	144
SBSP3		8.959,09	13.389,12	49,45	144
VALE3		10.390,67	8.464,32	-18,54	144
ITUB4	**Mensal**	866,55	978,00	12,86	33
PETR4		1.088,89	1.283,70	17,89	33
SBSP3		2.069,82	3.068,34	48,24	33
VALE3		2.369,01	1.939,74	-18,12	33

Fonte: Elaborada pelo autor.

Em termos descritivos, na tabela acima (4.1), de investimentos no formato livre, a coluna *Ativo* apresenta os ativos-alvo pré-selecionados na seção de 3.2 e utilizados como objetos de investigação do presente trabalho. Também são apresentadas as escalas (semanais e mensais) nas quais os investimentos foram realizados, coluna *Escala*. Em seguida, verifica-se a coluna *Invest. [R$]*, a qual representa os valores totais aportados por ativo ao longo das 144 semanas (escala semanal) ou dos 33 meses (escala mensal) de investimento (*Nº de compradas*). Ainda sobre a coluna *Invest. [R$]*, observa-se que os valores aplicados variam de ativo para ativo. Essa diferença decorre essencialmente da variação do preço de fechamento de cada ativo na data de obtenção do mesmo, de modo que ativos com preços diferentes naturalmente demandam montantes distintos de investimento.

Já a coluna *Valor da carteira [R$]*, representa o valor monetário final associado ao acumulo das ações, sendo 144 ações para a escala semanal e 33 para a mensal, até novembro de 2024, após investimentos recorrentes e consecutivos de 3 anos (*Invest. [R$]*). Sob a perspectiva de aquisão de ativos, vale ressaltar que, seguiu-se na presente análise a regra de aquisição de apenas uma ação por vez a cada período de investimento, de modo que, o número de operações de compra representa exatamente a mesma quantidade de ações acumuladas.

Particularizando a análise aos resultados quantitativos dos investimentos de característica livre, fica evidente o comportamento do conjunto. Em termos de retornos percentuais, os ativos equivalentes têm praticamente o mesmo desempenho tanto na escala semanal como mensal, diferindo em média de 0,53 pontos percentuais positvios para a escala semanal. Todavia, a diferença acentua-se quando comparado ativos distintos. Assim, escolhendo a escala mensal como referência em função do menor aporte (*Invest. [R$]*) realizado por ativo – fato que minimiza o risco de exposição ao mercado, verifica-se que SBSP3 (48,24%) apresenta o melhor desempenho dentre os quatro ativos, seguida por PETR4 (17,89%) e ITUB4 (12,86%), ao passo que VALE3 registrou resultado negativo, com perda de 18,12% do valor aplicado. Por fim, observa-se ainda que, embora o volume investido varie conforme a quantidade de operações realizadas em cada escala, a relação entre investimento e retorno manteve-se consistente, reforçando a robustez das tendências identificadas e apresentadas na sub-seção 3.2.

Tendo, então, sido estabelecido o referencial de desempenho para o sistema de investimento em ações da B3, torna-se possível avaliar de forma comparativa a efetividade do emprego de técnicas de aprendizado de máquina neste contexto. Assim, a Tabela 4.2 é apresentada. Nela os resultados dos investimento orientados por modelos de aprendizado de máquina, tais como: k-NN, árvore de decisão e floresta aleatória nas escalas semanais e mensais são expostos.

Em termos gerais, a Tabela 4.2, a seguir, mantém a lógica de exposição dos resultados introduzidos pela Tabela 4.1, de modo que são apresentados os ativos avaliados (*Ativo*), o valor total investido (*Invest. [R$]*), o valor final acumulado da carteira (*Valor da carteira [R$]*), o retorno percentual (*Retorno [%]*) e o número de compras efetuadas (*N° de compras*), mantendo também a regra de aquisição de uma ação a cada período de investimento. De forma adicional, na coluna *Modelo* são apresentados os modelos de aprendizado de máquina correspondentes a cada conjunto de resultado.

Tabela 4.2: Resultados semanais e mensais das compras realizadas utilizando os sinais dos modelos de aprendizado de máquina.

Ativo	Modelo	Operações semanais				Operações mensais			
		Invest. [R$]	Valor da carteira [R$]	Retorno [%]	Nº de compradas	Invest. [R$]	Valor da carteira [R$]	Retorno [%]	Nº de compradas
ITUB4	**KNN**	1.692,45	1.956,00	15,57	66	401,70	474,18	18,04	16
PETR4		2.025,71	2.279,34	12,52	63	509,83	574,56	12,70	16
SBSP3		4.435,80	6.694,56	50,92	72	868,84	1.394,7	60,52	15
VALE3		5.494,92	4.479,86	-18,47	77	1.100,42	992,96	-9,77	16
ITUB4	**Árvore de Decisão**	1.747,55	1.985,64	13,62	67	276,06	326,00	18,09	11
PETR4		2.182,60	2.460,24	12,72	68	600,63	682,29	13,60	19
SBSP3		3.566,48	5.485,82	53,82	59	1.061,30	1.580,66	48,94	17
VALE3		4437,26	3.580,92	-19,30	63	1.161,22	999,26	-13,95	17
ITUB4	**Floresta Aleatória**	2.019,58	2.282,00	12,99	77	348,30	414,91	19,12	14
PETR4		2.193,01	2.460.24	12,19	68	421,77	466,83	10,68	13
SBSP3		4.077,49	6.136,68	50,50	66	1.061,30	1.580,66	48,94	17
VALE3		4.942,02	3.978,80	-19,49	70	1.312,51	1.116,82	-14,91	19

Fonte: Elaborada pelo autor.

Com base nos resultados da tabela acima, e considerando inicialmente a escala semanal (operações semanais), é possível notar que o *rank* de desempenho dos ativos está disposto da seguinte forma. Para o modelo k-NN, temos: VALE3 (-18,47%) na última posição apresentando desempenho negativo, seguida por PETR4 (12,52%) na terceira colocação, ITUB4 (15,57%) na segunda e SBSP3 (50,92%) na primeira posição, com desempenho aproximadamente 3,27 vezes maior que o segundo colocado (ITUB4). Para o modelo da floresta aleatória, os valores permaneceram próximos aos do k-NN, com uma leve alteração negativa para ITUB4 e VALE3, que valendo-se do referido modelo apresentaram redução em suas eficiências percentuais de aproximadamente 2,58% e 1,02% respectivamente. Da mesma forma, a sequência de desempenho permaneceu inalterada para a floresta aleatória, estando VALE3 (-19,49%) na última posição, PETR4 (12,19%) na terceira colocação, ITUB4 (12,99%) na segunda e SBSP3 na primeira com 50,5% de retorno. Já para a árvore de decisão, o que mais se destacou foi o incremento médio de 3,11% no desempenho de SBSP3 em relação ao obtido com os dois primeiros modelos apresentados, tendo esse ativo atingido incríveis 53,82 pontos percentuais de retorno em relação ao capital investido.

Tais resultados, quando comparados aos obtidos via investimentos livres e recorrentes (Tabela 4.1) apresentam dois pontos que merecem destaque. O primeiro relacionado à PETR4, que no cenário utilizando modelos de aprendizado de máquina apresentou desempenho reduzido de 5,87%, em média. O segundo associado a SBSP3, que registrou aumento médio de 2,3% ao utilizar os modelos de aprendizado (*k-NN*, árvore de decisão e floresta aleatória) em relação ao formato livre de investimento. Sendo importantate ressaltar que, tal incremento no retorno de Sabesp deu-se com um número médio e aproximado de apenas 66 operações de compra de ações, valor 2,19 vezes menor em comparação ao registrado para o mesmo ativo no formato livre de investimento. Tal resultado, evidencia a eficácia da modelagem para a SBSP3, ao proporcionar

maior retorno com menor risco de exposição ao mercado financeiro.

Na escala mensal, os ativos, que tiveram suas operações sinalizadas pelos modelos de aprendizado de máquina, mantiveram a mesma ordem da observada na escala semanal, diferindo apenas nos percentuais de retorno. De modo que, mesmo com retorno ainda negativo, VALE3 apresentou melhora de 9,53% na escala mensal utilizando o modelo k-NN como sinalizador de operação, de 5,54% com a árvore de decisão e de 3,56% utilizando o modelo de floresta aleatória. PETR4, por sua vez, apresentou leves oscilações positivas na mudança de escala, tanto para o modelo k-NN (0,18%) e como para o da árvore de decisão (0,88%), com queda de 1,51% do retorno ao utilizar o modelo de floresta aleatória, todos tendo como referência comparativa os valores da escala semanal. Já ITUB4, com desempenho médio de 18,42% para a associação dos três modelos, apresentou incremento interessante no retorno percentual com a mudança para a escala mensal, tendo um aumento de 2,47% ao utilizar o modelo k-NN, 4,47% ao utilizar a árvore de decisão e de 6,13% na utilização da floresta aleatória como sinalizador.

O ativo SBSP3, todavia, apresentou um cenário contrastante. Uma vez que, ao utilizar o modelo k-NN como sinalizador de operação, o referido ativo teve um incremento de 6,7% em relação ao valor apresentado na escala semanal, atingindo incríveis 60,52% de retorno financeiro na escala mensal. Contudo, utilizando as sinalizações de compras provenientes da árvore de decisão e da floresta aleatória, o ativo teve uma redução de desempenho, ainda em comparação a escala semanal, de 1,56% e 1,98% respectivamente.

Por fim, ao compararmos os resultados da modelagem numérica (Tabela 4.2) aos obtidos nas compras livres e recorrentes (Tabela 4.1), observa-se que, para ITUB4, tanto o modelo k-NN (18,04%), como o da árvore de decisão (18,09%) e o da floresta aleatória (19,12%) superaram os resultados livres na escala mensal.

No caso da PETR4, os modelos não se mostraram mais eficientes do que a estratégia de referência, uma vez que, os retornos percentuais dos investimentos realizados foram de 12,7% para o k-NN, 13,6% para a árvore de decisão e de 10,68% para a floresta aleatória, todos abaixo do observado na estratégia aleatória (17,89%). Já em relação à VALE3, embora os resultados tenham permanecido negativos, como já visto para todas as formas e escalas de investimentos deste ativo, os modelos de aprendizado de máquina ajudaram a reduzir parcialmente as perdas, passando de -18,12% na estratégia aleatória para -9,77% na sinalização de compra utilizando o modelo k-NN, -13,95% na árvore de decisão e de -14,91% na floresta aleatória. Para SBSP3, os modelos proporcionaram ganhos expressivos, alcançando uma média de 52,80% (KNN – 60,52% e árvore de decisão e floresta aleatória – 48,94%) em oposição aos 48,24% do investimento livre, evidenciando aqui a maior vantagem da modelagem preditiva.

4.1.3 MÉTRICAS DE AVALIAÇÃO

À luz dos resultados operacionais, apresentados na sub-seção anterior (4.1.2), a análise subsequente concentrar-se nos casos específicos em que os resultados utilizando algoritmos de aprendizado de máquina se destacaram em termos de desempenho. Deste modo, tais casos (modelo associado a ativo) foram selecionados tendo como critério os quatro pontos a seguir:

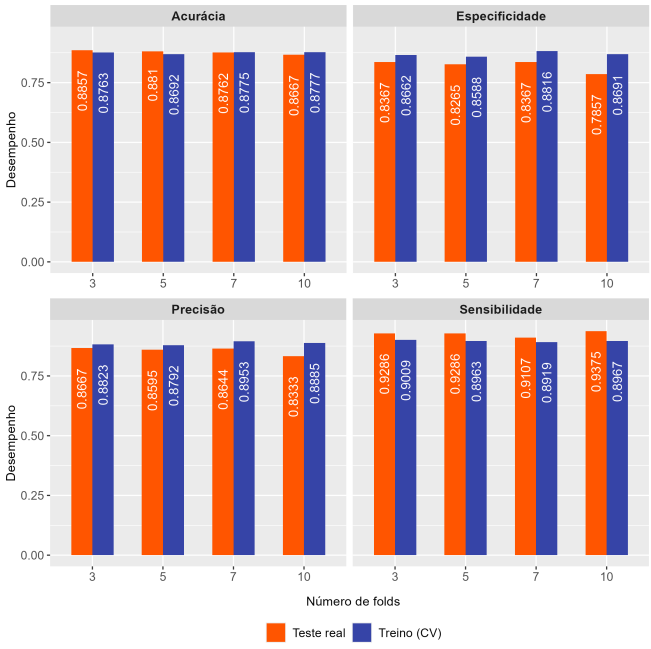
- ganho relativo dos modelos em relação ao investimento livre (prioridade alta);
- consistência entre modelos (k-NN, árvore, RF) na mesma escala, e;
- risco operacional (menos operações para igual/maior retorno), e;
- unicidade na associação ativo-modelo.

A partir desses critérios, foram escolhidos para análise os seguintes casos: SBSP3, modelada com a árvore de decisão na escala semanal e com o modelo k-NN na escala mensal; e ITUB4, com sinalizações de compra via floresta aleatória com periodicidade mensal. PETR4 e VALE3 por não atenderem simultâneamente aos critérios: ganho relativo dos modelos utilizados e robustez por escala, foram desclassificadas. Neste sentido, tais ativos e modelos serão discriminados segundo as métricas de avaliação e desempenho típicas dos algoritmos de aprendizado de máquina, descritas *a priori* em termos teóricos na seção 2.4.

SABESP - SBSP3

Sabesp, como apresentado na Tabela 4.2, utilizando o modelo de árvore de decisão na escala semanal, apresentou desempenho considerável ao atingir retorno financeiro de 53,82% sob o capital investido em três anos. Tal performance, fica ainda mais evidente ao se analisar as principais métricas de avaliação deste trinômio (ativo–escala–modelo), assim como apresentado na Figura 4.2, a seguir.

Figura 4.2: Desempenho semanal de SBSP3 aliado por métrica, considerando quatro diferentes *folds* de teste-treino e utilizando o modelo de árvore de decisão.



Fonte: Elaborada pelo autor.

A Figura acima (4.2), apresenta de forma sequêncial a acurácia, especificidade, precisão e a sensibilidade do referido sistema. Sendo, os valores apresentados nos gráficos, as médias obtidas em cada um dos *folds* para os conjuntos de teste (barras laranja) e treino (barras azuis). Aqui

há dê-se salientar que, a performance do teste reflete o desempenho do modelo treinado com o respectivo número de *fold*, utilizando o conjunto de treinamento de 2022 a 2024 (20% dos dados) previamente separado para este fim. Neste contexto, é possível verificar a estabilidade do modelo em ambos os cenários e para cada uma das métricas, uma vez que, em todas as medidas (métrica associada um número de *fold*) registradas pouca oscilação é notada.

Essa estabilidade descaracteriza qualquer possbilidade de *overfitting*, um vez que, é sinal de robustez e boa generalização. Isto tendo em vista que, o *overfitting* se manifesta tipicamente por um desempenho muito superior no treino em relação ao teste, especialmente com alta variância entre *folds*. No caso de SBSP3 semanal, observa-se o oposto, de modo que, as métricas de teste são comparáveis ou até superiores às de treino, como no caso da sensibilidade que tem valores mínimos de 91,07% no teste contra 89,10% no treino, de acordo com a Figura 4.2. Esse comportamento é incompatível com *overfitting* e sugere que o modelo foi capaz de capturar regularidades/padrões reais do fenômeno de compra estudado.

A Tabela 4.3, que apresenta as métricas de avaliação para SBSP3 semanal em termos médios, corrobora com o mencionado anteriormente. Nesta, é possível verificar percentualmente que a oscilação média dos desempenhos entre as métricas de avaliação fica em torno de 1,14%, estando a média das métricas de treinamento próxima da média de teste (88,17% e 87,03%). Essa avaliação, apesar de pouco convencional, apresenta o comportamento global entre teste e treino.

Tabela 4.3: Resultados percentuais médios das métricas de avaliação de SBSP3 na escala semanal usando o modelo de árvore de decisão.

	Média Treino [%]	**Média Teste [%]**	**Diferença [%]**
Acurácia	87,52	87,74	-0,22
Especificidade	86,89	82,14	4,75
Precisão	88,63	85,60	3,04
Sensibilidade	89,65	92,63	-2,99
Média	88,17	87,03	1,14

Fonte: Elaborada pelo autor.

Nesse sentido, é possível notar pela tabela acima que as diferenças percentuais entre treino e teste para a especificidade e precisão apresentam uma variação superior a 3,00 pontos percentuais, fato que em muito casos poderia ser considerado como de atenção. No entanto, tais diferenças, não devem ser tratadas como um alarme, visto que estão basicamente ligadas ao *fold* 10. No qual, a média de detecção de falsos positivos (classificação de preços elevados como sendo baixos) utilizando os dados de teste teve leve redução no desempenho, o que pode ser entendido por um possível desbalanceamento da base de teste que eventualmente, no período selecionado (2022 - 2024), apresentou mais casos verdadeiros positivos (sensibilidade alta) do que pontos errôneos de compra do ativo.

De forma direcionada, nota-se que a acurácia dos testes (87,74%) demonstra alta capacidade de predição ao classificar corretamente tanto os momentos de preço realmente baixo (verdadeiros positivos) quanto os de preço "não baixo" (verdadeiros negativos) em relação ao total das predições realizadas. Por conseguinte, a especificidade, dentre as quatro métricas avaliadas, é a

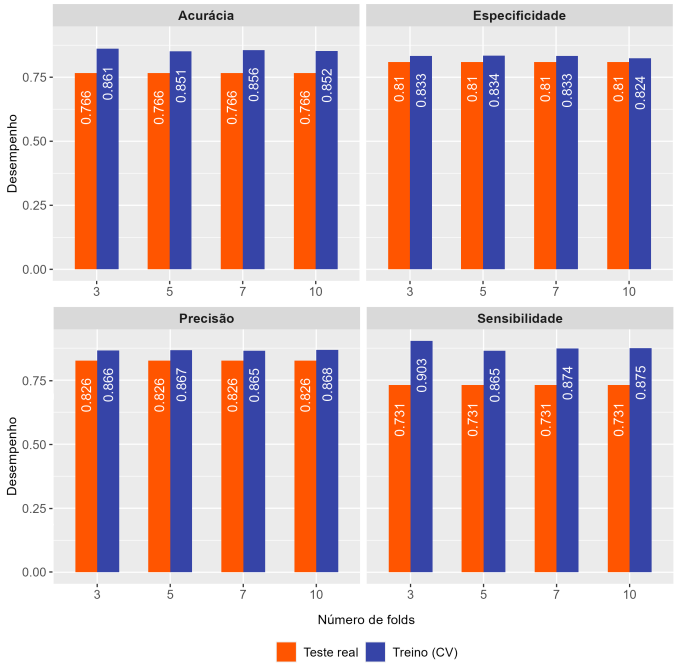
de menor taxa percentual (82,14%). No entanto, apesar de reduzida, apresenta boa capacidade em discriminar falsos sinais de compra, visto que a performance global do modelo em termos de retorno financeiro atingiu alto desempenho. Em outras palavras, o modelo foi eficaz em não comprar em momentos ruins, registrando cerca de 82% de acertos nesse aspecto, o que impediu o investimento de capital em momento não tão propício.

A precisão, por sua vez, atingiu valor médio no teste de 85,6%, indicando que, em aproximadamente 85% das vezes em que o modelo recomendou uma compra, essa recomendação correspondeu a um momento verdadeiramente favorável (preço realmente baixo). Assim, tal métrica é de extrema importância aos propósitos do presente trabalho, pois, como já apresentado, ao utilizar modelos de aprendizado de máquina o processo de investimento fica restrito a poucas operações de compra quanto comparado ao formato livre, o que torna ainda mais crucial a assertividade das entradas realizadas.

Por fim, a métrica da sensibilidade que representa o propósito do trabalho teve valor médio na etapa de teste de 92,63%, o que corrobora com o excelente retorno financeiro dos investimentos em Sabesp na escala semanal usando o modelo de árvore de decisão. Deste modo, o referido valor indica que em que em aproximadamente 92% dos momentos que o preço de SBSP3 estava realmente baixo, o modelo acertou ao recomendar uma compra. Descrevendo o que é considerado como um baixo custo de oportunidade.

Partindo para a escala mensal, e utilizando o modelo k-NN, Sabesp atingiu o mais alto valor de retorno percentual dentre todas as combinações e contextos avaliados, cerca de 60%. Desse modo, é razoável a avaliação das métricas de desempenho para o referido cenário, tal como feito anteriormente para a escala semanal, buscando validar a modelagem utilizada ou refutá-la frente a possíveis vícios. Assim, a Figura 4.3, a seguir, apresenta as métricas de avaliação de SBSP3 na escala mensal usando o modelo dos k-vizinhos mais próximos.

Figura 4.3: Desempenho mensal de SBSP3 aliado por métrica, considerando quatro diferentes *folds* de teste-treino e utilizando o modelo k-NN.



Fonte: Elaborada pelo autor.

Pela avaliação da figura acima, duas questões surgem de forma explícita. As quais são: a invariabilidade dos resultados dos testes em relação ao treino e a diferença marcante dos desempenhos associados à acurácia e a sensbilidade. Em relação a invariabilidade dos resultados dos testes, observa-se que os valores por *fold* em cada uma das métricas não se alteram dentro do mesmo conjunto de avaliação. De forma que, o desempenho da precisão nos testes permaneceu constante em 82,6%, independentemente do número de *folds* (3, 5, 7 ou 10) adotado no treinamento do modelo testado. A especificidade em aproximadamente 81,0%, seguida pela acurácia que também manteve constante o desempenho de 76,6% nos testes. Por fim, verifica-se que a sensibilidade teve cerca de 73,1% de desempenho nos testes ao longo dos quatro cenários avaliados (*folds*). Tal invariabilidade, em geral, não representa um problema se avaliada de forma restrita aos testes, pois como os 20% dos dados separados para essa etapa (dados de 2022 - 2024) se mantém inalterados, então ao aplicá-los a um mesmo modelo nenhuma variação deveria ser observada, mantendo assim o resultados constantes.

Todavia, ao considerar que a avaliação dos testes também deve levar em conta a variação entre os modelos treinados com diferentes *folds*, é razoável esperar que, em pelo menos um dos casos, o desempenho no conjunto de teste real reflita, de forma proporcional, as variações do treinamento. Isso porque, em um modelo bem ajustado, independentemente de como os dados de treinamento são distribuídos entre os *folds*, o desempenho em dados não vistos deve reproduzir de forma consistente as variações internas do processo de aprendizado. Assim, essa invariabilidade dos desempenhos associados aos testes aproxima o presente modelo do diagnóstico de sobreajuste na modelagem (*overfitting*).

Tabela 4.4: Resultados percentuais médios das métricas de avaliação de SBSP3 na escala mensal usando o modelo de k-NN.

	Média Treino [%]	**Média Teste [%]**	**Diferença [%]**
Acurácia	85,49	76,60	8,90
Especificidade	83,10	80,95	2,15
Precisão	86,65	82,61	4,04
Sensibilidade	87,92	73,08	14,84
Média	85,79	78,31	7,48

Fonte: Elaborada pelo autor.

A ideia de sobreajuste, no entanto, não pode ser estruturada apenas sobre a observação de invariabilidade nos desempenhos. Mas deve ser apoiada por uma marcante defasagem entre os resultados do modelo treinado em relação aos alcançados na etapa de teste. Desse modo, o segundo ponto suscitado no início da discussão, aqui realizada, vem a corroborar de forma decisiva para a classificação de *overfitting* do modelo avaliado. Uma vez que, a defasagem é claramente evidenciada nas métricas de acurácia e sensibilidade, nas quais diferenças médias entre treino e teste atingem aproximadamente 8,90% e 14,89%, como pode ser comprovado pela análise da Figura 4.3 e da Tabela 4.4. Esses valores evidenciam a perda de desempenho em relação a dados não vistos durante a modelagem, mas também mostram uma falha considerável do modelo em reconhecer corretamente a classe positiva (erro do tipo I), o que é especialmente

problemático em situações onde se deseja detectar a referida classe, como é o caso do presente trabalho.

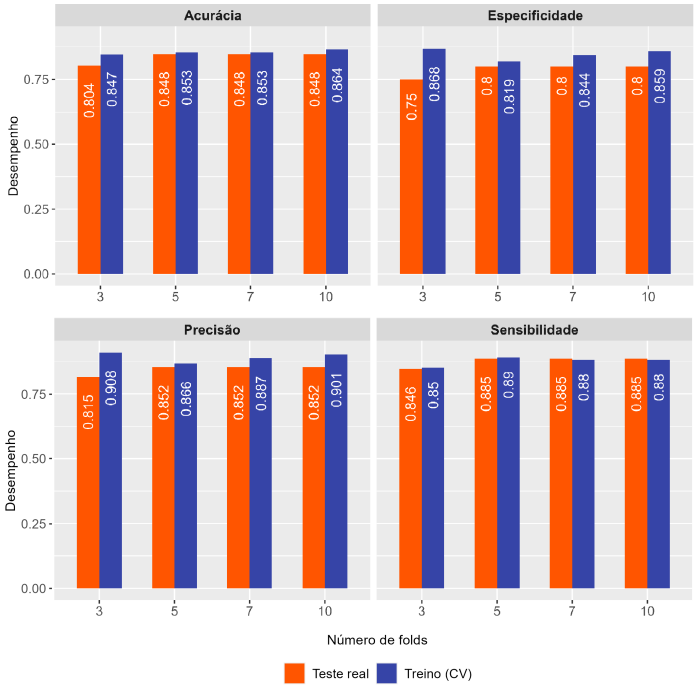
Assim, mesmo tendo atingido o elevado valor de 60,52% de retorno financeiro, com apenas R$ 868,84 investidos e com o quarto menor número de compras (15) dentre todas as avaliações realizadas. O ativo SBSP3, na escala mensal utilizando o algoritmo de aprendizado de máquina k-NN com sinalizador de entradas/compras deve ser desconsiderado das comparações futuras.

ITAÚ UNIBANCO - ITUB4

Do trinômio ativo–escala–modelo, ITUB4 na escala mensal modelado com o algoritmo de floresta aleatória, foi dentre todos os sistemas avaliados, a terceiro opção que melhor se adequou aos critérios descritos no início da presente análise. Ao longo do período dos três anos avaliados (2022 - 2024), esse ativo recebeu um total de R$348,30 em aportes orientados por algoritmo de aprensizado de máquina, distribuídos em 14 operações no mercado de ações. De tal sorte que, o valor da carteira nas referidas condições atingiu R$ 414,91, o que descreve um retorno acumulado de 19,12%.

Em relação aos desempenhos das métricas de avaliação, apresentados pela Figura 4.4. ITUB4 teve comportamento característico de um sistema bem modelado, sem indícios de sobreajuste (*overfitting*). Foi observado regularidade nos resultados, com os desempenhos de treino e teste permanecendo próximos entre si sob diferentes configurações da validação cruzada. Não obstante, para a maioria das métricas, houve uma leve melhora no desempenho à medida que o número de partições (*folds*) da validação cruzada aumentou.

Figura 4.4: Desempenho mensal de ITUB4 aliado por métrica, considerando quatro diferentes *folds* de teste-treino e utilizando o modelo de floresta aleatória.



Fonte: Elaborada pelo autor.

A acurácia, apresentada no gráfico acima, é estavél e progressiva. De modo que, o treino

tem desempenho inicial de 84,7% e o teste de 80,4% com três *folds*, atingindo o ponto máximo em dez *folds* com 86,4% para o treino e de 84,8% para o teste. Isto mostra que, a proporção de previsões corretas (de compra e não-compra) é alta mas não chega ao patamar de 100%, o que é esperado em contextos reais de predição. No que concerne a sensibilidade, o melhor desempenho de treinamento foi atingido ao se utilizar cinco *folds*, 89,0%. Para a etapa de teste, a estabilidade foi atingida também no quinto *fold*, com 88,5% de desempenho. Esses valores mostram que o modelo é capaz de detectar, com alta confiabilidade, as situações de interesse, ou seja, de pontos de compra, o que reforça sua adequação ao objetivo do estudo.

Para a precisão, verifica-se que na etapa de treinamento os *folds* três (90,8%) e dez (90,1%) têm maior desempenho que os *folds* cinco (86,6%) e sete (88,7%). Essa diferença no desempenho dos *folds* mencionados, contribui para o aumento da média geral de treino, que atinge 89,05%, conforme apresentado na Tabela 4.5. Já para a etapa de teste da referida métrica, temos o primeiro *fold* com 81,5% de desempenho e os seguinte com 85,2%, valores que produzem um resultado médio de aproximadamente 84,25% de desempenho, e gera uma diferença final de apenas 4,79 pontos percentuais em relação ao treinamento. Essa diferença, apesar de ser considerada moderada para os padrões do presente trabalho, mostra que o modelo tem boa capacidade de generalização. De modo que, é possível inferir que a cada dez recomendações de compra oito realmente são compras interessantes.

Tabela 4.5: Resultados percentuais médios das métricas de avaliação de ITUB4 na escala mensal usando o modelo de floresta aleatória.

	Média Treino [%]	**Média Teste [%]**	**Diferença [%]**
Acurácia	85,44	83,70	1,74
Especificidade	84,75	78,75	6,00
Precisão	89,05	84,26	4,79
Sensibilidade	87,51	87,50	0,01
Média	86,69	83,55	3,13

Fonte: Elaborada pelo autor.

Já, a especificidade dentre as métricas avaliadas é a que apresentou maior discrepância entre a média do treinamento em relação a de teste, como pode ser verificado na tabela acima a mesma atingiu o valor de 6,0 pontos percentuais. Tal diferença, está associada primordialmente ao perfil da série histórica do referido ativo, que como apresentado na análise descritiva dos dados (Capítulo 3) tem tendência crescente com período de forte descontinuidade. De modo que, tais descontinuidades não são interpretadas corretamente pelo modelo e geram um número elevado de falsos positivos, prejudicando o desempenho da referida métrica.

Por fim, utilizando todo o conhecimento gerado pelos resultados apresentados, e os três sistemas que atenderam simultaneamente aos critérios de: ganho relativo, consistência entre modelos, risco operacional e unicidade na associação ativo-modelo, chegamos a seguinte classificação de importância da associação ativos–escalas–modelos:

1ª posição: SBSP3 semanal usando árvore de decisão;

2ª posição: ITUB4 mensal usando floresta aleatória, e na;

3ª posição: SBSP3 semanal usando k-NN (desconsiderado devido a suposição de *overfitting*).

4.1.4 SISTEMA ESCOLHIDO

A presente seção tem por objetivo apresentar o sistema de investimento que, após todas as avaliações e critérios estabelecidos, foi definido como o "campeão", sendo considerado o sistema balizador para os investimentos orientados por modelos de aprendizado de máquina no contexto do presente trabalho.

Neste sentido, tomou-se como referência o *rank* exposto na subseção anterior (4.1.3), bem como os resultados semanais e mensais das compras realizadas a partir das sinalizações dos modelos (Tabela 4.2). De modo que, a escolha final recaiu sobre o ativo Sabesp (SBSP3) na escala semanal utilizando o modelo de árvore de decisão, visto que, dentre os modelos de melhor desempenho, este foi o único sistema que, simultaneamente, superou os retornos da modalidade de investimento livre, atendeu aos critérios estabelecidos na etapa de análise das métricas de avaliação e ainda apresentou retorno financeiro percentual superior ao acumulado pela taxa Selic (33,76%) [39] nos três anos avaliados, reforçando sua relevância como referência central neste estudo. Assim, na Tabela 4.6 são apresentados, de forma resumida, o resultado financeiro e os principais parâmetros do sistema descrito.

Tabela 4.6: Resultado financeiro e parâmetros do sistema de investimento SBSP3 semanal usando árvore de decisão.

Resultado financeiro		**Parâmetros e atributos do modelo**	
Ativo	SBSP3	**Modelo**	Árvore de decisão
Escala	Semanal	**Custo de complexidade**	0.0001
Investimento	R$ 3.566,48	**Profundidade máxima**	3
Valor da carteira	R$ 5.485,82	**Mínimo de observações por nó**	5
Retorno	53,82%	**Validação cruzada**	10
Número de Compradas	59	**Métrica de teste**	Sensibilidade
Período inicial de teste	2022	**Média do treino**	89,67%
Período final de teste	2024	**Média do teste**	93,75%

Fonte: Elaborada pelo autor.

Na Tabela 4.6, observada acima, verifica-se que o sistema de investimento para SBSP3 operando em escala semanal e orientado pelo modelo de árvore de decisão, apresentou desempenho significativo ao longo do período avaliado (2022 - 2024). Tendo este, valorização de 53,82%, passando dos R$ 3.566,48 de investimento inicial ao valor acumulado de R$ 5.485,82 (carteira

de investimento). Tal incremento financeiro, foi obtido a partir de 59 operações sinalizadas pelo modelo. O qual foi parametrizado com árvores de profundidade máxima de três níveis, com um mínimo de cinco observações por nó, com custo de complexidade de 10^{-4} e dez *folds* de validação cruzada.

Por fim, todo o sistema foi calibrado e processado tendo como base a métrica de sensibilidade, a qual possibilitou a avaliação e redução do custo de oportunidade na sinalização de pontos de compra.

4.2 Apostas na roleta de cassino

Tal como descrito na subseção 3.4, a simulação do sistema de aposta foi realizada utilizando, como parâmetros financeiros de entrada, os dados finais do sistema de investimento modelado para SBSP3. Isto pois, sob as mesmas condições de aporte financeiro torna-se possível a comparação de desempenho entre os sistemas.

Deste modo, o capital inicial definido no código escrito em linguagem de programação *R* para a roleta de cassino foi de R$ 3.566,48. Como critério de parada, a restrição de redução do capital a zero e o limite de 59 rodadas (apostas) foram adicionados. Uma vez que, não se admite saldo negativo acumulado nesse tipo de sistema e, além disso, não se deve ter mais de 59 oportunidades de crescimento de capital, haja vista que o modelo apresentado na subseção 4.1.4 dispendeu exatamente essa quantidade de operações para atingir o retorno percentual apresentado de 53,82%. O aporte inicial da primeira rodada foi fixado em R$ 68,00. Esse valor foi escolhido por três motivos: primeiro, por representar o mínimo necessário para que, em uma simulação ideal de 59 ganhos consecutivos utilizando a estratégia *Streets of Gold*, fosse possível superar o retorno percentual alcançado pelos modelos de aprendizado de máquina; segundo, por ser o menor valor inteiro que permite a divisão exata entre as quatro *double streets* dessa estratégia, e por fim, por se tratar de um capital acessível ao grande público.

Assim, sob tais condições e utilizando a estratégia de aposta *Streets of Gold* (2.1.2), as simulações para a roleta de cassino obtiveram os resultados por rodada discriminados nas tabelas do Apêndice D (Tabelas D.1, D.2 e D.3). De forma consolidada, tais resultados encontram-se resumidos na Tabela 4.7, apresentada a seguir.

Tabela 4.7: Resumo do resultado financeiro da simulação de aposta na roleta de cassino, apresentando os valores base, os valores médios e os valores extremos.

Investimento nas rodadas [R$]		Ganhos/perdas da rodada [R$]		Resultado da rodada[R$]		Capital acumulado [R$]			Nº rodadas	
1º e menor	59º e maior	Menor	Maior	Menor	Maior	Menor	Maior	Final	(-)	(+)
64,00	160,00	-160,00	128,00	0,0	288,00	3.534,00	4.302,00	3.662,00	25 (42,37%)	34 (57,63%)

Fonte: Elaborada pelo autor.

Na Tabela 4.7, acima, os resultados da roleta de cassino são apresentados em cinco blocos distintos. Os valores dos investimentos, os ganhos máximo e mínimos registrados, os resultados

extremos obtidos, a evolução do capital acumulado e o número de rodadas realizadas.

Em relação aos valores dos investimentos, obeserva-se que o 1º aporte também representa o menor valor aplicado. Isto ocorre essencialmente, pois a primeira rodada trata-se de um investimento sem correção de perdas. De modo que, o aporte máximo somente pode ser atingindo, de acordo com a estratégia *Streets of Gold* (2.1.2), a partir da segunda rodada caso tenha ocorrido perda na etapa de abertura. Sendo o referido valor R$ 160,00. Assim, concluí-se que a última rodada (59º) foi precedida de uma aposta deficitária, ao passo que a mesma registrou aporte máximo.

Quanto aos ganhos, verifica-se que os maiores resultados ocorreram justamente nas rodadas de investimento máximo, de modo que, o melhor desempenho em uma rodada atingiu R$ 288,00, com saldo positivo de R$ 128,00, enquanto o pior resultado foi de R$ 0,00, implicando a perda integral do valor aportado (R$ 160,00). Em termos de ganhos consecutivos, a maior sequência atingida foi de nove rodadas, começando na 9º a partir de um aporte de valor máximo, até a 17º rodada com aporte mínimo, acumulando R$ 384,00 de ganho financeiro via R$1056,00 de resultado acumulado, mediante a R$ 672,00 aportados.

No tange a variação do capital acumulado ao longo das rodadas, nota-se que o menor valor atingido foi o de R$3.534,00, valor esse inferior ao capital inicial disponível. Já, o maior capital acumulado alcançou R$ 4.302,00, representando um incremento de 20,62% sobre o capital inicial. Ao término da simulação, o capital final acumulado foi de R$ 3.662,00 (2,68% de retorno), valor próximo ao mínimo observado, evidenciando que as perdas superaram os ganhos no fechamento do ciclo das 59 rodadas. Por fim, observa-se que nas simulações da roleta, cerca 42,37% das rodadas tiveram resultados negativos e 57,63% positivos, evidenciando o baixo retorno percentual deste sistema.

Neste contexto, fica evidente que os aportes financeiros na roleta de cassino, mesmo com o aúxilio de uma estratégia elaborada de execução, como a apresentada, não configura uma alternativa viável de aplicação de capital em cenários nos quais o objetivo final seja a evolução patrimonial. Isto porque, conforme verificado na análise acima, após 59 rodadas, esse sistema apresentou desempenho aproximadamente 20 vezes menor que o registrado para o sistema selecionado na subseção 4.1.4 (SBSP3–semanal-árvore de decisão, 53,82%), de investimento orientado por modelos de aprendizado de máquina. Sendo assim, a roleta de cassino é uma alternativa restrita ao entretenimento dos apostadores e daqueles que desejam compreender os efeitos da aleatoriedade em sistemas estocásticos.

5. Conclusões e sugestões

O presente trabalho teve como objetivo principal a criação e comparação de dois sistemas de aporte de capital em busca de lucros de curto e médio prazo. Para tanto, foi desenvolvido em linguagem de programação *R* um ecossistema composto por modelos de aprendizado de máquina direcionado aos investimentos na Bolsa de Valores do Brasil – B3, e um sistema de simulação computacional de apostas do tipo roleta de cassino. De modo que, fosse possível a comparação quantitativa de desempenho destes sob circunstânceas bem definidas de performance financeira e processamento computacional.

No que diz respeito à modelagem de aprendizado de máquina, foram avaliados, em nível experimental e teórico, dois algoritmos clássicos e um método combinado (*ensemble*), todos devidamente descritos no presente texto. Sendo estes: os k-vizinhos mais próximos (k-NN), a árvore de decisão e a floresta aleatória, aplicados em experimentos computacionais envolvendo à evolução temporal dos preços dos ativos analisados, em busca de sinalizações de compra que produzisse a maior evolução patrimonial possível frente ao investimento realizado.

Neste contexto, as simulações envolvendo algoritmos de aprendizado de máquina foram conduzidas com quatro dos ativos mais representativos, do índice Ibovespa: PETR4, VALE3, ITUB4 e SBSP3. Os quais experimentaram as múltiplas combinações de escala (semanal e mensal) junto aos modelos apresentados acima. Destas combinações, o modelo de árvore de decisão aplicado à SBSP3 na escala semanal apresentou o melhor resultado geral. Atingindo, esse sistema, o incrível retorno acumulado de 53,82% no período avaliado de 2022 a 2024, superando tanto os investimentos livres e recorrentes (sem sinalização de modelos) quanto o retorno acumulado da taxa de juros Selic no mesmo período (33,76%). Demonstrando, assim, a resiliência da modelagem preditiva na detecção de oportunidades/pontos de compra com elevado potencial de valorização. Cabe, ainda, destacar, que ITUB4 na escala mensal modelado pelo algoritmo de floresta aleatória também apresentou resultados interessantes frente ao seu próprio perfil de grande volatilidade; no entanto, não superou a característica crescente de Sabesp.

Sob tal perspectiva, o sistema de aposta do tipo roleta de cassino não conseguiu atingir sequer metade do desempenho apresentado pelo trinômio SBSP3-semanal-árvore de decisão. Chegando a apenas 2,68% de retorno financeiro, isto, mesmo sendo otimizado por meio da estratégia de perfil progressista conhecida como *Street of Gold*. Mostrando, desta forma, que a intrínseca aleatoriedade das apostas impede o crescimento e manutenção consistente do patrimônio aportado, ao passo que os perídos/rodadas lucrativas ocorrem pontualmente ao longo da execução

deste sistema. Assim, o presente estudo confirma a superioridade das estratégias orientadas por modelos estatístico-computacionais em detrimento aos sistemas puramente estocásticos.

Em relação as métricas de avaliação dos modelos de aprendizado de máquinas, estas desempenharam papel essencial na análise e validação dos sistemas propostos. Permitindo que o equilíbrio entre risco, retorno e robustez fosse identificado, evitando, assim, que sistemas sobreajustados fossem selecionados e que a adequação paramétrica dos modelos fosse realizado de forma pertinente e de acordo com o que estabelece a literatura vigente.

Do ponto de vista incremental, como continuidade deste estudo, sugere-se que três aspectos sejam explorados, todos no contexto dos investimentos em ações. O primeiro consiste na adição de novos modelos de inteligência artificial, tais como os pertencentes a família dos dos métodos de *boosting* (XGBoost e LightGBM), bem como na implementação de redes neurais recorrente, em especial a LSTM. O segundo aspecto, refere-se à modificação do sistema de pré-processamento e filtragem dos dados, que em avaliação preliminar tem potencial para incrementar o desempenho dos modelos. E por último, a inclusão de variáveis macroeconômicas associadas a volatilidade e a dinâmica dos mercados.

Por fim, o presente trabalho reforça a relevância dos modelos de aprendizado de máquina na descrição e avaliação do mercado financeiro, contribuindo para os avanços acadêmicos e, sobretudo, como ferramenta de elucidação de alternativas de menor risco voltadas à evolução patrimonial.

Referências Bibliográficas

[1] A. Ali and W. K. Mashwani. A supervised machine learning algorithms: applications, challenges, and recommendations. *Proceedings of the Pakistan Academy of Sciences: A. Physical and Computational Sciences*, 60(4):1–12, 2023.

[2] H. Alkhazzar and B. Moslem. Thyroid disease prediction using machine learning model: Decision tree. *EDRAAK*, 2024:84–100, Jul. 2024.

[3] F. Arden and C. Safitri. Hyperparameter tuning algorithm comparison with machine learning algorithms. In *2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 183–188. IEEE, 2022.

[4] B3. Número de investidores na b3 cresce 34% em renda fixa e 23% em renda variável em 12 meses, 2023. Disponível em: https://www.b3.com.br/pt_br/noticias/numero-de-investidores-na-b3-cresce-34-em-renda-fixa-e-23-em-renda-variavel-em-12-meses.htm. Acesso em: 9 nov. 2024.

[5] B3 - Brasil, Bolsa, Balcão. Índice ibovespa – composição da carteira. https://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/indice-ibovespa-ibovespa-composicao-da-carteira.htm, 2025. Acessado em: 15 ago. 2025.

[6] Banco Central do Brasil. Rendimento dos depósitos de poupança, 2023. Disponível em: https://www.bcb.gov.br/estatisticas/remuneradepositospoupanca. Acesso em: 5 out. 2025.

[7] Banco Central do Brasil. Relatório de inflação - março de 2024, março 2024. Relatório detalhado sobre os componentes da inflação e a evolução do IPCA em 2023.

[8] Banco Central do Brasil. Operações de mercado aberto (open market). https://www.topinvest.com.br/politica-monetaria/, 2025. Acesso em: 20 ago. 2025.

[9] F. G. Blanchet, P. Legendre, and D. Borcard. Forward selection of explanatory variables. *Ecology*, 89(9):2623–2632, 2008.

[10] J. Brownlee. Machine learning algorithms from scratch. *Machine Learning Mastery*, 2019.

[11] L. F. R. Chaves. Repositório do projeto de conclusão de curso: Finalundergradproject_r, 2025. Disponível em: https://github.com/LucasFRChaves/FinalUnderGradProject_R. Acesso em: 16 ago. 2025.

[12] Comissão de Valores Mobiliários (CVM). *TOP: Mercado de Valores Mobiliários Brasileiro*. 5: ed. edition, 2024. Acesso em: 20 de agosto de 2025.

[13] A. Damodaran. *Investment Valuation: Tools and Techniques for Determining the Value of Any Asset*. Wiley, 3 edition, 2012.

[14] A. B. das Entidades dos Mercados Financeiro e de Capitais (ANBIMA). Raio x do investidor brasileiro – 7ª edição, 2023. Relatório sobre o perfil do investidor brasileiro com dados de 2023.

[15] K. Faceli, A. C. Lorena, J. Gama, T. A. d. Almeida, and A. C. P. d. L. F. d. Carvalho. Inteligência artificial: uma abordagem de aprendizado de máquina. 2021.

[16] H. Ferreira, E. Almeida, W. Espinosa-García, E. Novais, J. Rodrigues, and G. Dalpian. Introduzindo aprendizado de máquina em cursos de física: o caso do rolamento no plano inclinado. *Revista Brasileira de Ensino de Física*, 44:e20220214, 2022.

[17] E. Fortuna. *Mercado Financeiro: Produtos e Serviços*. Qualitymark, Rio de Janeiro, 21 edition, 2020.

[18] L.-Z. Gao, C.-Y. Lu, G.-D. Guo, X. Zhang, and S. Lin. Quantum k-nearest neighbors classification algorithm based on mahalanobis distance. *Frontiers in Physics*, 10:1047466, 2022.

[19] R. Gupta et al. Feature selection techniques and its importance in machine learning: a survey. In *2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pages 1–6. IEEE, 2020.

[20] D. A. Hadi and D. A. N. Sirodj. Metode random forest untuk klasifikasi penyakit diabetes. In *Bandung Conference Series: Statistics*, volume 3, pages 428–435, 2023.

[21] Y. Huang and X. Zhou. Artificial intelligence random forest algorithm and the application. In *2020 International Conference on Data Processing Techniques and Applications for Cyber-Physical Systems: DPTA 2020*, pages 205–213. Springer, 2021.

[22] R. Izbicki and T. M. dos Santos. *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki, 2020.

[23] N. Jain and P. K. Jana. Lrf: A logically randomized forest algorithm for classification and regression problems. *Expert Systems with Applications*, 213:119225, 2023.

[24] J. Jo. Ensemble learning: bagging / bootstrap aggregation, 7 2022. Disponível em: https://medium.com/@wjj1019/ensemble-learning-bagging-bootstrap-aggregation-b7900e58caa2. Acesso em: 28 ago. 2025.

[25] J. Kazemitabar, A. Amini, A. Bloniarz, and A. S. Talwalkar. Variable importance using decision trees. *Advances in neural information processing systems*, 30, 2017.

[26] Kevin. Part 7: random forests | bs0004 code annotations, 2025. Disponível em: https://bookdown.org/jcog196013/CancerMe/ranger.html. Acesso em: 27 ago. 2025.

[27] S. Lee, C. Lee, K. G. Mun, and D. Kim. Decision tree algorithm considering distances between classes. *IEEE Access*, 10:69750–69756, 2022.

[28] R. Lessa. Direito autoral brasileiro e a inteligência artificial (ia), 2025. Disponível em: https://www.jusbrasil.com.br/artigos/direito-autoral-brasileiro-e-a-inteligencia-artificial-ia/2055309721. Acesso em: 24 ago. 2025.

[29] W. Lestari, S. Sumarlinda, and A. B. Rahmat. Improvement of prediction model using k-nearest neighbors (knn) and k-means in medical data. In *Proceeding of International Conference on Science, Health, And Technology*, pages 200–207, 2024.

[30] G. F. Luger. *Artificial Intelligence*. Pearson, 2009.

[31] S. Merugu and M. Priya. Role of statistics in artificial intelligence. *International Journal of Engineering Applied Sciences and Technology*, 8:96–98, 04 2023.

[32] F. S. Mishkin. *The Economics of Money, Banking, and Financial Markets*. Pearson, 12 edition, 2018.

[33] C. Mwangi. Como jogar roleta: regras da roleta para iniciantes, 2025. Disponível em: https://roleta77brazil.com/como-jogar. Acesso em: 14 ago. 2025.

[34] A. A. Neto. *Mercado Financeiro*. Atlas, São Paulo, 13 edition, 2022.

[35] I. Parmar, N. Agarwal, S. Saxena, R. Arora, S. Gupta, H. Dhiman, and L. Chouhan. Stock market prediction using machine learning. In *2018 first international conference on secure cyber computing and communication (ICSCCC)*, pages 574–576. IEEE, 2018.

[36] Pinnacle. A história da roleta, 2025. Disponível em:https://www.pinnacle.com/betting-resources/pt/casino/the-history-of-roulette/l9tjduaqjavmbnrl. Acesso em: 14 ago. 2025.

[37] F. Rahmat, Z. Zulkafli, A. J. Ishak, R. Z. Abdul Rahman, S. D. Stercke, W. Buytaert, W. Tahir, J. Ab Rahman, S. Ibrahim, and M. Ismail. Supervised feature selection using principal component analysis. *Knowledge and Information Systems*, 66(3):1955–1995, 2024.

[38] Y. I. P. Ramania et al. Keberhasilan harga saham dan aksi stock split. *Perspektif Akuntansi*, 7(1):38–57, 2024.

[39] Receita Federal do Brasil. Cálculo de taxa selic, 2025. Disponível em: https://sicalc.receita.economia.gov.br/sicalc/selic/consulta. Acesso em: 5 abr. 2025.

[40] Rondoniagora.com. Flat betting pré-apostas em esportes: tudo o que você precisa saber, 4 2023. Disponível em: https://www.rondoniagora.com/negocios/flat-betting-pre-apostas-em-esportes-tudo-o-que-voce-precisa-saber. Acesso em: 19 ago. 2025.

[41] M. Rosa. Sistema para roleta: Estratégia streets of gold, 2023. Disponível em: https://casasdeapostasonline.pt/estrategia-streets-of-gold/. Acesso em: 19 ago. 2025.

[42] M. Schonlau and R. Y. Zou. The random forest algorithm for statistical learning. *The Stata Journal*, 20(1):3–29, 2020.

[43] M. Shahhosseini and G. Hu. Improved weighted random forest for classification problems. In *International Online Conference on Intelligent Decision Science*, pages 42–56. Springer, 2020.

[44] S. Sharma. Supervised learning: An indepth analysis. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 2024.

[45] K. Sindhu Meena and S. Suriya. A survey on supervised and unsupervised learning techniques. In *International Conference on Artificial Intelligence, Smart Grid and Smart City Applications*, pages 627–644. Springer, 2019.

[46] K. Sindhu Meena and S. Suriya. A survey on supervised and unsupervised learning techniques. *Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications*, pages 627–644, 2020.

[47] R. Souza. A psicologia financeira em jogos de apostas: desmistificando o conceito de investimento e alertando sobre os riscos, 2025. Disponível em: https://www.gov.br/investidor/pt-br/penso-logo-invisto/a-psicologia-financeira-em-jogos-de-apostas-desmistificando-o-conceito-de-investimento-e-alertando-sobre-os-riscos. Acesso em: 11 ago. 2025.

[48] A. Thevapalan and J. Le. Tutorial de árvores de decisão do r: Exemplos e código em r para regressão e classificação | datacamp, 4 2024. Disponível em: https://www.datacamp.com/pt/tutorial/decision-trees-R. Acesso em: 26 ago. 2025.

[49] X. Wu, X. Liu, and Y. Zhou. Review of unsupervised learning techniques. In *Proceedings of 2021 Chinese Intelligent Systems Conference: Volume II*, pages 576–590. Springer, 2021.

[50] C. Zhang, R. Dong, X. Zhang, and W. Li. A comparative analysis of supervised/unsupervised algorithms in phm application. In *2022 Global Reliability and Prognostics and Health Management (PHM-Yantai)*, pages 1–5. IEEE, 2022.

A. Ativos Listados na B3

Tabela A.1: Empresas com participação percentual igual ou superior a 2% no índice ibovespa (ibov) no ano corrente de 2025 - carteira teórica do ibovespa.

Ativo/Ticker	Descrição	Participação [%]
VALE3	Vale S.A.	10,81
ITUB4	Itaú Unibanco Holding S.A.	8,23
PETR4	Petróleo Brasileiro S.A. – Petrobras	6,496
PETR3	Petróleo Brasileiro S.A. – Petrobras	4,546
BBDC4	Banco Bradesco S.A.	3.907
SBSP3	Companhia de Saneamento Básico do Estado de São Paulo – Sabesp	3,692
B3SA3	B3 S.A. – Brasil, Bolsa, Balcão	3,508
ELET3	Centrais Elétricas Brasileiras S.A. – Eletrobras	3,397
ITSA4	Itaúsa – Investimentos Itaú S.A.	2,986
BBAS3	Banco do Brasil S.A.	2,924
WEGE3	WEG S.A.	2,922
EMBR3	Embraer S.A.	2,812
ABEV3	Ambev S.A.	2,724
BPAC11	Banco BTG Pactual S.A.	2,505
EQTL3	Equatorial Energia S.A.	2,058
RDOR3	Rede D'Or São Luiz S.A.	1,835
RENT3	Localiza Rent a Car S.A.	1,718
PRIO3	PetroRio S.A.	1,523
SUZB3	Suzano S.A.	1,503
ENEV3	Eneva S.A.	1,216
VBBR3	Vibra Energia S.A.	1,154
VIVT3	Telefônica Brasil S.A.	1,136
BBSE3	BB Seguridade Participações S.A.	1,067
TOTS3	Totvs S.A.	1,051
RAIL3	Rumo S.A.	1,037
GGBR4	Gerdau S.A.	1,006

Fonte: Extraída de [5] - Alterada.

Tabela A.2: Empresas com participação percentual inferiores 1% e superiores a 0.372% no índice ibovespa (ibov) no ano corrente de 2025 - carteira teórica do ibovespa.

Ativo/Ticker	Descrição	Participação [%]
BBDC3	Banco Bradesco S.A.	0,971
CMIG4	Companhia Energética de Minas Gerais – CEMIG	0,953
CPLE6	Companhia Paranaense de Energia – Copel	0,946
LREN3	Lojas Renner S.A.	0,914
RADL3	Raia Drogasil S.A.	0,91
UGPA3	Ultrapar Participações S.A.	0,898
TIMS3	TIM S.A.	0,831
BRFS3	BRF S.A.	0,758
ENGI11	Energisa S.A.	0,722
KLBN11	Klabin S.A.	0,683
ASAI3	Sendas Distribuidora S.A. – Assaí	0,636
MOTV3	Motiva S.A.	0,627
ELET6	Centrais Elétricas Brasileiras S.A. – Eletrobras	0,558
EGIE3	Engie Brasil Energia S.A.	0,542
HAPV3	Hapvida Participações e Investimentos S.A.	0,508
ALOS3	Allos S.A.	0,5
SANB11	Banco Santander (Brasil) S.A.	0,479
PSSA3	Porto Seguro S.A.	0,474
HYPE3	Hypera S.A.	0,458
ISAE4	ISA CTEEP – Companhia de Transmissão de Energia Elétrica Paulista	0,423
CXSE3	Caixa Seguridade Participações S.A.	0,41
NATU3	Natura & Co Holding S.A.	0,408
CMIN3	CSN Mineração S.A.	0,395
MULT3	Multiplan Empreendimentos Imobiliários S.A.	0,392
CSAN3	Cosan S.A.	0,38

Fonte: Extraída de [5] - Alterada.

Tabela A.3: Empresas com participação percentual inferiores 0.375% no índice ibovespa (ibov) no ano corrente de 2025 - carteira teórica do ibovespa.

Ativo/Ticker	Descrição	Participação [%]
BRAV3	Bravante Participações S.A.	0,372
CPFE3	CPFL Energia S.A.	0,353
SMFT3	Smart Fit Escola de Ginástica e Dança S.A.	0,352
TAEE11	Transmissora Aliança de Energia Elétrica S.A – Taesa	0,351
CYRE3	Cyrela Brazil Realty S.A. Empreendimentos e Participações	0,307
CSNA3	Companhia Siderúrgica Nacional	0,274
FLRY3	Fleury S.A.	0,272
GOAU4	Metalúrgica Gerdau S.A.	0,268
STBP3	Santos Brasil Participações S.A.	0,258
AZZA3	Azzas 2154 S.A.	0,257
COGN3	Cogna Educação S.A.	0,25
POMO4	Marcopolo S.A.	0,249
MRFG3	Marfrig Global Foods S.A.	0,242
IGTI11	Iguatemi S.A.	0,208
DIRR3	Direcional Engenharia S.A.	0,205
YDUQ3	Yduqs Participações S.A.	0,195
BRAP4	Bradespar S.A.	0,19
RECV3	PetroReconcavo S.A.	0,185
IRBR3	IRB Brasil Resseguros S.A.	0,175
SLCE3	SLC Agrícola S.A.	0,161
VIVA3	Vivara Participações S.A.	0,154
MGLU3	Magazine Luiza S.A.	0,15
AURE3	Auren Energia S.A.	0,145
BRKM5	Braskem S.A.	0,112
MRVE3	MRV Engenharia e Participações S.A.	0,108
USIM5	Usinas Siderúrgicas de Minas Gerais S.A – Usiminas	0,106
SMTO3	São Martinho S.A.	0,105
BEEF3	Minerva S.A.	0,103
RAIZ4	Raízen S.A.	0,095
VAMO3	Vamos Locação de Caminhões, Máquinas e Equipamentos S.A.	0,088
PCAR3	Companhia Brasileira de Distribuição – Grupo Pão de Açúcar	0,068
PETZ3	Petz S.A.	0,052
CVCB3	CVC Brasil Operadora e Agência de Viagens S.A.	0,051

Fonte: Extraída de [5] - Alterada.

B. MEDIDAS DE TENDÊNCIA CENTRAL

Tabela B.1: Distribuição percentual das posições compradas e não compradas associadas a análise descritiva (mínimo, 1º quartil e mediana) dos retornos dos preços de fechamento.

Ativo	Escala	Mínimo [%]		1º Quartil [%]		Mediana [%]	
		Compra	Sem Compra	Compra	Sem Compra	Compra	Sem Compra
PETR4	**Semanal**	47,58	52,42	36,33	63,67	23,42	76,58
VALE3		47,75	52,25	37,50	62,50	25,17	74,83
SBSP3		46,92	53,08	36,58	63,42	24,00	76,00
ITUB4		47,92	52,08	36,83	63,17	24,75	75,25
PETR4	**Mensal**	47,04	52,96	33,70	66,30	23,70	76,30
VALE3		48,89	51,11	34,44	65,56	22,22	77,78
SBSP3		44,07	55,93	31,48	68,52	20,00	80,00
ITUB4		43,24	56,76	35,14	64,86	21,62	78,38

Fonte: Elaborada pelo autor.

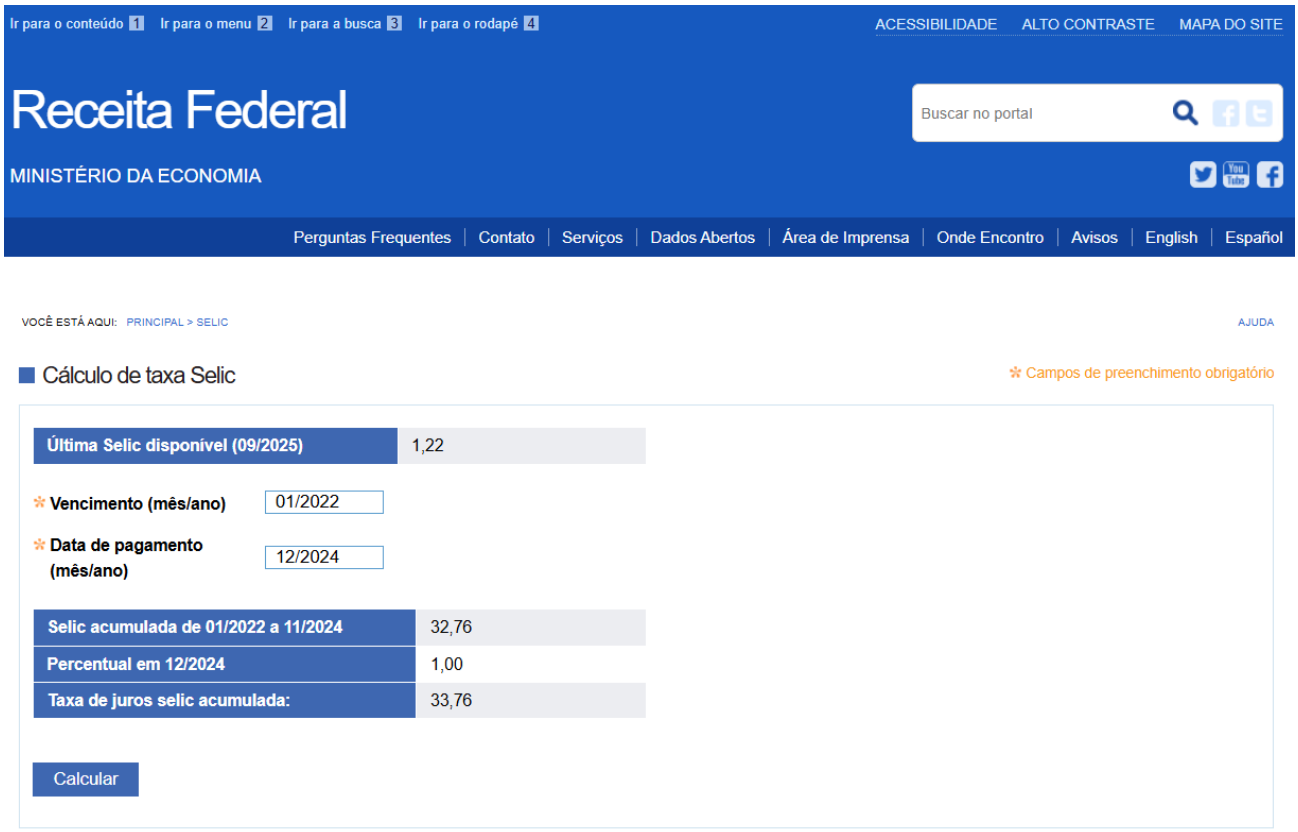
Tabela B.2: Distribuição percentual das posições compradas e não compradas associadas a análise descritiva (média, 3º quartil e máximo) dos retornos dos preços de fechamento.

Ativo	Frequência	Média [%]		3º Quartil [%]		Máximo [%]	
		Compra	Sem Compra	Compra	Sem Compra	Compra	Sem Compra
PETR4	**Semanal**	18,25	81,75	12,83	87,17	0,00	100,00
VALE3		19,42	80,58	12,67	87,33	0,00	100,00
SBSP3		19,83	80,17	12,50	87,50	0,08	99,92
ITUB4		18,50	81,50	12,58	87,42	0,00	100,00
PETR4	**Mensal**	18,52	81,48	12,59	87,41	0,37	99,63
VALE3		19,26	80,74	9,63	90,37	0,74	99,26
SBSP3		15,93	84,07	10,37	89,63	0,74	99,26
ITUB4		17,37	82,63	11,20	88,80	0,77	99,23

Fonte: Elaborada pelo autor.

C. Taxa selic acumulada

Figura C.1: Taxa selic acumualda de 2022 até 2024.



Fonte: Extraída de [39].

D. RESULTADO DA ROLETA DE CASSINO

Tabela D.1: Resultados das simulações da roleta de cassino utilizando a estratégia *Street of Gold*, rodada 1-19.

Rodada	Valor investido na rodada [R$]	Resultado da roda [R$]	Retorno financeiro da roda [R$]	Capital Acum. [R$]	Nº sorteado	Streets Modo A	Streets Modo B	Retorno [%]
1	64,00	96,00	32,00	3.598,00	19	13-18; 19-24; 25-30; 31-36	-	0.88%
2	64,00	0,0	-64,00	3.534,00	3	13-18; 19-24; 25-30; 31-36	-	-0.91%
3	160,00	288,00	128,00	3.662,00	29	13-18; 19-24; 25-30; 31-36	16-21; 22-27; 28-33	2.68%
4	64,00	0,0	-64,00	3.598,00	6	7-12; 13-18; 19-24; 25-30	-	0.88%
5	160,00	288,00	128,00	3.726,00	22	7-12; 13-18; 19-24; 25-30	10-15; 16-21; 22-27	4.47%
6	64,00	96,00	32,00	3.758,00	30	7-12; 13-18; 19-24; 25-30	-	5.37%
7	64,00	96,00	32,00	3.790,00	9	7-12; 13-18; 19-24; 25-30	-	6.27%
8	64,00	0,0	-64,00	3.726,00	1	7-12; 13-18; 19-24; 25-30	-	4.47%
9	160,00	288,00	128,00	3.854,00	19	7-12; 13-18; 19-24; 25-30	10-15; 16-21; 22-27	8.06%
10	64,00	96,00	32,00	3.886,00	7	1-6; 7-12; 13-18; 19-24	-	8.96%
11	64,00	96,00	32,00	3.918,00	23	1-6; 7-12; 13-18; 19-24	-	9.86%
12	64,00	96,00	32,00	3.950,00	4	1-6; 7-12; 13-18; 19-24	-	10.75%
13	64,00	96,00	32,00	3.982,00	24	1-6; 7-12; 13-18; 19-24	-	11.65%
14	64,00	96,00	32,00	4.014,00	10	1-6; 7-12; 13-18; 19-24	-	12.55%
15	64,00	96,00	32,00	4.046,00	11	1-6; 7-12; 13-18; 19-24	-	13.45%
16	64,00	96,00	32,00	4.078,00	15	1-6; 7-12; 13-18; 19-24	-	14.34%
17	64,00	96,00	32,00	4.110,00	1	1-6; 7-12; 13-18; 19-24	-	15.24%
18	64,00	0,0	-64,00	4.046,00	25	1-6; 7-12; 13-18; 19-24	-	13.45%
19	160,00	288,00	128,00	4.174,00	13	1-6; 7-12; 13-18; 19-24	4-9; 10-15; 16-21	17.03%

Fonte: Elaborada pelo autor.

Tabela D.2: Resultados das simulações da roleta de cassino utilizando a estratégia *Street of Gold*, rodada 20-41.

Rodada	Valor investido na rodada [R$]	Resultado da roda [R$]	Retorno financeiro da roda [R$]	Capital Acum. [R$]	Nº sorteado	Streets Modo A	Streets Modo B	Retorno [%]
20	64,00	0,0	-64,00	4.110,00	6	7-12; 13-18; 19-24; 25-30	-	15,24%
21	160,00	288,00	128,00	4.238,00	23	7-12; 13-18; 19-24; 25-30	10-15; 16-21; 22-27	18,83%
22	64,00	96,00	32,00	4.270,00	10	7-12; 13-18; 19-24; 25-30	-	19,73%
23	64,00	96,00	32,00	4.302,00	27	7-12; 13-18; 19-24; 25-30	-	20,62%
24	64,00	0,0	-64,00	4.238,00	4	7-12; 13-18; 19-24; 25-30	-	18,83%
25	160,00	0,0	-160,00	4.078,00	2	7-12; 13-18; 19-24; 25-30	10-15; 16-21; 22-27	14,34%
26	160,00	96,00	-64,00	4.014,00	29	7-12; 13-18; 19-24; 25-30	10-15; 16-21; 22-27	12,55%
27	160,00	96,00	-64,00	3.950,00	7	7-12; 13-18; 19-24; 25-30	10-15; 16-21; 22-27	10,75%
28	160,00	288,00	128,00	4.078,00	20	7-12; 13-18; 19-24; 25-30	10-15; 16-21; 22-27	14,34%
29	64,00	96,00	32,00	4.110,00	23	13-18; 19-24; 25-30; 31-36	-	15,24%
30	64,00	96,00	32,00	4.142,00	13	13-18; 19-24; 25-30; 31-36	-	16,14%
31	64,00	0,00	-64,00	4.078,00	5	13-18; 19-24; 25-30; 31-36	-	14,34%
32	160,00	0,0	-160,00	3.918,00	8	13-18; 19-24; 25-30; 31-36	16-21; 22-27; 28-33	9,86%
33	160,00	288,00	128,00	4.046,00	27	13-18; 19-24; 25-30; 31-36	16-21; 22-27; 28-33	13,45%
34	64,00	0,0	-64,00	3.982,00	33	7-12; 13-18; 19-24; 25-30	-	11,65%
35	160,00	0,0	-160,00	3.822,00	5	7-12; 13-18; 19-24; 25-30	10-15; 16-21; 22-27	7,16%
36	160,00	96,00	-64,00	3.758,00	28	7-12; 13-18; 19-24; 25-30	10-15; 16-21; 22-27	5,37%
37	160,00	96,00	-64,00	3.694,00	8	7-12; 13-18; 19-24; 25-30	10-15; 16-21; 22-27	3,58%
38	160,00	0,0	-160,00	3.534,00	3	7-12; 13-18; 19-24; 25-30	10-15; 16-21; 22-27	-0,91%
39	160,00	288,00	128,00	3.662,00	25	7-12; 13-18; 19-24; 25-30	10-15; 16-21; 22-27	2,68%
40	64,00	0,0	-64,00	3.598,00	35	10-15; 16-21; 22-27; 28-33	-	0,88%
41	160,00	288,00	128,00	3.726,00	25	10-15; 16-21; 22-27; 28-33	13-18; 19-24; 25-30	4,47%

Fonte: Elaborada pelo autor.

Tabela D.3: Resultados das simulações da roleta de cassino utilizando a estratégia *Street of Gold*, rodada 42-59.

Rodada	Valor investido na rodada [R$]	Resultado da roda [R$]	Retorno financeiro da roda [R$]	Capital Acum. [R$]	Nº sorteado	Streets Modo A	Streets Modo B	Retorno [%]
42	64,00	96,00	32,00	3.758,00	12	1-6; 7-12; 13-18; 19-24	-	5,37%
43	64,00	0,0	-64,00	3.694,00	30	1-6; 7-12; 13-18; 19-24	-	3,58%
44	160,00	288,00	128,00	3.822,00	17	1-6; 7-12; 13-18; 19-24	4-9; 10-15; 16-21	7,16%
45	64,00	96,00	32,00	3.854,00	29	13-18; 19-24; 25-30; 31-36	-	8,06%
46	64,00	96,00	32,00	3.886,00	18	13-18; 19-24; 25-30; 31-36	-	8,96%
47	64,00	96,00	32,00	3.918,00	13	13-18; 19-24; 25-30; 31-36	-	9,86%
48	64,00	96,00	32,00	3.950,00	29	13-18; 19-24; 25-30; 31-36	-	10,75%
49	64,00	96,00	32,00	3.982,00	28	13-18; 19-24; 25-30; 31-36	-	11,65%
50	64,00	0,0	-64,00	3.918,00	3	13-18; 19-24; 25-30; 31-36	-	9,86%
51	160,00	288,00	128,00	4.046,00	31	13-18; 19-24; 25-30; 31-36	16-21; 22-27; 28-33	13,45%
52	64,00	0,0	-64,00	3.982,00	7	10-15; 16-21; 22-27; 28-33	-	11,65%
53	160,00	96,00	-64,00	3.918,00	32	10-15; 16-21; 22-27; 28-33	13-18; 19-24; 25-30	9,86%
54	160,00	0,0	-160,00	3.758,00	2	10-15; 16-21; 22-27; 28-33	13-18; 19-24; 25-30	5,37%
55	160,00	0,0	-160,00	3.598,00	9	10-15; 16-21; 22-27; 28-33	13-18; 19-24; 25-30	0,88%
56	160,00	288,00	128,00	3.726,00	24	10-15; 16-21; 22-27; 28-33	13-18; 19-24; 25-30	4,47%
57	64,00	96,00	32,00	3.758,00	27	10-15; 16-21; 22-27; 28-33	-	5,37%
58	64,00	0,0	-64,00	3.694,00	3	10-15; 16-21; 22-27; 28-33	-	3,58%
59	160,00	0,0	-160,00	3.534,00	7	10-15; 16-21; 22-27; 28-33	13-18; 19-24; 25-30	-0,91%

Fonte: Elaborada pelo autor.