UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Instituto de Ciência e Tecnologia
Câmpus de Sorocaba

VINÍCIUS GOMES RODELLA

ESTUDO DE CASO: APLICAÇÃO DE *MACHINE LEARNING* PARA A PREVISÃO DE TENDÊNCIAS DAS AÇÕES DAS BOLSAS DE VALORES BRASILEIRA E NORTE AMERICANA

Sorocaba

2023

Vinícius Gomes Rodella

ESTUDO DE CASO: APLICAÇÃO DE *MACHINE LEARNING* PARA A PREVISÃO DE TENDÊNCIAS DAS AÇÕES DAS BOLSAS DE VALORES BRASILEIRA E NORTE AMERICANA

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do título de Bacharel em Engenharia de Controle e Automação pela Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Ciência e Tecnologia.

Orientador: Prof. Dr. Márcio Alexandre Marques

Sorocaba

2023

R687e

Rodella, Vinícius Gomes

Estudo de caso: aplicação de machine learning para a previsão de tendências das ações das bolsas de valores brasileira e norte americana / Vinícius Gomes Rodella. -- Sorocaba, 2023

78 f. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Engenharia de Controle e Automação) - Universidade Estadual Paulista (Unesp), Instituto de Ciência e Tecnologia, Sorocaba

Orientador: Márcio Alexandre Marques

1. Aprendizado do computador. 2. Inteligência artificial. 3. Mercado de capitais. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Ciência e Tecnologia, Sorocaba. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Instituto de Ciência e Tecnologia
Câmpus de Sorocaba

ESTUDO DE CASO: APLICAÇÃO DE MACHINE LEARNING PARA A PREVISÃO DE TENDÊNCIAS DAS AÇÕES DAS BOLSAS DE VALORES BRASILEIRA E NORTE AMERICANA

VINICIUS GOMES RODELLA

ESTE TRABALHO DE GRADUAÇÃO FOI JULGADO ADEQUADO COMO PARTE DO REQUISITO PARA A OBTENÇÃO DO GRAU DE **BACHAREL EM ENGENHARIA DE CONTROLE E AUTOMAÇÃO**

Profº. Drº. Everson Martins

Coordenador

BANCA EXAMINADORA:

Prof. Dr. Márcio Alexandre Marques

Orientador/UNESP – Campus de Sorocaba

Profa. Dra. Luiza Amalia Pinto Cantão

UNESP – Campus de Sorocaba

Profa. Dra. Maria Lúcia Pereira Antunes

UNESP – Campus de Sorocaba

Janeiro de 2023

Este trabalho é dedicado à minha família e amigos.

AGRADECIMENTOS

Aos meus pais, **Marcos Antônio** e **Silvia Helena**, que sempre preferiram desfazer dos seus próprios sonhos em função dos meus, sempre incentivando o estudo e a busca pela evolução moral e espiritual. Também por sempre estarem ao meu lado em todos os momentos, pelo incentivo e apoio em minhas escolhas e decisões.

Aos meus avós, pelo cuidado e carinho e por incentivarem meus pais nos estudos, incentivo que depois estendeu-se até mim. Em especial a minha vó **Diva**, que é uma segunda mãe.

Ao meu irmão, **Leonardo** e vários amigos que passaram, que estão neste momento e os quais no futuro estarão, pela parceria, conselhos, aprendizados e pelos momentos de distração durante a caminhada.

A minha namorada **Stefanie**, pelo amor, companheirismo, incentivo, aprendizados e por me mostrar as coisas importantes com um outro olhar.

Ao meu orientador **Prof. Dr. Márcio Alexandre Marques** pelo auxílio, paciência, confiança e provocações pertinentes que agregaram a este trabalho.

Aos funcionários da Unesp Sorocaba, por permitirem o funcionamento de toda a estrutura do campus.

Por fim, agradeço a todos os professores que um dia me ensinaram e contribuíram para minha formação profissional e pessoal. Em especial aos professores do Colégio London e Cursinho Kelvin de São José do Rio Preto, por desenvolverem em mim a vontade em buscar conhecimento.

“Ideias e somente ideias podem iluminar a escuridão."

(Ludwig von Mises (1881 – 1973), economista austríaco)

RESUMO

A predição da tendência no mercado acionário é um trabalho difícil, uma vez que os bens negociados tendem a sofrer diversas influências, como a política internacional, o movimento do câmbio e a política interna do país em que o ativo está sendo negociado. Apesar disso, o profissional *trader* aplica técnicas gráficas e de indicadores para a detecção de padrões e assim efetuar sua análise buscando operar com alta frequência na bolsa na tentativa de obter ganhos significativos, contrariando a hipótese do mercado eficiente. Esta detecção de padrões e a possibilidade da análise de séries temporais, também permitem a realização de diversos estudos que buscam modelar possíveis tendências dos ativos, sendo impulsionado pelo avanço da inteligência artificial, possibilitando que um hardware mais robusto execute modelos robustos e complexos. Neste contexto, este trabalho busca uma nova abordagem para este campo de pesquisa apresentando a aplicação dos métodos de *machine learning*: *Random Forest* e *Adaboost*, capazes de seguirem uma determinada estratégia de investimento utilizando análise gráfica das ações, buscando prever a tendência de alta ou de baixa dos ativos nas bolsas de valores brasileira e norte americana em diferentes períodos. Desta maneira, foi desenvolvido uma metodologia utilizando 6 ações de cada região em estudo, onde foi escolhida a seguinte combinação dos indicadores para os modelos: média móvel, índice de força relativa, divergência e convergência da média móvel e taxa de variação. Durante o desenvolvimento, métodos diferentes para o cálculo dos indicadores e a captura dos dados foram testados, verificando também, a necessidade do balanceamento dos dados e uma melhor escolha dos períodos de análise para cada ação com a finalidade de atingir uma predição satisfatória, ou seja, onde a inteligência artificial (IA) tenha aprendizado real, evitando o *overffiting*. Assim, os resultados indicam que as ações norte americanas variaram menos do que as brasileiras e isso pode ser justificado pela maior volatilidade da bolsa brasileira. Por fim, o uso desses modelos de aprendizado de máquina aplicados para previsão das tendências das ações nas bolsas de valores mostrou-se inviável na prática devido a necessidade de balancear os dados, mas com um potencial muito promissor.

Palavras-chave: Inteligência Artificial, Hipótese do Mercado Eficiente, *Random Forest*, *Adaboost*.

ABSTRACT

Prediction of trend in the stock market is a difficult task, as the traded goods tend to suffer various influences, such as international politics, exchange rate movement, and the domestic policy of the country in which the asset is being traded. Despite, the professional trader applies graphic and indicator techniques for pattern detection and thus performs their analysis seeking to operate with high frequency in the stock exchange in an attempt to obtain significant gains, contradicting the efficient market hypothesis. This pattern detection and the possibility of time series analysis also allow for the conduct of various studies that aim to model possible asset trends, being propelled by the advancement of artificial intelligence, allowing for more robust hardware to execute robust and complex models. In this context, this work seeks a new approach to this research field by presenting the application of the machine learning methods: Random Forest and Adaboost, capable of following a certain investment strategy using graphic analysis of stocks, seeking to predict the trend of assets in the Brazilian and North American stock exchanges at different periods. In this way, a methodology was developed using 6 actions from each region under study, where the following combination of indicators was chosen for the models: moving average, relative strength index, divergence and convergence of the moving average, and rate of variation. During development, different methods for calculating indicators and data capture were tested, also verifying the need for data balancing and a better choice of analysis periods for each action in order to achieve a satisfactory prediction, that is, where artificial intelligence (AI) has real learning, avoiding overfitting. Thus, the results indicate that North American actions varied less than Brazilian actions and this can be justified by the greater volatility of the Brazilian stock exchange. Finally, the use of these machine learning models applied to predict the trends of assets in stock exchanges proved unfeasible in practice due to the need to balance data, but with a very promising potential.

Keywords: Artificial Intelligence, Efficient Market Hypothesis, Random Forest, Adaboost.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

ML — *Machine Learning*

IA — Inteligência Artificial

HME — Hipótese do Mercado Eficiente

B3 — Brasil Bolsa Balcão

BOVESPA — Bolsa de valores do Estado de São Paulo

BM&F — Bolsa de Mercadoria & Futuros

BM&FBOVESPA — Bolsa de Mercadoria & Futuros e Bolsa de valores do Estado de São Paulo

CETIP — Central de Custódia e Liquidação Financeira de Títulos

CVM — Comissão de Valores Mobiliários

NYSE — *The New York Stock Exchange*

NASDAQ — *National Association of Securities Dealers Automated Quotations*

S&P500 — *Standard & Poor's 500*

RF — *Random Forest*

MMS — Média Móvel Simples

MME — Média Móvel Exponencial

RSI — Índice de Força Relativa

MACD — *Moving Average Convergence Divergence*

ROC — *Rate of Change*

SVM — *Support Vector Machine*

RNA — Redes Neurais Artificiais

AAPL — Apple

NFLX — Netflix

AMZN — Amazon

TWTR — Twitter

KO — Coca Cola

ITUBS4.SA — Itaú Unibanco Holdings S.A.

VALE3.SA — Vale

ABEV3.SA — Ambev

WEGE3.SA — Weg

CIEL3.SA — Cielo

BRFS3.SA — BRF

ITUBS4 — Itaú Unibanco Holdings S.A.

VALE3 — Vale

ABEV3 — Ambev

WEGE3 — Weg

CIEL3 — Cielo

BRFS3 — BRF

ROC — Receiver Operating Characteristic

AUROC — Area Under the Receiver Operating Characteristic

IPCA — Índice Nacional de Preços ao Consumidor Amplo

IGP — Índice Geral de Preço

PIB — Produto Interno Bruto

SELIC — Sistema Especial de Liquidação e de Custódia

SUMÁRIO

1. INTRODUÇÃO

1.1 Contextualização e desafios

A Inteligência Artificial (IA), é um assunto com grande cobertura da mídia e discussões públicas sobre o tema são praticamente impossíveis de evitar, até mesmo ocorrendo uma certa banalização do termo, já que até mesmo os pesquisadores de IA utilizam diversas definições. Desse modo, para exemplificar, a IA tem como a possibilidade de máquinas aprenderem com experiências e desempenharem tarefas como seres humanos (DATA SCIENCE ACADEMY 1, 2020).

Atualmente, devido ao avanço dos métodos matemáticos e computacionais, e do hardware, a IA é capaz de lidar com a incerteza, diferentemente da maioria dos métodos dos anos de 1960 e 1980, que por conta disso, é aplicável no mundo real. Algumas das aplicações mais vistas são em carros autônomos, recomendação de conteúdo nas redes sociais e processamento de imagem e vídeo, como o reconhecimento facial (DATA SCIENCE ACADEMY 2, 2020).

Apesar de avançar em todos os campos comentados, os avanços recentes mais empolgantes estão no campo de *deep learning*, onde contas e cálculos são possíveis pelo grande avanço do hardware, nas quais a previsão e classificação estão sendo realizadas sem o aprendizado supervisionado, ou seja, com novos métodos. Entretanto, por ser uma área muito recente e pouco explorada, diversas limitações ainda são encontradas (CHUI, MANYIKA, MIREMADI, 2018).

Dentre elas destacam-se a necessidade de grandes volumes de dados para executar uma tarefa complexa com precisão e a dificuldade em discernir como um modelo matemático com aprendizado profundo chega a uma previsão ou recomendação. Sendo assim, ainda é muito utilizado *o machine learning* ao invés do recente *deep learning*.

Nesse sentido, atualmente, o mercado de ações no nível mais básico de descrição, se refere ao mercado abstrato que comercializa a propriedade das empresas. Entretanto, as primeiras bolsas, criadas no século XV, conectavam pessoas com base na troca de moedas, letras de câmbio e até metais preciosos, sendo assim até o século XVII. Desta forma, desde os primórdios é notório uma função crucial deste sistema, que além de estabelecer um local padrão, conecta necessidades, facilitando o elo mais importante do sistema econômico em que vivemos hoje, conhecido como troca, conectando poupadores e devedores (BIANCA, 2019).

Observando este sistema de negociação ao longo dos anos, nota-se um grande avanço na quantidade de transações e no número de investidores, com o primeiro avanço, ocasionado pela evolução da tecnologia, já o segundo, pelo acesso à informação e a facilidade de investir. Assim, com cada vez mais transações por dia, proporcionando uma grande liquidez nestes mercados, a prática de *trading* se tornou cada vez mais comum, por exemplo, a quantidade de negociações realizadas na Bolsa de Nova York (NYSE) em 2009 superou em mil vezes as negociadas em 1967 (MADEO, FERREIRA, RAMALHO, FANTINATO, 2012).

Dessa forma, com a prática cada vez mais popular e até enganosa do *trading*, uma vez que muitas pessoas acreditam que este ramo traz dinheiro fácil, onde na verdade é uma área extremamente difícil, de alto risco e que necessita de muita experiência. Dado que mesmo com estudo técnico é difícil prever o movimento dos ativos em um curto espaço de tempo, como o *trading* sugere. Concomitantemente, bons *traders* conseguem bons rendimentos, gerando maiores ganhos do que apenas realizando o *holding* das ações, porém um fator acaba pesando na maioria que se arriscam nesta área, o fator emocional.

Com relação a este último ponto, sabe-se que a maioria das pessoas que entram nesta área, inicialmente deixam-se afetar pela emoção, o que muitas vezes resultam em prejuízos financeiros expressivos, já que é necessário a utilização da razão para tomar decisões. Também, até mesmo pessoas mais experientes e com uma boa inteligência emocional, podem acabar deixando ser levadas algumas vezes pela emoção, tornando interessante o estudo da previsão das tendências de maneira computacional, ou seja, sem o fator emoção para operar os ativos.

Posto isto, a análise técnica para previsão de preços encontra-se em um grande desafio. As ações são influenciadas por fatores sociais, econômicos e políticos, possuindo características de imprevisibilidade. Entre os negociantes do mercado, é conhecida a Hipótese do Mercado Eficiente (HME), apontado primeiramente por COOTNER, 1964 e FAMA, 1969, onde afirmam que qualquer técnica de previsão de preços é, a princípio, inútil, uma vez que a variação de preços é aleatória, independente do mercado (COOTNER, 1964) (FAMA, 1969).

Entretanto, mesmo com uma grande força inicial acerca desta teoria, outros autores não compactuaram com a eficiência do mercado e esta teoria passou a ser amplamente questionada. Entre os estudos, notava-se uma assimetria nas distribuições de retornos (MANDELBROT, RICHARD, 2004) e padrões que influenciavam os movimentos de mercado (BORGES, 2010) e (COVA, 2011), possibilitando a previsão dos ativos, contrariando a teoria do mercado eficiente.

Em vista disso, diversos métodos foram criados e em muitas vezes a aplicação é útil e gera resultados, contudo, é importante ressaltar que nenhum dos métodos são perfeitos e devido a imprevisibilidade a tarefa se torna complexa, não conseguindo explicar os movimentos dos preços com total acurácia.

1.2 Objetivos e justificativa

Dessa maneira, este trabalho busca se aprofundar na utilização de métodos de aprendizado de máquina (AM) (*machine learning* (ML)) na bolsa de valores, tendo como objetivo a utilização de dois métodos diferentes de classificação de ML para treinar os algoritmos acerca da tendência das ações: *Adaboost* e *Random Forest*.

Propõe-se treinar dois métodos de *machine learning* capazes de seguirem uma determinada estratégia de investimento, buscando prever a tendência de alta ou de baixa das ações nas bolsas de valores brasileira e norte americana. Tem-se então como objetivos: (1) estudar a aplicação de técnicas de *machine learning* no contexto do mercado financeiro; (2) estudar e definir a estratégia de investimento; (3) definir indicadores que possuam boa sinergia para atuarem na estratégia; (4) construir um modelo computacional treinado para que obtenha lucro no mercado de ações e (5) comparar e discutir os resultados obtidos.

1.3 Estrutura deste trabalho

Esse Projeto Final de Curso está dividido em 6 capítulos:

O capítulo 1 contextualiza o tema do trabalho mostrando os desafios da área que abrange a pesquisa de algoritmos de *machine learning* aplicados à previsão do mercado de ações. Também, apresenta os objetivos e justificativa do estudo.

O capítulo 2 introduz a teoria do trabalho e aborda os trabalhos relacionados da área, realizando a revisão bibliográfica deste campo do estudo, descrevendo como funciona o mercado acionário e a inteligência artificial, bem como as principais bolsas dos Estados Unidos e a bolsa Brasileira, explica a importante hipótese do mercado eficiente, esclarece os métodos de análise do mercado assim como o funcionamento dos indicadores básicos e avançados (técnicos), para no final expor sobre *machine learning* e os modelos utilizados no trabalho *(Random Forest e AdaBoost).*

No capítulo 3 tem-se a exposição da metodologia do trabalho para a criação dos algoritmos de previsão das ações, descrevendo as etapas de coleta de dados e tratamento inicial,

extração e utilização dos indicadores e por fim como foi realizada a avaliação do desempenho dos modelos.

O capítulo 4 demonstra todos os resultados obtidos pelos modelos de *machine learning* através dos dados da avaliação de desempenho como a acurácia, precisão, *recall, F1 score* e curva ROC.

Por fim, o capítulo 5 apresenta a conclusão deste trabalho, elencando as contribuições realizadas e possíveis propostas para trabalhos futuros.

2. REVISÃO BIBLIOGRÁFICA e ASPECTOS TEÓRICOS

Neste capítulo serão apresentados os conceitos e a teoria do mercado de ações brasileiro e americano, como funcionam, bem como o que é a hipótese do mercado eficiente e as análises técnicas utilizadas no mercado financeiro. Posteriormente, também serão abordados conceituação da inteligência artificial e de *machine learning* aplicadas.

2.1 O mercado acionário

O mercado acionário é um ambiente público onde as empresas de capital aberto negociam títulos mobiliários, imobiliários e frações de seu patrimônio. As negociações ocorrem na bolsa de valores ou nos mercados de balcão.

A movimentação do mercado financeiro e de capitais são influenciadas por mudanças na política e economia, como por exemplo o aumento recente da taxa básica de juros pelos bancos centrais de grande parte dos principais países, ocasionando uma fuga do mercado de ações para o investimento em renda fixa. Também por acontecimentos geopolíticos, como a invasão recente da Ucrânia pela Rússia, que ocasionaram uma queda generalizada nas ações no mundo todo, mas principalmente na Rússia, onde a bolsa chegou a cair certa de 50% (PODER360, 2022).

Os resultados e notícias das empresas também influenciam na movimentação, como por exemplo, no Brasil, o Itaú que obteve um balanço do 4º trimestre acima do esperado pelo mercado, subindo o lucro de 2021 em 45% com relação a 2020, ocasionando o avanço de 6,8% da ação (ITUB4) na manhã da divulgação dos resultados da empresa (RIZÉRIO, FÁBIO, 2022).

Antigamente, o mercado acionário possuía pregões presenciais e operadores anunciavam as ações e ficavam no telefone comprando e vendendo. Entretanto, hoje todo o processo ocorre digitalmente. No Brasil as operações de compra e venda ocorrem na B3, bolsa de valores do país, já nos Estados Unidos as operações ocorrem na NYSE e NASDAQ, bolsa de valores de Nova York e a Associação Nacional de Corretores de Títulos de Cotações Automáticas (*National Association of Securities Dealer Automated Quotations*) (INFOMONEY, 2022).

2.1.1 B3

A B3 é resultado de duas fusões, uma que iniciou em 26 de março de 2008 entre a Bovespa (Bolsa de valores do Estado de São Paulo) e a BM&F (Bolsa de Mercadoria & Futuros) e outra entre a BM&FBovespa e a CETIP (Central de Custódia e Liquidação Financeira de Títulos) em março de 2017 (COMOINVESTIR, 2019).

A sigla 'B3' significa Brasil, Bolsa e Balcão e atualmente é a oitava maior bolsa de valores do mundo, sediada na cidade de São Paulo, possuindo 388 companhias brasileiras e um total de 4,2 trilhões de reais ou 824 bilhões de dólares em julho de 2022. Ela é regulada pela Comissão de Valores Mobiliários (CVM).

O índice que indica o desempenho médio das cotações das ações da B3 chama-se Índice Bovespa e é formado pelas ações com maior volume negociado nos últimos meses. Assim, quando ocorre uma valorização na média dos preços dessas empresas o índice aumenta.

2.1.2 NYSE E NASDAQ

A bolsa de valores de Nova York, do inglês, *The New York Stock Exchange*, está localizada em Manhattan, na Wall Street, foi criada em 1792 e é a maior bolsa de valores do mundo. Para efeitos de comparação com a B3, são mais de 2780 empresas listadas e o valor das empresas listadas na NYSE somam 21 trilhões de dólares. O índice de desempenho é o *NYSE Composite* (PINTO, 2020).

Já a NASDAQ é a segunda maior bolsa do mundo, estando listadas mais de 2800 empresas e possuindo mais de 7 trilhões de dólares em valor de mercado, sendo em sua maioria empresas de pequena ou média capitalização. O índice de desempenho é a *NASDAQ Composite.*

Apesar dos diferentes índices, os mais utilizados para acompanhar o mercado americano são conhecidos como Dow Jones e S&P500, ambos incluem papéis que são negociados em outras bolsas americanas. O índice S&P500 é composto pelas 500 maiores ativos da NYSE E NASDAQ, já o índice Dow Jones avalia as 30 maiores ações que estão na NYSE (PINTO, 2020).

2.2 Hipótese do Mercado Eficiente

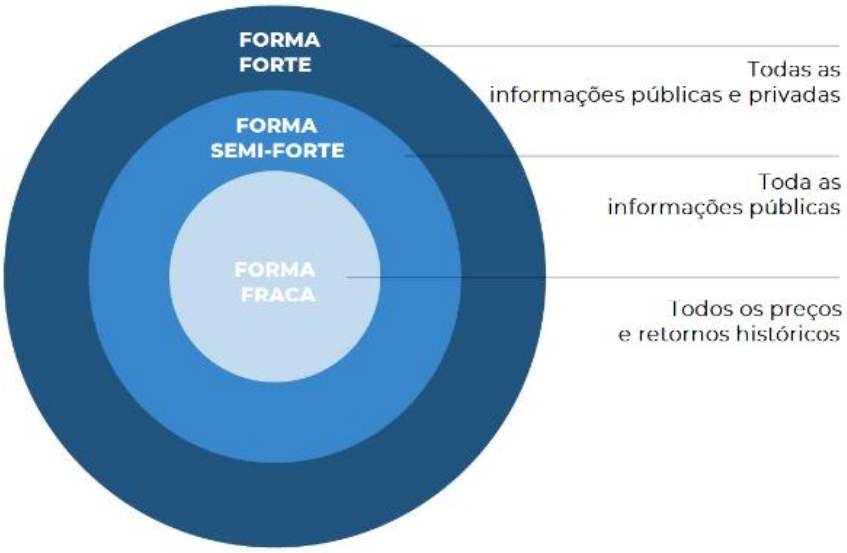
Como descrito anteriormente, o mercado de ações é influenciado por fatores sociais, econômicos e políticos e por conta disso uma das visões do meio científico afirma que as ações possuem grande imprevisibilidade. É conhecida a Hipótese do Mercado Eficiente (HME) (MUSSA *et. al*), ela se baseia em dois componentes centrais que estão relacionados: a racionalidade dos preços e a possibilidade de vencer o mercado.

A primeira fixa-se na afirmação que o preço está sempre certo, ou seja, que qualquer ativo está sendo negociado pelo seu valor intrínseco verdadeiro pois o preço representa todas as informações existentes sobre as empresas emissoras e todas as vezes as informações são públicas. Já a segunda, fixa-se na afirmação de que não há maneira de vencer o mercado através da análise dos ativos pois os preços já refletem as informações disponíveis.

Desta maneira, dentro da hipótese existem três níveis diferentes de eficiência, como mostra a figura 1.

- Eficiência fraca: O mercado é eficiente em refletir todas as informações públicas disponíveis e os retornos no mercado são independentes, ou seja, resultados passados não auxiliam em prever resultados futuros, invalidando análises gráficas e metodologias de *traders*.
- Eficiência Semi-Forte: Engloba a hipótese fraca e sugere que novas informações são absorvidas instantaneamente pelo mercado, desta forma, investidores não conseguem resultados acima do mercado com informações conhecidas, invalidando a utilização de análises fundamentalistas.
- Eficiência Forte: Engloba as outras hipóteses e sugere que os preços também refletem informações privadas, assim mesmo com informações privadas não seria possível um retorno acima do mercado.

Figura 1- Representação dos níveis de eficiência.



Fonte: GUTIERRI, 2021.

Assim, a única maneira de lucrar em um mercado eficiente é fazer apenas investimentos de forma passiva, como por exemplo realizar investimentos em fundos com índice como o IBOVESPA e o S&P500, uma vez que representam uma média ampla do mercado, possibilitando investir em 'todo mercado' ao mesmo tempo.

Deste modo, quando comparar os rendimentos passivos com os rendimentos ativos de gestores, o resultado seria próximo, pois os rendimentos ativos convergem para a média do mercado devido a eficiência, e assim um rendimento passivo seria mais vantajoso por possuir menos custos e taxas, e o mesmo desempenho.

Contudo, investidores que operam diariamente, grandes investidores e outros estudiosos, não concordam com esta hipótese. Um famoso investidor que é contra esta hipótese é Warren Buffett, ele é conhecido por ser o maior investidor do mundo e possuir rendimentos consistentes bem acima de índices como o S&P500. Uma citação famosa dele é:

> "Eu seria um mendigo vagando pelas ruas com uma caneca nas mãos se os mercados fossem sempre eficientes" (LOWE, 2013).

Bem como Buffett, Howard Marks, escritor e investidor, diz que a teoria não está completamente equivocada, mas que a ineficiência do mercado é uma condição necessária para

que investidores consigam retornos consistentes acima da média, fato alcançado por diversos investidores (MARKS, 2011).

Burton diz que existem evidências estatísticas para a HME (Hipótese do mercado eficiente), porém elas são inconclusivas (MARKIEL, 2011). Também, o economista, John Quiggin relata que o Bitcoin é uma prova viva que a hipótese é falha, uma vez que quase tudo que hipótese diz não se aplica a este ativo, que possui liquidez e investidores racionais (QUIGGIN, 2018).

Paul Samuelson, vencedor do prêmio Nobel de economia em 1970, afirma que o mercado pode ser micro eficiente, ou seja, para alguns tipos de ativos ou situações a hipótese pode ser verdadeira, entretanto, para o mercado como um todo, a hipótese será falsa (JUNG, SHILLER, 2006).

2.3 Métodos de análise de mercado

Em geral, os métodos para análise do mercado de ações dividem-se em 3 campos (PIMENTA, 2014):

- Métodos convencionais, utilizam estatísticas de séries temporais e econometria;
- Métodos de mercado, utilizam análise fundamentalista, análise técnica, *candlestick*, *tape reading* etc.
- Métodos computacionais, fazem uso de algoritmos com inteligência artificial, utilizando ou não técnicas convencionais e de mercado para realizar as previsões.

Recentemente, outras abordagens também ganharam relevância na comunidade científica, entre elas constam o estudo de finanças comportamentais, hipótese de mercados fractais, sentimento de mercado e entre outras (ALMEIDA, 2019).

Independentemente do método, existe uma classificação em 3 diferentes modalidades com base na frequência das transações do *trade* (XAVIER, 2009):

- *Position trade:* Representa uma operação em longo prazo, normalmente, meses ou anos. Está relacionada a prática de *Buy and Hold* e análises fundamentalistas.
- *Swing trade:* Operação realizada em um período de poucos dias, geralmente entre 2 e 7 dias. O investidor busca oportunidades por índices, fontes, gráficos e notícias

para operar.

- *Day trade*/*Intraday:* Operações realizadas dentro de um mesmo pregão, ocorrendo em um intervalo de horas ou até minutos. Utiliza-se análise técnica, como a *tape reading*, e quando se utiliza robôs, pode-se operar até em milissegundos.

Como mostrado na figura 2, além destas, também existem os estilos de negociação menos arriscados, ou seja, que possuem menor frequência: *deep value investing*, *value investing*, *buy & hold*, que são melhores para o longo prazo. Dentre elas, a mais utilizada é o *buy & hold*, que significa comprar e segurar, ou seja, comprar a ação e permanecer comprado por um longo período com a finalidade do rendimento acompanhar o crescimento daquele ativo.

Figura 2 - Representação da frequência de cada estilo de negociação.

Fonte: MARCELO, 2020.

Agora, independente da frequência de transação (figura 2) e do método de análise, existem indicadores que sempre são avaliados, entre eles pode-se citar: o preço de abertura, preço de fechamento, preço mínimo, preço máximo e o volume, sendo todos indicadores diários de uma ação.

No caso do indicador de volume existem duas variações, o volume financeiro, que representa o valor total das negociações daquele dia, e o volume de transações, que como o nome diz, representa o valor total de transações do dia em análise.

Desse modo, estes indicadores (indicadores básicos, seção 2.4) também servem como base para o cálculo de diversos outros indicadores utilizados na análise das ações (indicadores técnicos na seção 2.5), que indicam a tendência e se o mercado está ‘sobrecomprado’ ou ‘sobrevendido’, ou seja, se está acima ou abaixo do valor ideal daquele ativo.

2.4 Indicadores básicos

- **Preço de abertura do dia (*Open*):**

Conhecido também como *Open*, indica o preço de início da ação no dia em análise, ou seja, normalmente será o mesmo preço de fechamento do dia anterior. Uma particularidade do preço de abertura é sua grande movimentação conforme as notícias que ocorreram durante o período em que o mercado estava fechado, ou seja, notícias positivas ou negativas, ocasionam a subida ou queda das ações na sua abertura.

- **Preço máximo do dia (*High*):**

Indica o preço máximo alcançado pelo ativo no dia em análise, serve como comparação ao preço mínimo e de fechamento, demonstrando ao investidor a variação daquele dia, e também, o possível valor que aquele ativo pode atingir novamente no curto prazo.

- **Preço mínimo do dia (*Low*):**

Indica o preço mínimo alcançado pelo ativo no dia em análise, assim como o preço máximo, serve como comparação aos outros indicadores de preço do ativo, fornecendo ao investidor a variação completa daquele dia (quando analisado com os outros indicadores de preço). Também é possível avaliar o potencial de queda que o ativo pode atingir novamente no curto prazo.

- **Preço de fechamento do dia (*Close*):**

Representa o preço com que o ativo estava no momento que ocorreu o fechamento da bolsa de valores da ação em análise. É o indicador mais utilizado pelos investidores para comparar o desempenho do histórico de preços, indicando o sentimento do mercado de acordo com o tempo.

- **Preço de fechamento ajustado (*Ajusted Close*):**

Indica o preço de fechamento após ajustes considerando a divisão de ações, dividendos, ofertas e direitos. Por exemplo, o ajuste por dividendos ocasiona uma redução do preço da ação de acordo com o dividendo distribuído. Ótimo indicador para análise do histórico da empresa, porém menos atrativo para quem atua com *day trading*.

- **Volume negociado no dia (*Volume*):**

Representa o volume financeiro transacionado no ativo no dia em análise, o fluxo de negociações. É utilizado pelos investidores como um validador de operações, ou seja, verifica-se o volume transacionado.

- **Data do dia (*Date*).**

Indica o dia em que estão sendo avaliados os indicadores.

2.5 Indicadores técnicos

2.5.1 Média Móvel Simples (MMS)

É caracterizada pela média dos valores de fechamento das ações durante um certo período em análise. Por exemplo, uma média móvel simples de 10 dias é calculada somando-se todos os valores de fechamento dos últimos 10 dias, dividindo pelo número de períodos em análise, nesse caso, 10. A equação (1) representa a média móvel simples.

$$M_\tau = \sum_{i=0}^{\tau-1} \frac{x_{t-i}}{\tau} \tag{1}$$

Onde:

- M é a média móvel simples;
- x é a série temporal;
- τ é o período em análise;
- t é o instante observado.

O termo ‘móvel’ é dado devido a frequente alteração dos dados para o cálculo da média, uma vez que assim que um novo valor de fechamento da ação surge, o último período da média é retirado para manter o número de períodos constante, desta maneira uma média móvel de 5 dias sempre possuirá 5 dias em análise.

Assim, este indicador atua como um rastreador, seguindo o comportamento do preço e confirmando a direção e tendência do preço. São considerados indicadores atrasados, pois necessitam de confirmação na tendência para dar sinal de entrada ou saída da ação, porém são ótimos para representar o valor médio real de preços.

2.5.2 Média Móvel Exponencial (MME)

A média móvel exponencial diferencia-se da simples pelo fato de proporcionar uma maior importância ao preço mais recente, desta maneira, adicionando um peso ao preço recente. Possui a função de acompanhar de forma ágil a mudança do preço de uma ação, facilitando a compreensão dos momentos de compra e venda, além de reduzir o atraso ao considerar mais os dados recentes. A equação 2 demonstra o cálculo desta variável.

$$M_{\propto}(t) = \propto \sum_{i=0}^{\tau} (1-\propto)^{i} X_{(i+1)} \tag{2}$$

Onde:

- M é a média móvel exponencial;
- x é a série temporal;
- τ é o período em análise;
- t é o instante observado;
- $\propto$ é o peso atribuído.

Uma estratégia de análise técnica, que é tanto utilizada pela média móvel simples, mas também pela média móvel exponencial, é o cruzamento das médias de curto prazo e longo prazo. Dessa forma, caso a média curta cruze para cima a média de longo prazo, pode ser um bom sinal de entrada no ativo, enquanto caso o contrário, cruze para baixo, pode ser um bom indicativo para sair do ativo.

Também existe a possibilidade de analisar o cruzamento da média móvel com o preço do ativo.

A determinação do número de períodos das médias vai de acordo com o perfil do investidor, para uma transação mais curta, analisar tendências de curto prazo, 5 e 20 períodos

são mais indicados, já para médio prazo, médias entre 20 e 60 períodos e para o longo prazo 100 ou mais. Também são utilizadas as variações, ou seja, o trabalho de uma média com curto prazo e outra com longo prazo, com a finalidade de analisar tendências e o cruzamento das médias.

2.5.3 Índice de Força Relativa (RSI)

É um indicador de *momentum*, ou seja, mede a velocidade e a mudança dos movimentos dos preços dos ativos, assim medindo a diferença de preço de fechamento atual para com o preço de fechamento de alguns dias atrás (o investidor decide o período a ser analisado), como pode ser visto na equação (3).

$$IFR = 100 - \left(1 + \left(\frac{100}{1 + \frac{U}{D}}\right)\right) \quad (3)$$

Onde:

- IFR (RSI) é o índice de Força Relativa;
- U é a média das cotações dos últimos ***n*** dias que a cotação da ação subiu.
- D é a média das cotações dos últimos ***n*** dias que a cotação da ação caiu.
- ***n*** é o número de dias em análise. Normalmente o valor utilizado é 14.

Como mostra na figura 3, este indicador varia de 0 a 100, acima de 70 indica que o ativo está supervalorizado, ou seja, o valor dele está acima do esperado e pode ocorrer uma reversão de tendência ou retração do preço. Também quando ele se encontra abaixo de 30, indica uma subvalorização, ou seja, o preço está abaixo do esperado, podendo ocorrer uma reversão de tendência ou subida do preço.

Figura 3 - Representação gráfica do indicador RSI.

Fonte: Baseado em (TRADEVIEW, 2022).

2.5.4 Williams %R

Também conhecido como um indicador de *momentum*, compara-se o preço de fechamento de um ativo com relação ao valor mais alto e mais baixo dos últimos *n* dias, demonstrado na equação (4).

$$\%R_i = \frac{máx_n - fechamento_i}{máx_n - mín_n} \times -100 \quad (4)$$

Onde:

- %Ri é o valor do indicador Williams %R para o dia i;
- n é o período de cálculo do indicador;
- máx é o valor máximo para a cotação do ativo nos últimos n períodos;
- mín é o valor mínimo para a cotação do ativo nos últimos n períodos;
- fechamento é o valor do fechamento do ativo nos últimos n períodos.

Conforme a figura 4, o indicador produz valores de -100 a 0. Quando o indicador estiver entre -80 e -100 indica sobrevenda, ou seja, sinal de compra. Já quando estiver entre -20 e 0, mostra sobrecompra, ou seja, sinal de venda.

Figura 4 - Representação gráfica do indicador Williams.

Fonte: Baseado em (TRADEVIEW, 2022).

2.5.5 MACD

Moving Average Convergence Divergence (MACD), também conhecido como Convergência/Divergência da média móvel, é utilizada para detectar rapidamente fortes tendências de curto prazo. Sua principal função é monitorar tendências e indicar possíveis sinais de mudança, confirmação ou reversão.

A MACD é calculada pela diferença entre a média móvel curta e a média móvel longa. Este indicador possui também uma linha de sinal e um histograma, o primeiro é calculado com a média móvel, já o segundo é a diferença entre a MACD e a linha de sinal.

Como mostra a figura 5, a linha da MACD fica positiva quando a média móvel curta cruza a longa de baixo para cima, assim ao contrário, o indicador fica negativo quando a curta cruza a longa de cima para baixo.

Figura 5 - Representação gráfica do indicador MACD.

Fonte: Baseado em (TRADEVIEW, 2022).

O cruzamento da MACD e a linha de sinal, ocasiona o sinal positivo e negativo do histograma indicando os sinais de compra e venda. Assim, quando o histograma está negativo e inicia subida, indica sinal de compra, já quando ele está positivo e iniciando queda, pode ser um sinal de venda.

2.5.6 Taxa de Variação

A taxa de variação (*Rate of Change* (ROC)), é um indicador técnico de *momentum* que mede a variação do percentual entre o preço atual e o preço registrado em alguns períodos atrás, como mostra a equação (5). Assim, a taxa de variação acompanha a relação entre os preços de fechamento em dois períodos.

$$ROC = \left(\frac{Fechamento - fechamento_n}{fechamento_n}\right) \times 100 \quad (5)$$

Para este indicador existem alguns sinais de compra e venda, conforme mostram as figuras 6, 7 e 8.

- Cruzamento da linha zero

Este acontecimento indica uma nova tendência. Quando o cruzamento ocorre de baixo para cima, indica uma tendência de alta, já quando ocorre o oposto, indica uma tendência de baixa.

Figura 6 - Representação gráfica do cruzamento da linha zero do indicador MACD.

Fonte: Baseado em (TRADEVIEW, 2022).

- Divergências

Este sinal ocorre quando o preço e a linha da taxa de variação possuem tendências diferentes, ou seja, enquanto uma sobe a outra desce. Quando existe uma tendência de alta na máxima no gráfico de preço e uma baixa na taxa de variação, um possível fim da tendência de alta do preço pode ocorrer. Já uma tendência de baixa na mínima do preço e uma tendência de alta na mínima da taxa, pode indicar um início de tendência de alta do preço do ativo.

Figura 7 - Representação gráfica das divergências do indicador MACD.

Fonte: Baseado em (TRADEVIEW, 2022).

- Médias móveis

As médias podem ser usadas para testar a virada da tendência, uma vez que o cruzamento da linha zero da taxa de variação pode produzir sinais falsos de alternância de tendência.

Assim, quando a taxa de variação cruzar a linha zero é possível confirmar a tendência de queda, desse modo a média de curto prazo cruza a tendência de deslocamento da média de longo prazo.

Figura 8 - Representação gráfica da estratégia das médias com o indicador MACD.

Fonte: Baseado em (TRADEVIEW, 2022).

2.6 Inteligência Artificial

Esta área é um subcampo da Ciência da Computação e possui diversos campos relacionados a Inteligência Artificial, como por exemplo: o aprendizado de máquina (*machine learning*), a ciência de dados (*data science*) e o aprendizado profundo (*deep learning*) (DATA SCIENCE ACADEMY 3, 2020). O primeiro diz respeito a soluções de IA que são adaptativas, o segundo é um termo genérico recente que envolve diversas subdisciplinas (como matemática, estatística, computação e até mesmo *machine learning*), e o último é um subcampo do aprendizado de máquina onde possui uma maior profundidade do modelo matemático.

A seguir são apresentados alguns conceitos importantes de *machine learning* e métodos do mercado de ações.

Feature

São variáveis individuais e independentes que atuam como entrada no sistema. Os modelos de previsão utilizam este recurso.

Label

É o valor que o modelo tenta prever a partir das *features*. Corresponde a saída final do

modelo.

Na figura 9, a *label* é definida pelo valor do salário (*salary*). Já as *features*, são todas as outras colunas (*position, experience, skill, country, city*) que serão utilizadas para a definição da *label.*

Figura 9 - Exemplo das variáveis.

Features					Label
Position	**Experience**	**Skill**	**Country**	**City**	**Salary ($)**
Developer	0	1	USA	New York	103100
Developer	1	1	USA	New York	104900
Developer	2	1	USA	New York	106800
Developer	3	1	USA	New York	108700
Developer	4	1	USA	New York	110400
Developer	5	1	USA	New York	112300
Developer	6	1	USA	New York	114200
Developer	7	1	USA	New York	116100
Developer	8	1	USA	New York	117800
Developer	9	1	USA	New York	119700
Developer	10	1	USA	New York	121600

Fonte: I2TUTORIAL, 2019.

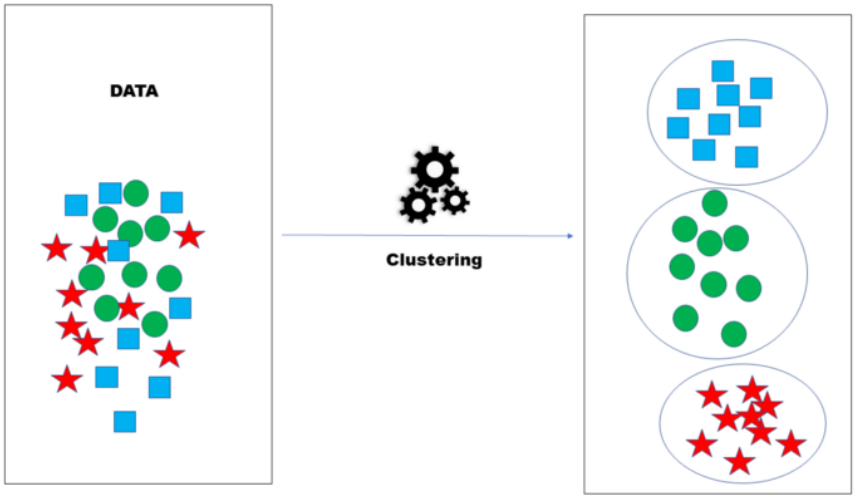
Treinamento não supervisionado

Algoritmos não supervisionados tem como objetivo encontrar padrões sobre um conjunto de dados sem a existência de qualquer tipo de classificação, ou seja, sem classe pré-definida. Assim, a classificação acontece na observação de entradas comuns de um subconjunto. Esta técnica necessita apenas do fornecimento das *features* e o algoritmo irá inferir os agrupamentos sem a utilização de *labels* (DIDATICA TECH, 2022).

A clusterização (*clustering*) é um exemplo de técnica não supervisionada, onde se tenta encontrar padrões e agrupamentos lógicos no conjunto de dados, sendo assim aplicável em diversos casos. Nessa situação, como dito anteriormente, você tem apenas os dados de entrada para buscar padrões e dividir em *clusters*, sem utilizar previsões de saída (*labels*).

A figura 10 mostra de maneira simplificada o resultado de uma clusterização de dados, separando logicamente os grupos dos dados conforme o seu formato.

Figura 10 - Exemplo de clusterização.



Fonte: ICHI PRO, 2020.

Outro exemplo de técnica de aprendizado não supervisionado é o por associação. Esta técnica é muito utilizada atualmente para prever vendas e descontos, analisar mercadorias compradas em conjunto, colocar produtos nas prateleiras e analisar padrões de navegação na *web*. Essa metodologia analisa a sequência de algo e procura encontrar padrões nela, ou seja, tenta-se uma associação e por isso funciona muito bem com os exemplos citados (GRANDO, 2022).

Também, existe o aprendizado por redução de dimensionalidade, conhecido também como generalização. Este método é muito utilizado hoje em dia para criar sistemas de recomendação, criar belas visualizações, modelar tópicos de documentos similares, análise de imagens falsas e gerenciamento de riscos. Assim, a generalização busca encontrar dependências "escondidas", pouco factíveis para a capacidade humana e por isso consegue encontrar, como no caso de recomendações de *marketing*, correlações bastante intrínsecas como: crianças que brincam de Minecraft, normalmente assistem mais desenhos animados (GRANDO, 2022).

Treinamento supervisionado

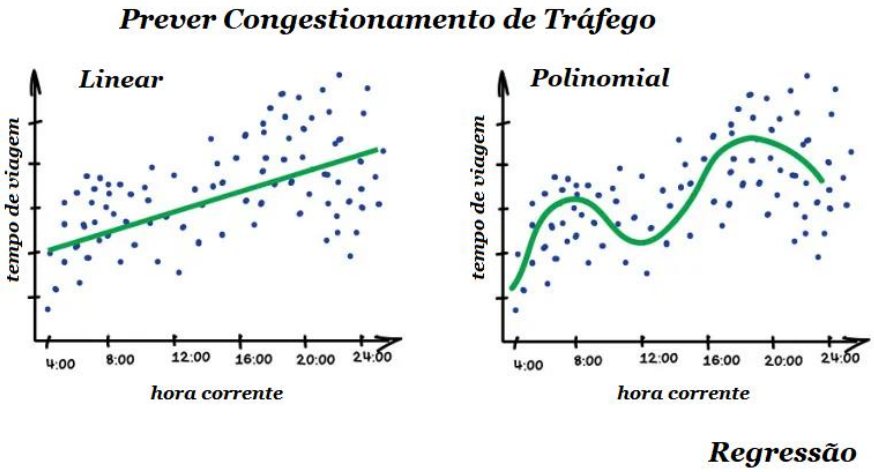
Já o treinamento supervisionado acontece quando o modelo aprende a partir de resultados pré-definidos (*labels*) e com a utilização das *features* (DIDATICA TECH, 2022). Um exemplo é um algoritmo que classifique carros de acordo com seu gasto de combustível. Para o treinamento serão necessários a informação dos dados de consumo de cada modelo e ano dos carros (*features*). Assim, carros que realizam mais de 8km por litro serão classificados como "bom consumo", enquanto aqueles que realizam menos, serão classificados como

“consumo ruim” (*labels*). Ao final do treinamento destes conjuntos o modelo estará preparado para classificar os carros.

Outro tipo de treinamento supervisionado é a regressão, este método busca prever um número, ao invés de uma categoria. Utilizando o exemplo do gasto de combustível citado anteriormente, ao contrário da categorização do carro como “bom consumo” ou “consumo ruim”, o modelo irá buscar prever o gasto dos carros, utilizando dados como o ano do carro, modelo e até a quilometragem para construir sua regressão (GRANDO, 2022).

Deste modo, a regressão irá buscar um padrão no valor de acordo com os dados, podendo resultar em uma linha reta com o método conhecido como regressão linear ou uma curva com o método de regressão polinomial, como pode ser visto na figura 11. Assim, uma equação resultante é gerada e conforme os dados inseridos nela, os valores de previsão do modelo são gerados.

Figura 11 – Exemplos gráficos de regressão linear e regressão polinomial.



Fonte: GRANDO, 2022.

2.7 Random Forest

O método *RandomForest* ou floresta aleatória faz parte dos métodos *ensemble*. Os métodos *ensemble* são construídos do mesmo modo que os algoritmos básicos, como regressão linear e árvore de decisão, entretanto, se diferenciam, pois são feitos através da combinação de diferentes modelos com o objetivo de encontrar um único resultado. (DIDATICA TECH 2, 2022)

O algoritmo foi desenvolvido por Leo Breiman e Adele Cutler (BREIMAN, 2001), e combina a ideia de ‘ensacamento’ de Breiman e a seleção aleatória de recursos introduzidos

por Tin Kam (HO, 1995) e Yali e Donald (AMIT; GEMAN, 1997), construindo um conjunto de árvores de decisão.

Neste método são criadas várias árvores de decisão (*decision trees*), que irão estabelecer regras para a tomada de decisão, ficando com uma estrutura parecida com um fluxograma. O seu modelo base, conhecido como árvore de decisão, quando possui uma árvore de aprendizado profundo, os seus resultados apresentam padrões altamente irregulares, ocasionando um *overfitting* no treinamento, possuindo assim uma variância muito alta.

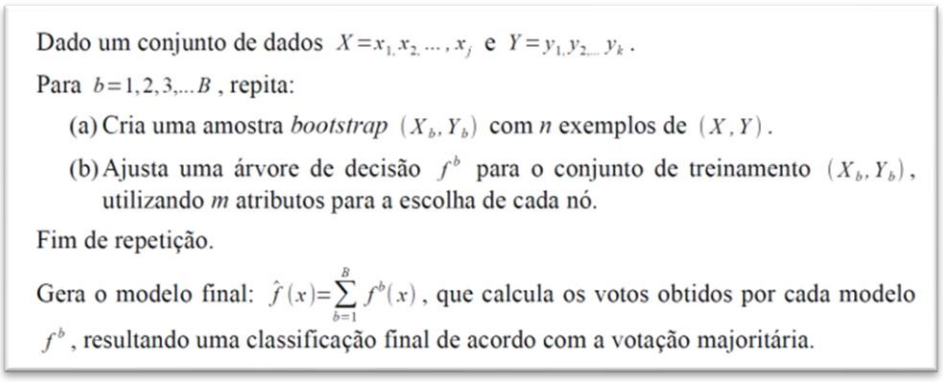
Desta maneira, o *Random Forest*, é uma forma de calcular a média dentre diversas árvores de decisão profundas, reduzindo assim a variância ao treinar diferentes partes do mesmo conjunto em treinamento.

Primeiramente, são selecionadas algumas amostras dos dados de treino, sendo utilizada uma técnica chamada de *bootstrap* para a seleção e assim é construída a primeira árvore de decisão (DIDATICA TECH 2, 2022).

Em seguida, tem-se a seleção aleatória das variáveis para cada nó do fluxo e apesar da possibilidade de se construir árvores com baixo acerto, ela se torna uma estratégia poderosa, pois acaba construindo diversas árvores, podendo evitar um *overfitting* (IBM CLOUD, 2020).

O funcionamento básico do algoritmo em problemas de classificação pode ser visto conforme os passos mostrados na figura 12.

Figura 12 - Algoritmo básico da técnica *Random Forest*.



Fonte: BREIMAN, 2001.

Segundo Breiman (BREIMAN, 2001) dentre as características que favorecem a utilização do algoritmo *Random Forest* tem-se:

- Desempenho satisfatório nas predições de conjuntos de dados com muitos

ruídos.

- Dificulta o *overfitting*, direcionando um classificador que não seja superajustado aos dados treinados.
- Permite o uso em situações com duas classes e em multiclasses.
- Pode ser aplicado em casos em que o número de atributos seja maior do que o de amostras (atributos).
- É possível operar com regressão ou classificação.
- Pode ser aplicado em problemas de grande escala.

2.8 AdaBoost

Outro algoritmo de aprendizado de máquina muito utilizado é o *Adaboost* (*Adaptive Boosting*), desenvolvido por Yoav Freund e Robert Schapire (FREUND; SCHAPIRE, 1996). É um método que aplica *Boosting*, que consiste em treinar preditores sequencialmente, tentando corrigir seu predecessor.

O *Adaboost* melhora um novo preditor ao dar atenção às instâncias onde ocorreram *underfitting* em seu predecessor, assim novos preditores são capazes de cobrirem os erros passados.

Segundo Pedro (AZAMBUJA, 2020), as principais vantagens de se utilizar *Adaboost* está na possibilidade utilizá-lo tanto na regressão quanto na classificação, por lidar bem com diferentes tipos de dados e ser um modelo resistente a *outliers*. Já entre as desvantagens é citado a pouca escalabilidade, possuir risco de *overfitting* e poder ser difícil e demorado para ajustar os hiperparâmetros.

Desta maneira, este algoritmo inicializa o processo com as amostras com pesos iguais, e conforme o algoritmo avança, os pesos das amostras classificadas incorretamente aumentam. Posteriormente, uma árvore de decisão é criada para cada *feature* com profundidade 1 e as árvores que erraram menos são selecionadas para compor a floresta.

Na figura 13, a árvore classifica as pessoas como “em forma” ou “não em forma” baseado no hábito de se exercitarem regularmente. Neste exemplo, a árvore classificou corretamente 3 pessoas como “fora de forma”, classificou também 5 pessoas como “em forma”, entretanto com uma classificada de maneira incorreta.

Figura 13 – Exemplo de um fluxograma para verificar se uma pessoa se exercita regularmente.

Fonte: AZAMBUJA, 2020.

Assim, calcula-se o peso ∝ da árvore (equação (6)) que foi selecionada para a floresta:

$$\propto = \eta \times log \frac{1 - errototal}{errototal} \tag{6}$$

Onde:

η: hiperparâmetro de taxa de aprendizagem (Valor padrão 1).

erro total: soma dos pesos das amostras classificadas incorretamente.

À vista disso, os pesos das amostras são atualizados, sendo que a próxima árvore de decisão usará os erros da árvore anterior. No caso, para classificação errada o peso será aumentado, já para classificações certas o peso será diminuído, de acordo com a equação 7:

$$novospeso = \begin{cases} pesoanterior \times e^{\propto} - se\ classificação\ errada \\ pesoanterior \times e^{-\propto} - se\ classificação\ certa \end{cases} \tag{7}$$

Por fim, um novo *dataset* é criado com o tamanho do original e os passos descritos anteriormente são refeitos até atingir o valor especificado pelo hiperparâmetro.

2.9 Ferramentas Utilizadas

Os algoritmos desenvolvidos neste trabalho foram implementados em Python, que é uma linguagem de programação interpretada, orientada a objetos e de alto nível. A plataforma utilizada para o desenvolvimento do código foi o Google Colaboratory, um serviço em nuvem gratuito hospedado pela Google, com a finalidade de incentivar a pesquisa de aprendizado de máquina e inteligência artificial.

2.9.1 Numpy

Este pacote é uma biblioteca de *Open Source* (código aberto) com a finalidade de realizar operações com *arrays*, que também oferece funções rápidas para tratamento e "limpeza" de dados, geração de subconjuntos e estatísticas descritivas (ver **Anexo 1**).

2.9.2 Pandas

Biblioteca *Open Source* (código aberto) que fornece ferramentas para análise e manipulação de dados. Facilita a leitura, manipulação e exibição dos dados de maneira rápida, flexível e expressiva.

2.9.3 Pandas DataReader

Biblioteca que possibilita a coleta de dados históricos de ações e criptomoedas de diversas bases de dados da internet, como por exemplo, o yahoo finance.

2.9.4 Datetime

Esse módulo permite a manipulação completa de datas e horas, possibilitando a transformação da data para um formato adequado.

2.9.5 Matplotlib

Biblioteca de código aberto que oferece diversas funcionalidades para a criação e manipulação de gráficos, sendo uma alternativa ao MATLAB.

2.9.6 Sklearn (Scikit-learn)

Módulo *open source* desenvolvido para aplicação de *machine learning*, possuindo ferramentas para a análise preditiva de dados.

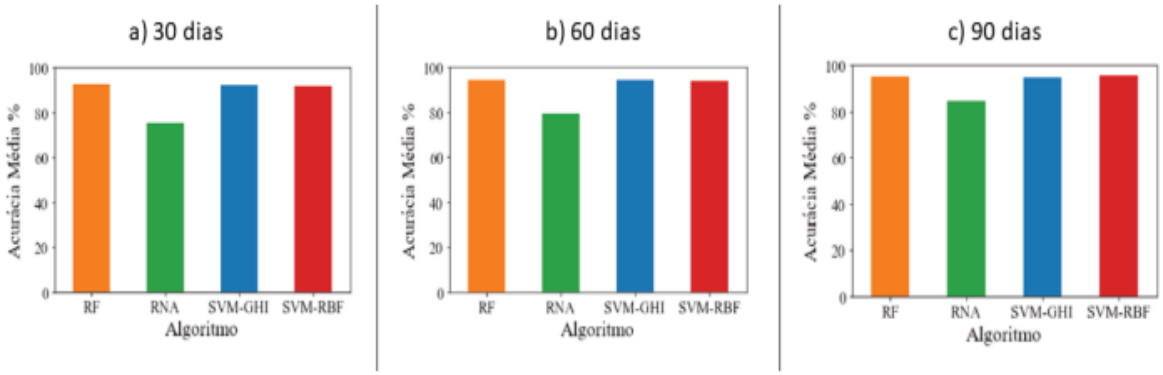
2.10 Inteligência artificial aplicada ao mercado

Inúmeros trabalhos da literatura apresentaram a previsão do movimento das ações da bolsa de valores com a utilização de inteligência artificial. A seguir, será apresentado um resumo dos trabalhos deste tema e que servirão de embasamento teórico para a presente pesquisa.

SANTOS (2020) apresentou a eficiência de quatro algoritmos de aprendizado de máquina destinados à predição das ações da B3, obtendo dados públicos das empresas da Vale (VALE3), Petrobras (PETR4) e Itaú Unibanco (ITUB4). Para isto, ele utilizou três diferentes métodos: *Random Forest*, SVM (*Support Vector Machine*) e RNA (Rede Neural Artificial), comparando também os resultados da predição das ações quando se utiliza divisão aleatória e divisão temporal.

A avaliação de desempenho dos modelos foi dada através da acurácia, ou seja, o desempenho é medido pela razão entre os acertos e todas as previsões (ver **seção 2.11**). O método SVM, possuiu mais de 96% de acurácia para as previsões de 60 e 90 dias como poder ser visto na figura 14.

Figura 14 - Resultados dos modelos utilizados



Fonte: SANTOS, 2020.

Também, o método *Random Forest* foi o que demonstrou maior acurácia dentre as previsões de 30 dias. No treinamento dos modelos o autor utilizou os indicadores técnicos de

índice de força relativa (RSI), média móvel exponencial, Williams %R, divergência e convergência da média móvel, variação de volume, oscilador estocástico e taxa de variação de preço.

ALCANTARA, *et al* (2018) apresentaram um estudo das técnicas de inteligência artificial para a análise temporal das ações. O algoritmo utilizado foi a Rede Neural Artificial, realizando, assim, uma aplicação em ambiente *web* para investimentos *online*. A interface da aplicação foi construída com o Electron e o React JS foi utilizado para o desenvolvimento das telas *web* e o MongoDB para o armazenamento dos dados. Segundo Alcantara este estudo serve de base para a criação de um algoritmo para investimento, pois demonstra de forma concisa o que é necessário para a construção. O trabalho também cita que as redes neurais se aplicam bem ao estudo devido a análise temporal dos dados das ações.

MITRE (2021) construiu um algoritmo com *Random Forest* para análise de tendências da bolsa de valores com gráfico *intraday* (gráfico de negociação com escala de 1 dia). Os resultados do modelo mostraram ganhos maiores do que *buy and hold* no período testado, chegando à 99,20% do valor investido com o método, enquanto a ação valorizou 24,34%. Assim, dentre 8 ativos analisados, em 7 o modelo superou a valorização da ação no período, demonstrando a consistência do método.

Diversas outras pesquisas foram realizadas com a finalidade da previsão de diferentes dados financeiros, como por exemplo, GUIMARÃES (2021), que aplicou o algoritmo C4.5 de árvores de classificação para auxiliar a previsão de *defaults* da dívida pública de 66 países, prevendo as crises de 1980 e de 1990. Guimarães concluiu que os algoritmos de aprendizado podem ser ferramentas promissoras, para quando não conseguirem prever crises, ao menos para detectarem sinais de crise. Também, citou que devido a natureza simples destes algoritmos, não se deve delegar a eles decisões estratégicas, mas sim como auxílio à tomada de decisão.

SANTOS (2017), construiu modelos de previsão baseados em processamento de linguagem natural e *machine learning* (*Random Forest e Support Vector Classifier)* para as mudanças da meta da taxa Selic que é decidida pelo Copom (Comitê de Política Monetária). A acurácia média do algoritmo foi de 83,3%, entretanto, segundo Paulo Santos, foi possível notar que quando a amostra de treino era reduzida, a acurácia caia consideravelmente para os modelos com Random Forest e *Support Vector Classifier*.

QUARESMA SANTOS (2022) procurou avaliar a aplicação de algoritmos de *machine learning* na elaboração de projeções da dívida ativa da União, onde a cada ano fiscal, previu-se o ano seguinte. Foram testados os algoritmos Regressão Linear, Árvore de Decisão, Floresta Randômica e Árvore de Decisão de Aumento de Gradiente, que foram “alimentados” por

indicadores macroeconômicos como IPCA, IGP-M, PIB, taxa de Câmbio e taxa SELIC, além de dados relativos a parcelamentos excepcionais e transações tributárias disponibilizadas aos contribuintes pela Fazenda Nacional. Rubens Quaresma concluiu em seu trabalho, que a árvore de decisão e a árvore de decisão com aumento de gradiente, obtiveram sucesso ao prever as projeções da dívida ativa da União.

2.11 Avaliação de desempenho

Diversas métricas de avaliação são utilizadas para avaliar os modelos de aprendizado de máquina. Existem métricas para modelos de regressão e para modelos de classificação. Como este trabalho utiliza modelos de classificação, as seguintes métricas são utilizadas:

- Acurácia (*Accuracy*)

Demonstra quantos dados foram classificados corretamente, independente da classe. Por exemplo, analisando os dados de uma ação por um período de 100 dias e aplicando um modelo de IA classificando assim 80 dias de forma correta, ou seja, acertando 80 vezes se a ação iria subir ou não, possuindo assim uma acurácia de 80%. A equação (8) demonstra o cálculo deste avaliador de desempenho.

$$Acurácia = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{8}$$

Onde,

t_p = Número de previsões positivas classificadas corretamente;

t_n = Número de previsões negativas classificadas corretamente;

f_p = Número de previsões positivas classificadas incorretamente;

f_n = Número de previsões negativas classificadas incorretamente;

Analisando a equação (7), a métrica é definida pela razão entre os acertos e todas as previsões.

- Precisão (*precision*)

Indica a razão entre a quantidade de exemplos classificados corretamente como positivos e o total de exemplos classificados como positivos. Assim, como mostra a equação 9, esta métrica avalia dos classificados positivos quantos realmente são positivos, dando ênfase aos falsos positivos (classificar como positivo, mas o correto seria negativo).

$$Precisão = \frac{t_p}{t_p + f_p} \tag{9}$$

- Revocação (*Recall*)

Parecido com a precisão, a revocação exprime a quantidade de exemplos positivos que foram classificados corretamente como positivos, dando assim, maior ênfase aos falsos negativos (classificar como negativo, mas o correto seria positivo). A equação 10 demonstra o seu cálculo.

$$Recall = \frac{t_p}{t_p + f_n} \tag{10}$$

- *F1 Score* (ou *F-measure*)

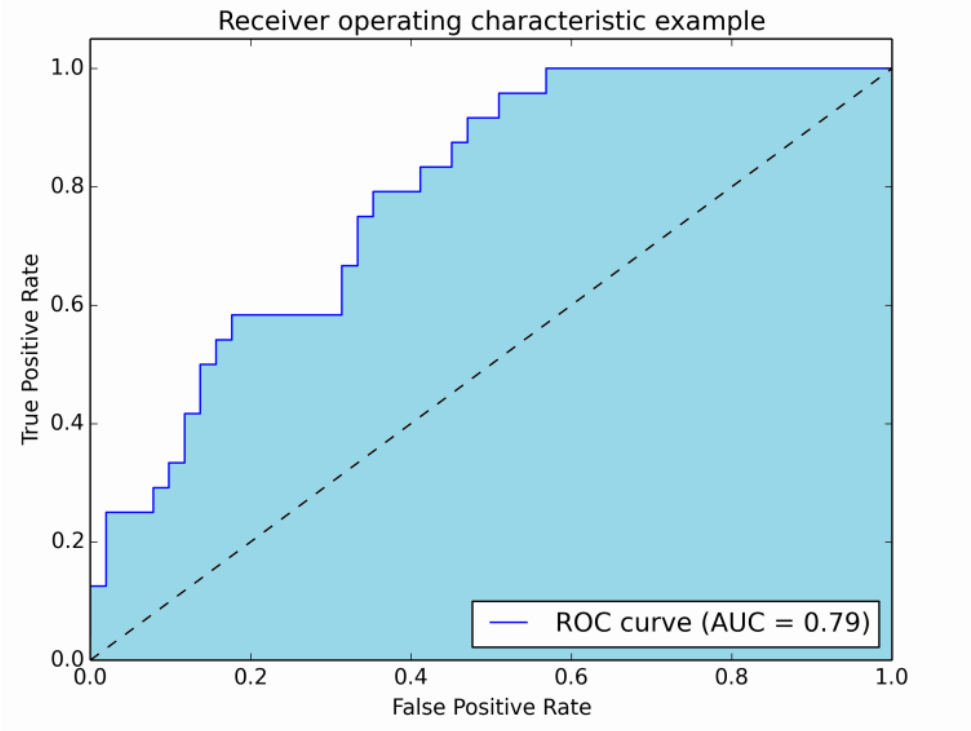
Mostra a harmônica entre a precisão e a revocação (*recall*), como mostra a equação 11. Assim, essa métrica atua como um resumo da qualidade do modelo, juntando as duas métricas relacionáveis.

$$F1\ Score = 2 \times \frac{precisão \times recall}{precisão + recall} \tag{11}$$

- Curva *ROC*

A curva *ROC* (*Receiver Operating Characteristic*), é utilizada para avaliar a *performance* de um classificador, medindo a taxa de falso positivo e a taxa de verdadeiro positivo, de acordo com diversos valores limites, como pode ser visto na figura 15.

Figura 15 - Representação de uma Curva *ROC*.



Fonte: RODRIGUES, 2018.

A linha tracejada na posição de 45 graus do espaço *ROC* representa a curva de um classificador que prevê de forma aleatória. Quanto mais a curva se aproximar do canto superior esquerdo e quanto maior a área sob curva, melhor será o modelo, ou seja, quanto mais se aproxima da diagonal tracejada menos preciso o teste será.

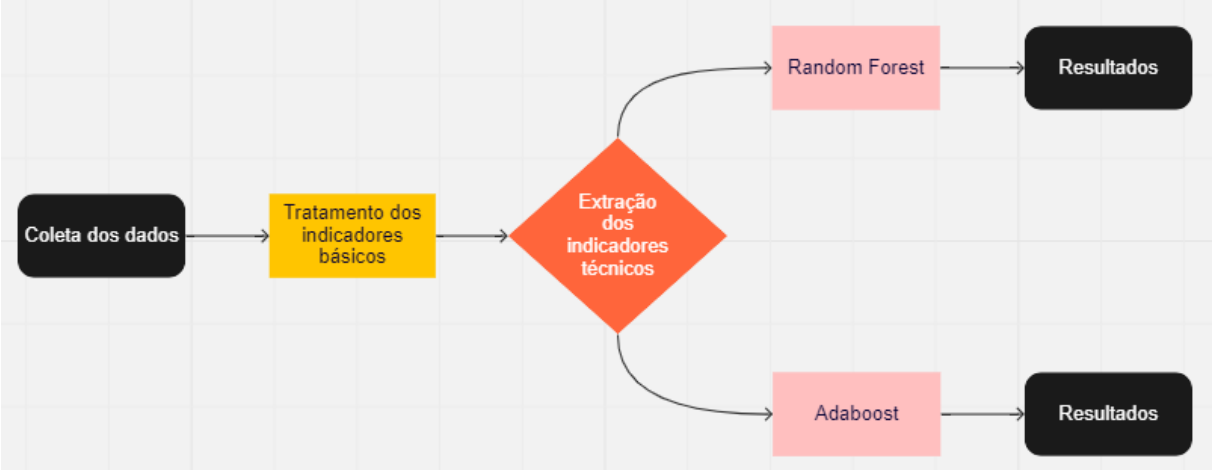
A curva *ROC* também demonstra a especificidade e a sensibilidade do teste, a primeira está relacionada a capacidade do teste detectar corretamente resultados positivos, já a segunda, de detectar corretamente resultados negativos.

3. METODOLOGIA

3.1 Introdução

Para desenvolver esse trabalho, inicialmente realizou-se a coleta dos dados, para posteriormente selecionar os atributos, ou seja, os indicadores básicos apresentados na **seção 2.4** foram transformados em indicadores técnicos (**seção 2.5**). Os dados públicos utilizados são de diversas cotações listadas na bolsa de valores. A figura 16 apresenta o fluxograma da forma como os modelos foram desenvolvidos.

Figura 16 - Fluxograma dos modelos desenvolvidos.



Fonte: Autoria própria

3.2 Coleta e tratamento dos indicadores básicos

Assim, o projeto iniciou-se com os dados sendo captados manualmente através do *site* oficial da B3, no caso de uma ação brasileira e para uma ação americana, no *site* oficial da NASDAQ ou NYSE. Na sequência, os arquivos com extensão .CSV (bloco de notas) foram gerados com os preços das ações (figura 17) de acordo com os meses solicitados e realizou-se o *upload* destes arquivos no Google Colab.

Figura 17 - Exemplo da representação dos dados do bloco de dados.

```
"Data","Último","Abertura","Máxima","Mínima","Vol.","Var%"
"01.07.2022","74,73","75,51","76,50","73,81","17,44M","-2,39%"
"30.06.2022","76,56","76,99","78,05","76,17","27,33M","-2,83%"
"29.06.2022","78,79","79,61","80,09","78,04","18,59M","-0,83%"
```

Fonte: Autoria própria.

Desta maneira, foi necessário que os dados fossem manipulados para se adequar a criação dos indicadores e para realizar os cálculos necessários. Primeiro, com a biblioteca Pandas, transformou-se os dados dos dias para o formato *datetime* e as vírgulas de todas as variáveis foram trocadas por pontos. Posteriormente, retirou-se o símbolo de porcentagem da variável de variância e a letra 'M' dos números que representam o volume que continham nestes bancos de dados.

Este método despendeu bastante tempo, pois para alterar a ação foi necessário trocar o arquivo e sempre que se reiniciou o uso do Google Colab, como é uma plataforma que salva temporariamente o banco de dados, perdia-se os dados em utilização.

Deste modo, durante o desenvolvimento do trabalho um método mais fácil e prático para coletar os dados foi implementado, pois além de possuir os formatos necessários para os cálculos, a troca do ativo a ser utilizado foi simplificado.

Assim, os dados utilizados para o treinamento e teste do modelo foram coletados utilizando a biblioteca *pandas datareader* (ver **Anexo 2**). Foram coletados os dados através da Yahoo Finance, uma plataforma midiática que fornece notícias financeiras, dados e comentários públicos, incluindo cotações de ações e relatórios financeiros.

Os campos recebidos pela extração com a biblioteca *pandas datareader* estavam em formato único e pertenciam a um período diferente. De acordo com a ação selecionada, os intervalos do período foram dados em dias.

Os dados recebidos neste método e representados na Tabela 1:

- Preço de abertura do dia (*Open*);
- Preço máximo do dia (*High*);
- Preço mínimo do dia (*Low*);
- Preço de fechamento do dia (*Close*);
- Preço de fechamento ajustado (*Ajusted Close*);
- Volume negociado no dia (*Volume*);
- Data do dia (*Date*).

Tabela 1 - Representação da tabela com os dados utilizados.

Date	High	Low	Open	Close	Volume	Adj Close
2014-01-02	15.897821	15.622339	15.797645	15.787628	25333788.0	11.079877
2014-01-03	15.917856	15.682444	15.777610	15.757575	40048592.0	11.058788
2014-01-06	15.877786	15.682444	15.717505	15.782619	25543620.0	11.076364
2014-01-07	15.942900	15.607312	15.767593	15.662409	25317616.0	10.992000
2014-01-08	15.907838	15.702479	15.702479	15.907838	21595142.0	11.164240
2014-01-09	15.822689	15.401953	15.802654	15.411970	30000606.0	10.816237

Fonte: Autoria própria.

Verificou-se também, que os tipos das variáveis que o método da biblioteca *datareader* retornou o Yahoo Finance, diferentemente do método antigo de coleta, todas as variáveis são float64, já no método antigo eram *objects*, sendo necessário realizar uma conversão para float64.

3.3 Construção dos Indicadores Técnicos

Agora para a criação dos indicadores técnicos utiliza-se um método manual de cálculo e um método automático com a biblioteca TALIB, uma biblioteca para análise técnica do mercado financeiro, este último método foi utilizado para verificar o valor encontrado manualmente.

Desta maneira, no Google Colab, plataforma onde o código do modelo é programado, foi realizada a instalação da biblioteca seguindo o passo a passo encontrado em um *notebook* público do Google Colab (COLAB, 2022).

Calculou-se todos os indicadores de acordo com as equações (1), (2), (3), (4) e (5). Assim, este método manual consistiu em um maior desafio e aprendizado devido ao cálculo ser realizado apenas com as funções matemáticas do Python.

3.4 Utilização dos Indicadores Técnicos

Agora, com os indicadores técnicos, criou-se as variáveis utilizadas nos modelos (ver **Anexo 3**), com objetivo de demonstrar os sinais de compra e venda dos ativos. Diversas alterações e randomizações da combinação dos indicadores foram necessárias para atingir o melhor resultado e as melhores previsões.

- **Média móvel de 10 e 30 dias**

O primeiro indicador criado é um sinal de venda ou compra de quando a média móvel de 10 dias for maior que a média móvel de 30 dias. O número 1 é colocado para os dias em que o sinal é positivo e -1 para os dias em que o sinal é negativo.

- **Média móvel acima de 10 dias**

Sinal de compra e venda conforme o preço for maior que a média móvel de 10 dias. Colocando o número 1 nos dias com sinal positivo e -1 nos dias de sinal negativo. Foram criadas variáveis das médias de 5, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150 e 200 dias, totalizando 14 variáveis.

- **Compra e venda indicados pelo RSI**

Duas variáveis com sinais de compra e venda de acordo com os valores encontrados do indicador técnico RSI. A variável '*OverboughtRSI'* avalia se o ativo está sobrevalorizado, já a variável '*OversoldRSI'*, avalia se o ativo está subvalorizado.

- **Compra e venda indicados pela Williams%R**

Assim como as variáveis anteriores, de acordo com o valor do indicador Williams%R, um sinal de sobrevalorização ou subvalorização é criado.

- **Divergência e Convergência (MACD) e Taxa de variação (PROC)**

Os valores da Convergência/Divergência da média móvel e da Taxa de variação também foram utilizados no modelo. Calculou-se a taxa de variação através da equação (5).

3.5 Divisão do *Dataset* e a variável de previsão

Inicialmente, dividiu-se o *dataset* de maneira randomizada. Entretanto, para o caso das séries temporais, uma divisão sem randomizar é o ideal, pois na divisão aleatória o modelo

utiliza os dados ‘futuros’ para previsão, quebrando a importante linha do tempo do histórico da ação.

Assim, a função de divisão do *dataset* foi escolhida sem randomizar, sendo os últimos 30% dos dados utilizados para a previsão (teste) e os outros 70% para o treino do modelo. Ao realizar cáclulos com 20% e 80% e 25% e 75%, verificou-se que a porcentagem escolhida (30% e 70%) obteve o melhor desempenho (ver **Anexo 4**).

A princípio, a variável de previsão era a subida ou a descida da ação, ou seja, quando a ação possuía uma variação positiva comparada ao dia anterior, o número 1 era colocado na variável ‘class’ do dia anterior, já quando era negativo, era colocado o número 0.

Entretanto, conforme variava-se os indicadores utilizados e buscava-se um resultado considerável, a precisão não ultrapassava dos 55% e possuía uma curva ROC ruim (avaliação da curva explicada na **seção 2.11**). Considerando que a probabilidade de acerto de uma ação subir ou não é de 50% por ser uma classificação binária, um modelo que combina uma curva ROC ruim e uma precisão próxima aos 50% é um modelo que não está possuindo aprendizado.

Desta maneira, uma forma encontrada para obter resultados com maior precisão na previsão dos dados foi o modelo parar de prever a subida ou descida da ação e começar a avaliar a tendência da ação. Assim, uma nova variável foi criada, indicando a tendência de baixa ou alta com os valores 0 ou 1. Esta variável comparou o preço do dia analisado com a média dos últimos 150 dias, ou seja, caso o preço atual estiver abaixo da média dos últimos 150 dias, a ação está em uma tendência de baixa (0) e caso o contrário, ela está em tendência de alta (1).

3.6 Escolha dos períodos das ações para aplicar ao modelo

Em princípio, escolheu-se períodos econômicos sem crises para aplicar no modelo, cada um de acordo com sua região (Estados Unidos e Brasil), ou seja, procurando evitar momentos de grandes interferências externas nas ações.

Entretanto, notou-se que o modelo possui um grande enviesamento dos resultados visto que a acurácia de todas as ações é extremamente alta e muitas vezes a curva ROC obteve um valor ruim (AUC abaixo de 0,6), demonstrando *overfitting* mesmo na utilização de modelos que reduzem a chance de isso acontecer.

Constatou que isso ocorre devido ao grande desbalanceamento dos dados das ações, ou seja, eram períodos em que as tendências de alta eram bem maiores que as tendências de queda, desta maneira o modelo possui dificuldade em “entender” os momentos de queda, possuindo grande aprendizado em tendências de alta, e baixo aprendizado em tendências de baixa. Outro ponto, era a baixa quantidade de dados aplicada ao modelo, que em alguns casos, ocasionava uma baixa acurácia.

Assim, com a finalidade de resolver este problema, o período de cada ação foi selecionado buscando maximizar a quantidade de dados com a finalidade de facilitar o encontro de padrões pelo modelo. O período com o melhor balanceamento dos dados para evitar o ‘*overfitting*’ também foi escolhido. Como utilizava-se dados reais e séries temporais, não foi possível balancear com perfeição todas as ações devido a necessidade de capturar um período contínuo. Desta maneira, foi escolhido o melhor balanceamento encontrado na ação escolhida, conforme a Tabelas 2 e a Tabela 3.

Tabela 2 - Número de dias em tendência de alta ou baixa e as datas inseridas para coleta dos dados de acordo com cada ação norte-americana.

	TSLA	**AAPL**	**NFLX**	**AMZN**	**TWTR**	**KO**
Número de dias com Tendência de Alta (Y=1)	460	3048	690	1268	1017	386
Número de dias com Tendência de Baixa (Y=0)	533	4004	961	1645	920	607
Data Inicial	01/01/2014	01/01/1981	01/01/2002	01/01/1997	01/01/2015	01/01/2014
Data Final	01/01/2018	01/01/2009	01/01/2009	01/01/2009	01/10/2022	01/01/2018

Fonte: Autoria própria.

Tabela 3 - Número de dias em tendência de alta ou baixa e as datas inseridas para coleta dos dados de acordo com cada ação brasileira.

	ITUB4	**VALE3**	**ABEV3**	**WEGE3**	**CIEL3**	**BRFS3**
Número de dias com Tendência de Alta (Y=1)	349	952	813	893	656	1633
Número de dias com Tendência de Baixa (Y=0)	634	1209	1476	1554	1190	1911
Data Inicial	01/01/2014	01/01/2014	01/01/2000	01/01/1997	01/01/2015	01/01/2014
Data Final	01/01/2019	01/10/2022	01/01/2009	01/01/2009	01/10/2022	01/01/2018

Fonte: Autoria própria.

Nas Tabelas 2 e 3, encontram-se a divisão dos dados de cada ação. A seguir, tem-se a estratégia adotada para algumas ações, e para as demais, apenas foi escolhido pelo balanceamento dos dados, outros motivos aparentes não foram encontrados.

Tesla (TSLA): Com ações desde 2010, Tesla possuiu grande crescimento no ano de 2013 e no final de 2019 em diante. Assim para excluir estes períodos de grande crescimento que podem enviesar o modelo utilizou-se os dados de 2014 até 2017 onde a empresa possuiu uma boa distribuição entre crescimentos e quedas, favorecendo a utilização dos indicadores do modelo.

Apple (AAPL): Com ações desde 1981, as ações da AAPL tiveram um grande crescimento após a crise de 2008, a ação disparou de 5 dólares em 2008 para 180 dólares em 2021, possuindo uma contínua tendência de alta, assim, escolheu-se os dados de 1981 até 2008, onde o balanceamento é bem maior.

Netflix (NFLX): Ações abertas desde 2002, assim como outras grandes empresas do mercado de 2022 ela possuiu crescimentos exponenciais, esta ação possuiu grande crescimento após a crise econômica desde 2008 e após 2019, além de uma queda de 75% entre 2021 e 2022.

Desta maneira, o período de 2002 até 2008, por ser um período de crescimento, mas contido, foi utilizado para testar e treinar o modelo.

Amazon (AMZN): Desde 1997 com ações na bolsa, as ações da Amazon tiveram um grande crescimento em 1998, mas uma queda proporcional em 2000, assim, igualmente as ações anteriores, após 2008 obteve um crescimento acentuado. Assim sendo, de 1997 até 2008 foi o período de análise.

Coca Cola. (KO): A combinação que atingiu o melhor desempenho foi do período de 2014 até o final de 2017, período escolhido devido ao seu momento de ciclos de altas e quedas equilibradas, possuindo um crescimento contido. Após 2008 a ação passou por um grande crescimento, até estabilizar-se próximo de 2014 e também possuindo um grande crescimento após 2018.

Itaú Unibanco Holdings S. A. (ITUB4.SA): Com ações desde 2000 nos dados da *Yahoo Finance*, a combinação que atingiu o melhor desempenho foi do período de 2014 até o final de 2018.

4. RESULTADOS e DISCUSSÕES

Este capítulo apresenta os resultados obtidos pelas técnicas empregadas para avaliar os modelos construídos (ver **Anexo 5**). Desta maneira, para cada modelo foram efetuados diversos treinos e testes com diferentes ações da bolsa brasileira (B3) e a bolsa americana (NYSE e NASDAQ) com a finalidade de comparação. O desempenho dos algoritmos é avaliado segundo as métricas discutidas na seção anterior.

4.1 Resultados com *Random Forest* (RF)

Para os ativos da bolsa de valores norte-americana o resultado da acurácia média do modelo *Random Forest* foi de 0,34% maior que para a acurácia dos ativos na bolsa brasileira com o mesmo modelo.

Nas Tabelas 4 e 5, tem-se os desempenhos do modelo RF nos ativos brasileiros e norte-americanos.

Tabela 4 - Desempenho do modelo RF nos **ativos brasileiros** do *dataset*.

Ativo	Acurácia	Precisão	Recall	F1 Score
ITUB4	64,06	0,61	0,64	0,62
VALE3	65,94	0,66	0,66	0,66
WEGE3	66,80	0,66	0,67	0,65
ABEV3	77,29	0,77	0,77	0,77
CIEL3	55,41	0,58	0,55	0,52
BRFS3	53,95	0,65	0,54	0,55
Média	**63,91**			

Fonte: O autor.

Tabela 5 - Desempenho do modelo RF nos **ativos norte-americanos** do *dataset* com a comparação da acurácia média dos ativos brasileiros.

Ativo	Acurácia	Precisão	Recall	F1 Score
TSLA	63,42	0,75	0,63	0,65
AAPL	63,42	0,71	0,74	0,72
NFLX	57,66	0,61	0,64	0,59
AMZN	67,39	0,67	0,67	0,67
TWTR	60,82	0,60	0,60	0,60
KO	72,81	0,72	0,73	0,72
Média	**64,25** (+0,34)			

Fonte: O autor.

Os ativos possuem uma variação considerável da acurácia, principalmente nos ativos brasileiros, atingindo desde 53,95% até 77,29% (BRFS3 e ABEV3), o que era esperado devido as diferenças do comportamento das diferentes ações.

Nas ações norte americanas os resultados têm menor variação quando comparado com a acurácia média dos ativos. A menor acurácia ficou em 57,66% (NFLX) e a máxima em 72,81% (KO), também todos os ativos se comportaram mais próximos a média.

Nos ativos brasileiros, duas ações (CIEL3 e BRFS3) se aproximaram de 50%, valor referencial para se acertar a tendência de subida ou descida de uma ação.

Os resultados mostram ainda, que as métricas de precisão, *recall* e *F1 score*, possuem valores que acompanharam a taxa de acurácia da ação e não apresentaram nenhum valor *outlier* (extremamente diferente dos demais), confirmando assim os valores da acurácia do modelo.

Por fim, os resultados do *Random Forest* indicam que o modelo possuiu aprendizado com as técnicas e dados adotados no modelo. Isso é verificado pela acurácia acima de 50% e valores semelhantes em precisão, *recall* e *F1 score*.

4.2 Resultados com *Adaboost*

Para os ativos da bolsa de valores brasileira o resultado da acurácia do modelo *Adaboost* foi 0,02% maior que para a acurácia dos ativos na bolsa norte americana com o mesmo modelo.

As Tabelas 6 e 7, apresentam os desempenhos do modelo *Adaboost* nos ativos brasileiro e norte-americanos.

Tabela 6 - Desempenho do modelo *Adaboost* nos ativos brasileiros comparando **acurácia** com o modelo RF.

Ativo	Acurácia	Precisão	Recall	F1 Score
ITUB4	61,79 (-2,27)	0,58	0,62	0,59
VALE3	62,39 (-3,55)	0,65	0,62	0,63
WEGE3	68,29 (+1,49)	0,68	0,68	0,67
ABEV3	78,02 (+0,73)	0,78	0,78	0,78
CIEL3	57,94 (+2,53)	0,60	0,58	0,57
BRFS3	52,72 (-1,23)	0,63	0,53	0,54
Média	**63,52** (-0,39)			

Fonte: O autor.

Tabela 7 - Desempenho do modelo *Adaboost* nos ativos norte-americanos comparando a **acurácia** com o modelo RF.

Ativo	Acurácia	Precisão	Recall	F1 Score
TSLA	59,06 (-4,36)	0,71	0,59	0,61
AAPL	67,43 (+4,01)	0,66	0,67	0,66
NFLX	58,26 (-0,60)	0,58	0,58	0,58
AMZN	70,59 (+3,21)	0,71	0,71	0,70
TWTR	57,21 (-3,61)	0,60	0,59	0,58
KO	68,45 (-4,36)	0,71	0,68	0,69
Média	**63,50** (-0,75)			

Fonte: O autor.

Assim como no *Random Forest*, os ativos possuem uma variação considerável da acurácia, principalmente nos ativos brasileiro, atingindo desde 52,72% até 78,02% (mesmos ativos do modelo *Random Forest*, BRFS3 e ABEV3).

Da mesma forma, nas ações norte americana, os resultados têm menor variação quando comparado com a acurácia média dos ativos. A menor acurácia ficou em 57,21% (TWTR) e a máxima em 68,45% (KO), sendo que todos os ativos se comportaram mais próximos a média.

Os ativos WEGE3, ABEV3, CIEL3, possuem um leve ganho na porcentagem da acurácia quando comparado com o resultado do modelo RF (em torno de 2%) e os ativos AAPL e AMZN têm um ganho considerável (em torno de 4%).

Já as ações ITUB4, BRFS3, NFLX, apresentam leve queda na acurácia, enquanto VALE3, TSLA, TWTR e KO, alcançaram quedas mais expressivas na acurácia (em torno de 4%).

Observando agora as médias do modelo de *Adaboos*t, a variação perante o modelo *Random Forest* foi pouco expressiva, com quedas de 0,39% para ativos brasileiros e 0,75% para norte-americanos.

Esse modelo indica um resultado próximo ao *Random Forest* e demonstra que o modelo obteve um aprendizado considerável ao possuir acurácia média acima de 50% e precisão, *recall* e *F1 score* de acordo com a taxa de acerto (acurácia*).* O aprendizado dos modelos será confirmado novamente na análise da curva ROC dos ativos na próxima seção (**Seção 4.3**).

Assim como no modelo de *Random Forest*, ambas as regiões dos ativos os valores da precisão, *recall* e *F1 score* apresentam valores que acompanharam a taxa de acurácia da ação e não apresentaram nenhum valor *outlier*, indicando assim, que ambos os modelos podem ser utilizados.

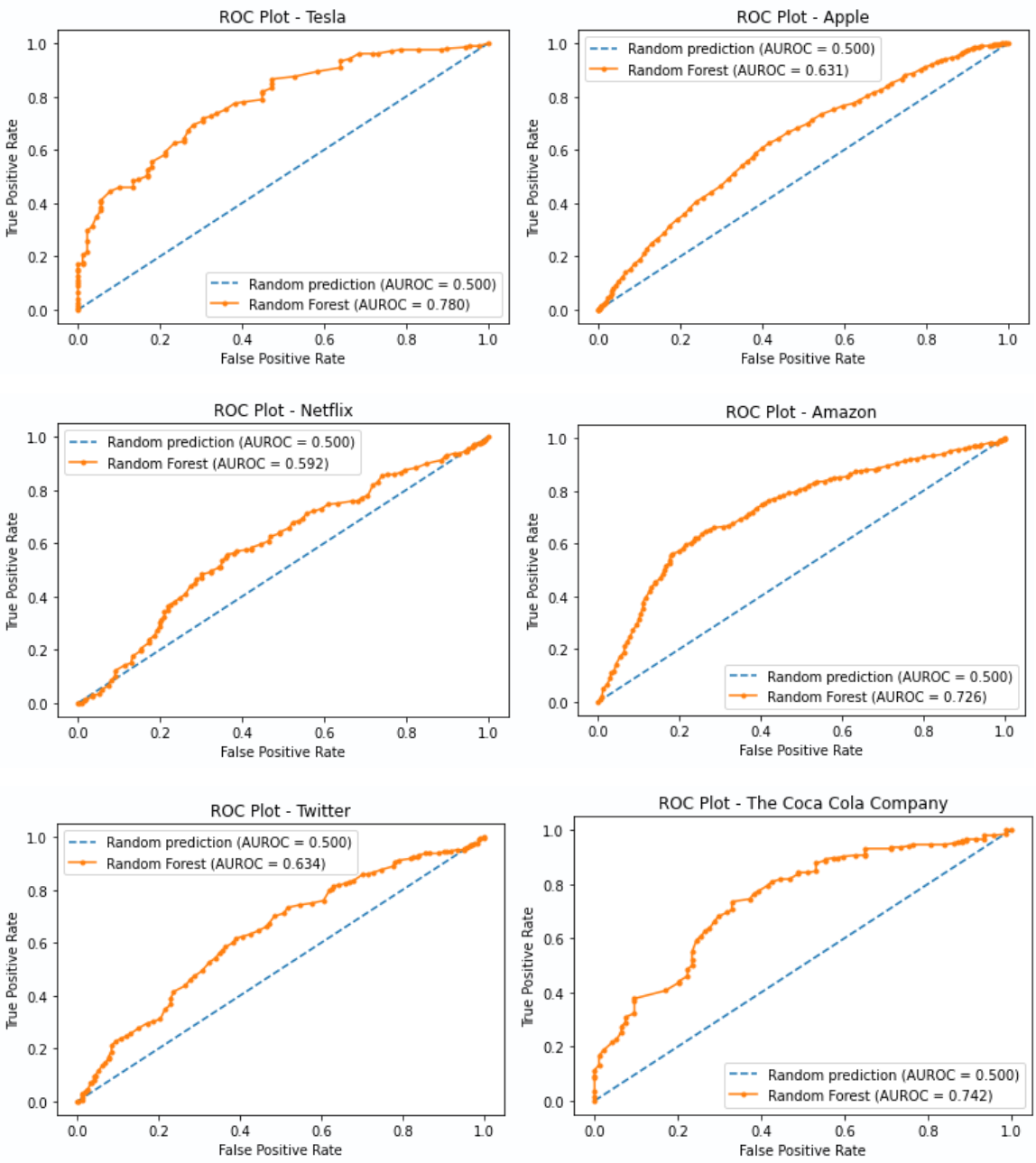
4.3 Curva ROC

A curva ROC é utilizada para confirmar a veracidade da acurácia calculada, uma vez que uma acurácia alta, porém com uma curva ROC ruim, indica o não aprendizado do modelo. A seguir, resultados dos cálculos da curva ROC são apresentados.

Random Forest – ativos Norte Americanos

Utilizando o método *Random Forest* nos ativos Norte Americanos, a média da área sobre a curva AUROC foi de 0,684. Os gráficos e valores encontram-se na figura 18.

Figura 18 - Gráficos da curva ROC para os **ativos norte-americanos** (*Random Forest*).



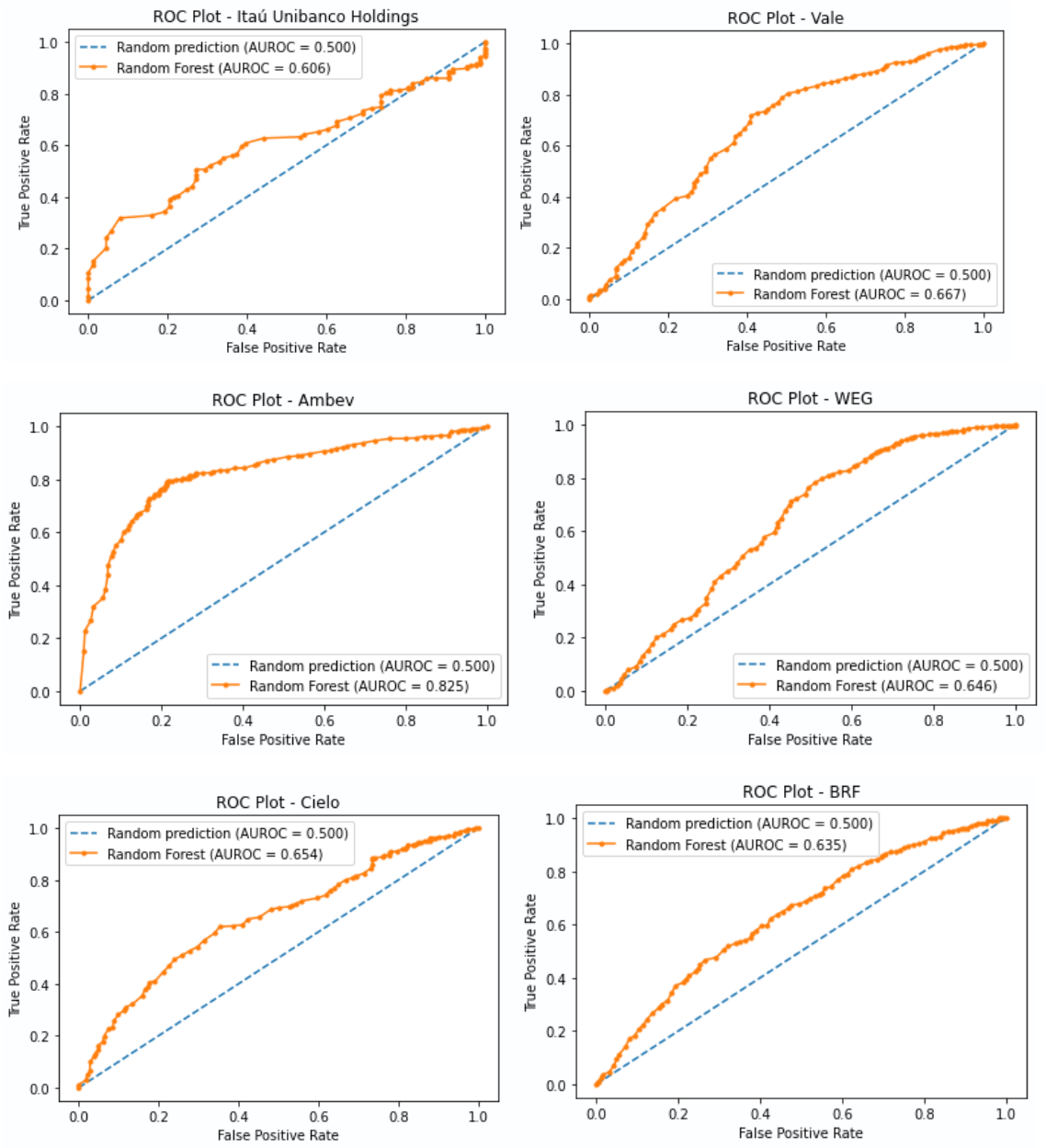
Fonte: Autoria própria.

Neste caso, a ação da TSLA apresentou uma curva ROC um pouco acima do esperado, isso é previsto em alguns resultados devido a aleatoriedade que envolve um modelo de *machine learning*.

Random Forest – ativos Brasileiros

Empregando o método *Random Forest* nos ativos Brasileiros, a média da área sobre a curva AUROC foi de 0,672. Os gráficos e valores são apresentados na figura 19.

Figura 19 - Gráficos da curva ROC para os **ativos brasileiros** (*Random Forest*).



Fonte: Autoria própria.

Itaú Unibanco Holdings possuiu o final de sua curva abaixo da reta de referência, demonstrando uma dificuldade para obter altas taxas de verdadeiros positivos em altas taxas de falsos positivos.

As ações da WEG também possuem uma queda na curva, porém no seu início, demonstrando uma maior dificuldade em obter taxas maiores de verdadeiro positivo em baixas taxas de falso positivo.

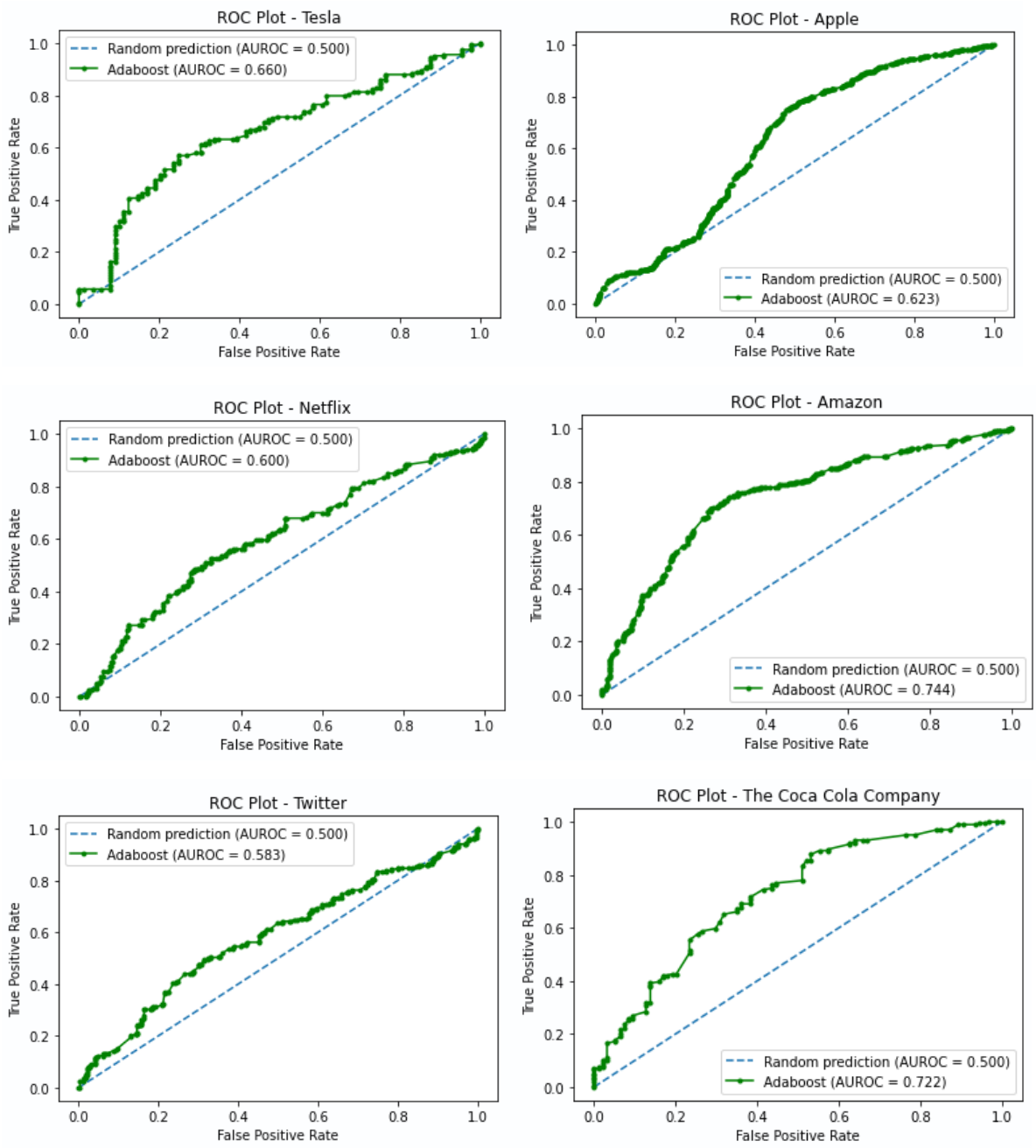
Já as ações BRFS3 e CIEL3 apesar de possuírem uma acurácia mais baixa, as curvas se aproximaram de curvas de ações com maior acurácia, demonstrando o aprendizado do modelo, apesar da menor acurácia.

Apesar destes pequenos detalhes, as suas respectivas curvas apresentaram uma certa característica de curva quando se observa toda a sua trajetória. Assim, estes e os demais ativos apresentam características normais da curva ROC e confirmam o aprendizado do modelo.

Adaboost – ativos Norte Americanos

Aplicando o método *Adaboost* nos ativos Norte Americanos, a média da área sobre a curva AUROC foi de 0,655, os gráficos e os valores encontram-se na figura 20.

Figura 20 - Gráficos da curva ROC para os ativos norte-americanos com *Adaboost*.



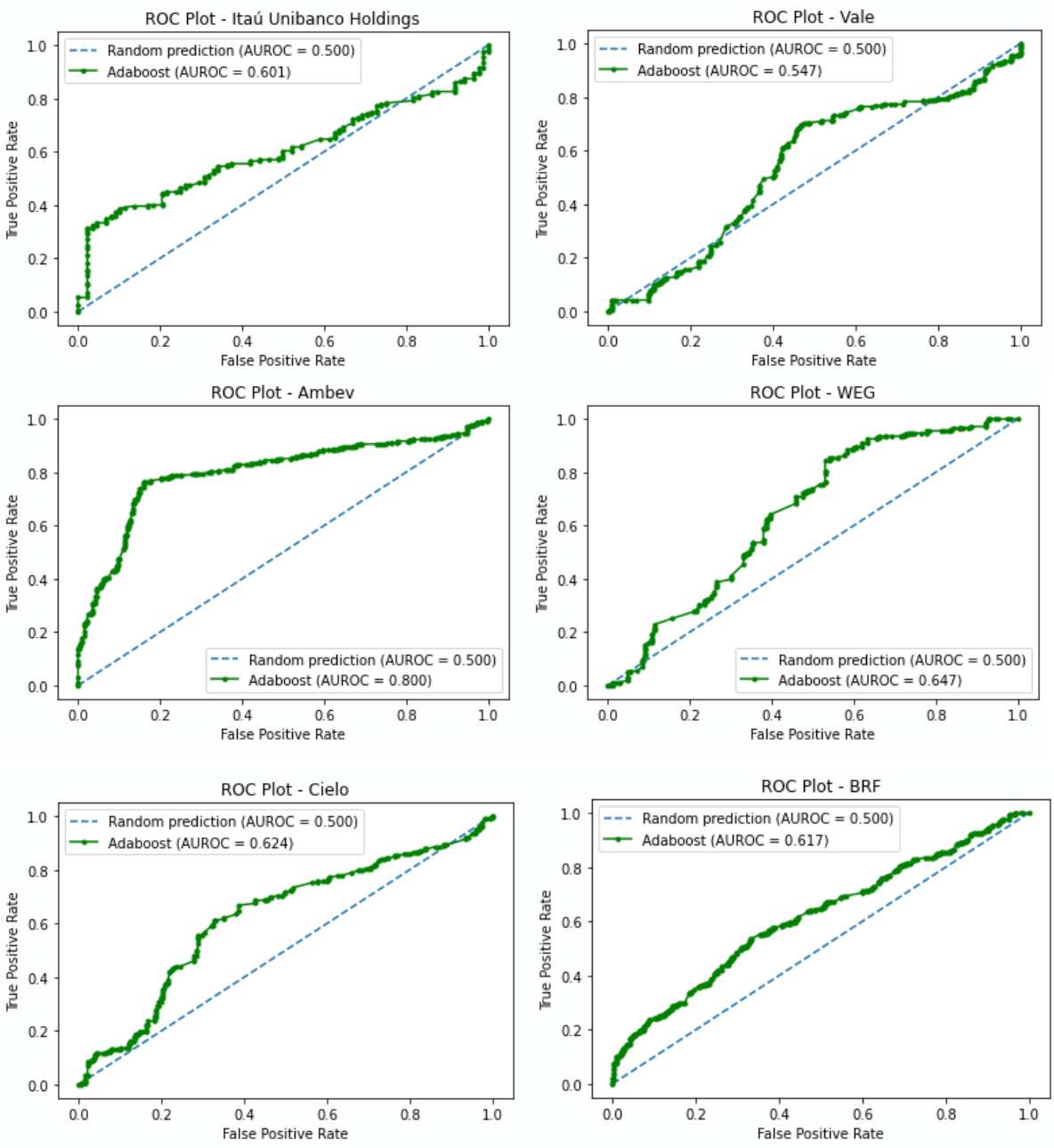
Fonte: Autoria própria.

Apple e Tesla apresentaram mais dificuldade em obter taxas maiores de verdadeiro positivo em baixas taxas de falso positivo (começo da curva). Netflix e Twitter demonstraram dificuldade para obter altas taxas de verdadeiros positivos em altas taxas de falsos positivos (final da curva).

Adaboost – ativos Brasileiro

Empregando o método *Adaboost* nos ativos Brasileiro, a média da área sobre a curva AUROC foi de 0,639, sendo que os gráficos e os valores podem ser observados na figura 21.

Figura 21 - Gráficos da curva ROC para os ativos brasileiros com *Adaboost.*



Fonte: Autoria própria.

Cielo, Vale e Itaú Unibanco Holdings, demonstraram dificuldade para obter altas taxas de verdadeiros positivos em altas taxas de falsos positivos. Também, novamente, a Cielo e a

Vale, apresentaram maiores contratempos para obter taxas maiores de verdadeiro positivo em baixas taxas de falso positivo (começo da curva).

Assim como para os ativos com *Random Forest*, os ativos aplicados *Adaboost* em ambas as regiões geográficas do estudo, apesar destes pequenos detalhes notados, no geral os gráficos apresentam comportamento normal da curva ROC, demonstrando o aprendizado do modelo.

Para todos os ativos, a área sobre a curva ROC (*AUROC*) seguiu a escala da acurácia, ou seja, quanto maior a acurácia, maior a área sobre a curva ROC.

Analisando cada método para cada região (Brasileira e Norte-Americana) é notável que para as ações com alta acurácia, como o caso da ABEV3 (Ambev), KO (Coca cola company) e AMZN (Amazon), elas possuem uma curva mais acentuada e próxima da curva ideal de unidade AUROC igual a 1.

Já para as demais ações, foi possível observar uma curva amena e mais próxima da reta referencial, demonstrando assim o aprendizado, mas com uma menor taxa.

Assim como na acurácia, a região norte-americana apresentou resultados melhores, com 0,655 de média da área de curva ROC para *Adaboost* e 0,684 para *Random Forest*, enquanto na brasileira, 0,639 e 0,672, respectivamente. Além disso, a AUROC do *Random Forest* possui uma área média sob a curva maior em ambas as regiões de estudo, com uma variação de 0,029 para a região dos Estados Unidos e 0,033 para o Brasil.

5. CONCLUSÃO

Neste trabalho, foi possível perceber uma menor variação na acurácia média dos resultados na bolsa norte-americana. Esse fato pode estar relacionado ao fato da maior estabilidade dos ativos norte-americanos quando comparados com os brasileiros, uma vez que as ações no Brasil sofrem os efeitos da bolsa norte-americana, como por exemplo, instabilidade política interna.

Notou-se também, que durante a implementação do trabalho e a separação dos dados, o método que respeita a função temporal, demonstrou ser mais eficaz do que os métodos de treinamento e teste randomizados que foram utilizados em outros trabalhos da literatura, pois os testes randomizados misturam os dados e ocasionam o *overfitting* do período em análise, sendo bastante ineficiente na análise de outros períodos.

A aplicação dos indicadores técnicos, ou seja, com a criação dos indicadores a partir dos indicadores básicos, demonstrou-se promissor devido aos resultados apresentados, como uma acurácia maior que 50%, precisão, *recall* e *F1 score* semelhantes entre si e próximos da acurácia, e também uma curva ROC normal em todos os ativos, confirmando o aprendizado do algoritmo construído.

Analisando os resultados com uma visão do mercado financeiro, em específico as métricas de acurácias encontradas, elas resultam em valores acima do referencial de 50%, sendo algumas bem expressivas, como por exemplo o caso da Ambev, onde com o modelo *Adaboost* conseguiu chegar em 78,02% na taxa de acertos das tendências da ação (acurácia). Deste modo, uma acurácia dessa precisão, contraria a hipótese do mercado eficiente discutida no meio acadêmico.

Não foi possível perceber diferenças claras e significativas dos resultados dos dois modelos de *machine learning* testados, entretanto, é possível apontar que os resultados semelhantes de ambos, podem ocorrer devido a algumas semelhanças dos modelos, como por exemplo, os algoritmos serem baseados na criação de uma floresta de árvores e ambos usarem *data sampling*. Assim, as variações dos resultados se devem apenas a aleatoriedade que existe no aprendizado de máquina.

Por fim, devido a necessidade de se balancear os dados, ou seja, de se escolher os momentos e situações que favoreçam o aprendizado da inteligência artificial, isso faz com que se evite o enviesamento do resultado e de *overfitting*. Conclui-se, que apesar dos acertos e

aprendizados dos modelos e métodos empregados, o uso desses modelos de aprendizado de máquina aplicados para previsão das tendências das ações nas bolsas de valores ainda é inviável na prática, mas muito promissor.

Trabalhos Futuros

Para trabalhos futuros, recomenda-se a aplicação de outros modelos e técnicas de aprendizado de máquina, como por exemplo, aplicar a técnica de análise de sentimento do público na internet, juntamente com a análise gráfica do ativo, a fim de encontrar ou confirmar novas hipóteses nesta área de pesquisa.

É interessante também, implementar um método em que o balanceamento dos dados não influencie tanto nos resultados, facilitando o seu uso no contexto real do mercado de ações, possibilitando assim, a aplicação deste modelo para gerenciamento de carteiras de ações visando o lucro.

REFERÊNCIAS BIBLIOGRÁFICAS

ALCANTARA et al. Estudo da estatística de análise temporal com inteligência artificial para aplicação em robôs de investimento no auxílio dos investidores para otimização de ganhos na bolsa de valores de são paulo. **Revista computação aplicada**. 2018.

ALMEIDA, S. **Finanças comportamentais: Quando as finanças encontram com a psicologia.** Dissertação para conclusão de graduação em Finanças – Universidade Federal do Ceará. Fortaleza. 2019.

AMIT, Yali; GEMAN, Donald. **Shape Quantization and Recognition with Randomized Trees**. Department of Statistics, University of Chicago, 1997.

AZAMBUJA, Pedro. [*S. l.*], 04 de dezembro de 2020. Disponível em < https://pedroazambuja.medium.com/adaboost-adaptive-boosting-dbbec150fced>. Acesso em 5 de julho de 2022.

BORGES, R**. Eficiência e o Mercado de Renda Variável Brasileiro**. Dissertação (Mestrado) — Universidade Cândido Mendes, Rio de Janeiro, Brasil, 2010.

BIANCA, Alvara. Conheça a origem da bolsa de valores. **Gorila Blog.** [s. l.], 2019. Disponível em: < https://gorila.com.br/blog/a-origem-da-bolsa-de-valores>. Acesso em 10 de agosto de 2022.

BREIMAN. **Random Forests.** Statistics Department, University of California, 2001.

B3: 3 coisas que você precisa saber sobre a antiga Bovespa. **ComoInvestir**. 24 de fevereiro de 2022. Disponível em: < https://www.poder360.com.br/economia/bolsas-desabam-com-ataque-da-russia >. Acesso em 10 de julho de 2022.

COOTNER, P. **The random character of stock market prices**. Cambrige: MIT Press, 1964.

COVA, C. J. G. **Finanças e mercado de capitais - mercados fractais: A nova fronteira das finanças**. In: São Paulo: Cengage Learning, 2011. cap. O declínio da hipótese de eficiência nos mercados financeiros e a emergência da hipótese fractal como novo paradigma descritivo do comportamento das séries temporais de retornos, p. 23–46.

CHUI, MANYIKA, MIREMADI. What ai can and can't do yet for your business. **McKinsey &Company** 11 de janeiro de 2018. Disponível em < https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/what-ai-can-and-cant-do-yet-for-your-business/pt-

br >. Acesso em 8 de abril de 2022.

COLAB. Install Ta-Lib on Google Colab. **Google Colab**. Disponível em < https://colab.research.google.com/drive/1xGx21E4oafx4WQbOCSPtQsxkD-ruMdK->. Acesso em 10 de Agosto de 2022.

COMOINVESTIR. B3: 3 coisas que você precisa saber sobre a antiga Bovespa. **ComoInvestir**. c2019. Disponível em: <https://comoinvestir.anbima.com.br/noticia/b3-antiga-bovespa/>. Acesso em 10 de julho de 2022.

DATA SCIENCE ACADEMY 1. O que é? E o que não é IA? **Data Science Academy: Inteligência Artificial Fundamentos 2.0**. c2020. Disponível em <https://www.datascienceacademy.com.br/path-player?courseid=inteligencia-artificial-fundamentos&unit=inteligencia-artificial-fundamentos_1532290095326_1Unit>. Acesso em 7 de abril de 2022.

DATA SCIENCE ACADEMY 2. Com devemos definir inteligência artificial? **Data Science Academy: Inteligência Artificial Fundamentos 2.0**. c2020. Disponível em <https://www.datascienceacademy.com.br/path-player?courseid=inteligencia-artificial-fundamentos&unit=inteligencia-artificial-fundamentos_fundamentos-de-inteligencia-artificial_1528640847328_0Unit>. Acesso em 7 de abril de 2022.

DATA SCIENCE ACADEMY 3. Campos relacionados com IA. **Data Science Academy: Inteligência Artificial Fundamentos 2.0**. c2020. Disponível em <https://www.datascienceacademy.com.br/path-player?courseid=inteligencia-artificial-fundamentos&unit=inteligencia-artificial-fundamentos_1532290149598_2Unit>. Acesso em 7 de abril de 2022.

DIDATICA TECH. Aprendizado Supervisionado ou Não Supervisionado. **Didática Tech.** [*S. l.*] c2022. Disponível em <https://didatica.tech/aprendizado-supervisionado-ou-nao-supervisionado>. Acesso em 7 de abril de 2022.

DIDATICA TECH 2. O que é e como funciona o algoritmo RandomForest. **Didática Tech.** [*S. l.*] c2022. Disponível em <https://didatica.tech/o-que-e-e-como-funciona-o-algoritmo-randomforest/>. Acesso em 7 de abril de 2022.

FAMA, E. F. Efficient capital markets: A review of theory and empirical work. **The Journal of Finance**, v. 25, n. 2, p. 383–417, 1969.

FREUND, SCHAPIRE. **A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.** AT&T Labs, 1996.

GRANDO, Nei. A Essência do Aprendizado de Máquina. **Blog do Nei.** C2022. Disponível em: < https://neigrando.com/2022/05/04/a-essencia-do-aprendizado-de-maquina/>. Acesso em 10 de agosto de 2022.

GUIMARÃES. **Previsão de default da dívida pública: uma aplicação de machine learning.** Dissertação (Mestrado em Economia) - Universidade Federal do Rio Grande do Sul. Uberlândia. 2021.

GUTIERRI, Ademir. Conheça as 3 Principais Formas de Investir em Ações. **Investing**. c2021. Disponível em: <https://br.investing.com/analysis/conheca-as-3-principais-formas-de-investir-em-acoes-200441124>. Acesso em 21 de julho de 2022.

HO, Tin Kam. **Random Decision Forest**. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, 1995.

IBM CLOUD. Random Forest. **IBM CLOUD.** [*S. l.*], 07 de dezembro de 2020. Disponível em <https://www.ibm.com/cloud/learn/random-forest>. Acesso em 5 de abril de 2022.

ICHI PRO. TÉCNICAS DE AGRUPAMENTO DE APRENDIZAGEM NÃO SUPERVISIONADA**. Ichi.pro**. 2020. Disponível em: <https://ichi.pro/pt/tecnicas-de-agrupamento-de-aprendizagem-nao-supervisionada-248180003997444>. Acesso em 29 de março de 2022.

INFOMONEY. O QUE É NYSE. **InfoMoney**. c2022. Disponível em: <https://www.infomoney.com.br/guias/o-que-e-nyse/>. Acesso em 21 de julho de 2022.

I2TUTORIALS. What do you mean by Features and Labels in a Dataset? **I2Tutorials**, Banglore, Karnataka, c2019. Disponível em: < https://www.i2tutorials.com/what-do-you-mean-by-features-and-labels-in-a-dataset/>. Acesso em 29 de março de 2022.

JUNG, SHILLER. **SAMUELSON'S DICTUM AND THE STOCK MARKET.** Cowles Foundation for Research in Economics Yale University, 2006.

LOWE. Who is right on the stock market. **The New York Times**. c2013. Disponível em:<https://www.nytimes.com/2013/11/15/opinion/rattner-whos-right-on-the-stock-market.html>. Acesso em 25 de julho de 2022.

MADEO, R. C. B.; FERREIRA, F.; RAMALHO, N.; FANTINATO, M. Papel estratégico e impacto dos sistemas de informação no mercado de ações: Um estudo envolvendo brasil e estados unidos. **Revista Eletrônica de Sistemas de Informação**, v. 11, n. 2, p. 1–22, 2012.

MARKIEL, Burton. **A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing**. c2011. W. W. Norton & Company, 2016.

MANDELBROT, B. B.; RICHARD, H. **Mercados financeiros fora de controle: A teoria dos fractais explicando o comportamento dos mercados** [Tradução por Afonso Celso da Cunha Cerra]. Rio de Janeiro: Elsevier, 2004.

MARCELO. ESPECULAR OU INVESTIR? DAY TRADE OU BUY & HOLD? CASSINO OU BOLSA? **Investidor no Japão**. C2020. Disponível em: < https://investidornojapao.com/especular-ou-investir-day-trade-ou-buy-hold/>. Acesso em 5 de agosto de 2022.

MARKS, Howard. The Most Important Thing: Uncommon Sense for the Thoughtful Investor. Columbia Business School Publishing, 2011.

MITRE, Rafael. **Aplicação de random forest para prever a tendência de ações na bolsa de valores**. Projeto de graduação de Engenharia Eletrônica e de Computação da Escola Politécnica, Universidade Federal do Rio de Janeiro, 2021.

MUSSA et al. **Hipótese de Mercados Eficientes e Finanças Comportamentais – As discussões persistem**. Pontifícia Universidade de São Paulo.

PIMENTA et al. Goldminer: A genetic programming based algorithm applied to brazilian stock market. In**: Proceedings of IEEE Symposium on Computational Intelligence and Data Mining (CIDM).** Orlando, EUA: IEEE, 2014. p. 9–24.

PINTO, Leonardo. O que é S&P 500? Saiba a importância do índice mais famoso do mundo. **XPI**. c2020. Disponível em:< https://conteudos.xpi.com.br/aprenda-a-investir/relatorios/o-que-e-sp-500-saiba-a-importancia-do-indice-mais-famoso-do-mundo/>. Acesso em 21 de julho de 2022.

PODER360. Bolsas desabam com ataque da Rússia. **Poder360**. c2022. Disponível em: < https://www.poder360.com.br/economia/bolsas-desabam-com-ataque-da-russia >. Acesso em 5 de julho de 2022.

QUARESMA SANTOS, R. Estimando a Arrecadação da Dívida Ativa da União com

Machine Learning: Uma análise baseada nos dados de arrecadação do período de 2015 a 2021. **Revista da CGU**, *[S. l.]*, v. 14, n. 26, 2022. DOI: 10.36428/revistadacgu.v14i26.529. Disponível em: https://revista.cgu.gov.br/Revista_da_CGU/article/view/529. Acesso em: 5 jan. 2023.

QUIGGIN, John. Bitcoin kills the efficient market hypothesis. **John Quiggin**. c2018. Disponível em: <https://johnquiggin.com/2018/02/09/bitcoin-kills-the-efficient-market-hypothesis/>. Acesso em 25 de julho de 2022.

RIZÉRIO, FÁBIO. Itaú (ITUB4): Resultado surpreende projeções já otimistas e analistas veem divisor de águas; ação fecha em alta. **InfoMoney**. c2022. Disponível em: <https://www.infomoney.com.br/mercados/itau-itub4-resultado-surpreende-projecoes-ja-otimistas-e-analistas-veem-divisor-de-aguas-acoes-disparam/>. Acesso em 10 de julho de 2022.

RODRIGUES, Vinicius. Entenda o que é AUC e ROC nos modelos de Machine Learning. **Medium.** c2018. Disponível em: < https://medium.com/bio-data-blog/entenda-o-que-%C3%A9-auc-e-roc-nos-modelos-de-machine-learning-8191fb4df772>. Acesso em 12 de agosto de 2022.

SANTOS, G. **Algoritmos de Machine Learning para previsão de ações da B3**. Dissertação (Mestrado em Engenharia Elétrica) - Faculdade de Engenharia Elétrica da Universidade Federal de Uberlândia. Uberlândia. 2020.

SANTOS, Paulo. **Modelos de previsão das mudanças na meta da taxa selic baseados em processamento de linguagem natural.** Projeto de graduação de Matemática Aplicada, Fundação Getúlio Vargas, 2017.

TRADEVIEW. **Acompanhe todos os mercados**. Disponível em: <https://br.tradingview.com/>. Acesso em 03 de agosto de 2022.

XAVIER, A. **Estratégias estatísticas em investimentos**. São Paulo: Novatec, 2009.

ANEXO 1

Bibliotecas

```
!pip install --upgrade pandas-datareader

import numpy as np
import pandas as pd
import datetime as dt
import pandas_datareader as pdr
import matplotlib.pyplot as plt

%matplotlib inline

from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, accuracy_score
from sklearn.ensemble import RandomForestClassifier
```

ANEXO 2

Extração dos dados

```
ticker = "ABEV3.SA"
df = pdr.get_data_yahoo(ticker, dt.datetime(2010,1,1), dt.datetime(2018
,1,1), interval='d')

#Visualizando o arquivo

df.head(5)
df.tail(5)
df.dtypes
```

ANEXO 3

Cálculo das variáveis

```
#Calculando na mão o sinal onde EMA10 > EMA 30

ema10 = df['Close'].ewm(span=10).mean()
ema30 = df['Close'].ewm(span=30).mean()

df['EMA10aboveEMA30'] = np.where(ema10 > ema30, 1, -1)


#Calculando sinais de ema

df['ema5'] = df['Close'].ewm(span=5).mean()
df['ema10'] = df['Close'].ewm(span=15).mean()
df['ema15'] = df['Close'].ewm(span=15).mean()
df['ema20'] = df['Close'].ewm(span=20).mean()
df['ema30'] = df['Close'].ewm(span=20).mean()
df['ema40'] = df['Close'].ewm(span=40).mean()
df['ema50'] = df['Close'].ewm(span=50).mean()
df['ema60'] = df['Close'].ewm(span=60).mean()
df['ema70'] = df['Close'].ewm(span=70).mean()
df['ema80'] = df['Close'].ewm(span=80).mean()
df['ema90'] = df['Close'].ewm(span=90).mean()
df['ema100'] = df['Close'].ewm(span=100).mean()
df['ema150'] = df['Close'].ewm(span=150).mean()
df['ema200'] = df['Close'].ewm(span=200).mean()

df['aboveEMA5'] = np.where(df['Close'] > df['ema5'], 1, -1)
df['aboveEMA10'] = np.where(df['Close'] > df['ema10'], 1, -1)
df['aboveEMA15'] = np.where(df['Close'] > df['ema15'], 1, -1)
df['aboveEMA20'] = np.where(df['Close'] > df['ema20'], 1, -1)
df['aboveEMA30'] = np.where(df['Close'] > df['ema30'], 1, -1)
df['aboveEMA40'] = np.where(df['Close'] > df['ema40'], 1, -1)

df['aboveEMA50'] = np.where(df['Close'] > df['ema50'], 1, -1)
df['aboveEMA60'] = np.where(df['Close'] > df['ema60'], 1, -1)
df['aboveEMA70'] = np.where(df['Close'] > df['ema70'], 1, -1)
df['aboveEMA80'] = np.where(df['Close'] > df['ema80'], 1, -1)
df['aboveEMA90'] = np.where(df['Close'] > df['ema90'], 1, -1)

df['aboveEMA100'] = np.where(df['Close'] > df['ema100'], 1, -1)
df['aboveEMA150'] = np.where(df['Close'] > df['ema150'], 1, -1)
df['aboveEMA200'] = np.where(df['Close'] > df['ema200'], 1, -1)

#Calculando na mão RSI
```

```
delta = df['Close'].diff()
up = delta.clip(lower=0)
down = -1*delta.clip(upper=0)
ema_up = up.ewm(com=13, adjust=False).mean()
ema_down = down.ewm(com=13, adjust=False).mean()
rs = ema_up/ema_down
df['RSI'] = 100 - (100/(1 + rs))
df['oversoldRSI'] = np.where(df['RSI'] < 30, 1, -1)
df['overboughtRSI'] = np.where(df['RSI'] > 70, 1, -1)

#Calculando na mão Williams%R

high14= df['High'].rolling(14).max()
low14 = df['Low'].rolling(14).min()
df['%R'] = -100*(high14 - df['Close'])/(high14 - low14)
df['oversold%R'] = np.where(df['%R'] < -80, 1, -1)
df['overbought%R'] = np.where(df['%R'] > -20, 1, -1)

#Calculando na mão MACD

exp1 = df['Close'].ewm(span=12).mean()
exp2 = df['Close'].ewm(span=26).mean()
macd = exp1 - exp2
macd_signal = macd.ewm(span=9).mean()
df['MACD'] = macd_signal - macd

#Calculando na mão Rate Of Change 'ROC'

days = 6
ct_n = df['Close'].shift(days)
df['PROC'] = (df['Close'] - ct_n)/ct_n

#Calculando a mão %K

df['%K'] = (df['Close'] - low14)*100/(high14 - low14)

df['oversold%K'] = np.where(df['%K'] < 20, 1, -1)
df['overbought%K'] = np.where(df['%K'] > 80, 1, -1)

#Classificando
df['Return'] = df['Close'].pct_change(1).shift(-1)
df['class'] = np.where(df['Return'] > 0, 1, 0)
```

ANEXO 4

Modelo

```
#Limpando dados vazios

df = df.dropna()

#Setando a variavel Y

df['target_cls'] = np.where(df['Close'].shift(-55) > df.ema150.shift(-
55), 1, 0)

#Setando a variavel X

predictors = ['EMA10aboveEMA30','MACD','RSI','%K','%R','PROC','overboug
ht%R','oversold%R','overboughtRSI', 'oversoldRSI','aboveEMA5','aboveEMA
10','aboveEMA15','aboveEMA20','aboveEMA30','aboveEMA40','aboveEMA50','a
boveEMA60','aboveEMA70','aboveEMA80','aboveEMA90','aboveEMA100']

X = df[predictors]
y = df[['target_cls']]

#Dividindo o dataset em treino e teste

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3
0, shuffle=False)

#Analisando o dataset preparado para o modelo
print('Y=0:', df['target_cls'].tolist().count(0))
print('Y=1:', df['target_cls'].tolist().count(1))
print('len X_train', len(X_train))
print('len y_train', len(y_train))
print('len X_test', len(X_test))
print('len y_test', len(y_test))
```

ANEXO 5

Resultados

```
#Treinando o modelo

rfc = RandomForestClassifier(random_state=0)
rfc = rfc.fit(X_train, y_train)

#Testando o modelo

y_pred = rfc.predict(X_test)

#Vendo a acurácia do modelo

report = classification_report(y_test, y_pred)
print('Model accuracy', accuracy_score(y_test, y_pred, normalize=True))
print(report)

#Curva ROC
r_probs = [0 for _ in range(len(y_test))]
rf_probs = rfc.predict_proba(X_test)
rf_probs = rf_probs[:, 1]

from sklearn.metrics import roc_curve, roc_auc_score

r_auc = roc_auc_score(y_test, r_probs)
rf_auc = roc_auc_score(y_test, rf_probs)

r_fpr, r_tpr, _ = roc_curve(y_test, r_probs)
rf_fpr, rf_tpr, _ = roc_curve(y_test, rf_probs)

import matplotlib.pyplot as plt

plt.plot(r_fpr, r_tpr, linestyle='--
', label='Random prediction (AUROC = %0.3f)' % r_auc)
plt.plot(rf_fpr, rf_tpr, marker='.', label='Random Forest (AUROC = %0.3
f)' % rf_auc)

# Title
plt.title('ROC Plot - BRF')
# Axis labels
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
# Show legend
plt.legend() #
# Show plot
plt.show()
```