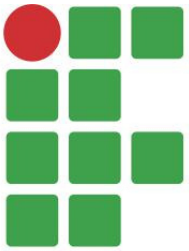


Instituto Federal de Educação, Ciência e Tecnologia da Paraíba

Campus Campina Grande

Coordenação do Curso Superior de Bacharelado em Engenharia de Computação

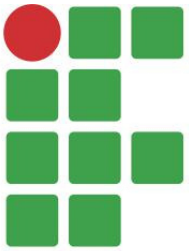
Avaliação de Algoritmos de Aprendizagem de Máquina para a Predição da Bolsa de Valores do Brasil

Gabriel de Lima e Silva

Rennan de Aguiar Ramos

Orientador: Prof. Igor Barbosa da Costa, D.Sc.

Campina Grande, Agosto de 2022



Instituto Federal de Educação, Ciência e Tecnologia da Paraíba

Campus Campina Grande

Coordenação do Curso Superior de Bacharelado em Engenharia de Computação

Avaliação de Algoritmos de Aprendizagem de Máquina para a Predição da Bolsa de Valores do Brasil

Gabriel de Lima e Silva

Rennan de Aguiar Ramos

Monografia apresentada à Coordenação do Curso Superior de Bacharelado em Engenharia de Computação do IFPB - Campus Campina Grande, como requisito parcial para conclusão do curso de Bacharelado em Engenharia de Computação.

Orientador: Prof. Igor Barbosa da Costa, D.Sc.

Campina Grande, Agosto de 2022

S586a Silva, Gabriel de Lima
Avaliação de algoritmo de aprendizagem de máquina para predição da Bolsa de Valores do Brasil / Gabriel de Lima e Silva, Rennan de Aguiar Ramos. - Campina Grande, 2022.
16 p.:il.

Trabalho de Conclusão de Curso - Artigo (Curso Superior de Bacharelado em Engenharia da Computação) - Instituto Federal da Paraíba, 2022.
Orientador: Prof.D.Sc. Igor Barbosa da Costa.

1.Engenharia da computação. 2. Aprendizado de máquina - análise de dados. 3. Inteligência artificial - algoritmo. I. Ramos, Rennan de Aguiar. II. Título.

CDU 004.421

Avaliação de Algoritmos de Aprendizagem de Máquina para a Predição da Bolsa de Valores do Brasil

Rennan de Aguiar Ramos e Gabriel de Lima e Silva
Discente
Instituto Federal de Educação, Ciência e Tecnologia
Campus Campina Grande
rennan.ramos@academico.ifpb.edu.br
lima.silva@academico.ifpb.edu.br

Igor Barbosa da Costa
Orientador
Instituto Federal de Educação, Ciência e Tecnologia
Campus Campina Grande
igor.costa@ifpb.edu.br

***Resumo*—Mercado de ações é onde diariamente ocorrem compras e vendas de porções de empresas que estão listadas nas bolsas de valores. Neste ambiente, o lucro é o resultado mais desejado pelos investidores. Nos últimos anos, o mercado vem se tornando cada vez mais atrativo e as estratégias de compra e venda estão se tornando cada vez mais complexas. Nesse sentido, o uso da inteligência artificial está cada vez mais comum, principalmente em ocasiões onde é possível obter uma vasta quantidade de dados. Este trabalho aplica técnicas de aprendizado de máquina com baixo poder computacional com intuito de processar e realizar predições das cinco ações mais líquidas da bolsa de valores brasileira. Especificamente, foi avaliado qual foi o desempenho dessas técnicas no cenário antes da pandemia e durante a pandemia. Ao final do primeiro experimento, o qual busca o algorítimo que obtém melhor retorno sobre o investimento, o *Gaussian Naive Bayes* e o *Random Forest* obtiveram os melhores resultados no período antes da pandemia e durante a pandemia foram o *Random Forest* e o *Decision tree*, ou seja, algoritmos baseados em árvores nesse experimento obtiveram melhores resultados, para o segundo experimento que realiza a predição do valor de fechamento das ações, o *Linear Regression* teve o melhor desempenho nos dois períodos.**

Abstract—The stock market is where purchases and sales of portions of companies that are listed on the stock exchanges are made. In this environment, profit is the most desired result by those who inhabit it. The market has become increasingly common and accessible, the same way that artificial intelligence has become in the last years, mainly in data analysis and process automation, especially in situations where there is a large amount of data involved.This article applies low computing power machine learning technique with objective of process and predict stocks variables in the five most liquid stocks on the Brazilian stock exchange, specifically, the performance during and before the pandemic was evaluated. At the end of the first experiment, which aims to detect the algorithm obtains the best return on investment, Gaussian Naive Bayes and Random Forest obtained the best results in the period before the pandemic and during the pandemic were the Random Forest and Decision tree, in other words, that is, tree-based algorithms in this experiment had better results, for the second experiment that predicts the closing value of stocks, Linear Regression had the best performance in both periods.

I. Introdução

Mercado financeiro é o ambiente onde ocorre a compra e venda de ações, títulos de renda fixa, fundos de investimento, câmbios de moedas estrangeiras e de *commodities*. Dentro desse ambiente existe quem empresta o dinheiro (pessoa ou instituição que empresta seu dinheiro para receber rendimentos) e quem recebe esse dinheiro (pessoa ou instituição que devolve o dinheiro recebido com juros). Para que ocorram essas negociações, existem os intermediários que fazem a ponte entre quem empresta e quem recebe o dinheiro, como por exemplo, a bolsa de valores que é responsável por negociar os títulos emitidos por empresas de capital aberto (ações) e as corretoras que são responsáveis por conectar a bolsa de valores aos investidores (RICO, 2021).

Em Março de 2022, o número de brasileiros que investem através da bolsa de valores atingiu a marca de 4,304 milhões, um crescimento de 43,7% frente a março de 2021 (MOREIRA, 2022). Esses investidores são pessoas físicas e jurídicas, que buscam com o auxílio de estratégias de investimentos, fazer operações de compra ou venda de ações com o intuito de obter lucro.

Existem várias estratégias de investimento no mercado financeiro. Esta pesquisa foca no *Buy and Hold*, que se trata de comprar determinada ação com o intuito de se manter com ela por bastante tempo, visando uma rentabilidade de longo prazo (estratégia mais segura e conservadora), onde o segundo experimento será para auxílio desta técnica, e o *Swing Trade* que se trata da compra e venda de ações em um curto período e em grande volume (estratégia que visa obter lucro num espaço mais curto de tempo) (NOMAD, 2022), onde o primeiro experimento visa auxiliar essa estratégia.

Para o *Swing Trade* são usados indicadores para auxiliar nas operações e existem diversas técnicas que auxiliam esses *trades*, como as análises gráficas (RICO, 2022).

Visto que o ambiente acima contém uma vasta gama de dados entra-se no ambito do aprendizado de máquina, que é uma subárea da inteligência artificial, que vai ser alimentada e treinada por esses dados, além disso ela é capaz aprender determinadas funções (SIMON, 2013), os algoritmos de aprendizado de máquina utilizados para a realização dos experimentos foram *Linear Regression*, *Gradient Boosting*, *Decision Tree*,

Decision Tree, *Gaussian Naive Bayes*, *K-nearest-neighbors*.

Nesse contexto, esta pesquisa visa realizar um estudo comparativo de algoritmos de aprendizagem de máquina tradicionais, como classificadores e algoritmos de regressão para construir um modelo de predição de preço de ações. Para os experimentos são usados dados da B3, Vale, Petrobras, Itaú e Bradesco, as cinco ações com maior liquidez na bolsa brasileira no ano de 2021 (BRASIL, 2021), todos os dados foram retirados do (FINANCE, 2022).

Na etapa de experimentos, foram avaliadas duas estratégias de predição. Na primeira, é realizada previsões dos valores de fechamento da ação para o dia seguinte. Na segunda, é realizada uma predição direta se deve ser realizada a compra ou venda da ação.

Além dessas abordagens, o *dataset* foi dividido em duas partes: pré-pandemia (até o final de 2019) e pandemia (até o final de 2021), para que seja possível gerar um resultado comparativo dos algoritmos em cenários diferentes. Assim, este trabalho pretende responder a seguinte questão de pesquisa: **Quais algoritmos têm melhor desempenho na predição de preço das ações no período pré-pandemia e durante a pandemia?**

Durante as próximas cinco seções, estão sendo abordados tópicos com o objetivo de tanto auxiliar o entendimento mínimo necessário para a compreensão para pesquisa, quanto os detalhes de desenvolvimento e conclusões. Na seção de fundamentação teórica está localizado as definições dos principais tópicos relacionados a pesquisa, em seguida há trabalhos relacionados onde expõe outras pesquisas do mesmo segmentos, e quais as diferenças entre elas e este trabalho, em metodologia demonstra cada etapa entre a idealização, desenvolvimento e conclusões dos experimentos, seguidamente por experimento e discussão de resultados onde detalhadamente mostra os resultados obtidos durante cada experimento, em conclusão e trabalhos futuros encontra a conclusão final sobre a pesquisa e os caminhos a seguir a partir destas conclusões, e por último encontra-se a bibliografia de referência que são referenciadas ao longo do texto.

II. Fundamentação teórica

Inicialmente são apresentados conceitos relacionados a bolsa de valores. Em seguida, discute-se os fundamentos importantes da área de inteligência de artificial.

A. Bolsa de valores

- **Ambiente de negociação**: A bolsa de valores é um ambiente que possibilita para os investidores, a negociação de ações de empresas. No Brasil, a única bolsa de valores em funcionamento é a B3, que é supervisionada pela Comissão de Valores Mobiliários. Esse órgão tem poder de fiscalização, visando garantir a existência de um processo seguro e disciplinado (REIS, 2020).
- **Ações**: Uma ação é uma pequena parcela de uma empresa. Uma vez que uma empresa decide ter seu capital aberto, qualquer investidor pode comprar essa parcela através da bolsa. A partir da compra de ações, o investidor passa a ter também os direitos e os deveres de um sócio.
- **Técnicas de negociação**: A técnica de *Swing Trade* é uma técnica no qual, onde em um curto período de tempo através de análise de gráficos o investidor especula a variação do preço dos ativos. A partir dessa análise, ele decide se é o momento de compra ou venda da ação. A técnica *Buy and Hold* é onde o investidor está pensando em longo prazo, e está disposto a comprar ativos e não movimentá-los por anos visando uma boa rentabilidade a longo prazo (REIS, 2020). As Figuras 1, 2 e 3 apresentam um exemplo de progressão do valor unitário de uma ação em um certo período de tempo.

Imagens retiradas do *Yahoo Finance*.

Figura 1: Primeiro cenário de compra.

Figura 2: Segundo cenário de compra.

Partindo da evolução dos preços demonstrados nas figuras 1, 2 e 3 pode-se simular:

- Um primeiro cenário onde houve uma compra de cinco ações ao valor de R$ 71,10 (figura 1) e uma venda das mesmas cinco ações por um valor de R$ 70,16 (figura 3), gerando uma perda de R$ 4,70.
- Um segundo cenário onde houve uma compra de cinco ações ao valor de R$ 69,43 (figura 2) e uma venda das

Figura 3: Cenário de venda.

mesmas cinco ações por um valor de R$ 70,16 (figura 3), gerando um ganho de R$ 3,65.

B. *Inteligência artificial*

A Inteligência Artificial (IA) é a capacidade de um computador ou robô controlado por computador de executar tarefas comumente associadas a seres inteligentes (COPELAND, 2020). Esse termo é usado frequentemente para descrever sistemas com processos intelectuais característicos dos humanos, como a capacidade de raciocinar, descobrir significados, generalizar ou aprender com experiências passadas. Desde o desenvolvimento do computador digital na década de 40, foi demonstrado que os computadores podem ser programados para realizar tarefas muito complexas como, por exemplo, descobrir provas de teoremas matemáticos ou jogar xadrez com grande proficiência. Ainda assim, apesar dos avanços contínuos na velocidade de processamento do computador e na capacidade de memória, ainda não existem programas que possam igualar a flexibilidade humana em domínios mais amplos ou em tarefas que exigem muito conhecimento cotidiano (COPELAND, 2020). Por outro lado, alguns programas atingiram níveis de desempenho igual ao de especialistas humanos na execução de determinadas tarefas específicas, de modo que a inteligência artificial mesmo sendo limitada, é encontrada em aplicações tão diversas como diagnóstico médico, mecanismos de busca de computadores e reconhecimento de voz ou caligrafia e outros (COPELAND, 2020).

C. *Aprendizado de Máquina*

Entre as sub-áreas da inteligência artificial, o *aprendizado de máquina* é um dos campos que mais avançou nas últimas décadas. O aprendizado de máquina é o método usado para treinar um computador para aprender com suas entradas de dados, mas sem programação explícita para todas as circunstâncias. O prendizado de máquina como é o campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados (SIMON, 2013).

De forma geral, pode-se considerar que o aprendizado de máquina é o processo no qual as máquinas são capazes de fazer previsões decorrentes do reconhecimento de padrões, padrões esses que surgem a partir da análise de dados realizadas pelos modelos, que a partir delas utiliza-se esses *insights* para fazer suas devidas previsões. Para realização deste trabalho utilizamos dois métodos diferentes para fazermos as análises, a primeira foi usar algoritmos de regressão para previsão dos valores de fechamento do dia seguinte de uma certa ação, e algoritmos de classificação para poder realizar uma decisão se é um bom momento para vender ou comprar certo ativo diariamente.

D. *Algoritmos de Aprendizado de Máquina*

Para este trabalho, foram escolhidos seis algoritmos para realização dos experimentos, sendo alguns de regressão e outros de classificação. São eles:

- **Linear Regression**: A análise de regressão linear é usada para prever o valor de uma variável com base no valor de outra. A variável que deseja prever é chamada de variável dependente. A variável que é usada para prever o valor de outra variável é chamada de variável independente. Essa forma de análise estima os coeficientes da equação linear, envolvendo uma ou mais variáveis independentes que melhor preveem o valor da variável dependente. A regressão linear se ajusta a uma linha reta ou superficial que minimiza as discrepâncias entre os valores de saída previstos e reais (IBM,).
- **Gradient Boosting**: O algoritmo *Gradient Boosting* é uma técnica de aprendizado de máquina para problemas de regressão e classificação, que produz um modelo de previsão na forma de um conjunto de modelos de previsão fracos, geralmente árvores de decisão. Ele constrói o modelo em etapas, como outros métodos de reforço, e os generaliza, permitindo a otimização de uma função de perda diferenciável arbitrária. O objetivo do algoritmo é criar uma corrente de modelos fracos, onde cada um tem como objetivo minimizar o erro do modelo anterior, por meio de uma função de perda. Aos ajustes de cada modelo fraco é multiplicado um valor chamado de taxa de aprendizagem. Esse valor, tem como objetivo determinar o impacto de cada árvore no modelo final. Quanto menor o valor, menor a contribuição de cada árvore (SILVA, 2020).
- **Decision Tree**: Uma árvore de decisão é um algoritmo de aprendizado de máquina que é utilizado para classificação e para regressão. Isto é, pode ser usado para prever categorias discretas (sim ou não, por exemplo) e para prever valores numéricos (o valor do lucro em reais). É um algoritmo que utiliza conceitos de recursividade. Ou seja, ele repete o mesmo padrão sempre na medida em que vai entrando em novos níveis de profundidade. É como se uma função chamasse a ela mesma como uma segunda função para uma execução paralela, da qual a primeira função depende para gerar sua resposta. O grande trabalho da árvore é justamente encontrar os nós que vão ser encaixados em cada posição. Quem será o nó raiz, quem será o nó da esquerda, e o da direita (SACRAMENTO, 2021).

- **Random Forest**: A floresta aleatória, é um grupo de árvores de decisões, porém diferente das árvores de decisões que criam regras para suas tomadas de decisões, a floresta aleatória escolherá os recursos utilizados de forma randômica, fará uma árvore de decisões, e em seguida calculada a média dos resultados (PESSANHA, 2019).
- **Gaussian Naive Bayes**: O classificador multinomial Naive Bayes é um dos modelos mais populares no aprendizado de máquina. Tomando como premissa a suposição de independência entre as variáveis do problema, o modelo de Naive Bayes realiza uma classificação probabilística de observações, caracterizando-as em classes pré-definidas. É interessante saber que o Naive Bayes é um dos modelos mais conhecidos a aplicar o conceito de probabilidade. Esse modelo, como o nome indica, faz uso do teorema de Bayes como princípio fundamental (HOUSE, 2021).
- **K-nearest-neighbors**: O KNN é um algoritmo não pramétrico, aonde a estrutura do modelo será determinada pelo dataset utilizado. Este algoritmo é categorizado como *lazy*, por ter o seu tempo de processamento maior que os demais. Esses algoritmos, não necessitam de dados de treinamento para se gerar o modelo, o que diminui em partes o processo inicial, mas em contrapartida gerará uma necessidade de analise posterior mais apurada. No caso de algoritmos que não necessitam de treinamento, todos os dados obtidos no dataset serão utilizados na fase de teste, resultando em um treinamento muito rápido e em um teste e validação lentos, momento o qual necessitamos estar bem atentos aos resultados gerados (LUZ, 2019).

E. Datasets

Datasets são conjuntos de dados tabulados que servem como entrada para serem utilizados por algoritmos de aprendizado de máquina. Um fator positivo dos *datasets* é que por serem dados tabulados, com linhas e colunas, temos dados bem organizados e com informações claras acerca de sua finalidade, mas isso também traz um ponto negativo é que uma quantidade vasta de dados podem ocasionar inconsistência que pode atrapalhar na análise dos dados. Dessa forma, é necessário realiza um processo de limpeza e tratamento antes dos algoritmos consumirem, para garantir que apenas dados úteis sejam considerados (HOPPEN, 2018).

Estes *datasets* foram extraídos de um período de onze anos, de janeiro de 2010 até dezembro de 2021, as características iniciais deles foram os valores de abertura, fechamento, pico e baixa das ações, sua data e o volume de ações manipuladas na mesma, totalizando por volta de 3 mil diferentes conjuntos de características para cada ação.

III. Trabalhos Relacionados

Nesta seção, são apresentados os trabalhos relacionados a esta pesquisa, classificando cada estudo de acordo com o objetivo, os algoritmos usados, se existe uma comparação entre o período pré-pandemia e durante a pandemia e também se é feito o uso de multiplas ações como pode ser visto na tabela 1.

Antes é importante saber que prever tendência se trata da criação de uma coluna auxiliar no estudo onde ela é preenchida de acordo com a análise que for realizada. Já prever valores se trata do ato de tentar prever o valor real de alguma coluna específica em um momento específico.

O objetivo de prever tendência aconteceu nos trabalhos de (SILVA, 2022) e (SANTOS, 2020) assim como em nosso primeiro experimento, já a previsão dos valores foi feita nos trabalhos de (OLIVEIRA, 2021), (STEFFEN, 2021) e (NOGUEIRA, 2021) assim como em nosso segundo experimento.

Nos algoritmos utilizados existe uma boa diversidade. (SILVA, 2022) faz o uso de LSTM com XGBoost, (OLIVEIRA, 2021) faz o uso de *MLP Regressor* com *GridSearchCV*, (STEFFEN, 2021) faz o uso do *LRS*, (NOGUEIRA, 2021) faz o uso do LSTM, (SANTOS, 2020) faz o uso de *Random Forest*, *SVM* e *RNA*, já este trabalho de pesquisa faz o uso do *GridSearchCV* e do Random Forest adicionando a eles *Linear Regression*, *Gradient Boosting*, *Decision Tree*, *Gaussian Naive Bayes* e *K-nearest-neighbors*.

O uso de múltiplas ações acontece apenas no trabalho de (STEFFEN, 2021) e de (SANTOS, 2020) e dentre todos os trabalhos temos em comum com o nosso o uso da B3SA3 no de (STEFFEN, 2021), VALE3 no de (SANTOS, 2020), PETR4 no de (SILVA, 2022), de (STEFFEN, 2021) e de (SANTOS, 2020), ITUB4 no de (OLIVEIRA, 2021) e de (SANTOS, 2020).

A comparação entre dados pré-pandemia e durante a pandemia é um diferencial adicionado à está pesquisa, visto que os trabalhos relacionados tiveram foco em demonstrar resultados nos períodos os quais trabalharam, mas nenhum teve foco em analisar os comportamentos durante e antes ou depois de um evento mundial que diretamente influenciou na bolsa.

IV. Metodologia

Nesta seção, serão descritos os passos para o desenvolvimento desta pesquisa.

A etapa inicial do processo foi definir qual seria o conjunto de dados. Durante essa etapa, foi decidido focar nas cinco ações com maior liquidez no mercado - B3, Vale, Petrobras, Itaú e Bradesco. Como fonte de dados foi utilizado o yahoo finanças (FINANCE, 2022). A extração foi feita dentro de um período de 11 anos (04/01/2010 à 30/12/2021), período que engloba tanto a pandemia, quanto também uma boa margem antecedente.

Após a definição do conjunto de dados, o próximo passo foi decidir quais algoritmos seriam utilizados. Nessa etapa foram filtrados algoritmos previamente utilizados no âmbito de finanças, e que também não fossem de alta complexidade e poder computacional, pois uma das premissas do estudo é analisar o desempenho de algoritmos que possam ser treinados em qualquer computador. Avaliando trabalhos relacionados e por tendências de mercado, foram determinados os seguintes algoritmos: *Linear Regression*, *Gradient Boosting*, *Decision Tree*, *Random Forest*, *Gaussian Naive Bayes*, *K-nearest-neighbors*.

Estudo	Objetivo	Algoritmos Usados	Comparação pré e durante pandemia?	Faz uso de multiplas ações?
(SILVA, 2022)	Prever tendência	*LSTM* com *XGBoost*	Não	Não
(OLIVEIRA, 2021)	Prever tendência	*MLP Regressor* com *GridSearchCV*	Não	Não
(STEFFEN, 2021)	Prever tendência	*LRS*	Não	Sim
(NOGUEIRA, 2021)	Prever valor	*LSTM*	Não	Não
(SANTOS, 2020)	Prever tendência	*Random Forest*, *SVM*, *RNA*	Não	Sim
Este trabalho	Prever valor e tendência	*Linear Regression*, *Gradient Boosting*, *Decision Tree*, *Random Forest*, *Gaussian Naive Bayes*, *K-nearest-neighbors*	Sim	Sim

Tabela 1: Tabela comparativa dos estudos

A etapa seguinte foi iniciar o processo de pré-processamento do dataset (engenharia de atributos). Os atributos foram criados com base na biblioteca finta (PEERCHEMIST, 2021), biblioteca que define um aglomerado de variáveis comuns utilizados por analistas de mercado. A Tabela 2 apresenta a listas de atributos criados.

Por fim, o *dataset* foi dividido para os cenários de estudo. Para o período pré-pandemia foi considerado o período de 04/01/2010 até 29/12/2017 para treino e de 30/12/2017 até 30/12/2019 para testes. Para o cenário pandêmico, ficou determinado o período de 04/01/2010 até 29/12/2019 para treino e de 30/12/2019 até 30/12/2021 para testes.

Com o conjunto de dados definido, iniciou-se o desenvolvimento dos experimentos. Os algoritmos escolhidos foram treinados e passaram por um processo de *tuning*, processo este onde se realiza melhorias manuais nos modelos de predição, cada modelo possui seus próprios parâmetros que podem ser customizados, dessa forma é possível dizer exatamente, ou um *range* de valores que esses algorítimos internamente irão utilizá-los para aprimorar seus desempenhos, visando tabelas 14 a 17. Durante a realização dos experimentos com os algoritmos, identificaram-se comportamentos diferentes para cada algoritmo utilizado.

Nome	Descrição
MA5	Média dos valores de fechamento da ação nos períodos de 5 dias.
MA15	Média dos valores de fechamento da ação nos períodos de 15 dias.
MA21	Média dos valores de fechamento da ação nos períodos de 21 dias.
MA50	Média dos valores de fechamento da ação nos períodos de 50 dias.
EMA5	Média exponencial dos valores de fechamento nos períodos de 5 dias.
EMA15	Média exponencial dos valores de fechamento nos períodos de 15 dias.
EMA21	Média exponencial dos valores de fechamento nos períodos de 21 dias.
EMA50	Média exponencial dos valores de fechamento nos períodos de 50 dias.
DIFF-HIGH-LOW	Diferença entre os valores de máxima e mínima do dia.
DIFF-OPEN-CLOSE	Diferença entre os valores de abertura e fechamento do dia.
NORM-VOL	Média móvel do volume nos últimos 5 dias.

Tabela 2: Tabela com colunas do *feature engineering*

V. Experimentos e Discussão dos Resultados

Nesta seção, são discutiros os resultados obtidos em cada experimento visando responder às questões de pesquisa definidas anteriormente.

A. QP: Quais algoritmos se portaram melhor no período pré-pandemia e durante a pandemia?

O primeiro experimento foi com os algoritmos *Decision Tree*, *K-Nearest Neighbors*, *Gaussian Naive Bayes*, e *Random Forest*, com o propósito de realizar uma classificação se o dia em questão é um bom momento para compra ou venda

das ações, realizando um experimento de tendência. Para esse experimento, além das *features* bases advindas do *dataset*, e das *features* advindas da *feature engineering*, foi criada uma nova coluna no *dataset*, que representam com 0 ou 1, onde 0 representa venda, e 1 representa compra. Assim trazia uma representação booleana se o dia era bom ou não para a compra do ativo, com auxílio de uma função onde verificava se a diferença entre o valor do fechamento do dia em análise e o fechamento do dia seguinte, em caso de positivo, era um cenário de lucro, então era classificado como compra, caso negativo, então era classificado como venda. Ou seja, a função verifica se de um dia para o outro, a ação iria gerar lucro ou uma perda, e toma a decisão do que deve ser feito no dia anterior, para evitar comprar uma ação quando a tendência for perder dinheiro, e auxiliar a comprar uma ação quando a tendência for lucrar. Dados estes que foram utilizados por cada um dos modelos durante a fase de treinamento, como variável alvo para predição. A partir disso, foram feitas as análises de cada algoritmo, para cada uma das 5 ações separadamente durante o período pré pandemia, e durante a pandemia. Com essas predições, inicia-se o segundo passo, que se trata da análise da lucratividade com as escolhas feita pelos algoritmos. Os resultados desses classificadores foram utilizados sobre a base de valores de fechamentos reais, para que fosse possível analisar a lucratividade que realmente teria sido acontecido caso os algoritmos tivessem sido utilizados no período. A função usada para calcular a lucratividade, analisa os indicadores de compra ou venda que os algoritmos previram anteriormente para calcular a diferença entre o valor que foi investido e o valor que foi vendido, porém no momento da indicação de venda predita, só é realizada a venda da ação caso a diferença entre o valor da compra e valor do fechamento atual foi maior ou igual que 5 reais, ou seja, o algoritmo sempre procurar por momentos que os modelos previram venda em que possui um lucro mínimo de 5 reais relacionado ao valor que a ação foi previamente comprado. Outros dois pontos que valem o destacar, é que para simulação está sendo considerado que no primeiro dia do teste, é simulado obrigatoriamente a compra de de 50 ações para cada ativo, e no último dia estão sendo vendidas todas ações que ainda restarem, mesmo que não seja um dia predito como momento ideal para venda, para dessa forma ter com exatidão qual foi o lucro, ou prejuízo durante o período.

Conclusões sobre o período pré pandemia

Na tabela 3 é possível notar que nas ações da Vale, todos algoritmos tiveram um *score* bem próximos. O *Gaussian Naive Bayes* foi o que obteve tanto o maior *score* quanto o maior *ROI* (Retorno sobre o investimento), e mesmo que três dos quatro algoritmos tenham tido o mesmo score, todos tiveram um resultado diferente para lucratividade, *ROI* e valor investido, concluindo que em diversos momento eles previram de forma diferente, e não foram exatamente a mesma previsão.

Na tabela 4 é possível notar que nas ações da Petrobras, houve valores totalmente diferentes para todas variáveis em análise. O maior *ROI* ficou com *Gaussian Naive Bayes* novamente, a maior lucratividade com *Decision Tree* e o melhor *score* com *Random Forest*, que foi também o que teve o segundo maior *ROI*.

Na tabela 5 é possível notar que nas ações do Itaú obteve-se um resultado parecido com a Vale, onde o mesmo algoritmo que teve o melhor *score* também teve o melhor *ROI*. Mas, desta vez foi a *Random Forest* que se destacou, e a melhor lucratividade ficou com o *K-Nearest Neighbors*, o mesmo que havia ocorrido no cenário da Vale.

Na tabela 6 é possível notar que nas ações do Bradesco, a *Random Forest* novamente atingiu o melhor *ROI*, e apesar de não ter tido o melhor *score*, que pela primeira vez foi do *K-Nearest Neighbors*, a *Random Forest* teve a segunda melhor nota de *score*, com uma diferença de 0.008 com o *K-Nearest Neighbors*.

Na tabela 7 é possível notar que nas ações da B3 foram obtidos os maiores índices de *ROI*, variando de 48.08% á 65.47%. O maior *score* e lucratividade ficaram com *Decision Tree*, já o maior *ROI* ficou com o *Gaussian Naive Bayes*.

Em geral, pode-se perceber que na maioria dos casos, o algoritmo que obteve o melhor *ROI*, foi o que também menos investiu, por isso sempre teve ou esteve entre os menores índices de lucratividade. Assim, é possível concluir que comprar menos ações em momentos específicos e vender também em momentos específicos, teve um retorno melhor em relação ao que foi investido do que aqueles que compraram mais vezes. Nesse cenário, o *Gaussian Naive Bayes*, foi o que obteve o melhor êxito, sendo o destaque em três das cinco ações avaliadas como melhor *ROI*, onde em uma delas também obteve o melhor *score*.

O *Decision Tree* e o *K-Nearest Neighbors* apesar de cada um, uma vez terem sido os destaques com *score*, nunca se destacaram com o *ROI*. Assim, pode-se concluir que mesmo com acerto em cerca de metade dos dias em que analisou, a outra metade que foram preditas erradas, foram erros bem mais significativos, diferente do *Gaussiann Naive Bayes* que por exemplo, na Petrobras, obteve o menor *score*, porém o maior *ROI*. Ou seja, apesar de ter errado mais ao classificar se deveria comprar ou vender, os momentos em que acertou foram mais significativos do que os que foram preditos de forma errada.

Conclusão sobre o período de pandemia

Na tabela 3 é possível notar que nas ações da Vale que todos os algoritmos obtiveram um score bem próximos, comportamento parecido com o que ocorreu no cenário pré pandemia. Desta vez, dois algoritmos se destacaram com o mesmo *score*, que foram o *Decision Tree* e o *Random Forest*. O *Random Forest* também obteve o maior *ROI*, seguindo o mesmo padrão no pré pandemia onde o mesmo algoritmo que obteve o maior *score*, também obteve o maior *ROI*.

Na tabela 4 é possível notar que nas ações da Petrobras, pela primeira vez aconteceu a situação do maior *ROI* coincidir com a maior Lucratividade, que foi com o *Gaussian Naive Bayes*. O mesmo que obteve o maior *ROI* no cenário pré pandemia, mas o maior score ficou com o *K-Nearest Neighbors*, que obteve o terceiro melhor resultado em relação ao *ROI* e a Lucratividade.

	Pré Pandemia				Durante Pandemia			
Algoritmo	***Score***	**Lucratividade**	***ROI***	**Valor investido**	***Score***	**Lucratividade**	***ROI***	**Valor investido**
Gaussian Naive Bayes	0.515	285.42	13.68%	2086.00	0.512	285.42	13.68%	2086.00
Decision Tree	0.514	1162.72	11.05%	10526.70	0.514	1107.36	8.16%	13576.62
K-Nearest Neighbors	0.514	1254.75	10.32%	12160.01	0.506	1445.12	7.39%	19546.27
Random Forest	0.514	936.29	9.78%	9574.16	0.514	700.61	13.01%	5383.40

Tabela 3: Primeiro experimento, Vale (VALE3)

	Pré Pandemia				Durante Pandemia			
Algoritmo	***Score***	**Lucratividade**	***ROI***	**Valor investido**	***Score***	**Lucratividade**	***ROI***	**Valor investido**
Gaussian Naive Bayes	0.475	272.43	32.92%	827.49	0.510	1062.96	14.87%	7149.93
Decision Tree	0.483	1144.61	24.45%	4681.45	0.483	696.93	13.88%	5021.65
K-Nearest Neighbors	0.508	961.07	23.18%	4145.97	0.514	524.20	12.70%	4126.95
Random Forest	0.520	840.27	25.80%	3257.34	0.497	338.99	11.69%	2900.48

Tabela 4: Primeiro experimento, Petrobras (PETR4)

	Pré Pandemia				Durante Pandemia			
Algoritmo	***Score***	**Lucratividade**	***ROI***	**Valor investido**	***Score***	**Lucratividade**	***ROI***	**Valor investido**
Gaussian Naive Bayes	0.491	364.80	15.77%	2312.93	0.528	-512.97	-15.13%	3390.58
Decision Tree	0.497	287.85	19.29%	1492.11	0.532	-73.63	-1.30%	5646.47
K-Nearest Neighbors	0.512	787.63	14.05%	5604.62	0.516	-436.71	-7.57%	5766.43
Random Forest	0.514	341.80	19.43%	1759.23	0.532	-165.94	-2.99%	5552.19

Tabela 5: Primeiro experimento, Itaú (ITUB4)

	Pré Pandemia				Durante Pandemia			
Algoritmo	***Score***	**Lucratividade**	***ROI***	**Valor investido**	***Score***	**Lucratividade**	***ROI***	**Valor investido**
Gaussian Naive Bayes	0.457	762.62	15.69%	4859.60	0.504	-496.29	-24.95%	1989.24
Decision Tree	0.495	656.38	18.70%	3509.47	0.489	-38.87	-0.83%	4697.02
K-Nearest Neighbors	0.508	970.35	16.42%	5910.29	0.506	-108.89	-3.07%	3542.55
Random Forest	0.500	531.64	23.19%	2292.87	0.489	-208.01	-4.56%	4560.06

Tabela 6: Primeiro experimento, Bradesco (BBDC4)

	Pré Pandemia				Durante Pandemia			
Algoritmo	***Score***	**Lucratividade**	***ROI***	**Valor investido**	***Score***	**Lucratividade**	***ROI***	**Valor investido**
Gaussian Naive Bayes	0.475	290.64	65.47%	443.95	0.552	259.63	17.13%	1515.36
Decision Tree	0.508	873.85	51.95%	1682.15	0.530	515.55	27.53%	1872.48
K-Nearest Neighbors	0.491	811.23	48.08%	1687.39	0.500	172.48	8.64%	1997.11
Random Forest	0.467	810.73	57.25%	1416.24	0.532	512.08	28.21%	1815.20

Tabela 7: Primeiro experimento, B3 (B3SA3)

Na tabela 5 é possível notar que nas ações do Itaú, aconteceu de todos os algoritmos coincidirem em presenciar um caso de prejuízo, todos algoritmos tiveram um *ROI* negativo, ou seja, não foi recuperado o dinheiro investido, o *Decision Tree* e o *Random Forest*, que ambos são árvores de decisões, foram os menos impactados, onde conseguiram manter o mesmo *score*, e as menores taxas de perda, vale ressaltar que *Random Forest* foi também no cenário pré-pandemia o que teve o melhor *score* e melhor *ROI*.

Na tabela 6 é possível notar que nas ações do Bradesco que também ocorre um cenário de perda, onde todos algoritmos tiveram resultados negativos, provocando prejuízo, o *Decision Tree* foi o menos impactado, onde obteve um *ROI* negativo de -0,83%, enquanto o *Gaussian Naive Bayes* teve de -24.95%, o melhor score foi do *K-Nearest Neighbors*, que coincidentemente foi o melhor *score* também no período pré

pandemia.

Na tabela 7 é possível notar que nas ações da B3, se obteve o melhor *ROI*, assim como havia acontecido pré pandemia. O *Gaussian Naive Bayes* teve o melhor *score*, enquanto o *Random Forest*, obteve o melhor *ROI*, apesar do melhor *score* não ter sido também do *Random Forest*, ele foi o segundo melhor.

Assim, nota-se que durante a pandemia não houve um bom lucro, porém os algoritmos nas ações da B3, conseguiram manter resultados aceitáveis, enquanto com Itaú e Bradesco tiveram resultados ruins.

Destaca-se de forma negativa o *K-Nearest Neighbors* que nenhum momento obteve um *ROI* satisfatório, e esteve entre os algoritmos que tiveram maior investimento, e menos lucratividade.

O *Decision Tree* e o *Random Forest*, para esse cenário durante pandemia foram o que melhor satisfizeram, visto que o *Random Forest* duas vezes obteve o melhor *ROI*, o *Decision tree* das cinco ações em duas não obteve o melhor *ROI* mas esteve próximo, e no caso do Itaú e Bradesco foi o que teve menos prejuízo relacionado aos demais.

O *Gaussian Naive Bayes* obteve uma vez o melhor *ROI*, ao mesmo tempo que também teve a melhor lucratividade, porém no caso Itaú e Bradesco que foram mais críticos, nas duas vezes ele teve o pior prejuízo relacionado aos outros três.

Para o segundo experimento, foram utilizados algoritmos de regressão para ver acurácia dos mesmos, realizando uma predição dos valores de fechamento, em períodos de dois anos, seja antes ou durante a pandemia. Dessa vez, foram escolhidos o *Linear Regression*, *Random Forest*, *Decision Tree* e *Gradient Boosting* para realizar a predição do valor de fechamento das ações no dia seguinte. Como métricas de avaliação de desempenho foram usadas MAE, MSE e MAD, (tabela 8), onde são métricas relativas a erros, ou seja, quão menores forem esses valores, mais assertivos esses algoritmos estão, todas as métricas seguem uma linearidade dentro do experimento, dessa forma não existe uma métrica mais importante que a outra, visto que diferente do experimento, o algorítimo que tiver a menor métrica em uma das três, também haverá as menores métricas na outras duas, ou aproximadamente.

Conclusão sobre o período pré-pandemia

Nitidamente ficou claro em todos os gráficos, figuras 4 a 43, quanto pelas tabelas de 9 a 13 que a *Linear Regression* se destacou em todas as métricas de avaliação.

Na vale, Petrobras, e B3, tabelas 9, 10, 13 e figuras 5, 13, 37, o *Random forest* obteve os segundos melhores resultados, e logo em seguida sempre acompanhado pelo *Decision Tree*.

Porém no Itaú e Bradesco (ver Tabelas 11 e 12), todos os algoritmos com exceção do *Linear Regression*, tiveram resultados bem discrepantes relacionados aos que deveriam ter se aproximado (ver Figuras 21, 22, 23, 29, 30, 31, 29)

Conclusão sobre o período de pandemia

O *Linear Regression* se destacou novamente, acompanhado do *Random Forest* como o segundo algoritmo com melhores taxas de assertividade. O *Decision Tree* em alguns momentos trocou de posição com *Gradient Boosting*, que por sua vez apenas durante a pandemia que alguma vez deixou de ser o algoritmo com pior métrica relacionado aos demais.

Ficou claro que o *Linear Regression* foi o único que tanto antes quanto durante a pandemia teve melhor pontuação, e por diversas vezes foi melhor que os outros algoritmos, analisando os gráficos, diversas vezes o *Random Forest*, *Decision Tree* e o *Gradient Boosting* chegaram a um teto onde não conseguiram mais ultrapassar, figuras 9, 10, 11, 21, 22, 23, 29, 30, 31, 37, 38, 39, 41, 42 e 43, fazendo com que não fosse atingido proximidade com os valores alvos, explorando o *dataset* utilizado durante o treino, percebe-se que esses valores máximos onde os algoritmos não conseguiram ultrapassar, estão iguais ou próximos dos picos que foram encontrados durante os treinos, ou seja durante os testes, as ações tiveram valores nunca antes vistas, fazendo com que os modelos não conseguissem prevê-los, pois eram desconhecidas aquelas possibilidades para os mesmo.

Este comportamento ocorre pois, por exemplo, o *Random Forest* nao consegue prever valores que extrapolam os valores obtidos durante o treinamento, isso ocorre pois ela trabalha com uma estrutura de árvores, onde o princípio diz que árvores possuem nós, onde nesses nós são tomadas decisões de ou seguir por um caminho, ou pelo outro, onde esses nós são medidas diretamente relacionadas ao valores obtidos durante o treino, desta a forma nunca existe um caminho para um valor além dos valores aprendidos durante o treinamento, o mesmo bloqueio acontece para os demais.

Devido a natureza do *Linear Regression*, que é gerar uma função qual busca uma linearidade entre o valor alvo e o conjunto de dados de entrada, conforme surgem novos valores, mesmo que não tenham sido presenciados durante a fase de de treino, a fórmula é capaz de mesmo assim encontrar essas linearidades.

Imagens dos experimentos.

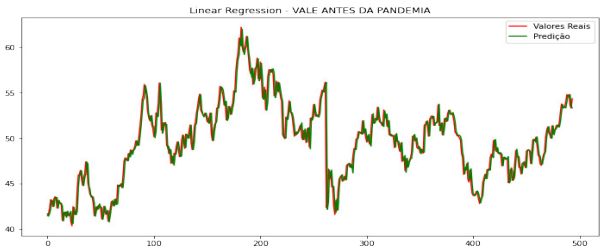


Figura 4: *Linear Regression* - Vale antes da pandemia

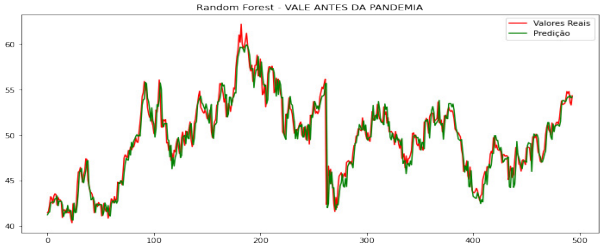


Figura 5: *Random Forest* - Vale antes da pandemia

Nome	Descrição
MSE	Provavelmente a métrica mais popular, de forma simplória este erro é similar ao erro absoluto onde é feita a soma cumulativa dos erros, porém, com uma única diferença, cada valor é elevado ao quadrado antes da soma. Desta forma é feita uma penalização em erros maiores.
MAD	O Erro mediano funciona como na estatística, dos erros computados, este erro é resistente a anomalias, ignora pontos mais extremos e privilegia manter a forma geral da distribuição.
MAE	É o erro mais básico e intuitivo quando lidamos com regressão, que nada mais é que a média do erro que cada ponto tem em relação à linha de regressão.

Tabela 8: Tabela com colunas das métricas de avaliação de desempenho

	Pré Pandemia			Durante Pandemia		
Algoritmo	**MSE**	**MAD**	**MAE**	**MSE**	**MAD**	**MAE**
Linear Regression	1.35	0.57	0.77	3.65	1.10	1.43
Random Forest	1.54	0.72	0.87	796.01	14.15	20.14
Decision Tree	3.62	1.36	1.52	794.38	14.13	20.11
Gradient Boosting	2.76	1.11	1.31	861.45	16.05	21.10

Tabela 9: Segundo experimento, Vale (VALE3)

	Pré Pandemia			Durante Pandemia		
Algoritmo	**MSE**	**MAD**	**MAE**	**MSE**	**MAD**	**MAE**
Linear Regression	0.39	0.36	0.46	0.62	0.36	0.52
Random Forest	0.55	0.44	0.56	0.68	0.40	0.57
Decision Tree	1.03	0.72	0.81	1.22	0.60	0.79
Gradient Boosting	2.28	1.31	1.33	0.68	0.43	0.55

Tabela 10: Segundo experimento, Petrobras (PETR4)

	Pré Pandemia			Durante Pandemia		
Algoritmo	**MSE**	**MAD**	**MAE**	**MSE**	**MAD**	**MAE**
Linear Regression	0.33	0.36	0.45	0.49	0.42	0.51
Random Forest	24.35	4.58	4.28	0.73	0.50	0.65
Decision Tree	23.93	4.48	4.29	1.36	0.73	0.89
Gradient Boosting	25.42	4.66	4.39	1.24	0.82	0.91

Tabela 11: Segundo experimento, Itaú (ITUB4)

	Pré Pandemia			Durante Pandemia		
Algoritmo	**MSE**	**MAD**	**MAE**	**MSE**	**MAD**	**MAE**
Linear Regression	0.28	0.34	0.41	0.44	0.39	0.51
Random Forest	26.87	4.06	4.07	0.84	0.57	0.69
Decision Tree	25.05	3.77	3.95	1.04	0.61	0.78
Gradient Boosting	30.21	4.49	4.36	2.34	2.34	2.36

Tabela 12: Segundo experimento, Bradesco (BBDC4)

	Pré Pandemia			Durante Pandemia		
Algoritmo	**MSE**	**MAD**	**MAE**	**MSE**	**MAD**	**MAE**
Linear Regression	0.06	0.14	0.19	0.26	0.29	0.38
Random Forest	11.25	1.16	2.34	5.66	1.24	1.76
Decision Tree	14.84	1.31	2.68	5.75	1.24	1.79
Gradient Boosting	32.63	1.49	3.80	5.87	1.23	1.80

Tabela 13: Segundo experimento, B3 (B3SA3)

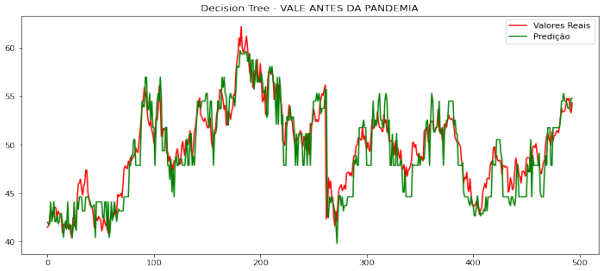


Figura 6: *Decision Tree* - Vale antes da pandemia

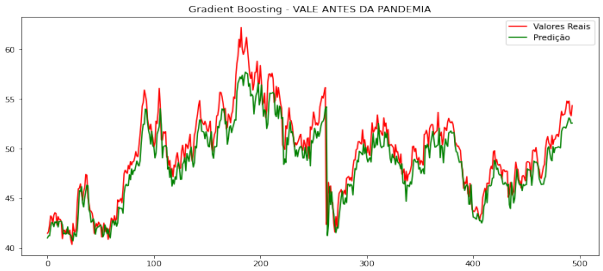


Figura 7: *Gradient Boosting* - Vale antes da pandemia

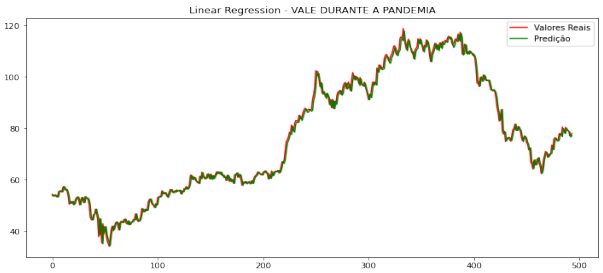


Figura 8: *Linear Regression* - Vale durante a pandemia

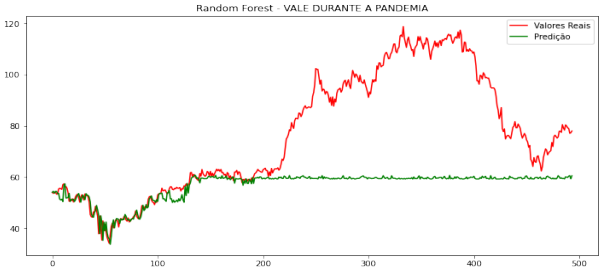


Figura 9: *Random Forest* - Vale durante a pandemia

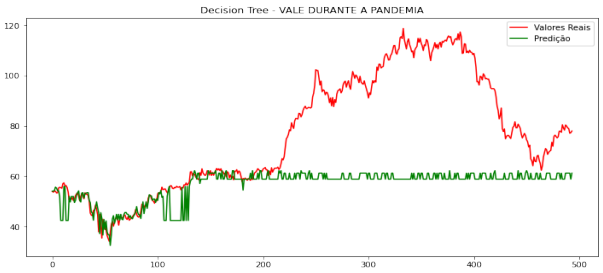


Figura 10: *Decision Tree* - Vale durante a pandemia

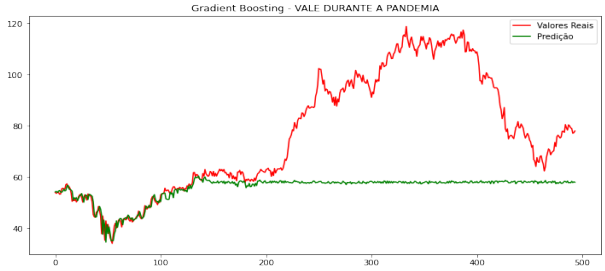


Figura 11: *Gradient Booster* - Vale durante a pandemia

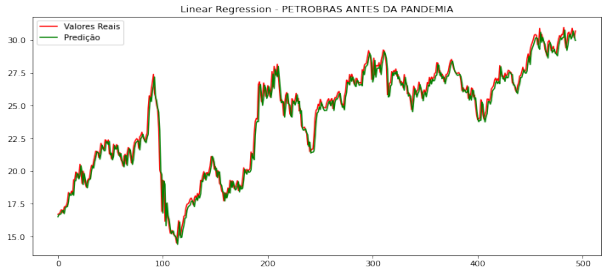


Figura 12: *Linear Regression* - Petrobras antes da pandemia

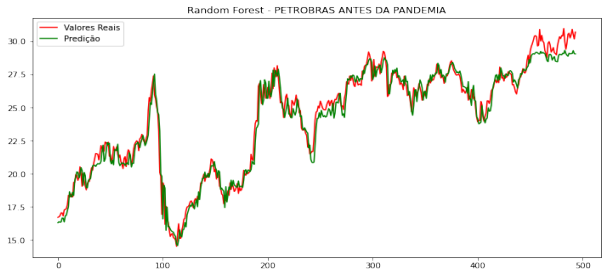


Figura 13: *Random Forest* - Petrobras antes da pandemia

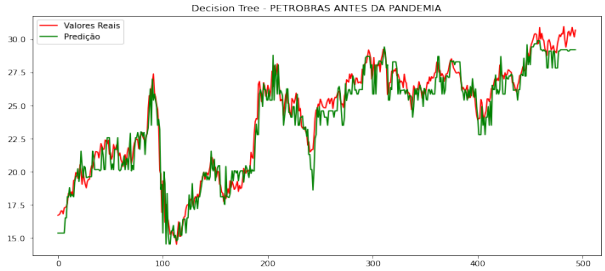


Figura 14: *Decision Tree* - Petrobras antes da pandemia

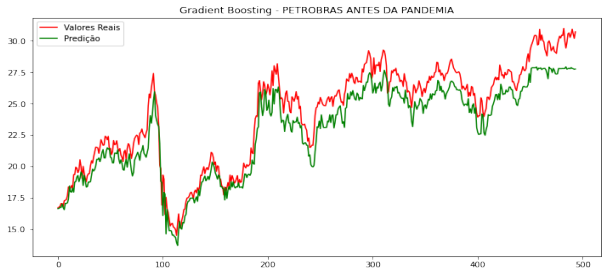


Figura 15: *Gradient Boosting* - Petrobras antes da pandemia

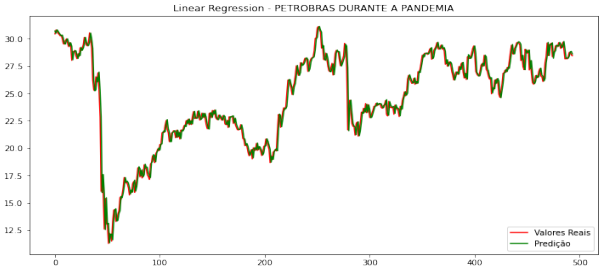


Figura 16: *Linear Regression* - Petrobras durante a pandemia

Figura 17: *Random Forest* - Petrobras durante a pandemia

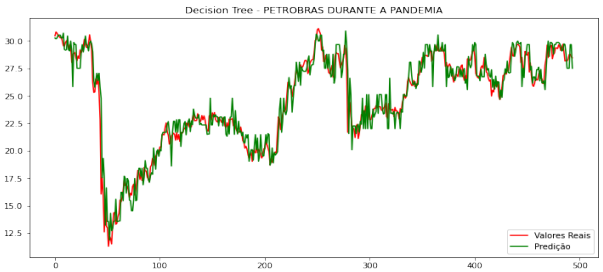


Figura 18: *Decision Tree* - Petrobras durante a pandemia

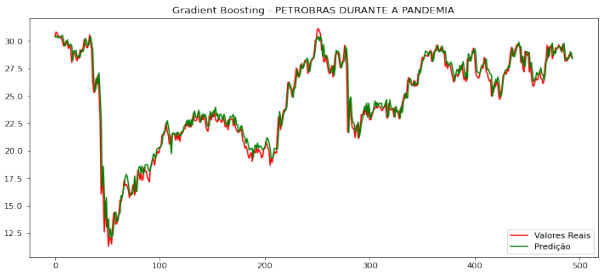


Figura 19: *Gradient Boosting* - Petrobras durante a pandemia

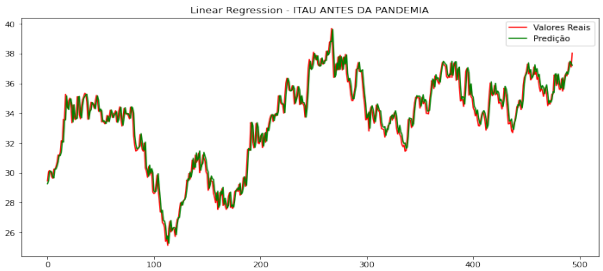


Figura 20: *Liner Regression* - Itaú antes da pandemia

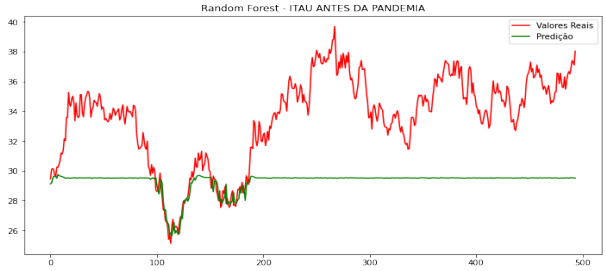


Figura 21: *Random Forest* - Itaú antes da pandemia

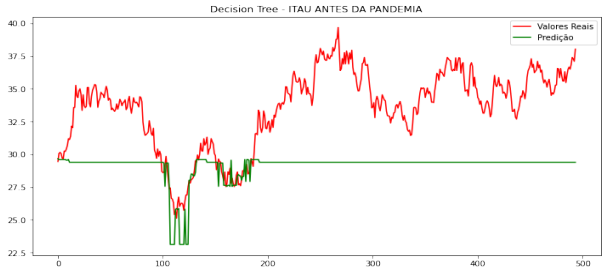


Figura 22: *Decision Tree* - Itaú antes da pandemia

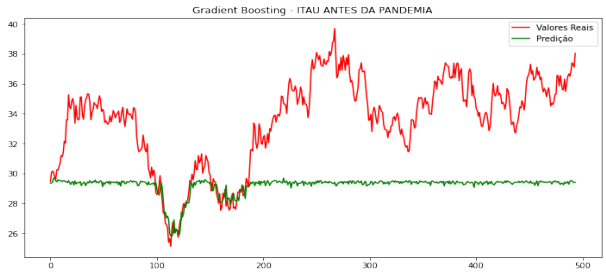


Figura 23: *Gradient Boosting* - Itaú antes da pandemia

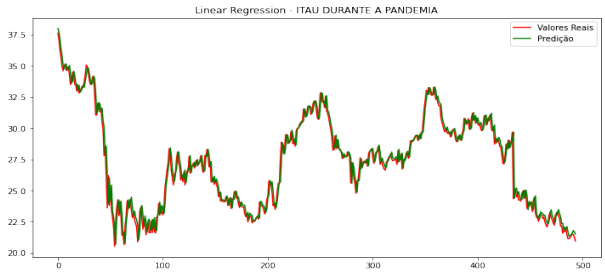


Figura 24: *Linear Regression* - Itaú durante a pandemia

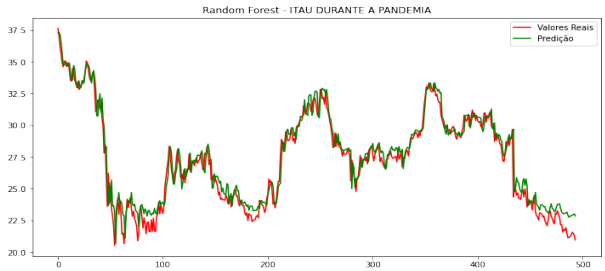


Figura 25: *Random Forest* - Itaú durante a pandemia

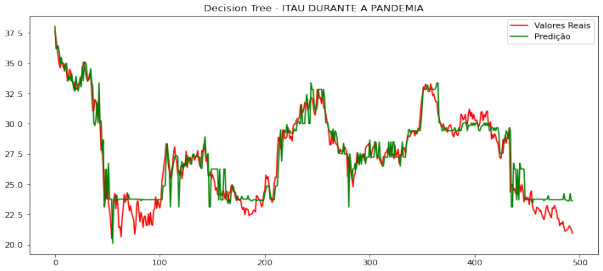


Figura 26: *Decision Tree* - Itaú durante a pandemia

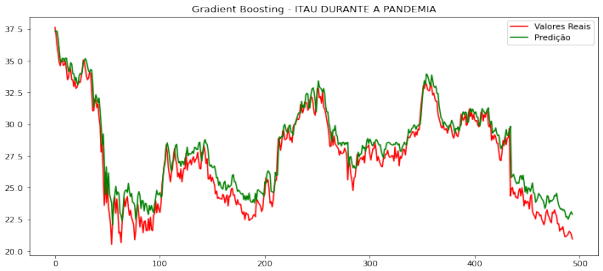


Figura 27: *Gradient Boosting* - Itaú durante a pandemia

Figura 28: *Linear Regression* - Bradesco antes da pandemia

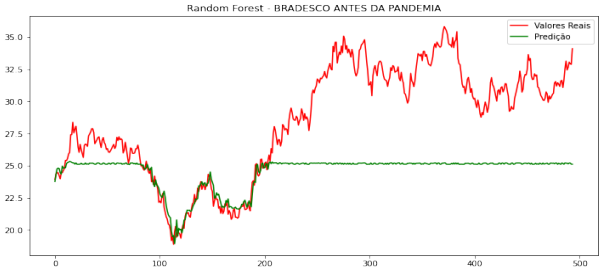


Figura 29: *Random Forest* - Bradesco antes da pandemia

Figura 30: *Decision Tree* - Bradesco antes da pandemia

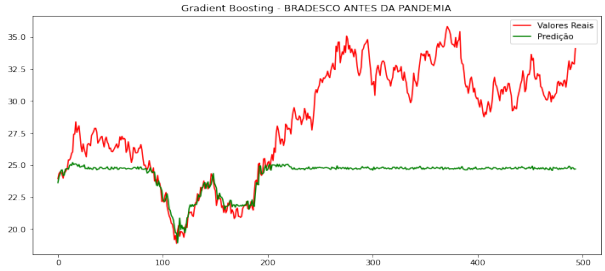


Figura 31: *Gradient Boosting* - Bradesco antes da pandemia

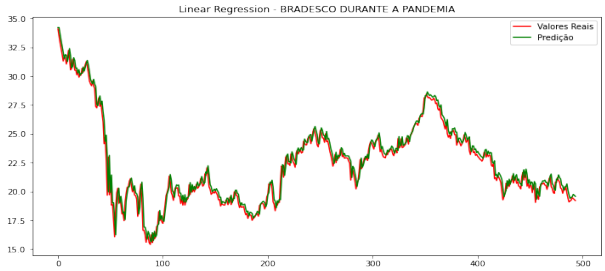


Figura 32: *Linear Regression* - Bradesco durante a pandemia

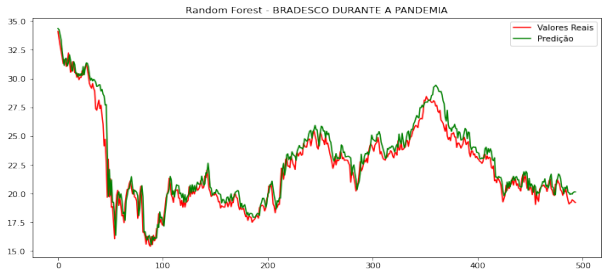


Figura 33: *Random Forest* - Bradesco durante a pandemia

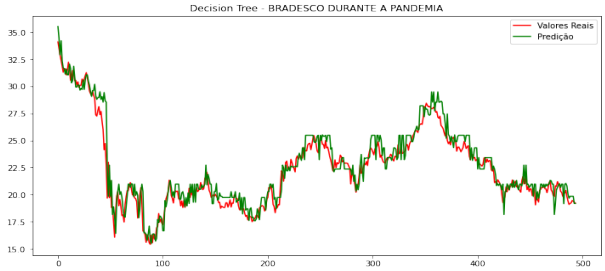


Figura 34: *Decision Tree* - Bradesco durante a pandemia

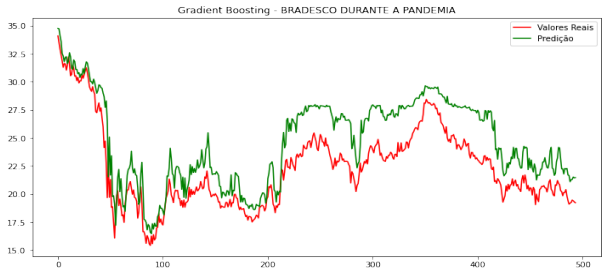


Figura 35: *Gradient Boosting* - Bradesco durante a pandemia

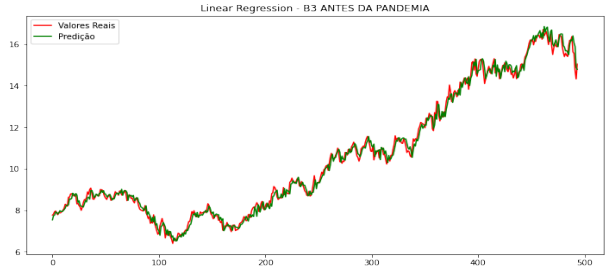


Figura 36: *Linear Regression* - B3 antes da pandemia

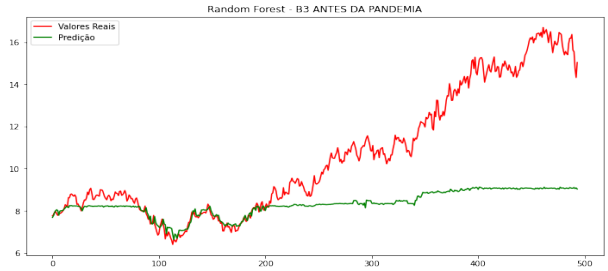


Figura 37: *Random Forest* - B3 antes da pandemia

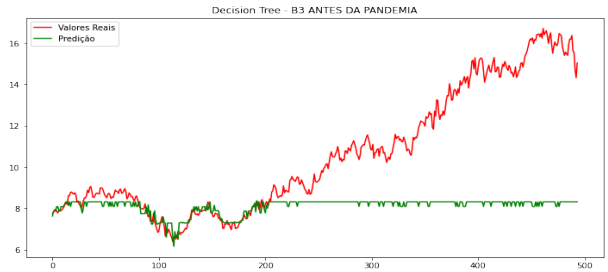


Figura 38: *Decision Tree* - B3 antes da pandemia

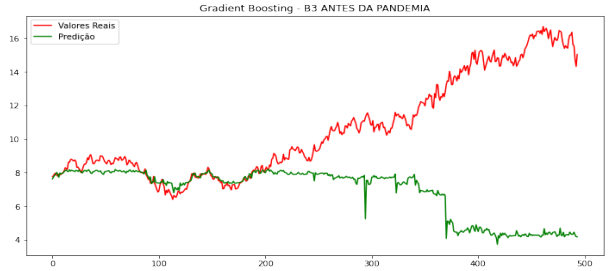


Figura 39: *Gradient Boosting* - B3 antes da pandemia

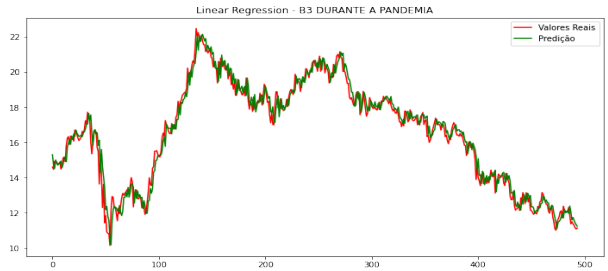


Figura 40: *Linear Regression* - B3 durante a pandemia

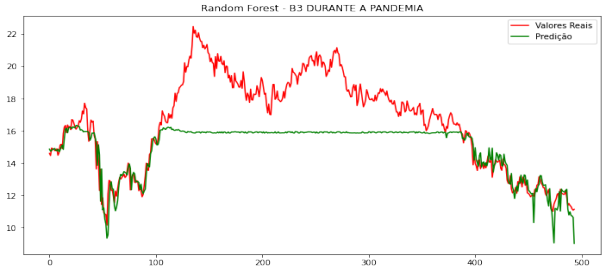


Figura 41: *Random Forest* - B3 durante a pandemia

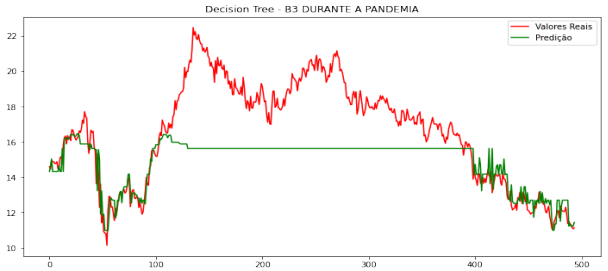


Figura 42: *Decision Tree* - B3 durante a pandemia

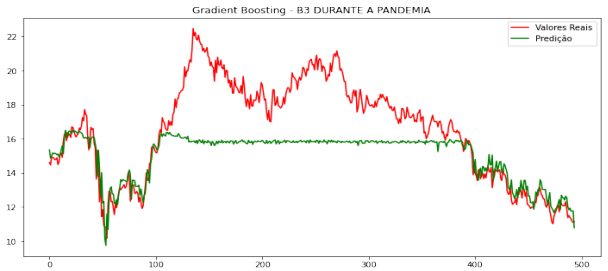


Figura 43: *Gradient Boosting* - B3 durante a pandemia

VI. Conclusões e Trabalhos Futuros

Este trabalho trouxe como intuito verificar o desempenho de seis algoritmos de baixo consumo computacional para realizar predições das seguintes ações da bolsa de valores B3 (B3SA3), Vale (VALE3), Petrobras (PETR4), Itaú (ITUB4) e Bradesco (BBDC4), no período pré-pandemia e durante a pandemia.

No primeiro experimento, três algoritmos se destacaram, o *Gaussian Naive Bayes*, *Random Forest* e o *Decision tree*. No período pré-pandemia, o *Gaussian Naive Bayes* e o *Random forest* apresentaram o melhor *ROI*, enquanto durante a pandemia, novamente o *Random Forest* e o *Decision Tree* se destacaram, onde ambos conseguiram minimizar o prejuízo com mais precisão. Pode-se concluir que os modelos baseados em árvores conseguiram ter um bom desempenho, e o Random Forest que tende de ser um avanço do *Decision Tree* conseguiu se destacar em ambos os cenários, interessante notar que o algoritmo *Gaussian Naive Bayes* que enquanto as ações seguiam uma linearidade e poucas oscilações, ele se comportou muito bem, mas ao colocá-lo durante a pandemia, onde ocorreram comportamentos fora do comum, ele não foi adaptável.

No segundo experimento ficou claro como o *Linear Regression* se destacou, pois teve bom desempenho em todas as

ações no período pré-pandemia. Os demais algoritmos, apesar de possuírem suas limitações, quando o valor do fechamento estava dentro de intervalo de preço existente no conjunto de treinamento, também obtiveram bons resultados. Porém, quando o preço da ação atingiu valores não existentes no conjunto de treinamento, o desempenho dos algoritmos foram insatisfatorios.

Como trabalho futuro, pode-se testar o desempenho de algoritmos mais robustos, como redes neurais profundas. Também pode-se observar o desempenho dos algorirmos "operando vendido", que é quando o investidor não tem a ação, adquire em forma de empréstimo e a vende, com expectativa de desvalorização para poder comprá-la mais barata, possibilitando a obtenção de lucros.

Outra possibilidade de trabalho futuro é fazer uma divisão de dados diferentes, aplicando o conceito de *growing window*. Nesse caso, os modelos são retreinados diariamente, recebendo os dados dados referente ao dia corrente para que possa fazer a predição do dia seguinte. Nessa pesquisa, foi utilizada a técnica *fixed window*, no qual foi definida uma janela de treinamento fixa. Com essa mudança, é possível que o problema que os algoritmos tiveram em não conseguir prever valores não existentes no conjunto de treinamento seja resolvido.

REFERÊNCIAS

BRASIL, E. 2021. Disponível em: ⟨https://investnews.com.br/cafeina/as-acoes-mais-negociadas-sao-as-que-rentabilizam-melhor⟩.
COPELAND, B. 2020. Disponível em: ⟨https://www.britannica.com/technology/artificial-intelligence⟩.
FINANCE, Y. 2022. Disponível em: ⟨https://finance.yahoo.com⟩.
HOPPEN, W. P. J. 2018. Disponível em: ⟨https://www.aquare.la/datasets-o-que-sao-e-como-utiliza-los⟩.
HOUSE, D. 2021. Disponível em: ⟨https://www.digitalhouse.com/br/blog/naive-bayes⟩.
IBM. Disponível em: ⟨https://www.ibm.com/br-pt/analytics/learn/linear-regression⟩.
LUZ, F. 2019. Disponível em: ⟨https://inferir.com.br/artigos/algoritimo-knn-para-classificacao⟩.
MOREIRA, F. 2022. Disponível em: ⟨https://www.infomoney.com.br/mercados/b3-b3sa3-numero-de-investidores-pessoa-fisica-sobe-437-em-marco-volume-medio-diario-em-acoes-cai-126⟩.
NOGUEIRA, M. O. d. L. Gabriel de O. G. 2021. Disponível em: ⟨https://arxiv.org/abs/2108.10065⟩.
NOMAD. 2022. Disponível em: ⟨https://nomadglobal.com/blog/estrategias-investimento⟩.
OLIVEIRA, A. C. d. 2021. Disponível em: ⟨https://repositorio.unisagrado.edu.br/jspui/handle/handle/166⟩.
PEERCHEMIST. 2021. Disponível em: ⟨https://pypi.org/project/finta/⟩.
PESSANHA, C. 2019. Disponível em: ⟨https://medium.com/cinthiabpessanha/random-forest-como-funciona-um-dos-algoritmos-mais-populares-de-ml-cc1b8a58b3b4⟩.
REIS, T. 2020. Disponível em: ⟨https://www.suno.com.br/guias/bolsa-de-valores⟩.
RICO, T. 2021. Disponível em: ⟨https://riconnect.rico.com.vc/blog/investir-na-bolsa-de-valores⟩.
RICO, T. 2022. Disponível em: ⟨https://riconnect.rico.com.vc/blog/swing-trade⟩.
SACRAMENTO, G. 2021. Disponível em: ⟨https://blog.somostera.com/data-science/arvores-de-decisao⟩.
SANTOS, G. C. 2020. Disponível em: ⟨https://repositorio.ufu.br/handle/123456789/29897⟩.
SILVA, J. 2020. Disponível em: ⟨https://medium.com/equals-lab/uma-breve-introduo-ao-algoritmo-de-machine-learning-gradient-boosting-utilizando-a-biblioteca-311285783099⟩.
SILVA, M. F. 2022. Disponível em: ⟨https://repositorio.ufu.br/handle/123456789/35138⟩.
SIMON, P. 2013. Disponível em: ⟨https://books.google.com.br/books?id=Dn-Gdoh66sgC&pg=PA89&redir_esc=y\#v=onepage&q&f=false⟩.
STEFFEN, A. A. 2021. Disponível em: ⟨http://repositorio.ufgd.edu.br/jspui/handle/prefix/4805⟩.

	Pré Pandemia	Durante Pandemia
Ações	***var smoothing***	***var smoothing***
Vale	0.351	0.811
Petrobras	0.811	0.811
Itaú	1.000	0.187
Bradesco	0.187	0.010
B3	1.000	0.010

Tabela 14: Hiperparâmetros - *Gaussian Naive Bayes*

	Pré Pandemia		Durante Pandemia	
Ações	***criterion***	***min samples leaf***	***criterion***	***min samples leaf***
Vale	gini	4	gini	1
Petrobras	entropy	1	gini	1
Itaú	entropy	2	gini	1
Bradesco	entropy	2	entropy	2
B3	gini	2	gini	4

Tabela 15: Hiperparâmetros - *Decision Tree*

	Pré Pandemia		Durante Pandemia	
Ações	***leaf size***	***n neighbors***	***leaf size***	***n neighbors***
Vale	1	1	1	1
Petrobras	1	8	1	8
Itaú	1	8	1	5
Bradesco	1	7	1	4
B3	1	6	1	2

Tabela 16: Hiperparâmetros - *K-Nearest Neighbors*

	Pré Pandemia		Durante Pandemia	
Ações	***min samples leaf***	***n estimators***	***min samples leaf***	***n estimators***
Vale	2	2	3	1
Petrobras	4	1	3	1
Itaú	2	1	2	3
Bradesco	2	1	1	2
B3	2	1	1	2

Tabela 17: Hiperparâmetros - *Random Forest*

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus Campina Grande
R. Tranqüilino Coelho Lemos, 671, Dinamérica, CEP 58432-300, Campina Grande (PB)
CNPJ: 10.783.898/0003-37 - Telefone: (83) 2102.6200

Documento Digitalizado Ostensivo (Público)

Trabalho de conclusão de curso

Assunto: Trabalho de conclusão de curso
Assinado por: Gabriel Lima
Tipo do Documento: Projeto
Situação: Finalizado
Nível de Acesso: Ostensivo (Público)
Tipo do Conferência: Cópia Simples

Documento assinado eletronicamente por:

- **Gabriel de Lima e Silva, ALUNO (201821250037) DE BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO - CAMPINA GRANDE,** em 23/09/2022 09:39:37.

Este documento foi armazenado no SUAP em 23/09/2022. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 633100
Código de Autenticação: eed93412aa