Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Departamento de Computação e Sistemas

Classificação de ações da bolsa de valores mediante técnicas de mineração de dados

Endhel Lopes de Freitas

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Janniele Aparecida Soares Araújo

Junho, 2022

João Monlevade–MG

Endhel Lopes de Freitas

Classificação de ações da bolsa de valores mediante técnicas de mineração de dados

Orientador: Janniele Aparecida Soares Araújo

Monografia apresentada ao curso de Sistemas de Informação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto, como requisito parcial para aprovação na Disciplina "Trabalho de Conclusão de Curso II".

Universidade Federal de Ouro Preto

João Monlevade

Junho de 2022

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

F866c Freitas, Endhel Lopes de.
Classificação de ações da bolsa de valores mediante técnicas de mineração de dados. [manuscrito] / Endhel Lopes de Freitas. - 2022.
66 f.: il.: color., gráf., tab..

Orientadora: Profa. Dra. Janniele Aparecida Soares Araujo.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Sistemas de Informação .

1. Ações (Finanças). 2. Bolsa de valores. 3. Classificação - Algoritmos computacionais. 4. Mercado de capitais. 5. Mineração de dados (Computação). I. Araujo, Janniele Aparecida Soares. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.62

Bibliotecário(a) Responsável: Flavia Reis - CRB6-2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

FOLHA DE APROVAÇÃO

Endhel Lopes de Freitas

Classificação de ações da bolsa de valores mediante técnicas de mineração de dados

Monografia apresentada ao Curso de Sistemas de Informação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação

Aprovada em 24 de junho de 2022

Membros da banca

Dra - Janniele Aparecida Soares Araujo - Orientadora (Universidade Federal de Ouro Preto)
Dr - Fernando Bernardes de Oliveira - (Universidade Federal de Ouro Preto)
Dra - Helen de Cassia Sousa da Costa Lima - (Universidade Federal de Ouro Preto)

Janniele Aparecida Soares Araujo, orientadora do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 02/09/2022

Documento assinado eletronicamente por **Janniele Aparecida Soares Araujo**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 02/09/2022, às 19:32, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0392984** e o código CRC **0C980702**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: (31)3808-0819 - www.ufop.br

Este trabalho é dedicado a minha mãe, Vanilza, e a minha namorada, Daniele. Elas são a minha base e minha fortaleza, e esta conquista também é delas.

Agradecimentos

Agradeço à minha mãe, Vanilza, pelo apoio, incentivo, e por toda a sua luta para me proporcionar uma melhor educação. Por todo o suporte financeiro e emocional que me manteve firme nesta luta e por todo o amor, carinho e criação, que foram cruciais para eu me tornar quem sou hoje. Sem ela eu nunca teria chegado até aqui.

Agradeço à minha namorada, Daniele, que me manteve acreditando neste sonho. Por ser a melhor companheira que eu poderia escolher, por sempre estar presente nos momentos em que precisei, por me incentivar, me suportar e me ajudar em todas as coisas. Por sempre acreditar em mim, por todo amor e carinho, assim como pela compreensão e paciência durante os momentos em que não pude me dedicar a ela.

Agradeço ao meu irmão, Enzo, por todos os conselhos, contribuições e palavras motivadoras. Pela amizade e cooperação que me permitiu continuar estudando e me dedicando a este curso.

Agradeço à minha orientadora Prof[a]. Dra. Janniele Araújo, por ter aceitado me guiar neste trabalho. Sou muito grato pelos ensinamentos, paciência e dedicação oferecidas durante este período.

Por fim, agradeço a todos os meus amigos e familiares que contribuiram de forma direta ou indireta para a realização deste trabalho de conclusão de curso.

"Science is more than a body of knowledge; it is a way of thinking."

— Carl Sagan (1934 – 1996),
in: The Demon-Haunted World: Science as a Candle in the Dark.

Resumo

Com a popularização do mercado de capitais e a crescente quantidade de empresas ofertando as suas ações na bolsa de valores, o número de investidores tem aumentado cada vez mais no Brasil. Devido à alta volatilidade das ações e à falta de conhecimento sobre o mercado e sobre os dados de balanço patrimonial e demonstrativo de resultados divulgados pelas empresas de capital aberto, muitas pessoas não realizam bons investimentos e adquirem prejuízos financeiros. Dado este problema, o objetivo deste trabalho é criar um *bot* no *Telegram* que preveja se uma ação irá render 3% ou mais até o próximo trimestre, a fim de apoiar a decisão dos investidores. Para tal fim, foram coletados dados de empresas listadas na bolsa de valores brasileira, do período de junho de 2011 a setembro de 2021, e foram treinados alguns algoritmos de classificação, a fim de encontrar o modelo com o melhor desempenho. O modelo com a melhor performance foi a *Random Forest*, tendo obtido 60% de precisão ao classificar ações que iriam render, conforme o critério estabelecido. Os resultados mostraram que a solução criada pode ser uma boa opção para auxiliar as pessoas a realizarem melhores investimentos.

Key-words: mineração de dados. classificação. mercado de ações.

Abstract

With the popularization of the capital market and the growing number of companies offering their shares on the stock exchange, the number of investors has been increasing more and more in Brazil. Due to the high volatility of stocks and the lack of knowledge about the market and the balance sheet and income statement data disclosed by publicly traded companies, many people do not make good investments and acquire financial losses. Given this problem, the objective of this work is to create a bot on Telegram that predicts whether a stock will yield 3% or more by the next quarter, in order to support investors' decision. To this end, data from companies listed on the Brazilian stock exchange were collected from the period June 2011 to September 2021, and some classification algorithms were trained in order to find the best performing model. The model with the best performance was the Random Forest, having obtained 60% accuracy when classifying stocks that would yield, according to the established criteria. The results showed that the solution created can be a good option to help people make better investments.

Key-words: data mining. classification. stock market.

Lista de ilustrações

Figura 1 – Hipótese básica da análise fundamentalista 20
Figura 2 – O Processo de Mineração de Dados CRISP-DM 25
Figura 3 – Exemplo de árvore de decisão 28
Figura 4 – Sigmoide (Função Logística) 29
Figura 5 – Rede neural de múltiplas camadas (a), sentido de propagação do sinal de entrada e retropropagação do erro (b) 31
Figura 6 – Representação do funcionamento de uma Floresta Aleatória 33
Figura 7 – Análise da Variável Resposta 41
Figura 8 – Análise de Correlações 42
Figura 9 – Método para classificar as ações 44
Figura 10 – Seleção por *Subset* 46
Figura 11 – Novo Conjunto de Dados Criado pelo Boruta 46
Figura 12 – Tabela de Acertos 47
Figura 13 – Resultado após 20 Repetições 47
Figura 14 – Matriz de Confusão 50
Figura 15 – Matriz de Confusão do *Dummy Classifier* 52
Figura 16 – Matriz de Confusão do *Logistic Reression* 53
Figura 17 – Matriz de Confusão do *Decision Tree* 53
Figura 18 – Matriz de Confusão do *Naive Bayes* 54
Figura 19 – Matriz de Confusão do *Random Forest* 55
Figura 20 – Conjuntos de Valores Escolhidos para os Parâmetros do *Random Search* 57
Figura 21 – Matriz de Confusão do Modelo Ajustado 58
Figura 22 – Comparação do Modelo de *Random Forest* Antes e Depois do Ajuste de Hiperparâmetros 59
Figura 23 – Gráficos de Avaliação do Modelo Final 59
Figura 24 – Performance do Modelo Final 60
Figura 25 – Arquitetura do Ambiente de Produção 61
Figura 26 – InvestBot - Bot no Telegram 62

Lista de tabelas

Tabela 1 – Indicadores Fundamentalistas Calculados 40
Tabela 2 – Variáveis Escolhidas pelo Boruta 48
Tabela 3 – Métricas do *Dummy Classifier*. 52
Tabela 4 – Métricas do *Logistic Reression*. 52
Tabela 5 – Métricas do *Decision Tree*. 53
Tabela 6 – Métricas do *Naive Bayes*. 54
Tabela 7 – Métricas do *Random Forest*. 54
Tabela 8 – Performance dos Modelos - *Cross Validation*. 56
Tabela 9 – Combinação de Hiperparâmetros. 57
Tabela 10 – Métricas do Modelo Ajustado. 58
Tabela 11 – Métricas do Modelo Ajustado - *Cross Validation*. 58

Lista de abreviaturas e siglas

API *Application Programming Interface*

B3 Brasil, Bolsa, Balcão

BH *Buy and Hold*

BMF Bolsa de Mercadorias e Futuros

CRISP-DM Processo Padrão de Indústria Cruzada para Exploração de Dados

Cetip Central de Custódia e Liquidação Financeira de Títulos

DY *Dividend Yield*

EBITDA *Earnings Before Interest, Taxes, Depreciation and Amortization*

IPO Oferta Inicial de Ações

KDD *Knowledge Discovery in Databases*

KNN *k-Nearest Neighbors*

LC Liquidez Corrente

LPA Lucro por Ação

ML Margem Líquida

MLP *Multilayer Perceptron*

NA Not Applicable

NB *Naive Bayes*

P/L Preço/Lucro

P/VP Preço/Valor Patrimonial

PAA-IDPSO-CD Aproximação por Valor Agregado de Segmento - Otimização por Enxame de Partículas Auto Adaptativa com detecção de mudança de conceito

PSR *Price Sales Ratio*

RPL Rentabilidade sobre Patrimônio Líquido

SAX-GA Aproximação por Valor Agregado Simbólico - Algoritmos Genéticos

VPA Valor Patrimonial da Ação

Sumário

1 INTRODUÇÃO 13
1.1 Problema 13
1.2 Objetivos 14
1.3 Metodologia 15
1.4 Organização do trabalho 15

2 REVISÃO BIBLIOGRÁFICA 17
2.1 Mercado de Capitais 17
2.1.1 Bolsa de Valores B3 17
2.1.2 Mercado de Ações 18
2.1.3 Análise Fundamentalista 19
2.1.3.1 Principais Componentes do Balanço Patrimonial e Demonstrativo de Resultados 20
2.1.3.2 Principais Indicadores Fundamentalistas 21
2.2 Mineração de Dados 24
2.3 Algoritmos de Classificação 27
2.3.1 *Decision Tree* (Árvore de Decisão) 28
2.3.2 *Logistic Regression* 28
2.3.3 *k-Nearest Neighbors* 30
2.3.4 *Naive Bayes* 30
2.3.5 *Multilayer Perceptron* 31
2.3.6 *Random Forest* 32
2.4 Trabalhos Relacionados 33

3 DESENVOLVIMENTO 36
3.1 Entendimento do Problema 36
3.2 Entendimento dos Dados 36
3.2.1 Balanços e Demonstrativos das Empresas 36
3.2.2 Cotações 37
3.2.3 Tratamento dos Dados 37
3.2.4 Criação dos Rótulos para as Ações 38
3.3 Preparação dos Dados 38
3.3.1 Descrição dos Dados 38
3.3.2 Engenharia de Variáveis 40
3.3.3 Filtragem de Variáveis 40
3.3.4 Análise Exploratória de Dados 41
3.3.4.1 Variável Resposta 41

3.3.4.2 Análise Multivariada . . . 41
3.3.5 Preparação dos Dados . . . 42
3.3.5.1 Transformações dos Dados para Variações Percentuais . . . 43
3.3.5.2 Divisão dos Dados em Treino e Teste . . . 43
3.3.5.3 *Rescaling* . . . 44
3.3.6 *Feature Selection* (Seleção de Variáveis) . . . 45
3.3.6.1 Boruta . . . 45

4 RESULTADOS . . . 49
4.1 Modelagem e Avaliação . . . 49
4.1.1 Métricas de Avaliação . . . 49
4.1.1.1 Matriz de Confusão . . . 49
4.1.1.2 Principais Métricas de Avaliação . . . 50
4.1.1.3 Métricas Utilizadas . . . 51
4.1.2 Modelagem . . . 51
4.1.2.1 *Dummy Classifier* . . . 51
4.1.2.2 *Logistic Regression* . . . 52
4.1.2.3 *Decision Tree* . . . 53
4.1.2.4 *K-Nearest Neighbors*, *Naive Bayes* e *Multilayer Perceptron* . . . 53
4.1.2.5 *Random Forest* . . . 54
4.1.2.6 *Cross Validation* . . . 55
4.1.3 Ajuste de Hiperparâmetros . . . 56
4.1.4 Avaliação do Modelo . . . 59
4.2 *Deployment* (Implementação) . . . 60

5 CONCLUSÃO . . . 63

REFERÊNCIAS . . . 65

1 Introdução

O mercado de capitais é um conjunto de instituições e de instrumentos que negociam com títulos e valores mobiliários, cujo objetivo é intermediar a negociação de ativos financeiros entre agentes compradores e os agentes vendedores. De acordo com Pinheiro (2019), o mercado de capitais representa um sistema de distribuição de valores mobiliários que tem o propósito de viabilizar a capitalização das empresas e dar liquidez aos títulos emitidos por elas.

Nos últimos anos, o número de investidores no Brasil cresceu exponencialmente. Segundo d'Ávila (2021), houve um crescimento de 92% no número de investidores pessoas físicas na bolsa de valores brasileira, aumentando em 1,5 milhão de investidores em 2020. Adicionalmente, o mercado de capitais é muito importante para a economia de um país, pois estimula a circulação de dinheiro, trazendo confiança ao mercado e segurança aos investimentos.

Dentre os tipos de investimentos, as ações são uma das principais formas de captação de recursos, visto que concede ao comprador a posse de uma fração de capital social da empresa, tornando-o sócio da mesma. Empresas que possuem capital aberto estão aptas a desenvolver projetos que viabilizam o seu crescimento. Nesse cenário, espera-se que a empresa obtenha bons resultados e o preço das ações se valorizem. Assim, o investidor terá lucro se vender um ativo em um valor mais caro do que comprou. Além disso, o acionista poderá obter rentabilidade ao receber proventos com as ações, como os dividendos, que são uma porcentagem dos ganhos da empresa, proporcinais ao número de ações que o comprador possui.

O mercado acionário, no entanto, é bastante volátil. Os preços variam frequentemente devido a influência que sofrem pelas transações de compra e venda, questões econômicas, notícias, dentre outros motivos. Elder (2016) é incisivo neste aspecto: "Os mercados estão cheios de cães e gatos, ações de empresas não lucrativas que, em algum momento, atravessam o telhado, desafiando a lei da gravidade". Consequentemente, o risco de perda afasta potenciais investidores, principalmente aqueles que possuem um perfil conservador. Por esse motivo, para investir em ações é necessário saber lidar com a volatilidade e analisar os ativos que podem trazer um retorno financeiro ao comprador.

1.1 Problema

As motivações por trás deste trabalho são os obstáculos que os investidores enfrentam ao entrarem no mercado de ações. A começar pela volatilidade dos papéis, que

faz com que os investidores inexperientes ou impacientes percam dinheiro na bolsa de valores. Também é importande enfatizar que para ter conhecimento sobre as ações que são mais confiáveis e rentáveis, é necessário estudar sobre o mercado de ações e utilizar um método eficiente de analisar ações, a fim de evitar a perda de dinheiro. A análise fundamentalista, por exemplo, consiste na avaliação de uma empresa de acordo com a sua situação financeira, mercadológica e até mesmo política, com o intuito de determinar o valor de suas ações. Além disso, os analistas fundamentalistas realizam a análise setorial e macroeconômico, ou seja, basicamente qualquer informação pública que os ajudem a investir com mais segurança. Logo, fica claro que estudar sobre o mercado de ações é uma tarefa que despende muito tempo e esforço, e nem todos os investidores estão dispostos a pagar este preço.

1.2 Objetivos

O objetivo deste trabalho é, portanto, classificar ações da bolsa de valores Brasil, Bolsa, Balcão (B3), como confiáveis ou não para investir. Para isso, será criado um modelo de classificação capaz de predizer se um papel irá render até o próximo trimestre com base em seu último relatório trimestral. É necessário destacar que determinar a confiabilidade de ações no longo prazo depende de uma série de fatores, e não só da valorização. Se uma ação foi bem até agora, não significa que seu resultado no futuro será o mesmo. Dessa maneira, o presente estudo visa criar uma ferramenta que garanta uma rentabilidade positiva sobre as ações apenas até o próximo trimestre. Ou seja, o modelo não é capaz de prever a confiabilidade de uma ação no longo prazo, afinal, quanto maior é o seu horizonte de previsão, maior é a incerteza associada a ele.

Ao final do trabalho, o modelo de classificação será inserido em um bot do Telegram, que será capaz de prever a confiabilidade das ações contidas na base de dados.

Para alcançar os resultados desejados, este trabalho possui os seguintes objetivos específicos:

- construir uma base sólida de dados de balanços patrimoniais, demonstrativos de resultados e indicadores fundamentalistas das empresas da B3;
- identificar um algoritmo de aprendizado de máquina que retorne os melhores resultados e definir os seus hiperparâmetros;
- desenvolver um *bot* no *telegram* que facilite à decisão do investidor.

1.3 Metodologia

Este trabalho visa coletar dados de balanços patrimoniais, demonstrativos de resultados e indicadores fundamentalistas das empresas listadas na bolsa de valores B3 e tratá-los de tal forma que possam ser submetidos a algoritmos de aprendizagem de máquina, com o intuito de classificar as empresas confiáveis para investimento.

Os passos para execução deste trabalho são:

- revisar a literatura relacionada a mineração de dados, mercado de ações e análise fundamentalista;
- desenvolver o trabalho com base na metodologia Processo Padrão de Indústria Cruzada para Exploração de Dados (CRISP-DM)
- levantar dados de balanços patrimoniais, demonstrativos de resultados e indicadores fundamentalistas das empresas;
- classificar empresas utilizando os algoritmos *Logistic Regression*, *Decision Tree*, *K-Nearest Neighbors*, *Naive Bayes*, *Multilayer Perceptron*, *Random Forest*;
- comparar os resultados dos algoritmos e escolher o algoritmo a ser usado no bot;
- escolher os melhores hiperparâmetros para o algoritmo;
- criar um modelo de classificação utilizando o algoritmo e seus hiperparâmetros escolhidos;
- analisar e discutir os resultados do modelo;
- desenvolver um bot no Telegram, que com o auxílio do modelo de classificação criado, seja capaz de classificar a confiabilidade das empresas da B3.
- concluir e avaliar a ferramenta construída para auxiliar a tomada de decisão dos investidores.

1.4 Organização do trabalho

O restante deste trabalho é organizado como se segue. O Capítulo 2 apresenta o referencial teórico utilizado no trabalho, no qual são definidos conceitos relacionados a mineração de dados, mercado de capitais e análise fundamentalista. O Capítulo 3 aborda todo o processo de preparação dos dados: o levantamento dos dados, a construção da base de dados, a limpeza e transformações dos dados. O Capítulo 4 é composto pela criação dos modelos de classificação, análise dos resultados obtidos, escolha do modelo

a ser utilizado no bot e a construção do bot do Telegram. Por fim, no Capítulo 5 são mostradas as conclusões do trabalho apresentado e sugestões para futuros trabalhos que venham a melhorar o presente estudo.

2 Revisão bibliográfica

Esta seção irá abordar todos os conceitos e informações necessárias para o entendimento deste trabalho. Além disso, serão mostrados alguns trabalhos relacionados ao presente estudo, evidenciando os problemas que buscam resolver, os objetivos e as soluções de cada um deles.

2.1 Mercado de Capitais

O mercado de capitais é um sistema criado para negociação de ativos financeiros, constituído por bolsa de valores, corretoras e instituições financeiras. A sua função é ligar os indivíduos, que tem dinheiro para investir e buscam rentabilizá-lo, com as instituições que procuram captar recursos para financiamento de projetos e pagamento de dívidas. Esse mercado pode ser entendido como um sistema de entidades e regras que incluem instituições, leis, normas, procedimentos e tecnologia, com o intuito de negociar documentos que representem investimentos em dinheiro ou bens que podem ser avaliados monetariamente, por meio de uma aquisição pública e, como consequência, forneça o recurso que irá financiar os projetos das empresas (SINATORA, 2016).

2.1.1 Bolsa de Valores B3

A bolsa de valores é o ambiente organizado para negociação de ações, e outros valores mobiliários, de empresas de capital aberto. Por meio de sistemas sofisticados e processos padronizados, ela assegura segurança e liquidez ao mercado, permitindo a compra e venda de ações.

> A bolsa de valores é o centro especialmente criado e mantido para negociação de valores mobiliários, em mercado livre e aberto, organizado e fiscalizado pelos corretores e pelas autoridades (PINHEIRO, 2019).

A bolsa de valores é responsável pela intermediação entre as entidades que precisam de financiamento e os investidores, tendo como objetivo facilitar a troca de fundos entre eles. Além disso, esta instituição deve fixar os preços dos títulos através da lei da oferta e demanda de forma transparente, afim de dar credibilidade e confiança ao mercado. Segundo Pinheiro (2019), as bolsas de valores cumprem os seguintes objetivos:

- facilitar a troca de fundos entre as entidades que precisam de financiamento e os investidores;

- proporcionar liquidez aos investidores em bolsa; assim, o investidor pode recuperar seu investimento quando precisar, utilizando a bolsa para vender seus ativos;
- fixar o preço dos títulos através da lei da oferta e demanda;
- dar informações aos investidores sobre as empresas que negociam em bolsa; por esse motivo, as empresas admitidas em bolsa devem informar periodicamente sua evolução econômica e cumprir uma série de requisitos;
- proporcionar confiança aos investidores, já que as compras e as vendas de valores estão garantidas juridicamente;
- publicar os preços e as quantidades negociadas a fim de informar aos investidores e às entidades interessadas.

A B3 é a única bolsa de valores existente no Brasil, resultado da fusão entre as antigas instituições Bovespa, Bolsa de Mercadorias e Futuros (BMF) e Central de Custódia e Liquidação Financeira de Títulos (Cetip). Como principal intermediadora do mercado de capitais brasileiro, a B3 disponibiliza plataformas para a negociação de ações, derivativos de ações, títulos de renda fixa, títulos públicos federais, derivativos financeiros, moedas à vista e *commodities* agropecuárias. As suas atividades são a criação e administração de sistemas de negociação e compensação, liquidação, depósito e registro para todas as principais classes de ativos, como ações, títulos de renda fixa corporativa, derivativos de moedas, taxas de juro e de *commodities* (B3, 2021).

Com sede em São Paulo e unidades em Rio de Janeiro e Alphaville, a B3 é uma companhia que oferece o ambiente e as condições necessárias para as negociações de compra e venda de títulos e valores mobiliários de forma transparente (PINHEIRO, 2019).

2.1.2 Mercado de Ações

O mercado de ações tem atraído cada vez mais investidores. No Brasil, muitas empresas estão abrindo o seu capital, e listando suas ações na bolsa de valores. De acordo com Alvarenga (2021), 13 companhias já fizeram o Oferta Inicial de Ações (IPO) no ano de 2021, e há ao menos 31 aguardando autorização para entrar na bolsa de valores B3.

Uma ação, ou papel, representa a menor parcela do capital de uma empresa, e quem compra uma ação se torna sócio desta empresa. Em outras palavras, as ações são títulos de participação negociáveis que representam parte do capital social de uma companhia, a qual garante ao seu comprador o direito de participação em seus resultados (PINHEIRO, 2019).

A ausência de um nível alto de burocracia em relação à compra e venda é uma característica do mercado de ações que o torna bastante atraente. Este fato é o que

configura a liquidez desse mercado, e permite ao investidor migrar entre negócios com muita facilidade (RASSIER; HILGERT, 2009).

A principal vantagem de se tornar acionista de uma empresa é adquirir partes dos ganhos que ela obtiver. Quando uma companhia aberta tem lucro, uma parcela dele é distribuído entre os sócios em forma de dividendos, na proporção do número de ações que cada um possui. Dividendos são o pagamento de uma parcela do lucro líquido de uma empresa a seus sócios.

No que tange o direito dos acionistas, existem dois tipos de ações, ordinárias e preferenciais:

- Ações Ordinárias - Este tipo de papel assegura ao investidor o direito a voto nas assembleias de acionistas, e cada ação deste tipo representa um voto;
- Ações Preferenciais - Este tipo de papel assegura ao investidor a prioridade de recebimento de dividendos sobre as ações ordinárias. Normalmente, este tipo de ação possui dividendos 10% maiores.

2.1.3 Análise Fundamentalista

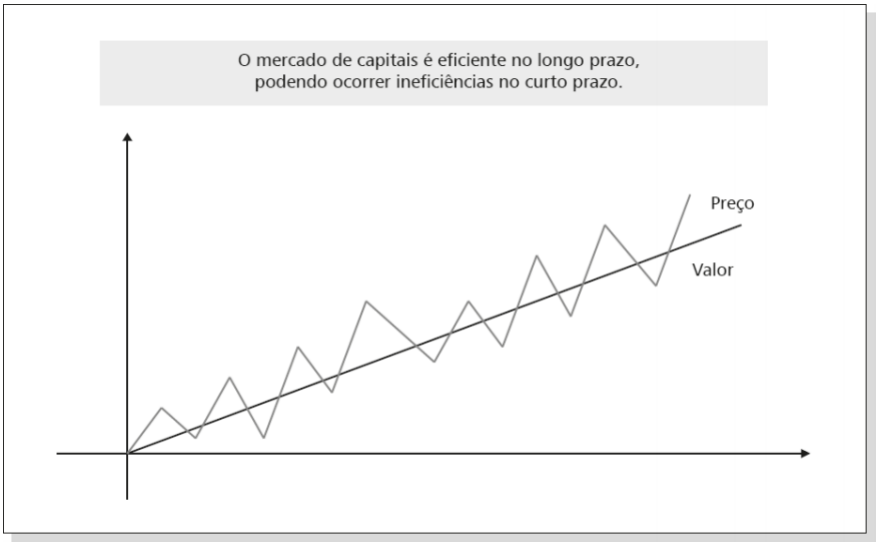
A análise fundamentalista tem por objetivo avaliar a situação financeira das empresas através de aspectos contábeis, macro e micro econômicos, visando projetar seus resultados futuros e determinar o valor intrínseco da empresa. Esse estudo é de suma importância no que diz respeito à análise de investimentos em ações, pois possibilita a identificação de empresas sólidas e rentáveis a longo prazo (INFOMONEY, 2021).

Os analistas fundamentalistas consideram que o mercado é eficiente no longo prazo, mas não necessariamente no curto, como pode ser visto na Figura 1. Logo, tentam mensurar o potencial de crescimento do lucro da empresa no futuro. Desse modo, é possível que, no momento atual, as cotações de uma ação não estejam refletindo o valor real da empresa, portanto, busca-se justamente antecipar o comportamento futuro dos papéis (INFOMONEY, 2021).

A análise de fundamentos tem por objetivo entender o potencial de crescimento da empresa, o nível de risco em relação ao negócio, a alavancagem financeira, a capacidade de captação e a flexibilidade financeira. Para isso, são calculados pelos analistas, uma série de indicadores que ajudam a avaliar o potencial de uma ação.

Indicadores fundamentalistas são dados calculados a partir do balanço da empresa e dados de mercado, que são comparadas entre si para serem capazes de demonstrar como foi o andamento das atividades. Permitem mensurar o crescimento ou a redução de fatores positivos ou negativos relacionados às atividades da empresa e podem indicar tanto o crescimento quanto o fracasso das vendas (DEBASTIANI; RUSSO, 2008).

Figura 1 – Hipótese básica da análise fundamentalista



Fonte: Pinheiro (2019, p. 402)

2.1.3.1 Principais Componentes do Balanço Patrimonial e Demonstrativo de Resultados

De acordo com Silva e Rodrigues (2018), estas são as principais informações encontradas no balanço patrimonial e demonstrativo financeiro das empresas:

- **Ativo**: são recursos que agregam algum valor futuro para determinada empresa;
- **Ativo Circulante**: são as disponibilidades da empresa que podem ser convertidos em dinheiro e serem usados dentro do período de um ano;
- **Ativo Não Circulante**: são as disponibilidades da empresa que não podem ser convertidos em dinheiro e serem usados dentro do período de um ano. O ativo não circulante é separado em: realizável a longo prazo, investimentos, imobilizado e intangível;
- **Realizável a Longo Prazo**: são ativos que serão convertidos em caixa após 1 ano da data da publicação do balanço;
- **Investimentos**: são os recursos que as empresas aplicam em ativos que não tenham relação com suas atividades operacionais.
- **Imobilizado**: são bens da empresa que não possuem grande liquidez, como por exemplo: fábricas, automóveis, instalações e equipamentos;
- **Intangíveis**: são ativos utilizados pelas empresas no exercício de suas atividades operacionais, como por exemplo: patentes, gastos com pesquisa e desenvolvimento;

- **Ativo Total**: é a composição dos Ativos Circulante e Não Circulante;
- **Passivo**: são as obrigações das empresas com terceiros;
- **Passivo Circulante**: são as obrigações de curto prazo da empresa, que deverão ser quitadas dentro de 1 ano;
- **Passivo Não Circulante**: são as obrigações que serão quitadas após o período de 1 ano;
- **Passivo Total**: é a composição dos Passivos Circulantes e Não Circulantes;
- **Lucro Líquido**: é o resultado apurado após computadas as receitas e despesas financeiras, o resultado dos impostos sobre os lucros e as participações no lucro. É o resultado que efetivamente pertence aos proprietários da empresa;
- **Patrimônio Líquido**: são todos os recursos que estão atrelados aos prorietários ou acionistas da empresa;
- **Receita Líquida**: são as vendas da empresa no período, não importando se são de curto ou longo prazo, deduzidos os impostos, eventuais devoluções e abatimentos concedidos após a entrega dos produtos;
- **Dividendos**: é a distribuição de lucros para os proprietários ou acionistas da empresa;
- **Receita Financeira**: representa os valores recebidos através de transações financeiras, tais como juros de uma aplicação, ou descontos obtidos na compra de um produto ou serviço;
- **Resultado antes das tributações e participações**: é o lucro ou prejuízo obtido antes do pagamento das tributações (impostos) e debêntures, empregados e administradores, partes beneficiárias e de fundos de assistência e previdência aos empregados;
- **Empréstimos e Financiamentos**: correspondem às dívidas que a empresa assumiu com uma instituição financeira. É uma conta que faz parte do passivo circulante.

2.1.3.2 Principais Indicadores Fundamentalistas

A seguir, serão explicados, segundo Debastiani e Russo (2008), os principais indicadores fundamentalistas, e que serão utilizados neste trabalho:

- **Liquidez Corrente (LC)**: Representa o quanto a empresa possui em relação a cada unidade monetária que deve no mesmo período. Isto quer dizer que o valor desse indicador deve ser sempre maior que 1, pois isso indicaria que a empresa

possui capacidade para pagamento de suas dívidas e disponibilidade financeira. Se a empresa demonstrar um acréscimo constante desse valor, ao longo do tempo, pode-se deduzir que a empresa tem aumentado a quantidade de dinheiro disponível em seu fluxo de caixa.

É calculada pela seguinte fórmula:

$$\text{LC} = \frac{\text{Ativo Circulante}}{\text{Passivo Circulante}}$$

- **Rentabilidade sobre Patrimônio Líquido (RPL)**: Representa a taxa de retorno dos acionistas no período de apuração. Mede a performance de geração de lucro que a empresa consegue produzir com capital próprio. Ou seja, quanto maior é o valor desse indicador, melhor.

  É calculado pela seguinte fórmula:

$$\text{RPL} = \frac{\text{Lucro Líquido}}{\text{Patrimônio Líquido}}$$

- **Lucro por Ação (LPA)**: Considera-se, nesse indicador, o Lucro Líquido num período de 12 meses e a quantidade atual de ações emitidas pela empresa. Ele representa quanto do lucro da empresa cabe a cada ação, e com base nele, o investidor pode apurar se o lucro gerado para cada uma de suas ações está diminuindo ou aumentando.

  É calculado pela seguinte fórmula:

$$LPA = \frac{\text{Lucro Líquido}}{\text{Quantidade de Ações}}$$

- **Valor Patrimonial da Ação (VPA)**: Este indicador representa o valor real de cada ação. Quando um papel é negociado acima de seu VPA, significa que o mercado acredita no potencial da empresa. Analogamente, se uma ação é negociada abaixo do seu VPA, demonstra que o mercado não acredita no crescimento da empresa.

  É calculado pela seguinte fórmula:

$$\text{VPA} = \frac{\text{Patrimônio Líquido}}{\text{Quantidade de Ações}}$$

- **Preço/Lucro (P/L)**: Ele utiliza, em sua fórmula de cálculo, o LPA. Portanto, para calculá-lo, é necessário obter primeiro o LPA da ação analisada. O P/L indica o tempo de retorno do investimento em anos, considerando que a empresa mantenha os atuais níveis de lucro. Para este indicador, quanto menor for o valor, melhor. Logo, uma redução contínua do P/L, indica a redução dos anos necessários, para se obter o retorno do investimento.

  Este é o indicador preferido entre os investidores. Sua fórmula de cálculo é:

$$P/L = \frac{\text{Cotação da Ação}}{\text{LPA}}$$

- **Preço/Valor Patrimonial (P/VP)**: Este indicador utiliza outro indicador fundamentalista, o VPA, para representar, quantitativavente, o ágio ou deságio que o mercado está disposto a pagar pela ação. O ideal é que este indicador seja maior do que 1, o que significaria que o mercado acredita no potencial da empresa e está pagando um valor maior do que o valor real. Todavia, pode ser arriscado para o investimento se o P/VP estiver muito elevado, visto que pode se tratar de uma valorização especulativa, sem amparo nos fundamentos da organização.

  Sua fórmula de cálculo é:

$$\text{P/VP} = \frac{\text{Cotação da Ação}}{\text{VPA}}$$

- ***Price Sales Ratio* (PSR)**: Este indicador é representado pela divisão entre o valor de mercado da ação e sua capacidade de geração de receita. Essa relação mostra o entusiasmo do mercado em relação ao papel, ou seja, quanto maior for o valor do PSR, mais arriscado é o investimento, pois a ação está supervalorizada, e passível de uma queda dos preços.

  Sua fórmula de cálculo é:

$$\text{PSR} = \frac{\text{Cotação da ação}}{(\text{Receita Líquida}/\text{Quantidade de Ações})}$$

- ***Earnings Before Interest, Taxes, Depreciation and Amortization* (EBITDA)**: O EBITDA se propõe a medir a capacidade de geração de caixa da empresa e a performance do administrador. Para isso, ele elimina fatores externos, que podem mascarar o lucro da empresa.

  Este é um indicador muito complexo e possui em sua fórmula muitas informações, conforme se vê a seguir:

$$\begin{aligned}\text{EBITDA} = \\ \text{Receita Líquida} - \text{Custo de Produtos Vendidos} - \text{Despesas da Atividade} + \\ \text{Depreciação Acumulada no Período} + \text{Amortização Acumulada no Período}\end{aligned}$$

- ***Pay-Out***: Na fórmula de cálculo desse indicador utiliza-se o LPA. Ele é expresso na forma de percentual e demonstra quanto do LPA está sendo distribuído aos acionistas na forma de dividendos. Logo, o foco do *Pay-Out* é o pagamento dos dividendos pela empresa, sendo possível analisar se a cota de distribuição está aumentando ou diminuindo ao longo do tempo. Contudo, por relacionar-se com o

LPA, caso a empresa melhore a sua performance, o *Pay-Out* pode diminuir. Isso, não é, necessariamente, um sinal ruim para os investidores.

Sua fórmula de cálculo é:

$$PayOut = \left(\frac{\text{Valor dos Dividendos}}{\text{LPA}}\right) \times 100$$

- ***Dividend Yield* (DY)**: O DY demonstra, de forma percentual, quanto do valor de mercado da ação está sendo distribuído aos investidores na forma de dividendos. Ele facilita a visualização da evolução da política de pagamentos praticada pela companhia, e demonstra se a direção tem disposição em remunerar melhor os seus acionistas.

  Este é um dos principais indicadores para investidores que visam receber dividendos. Sua fórmula de cálculo é:

$$\text{DY} = \left(\frac{\text{Valor dos Dividendos}}{\text{Cotação da Ação}}\right) \times 100$$

- **Margem Líquida (ML)**: Este indicador mostra o percentual de lucratividade de uma empresa. A ML é a porcentagem do lucro líquido obtido pela empresa em relação à receita total, sendo um indicador percentual que demonstra qual é o lucro líquido para cada unidade de venda da empresa. É calculado da seguinte fórmula:

$$\text{ML} = \left(\frac{\text{Lucro Líquido}}{\text{Receita Líquida}}\right) \times 100$$

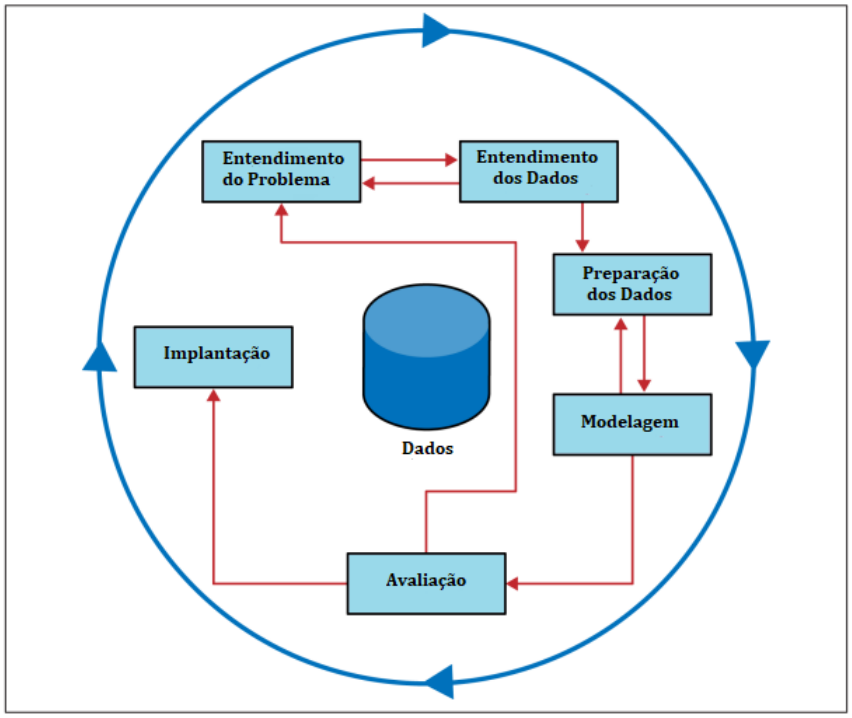
2.2 Mineração de Dados

Em virtude da democratização da internet e criação das redes sociais, quantidades massivas de dados começaram a ser geradas diariamente. Nas últimas décadas, avanços rápidos na tecnologia de coleta e armazenamento permitiram que as organizações acumulassem grandes quantidades de dados. Entretanto, extrair informações úteis a partir dos dados se provou um grande desafio. As técnicas e ferramentas tradicionais de análise de dados se tornaram obsoletas perante os enormes conjuntos de dados, e consequentemente, novas técnicas e ferramentas precisaram ser criados para extrair conhecimento dos dados.

A mineração de dados é um termo que se refere ao uso de técnicas e ferramentas para extração de conhecimento a partir de dados. De acordo com Tan, Kumar e Steinbach (2005), a mineração de dados é um conjunto de tecnologias que misturam os métodos tradicionais de análise de dados com algoritmos sofisticados para processar grandes volumes de dados.

Segundo Provost e Fawcett (2013), a mineração de dados envolve a aplicação substancial de ciência e tecnologia, e possui um processo bem definido, que estrutura o problema, permitindo consistência, repetitividade e objetividade. Desse modo, o CRISP-DM é uma codificação útil do processo de mineração de dados, pois explicita o fato de que a repetição é essencial para os projetos, pois não é possível encontrar a solução ideal para problema em apenas uma tentativa. A cada ciclo do projeto, é possível compreender melhor o problema e os dados, e encontrar novas ferramentas, técnicas e algoritmos que possam contribuir mais para a performance do modelo criado.

Figura 2 – O Processo de Mineração de Dados CRISP-DM



Fonte: Adaptado de Provost e Fawcett (2013, p. 27)

Esse método surgiu em 1996 como forma de apoio ao *Knowledge Discovery in Databases* (KDD). Ambos possuem fases muito semelhantes, contudo, o CRISP-DM possui uma etapa de entendimento do problema, que supre a necessidade de entender melhor o contexto e as questões de negócios, e assim desenvolver ferramentas de dados cada vez mais úteis para as empresas. A seguir, serão explicadas cada uma das etapas que compõem este processo:

- Entendimento do Problema - Esta etapa do processo é vital, pois compreender o problema a ser resolvido nem sempre está claro, e muitas vezes é necessário reformular o problema e as soluções. O diagrama mostrado na Figura 2 representa isso como ciclos dentro de um ciclo, em vez de um simples processo linear. Compreender o

problema de negócio e formular uma solução criativa, que faça uso das ferramentas que a mineração de dados possui, requer muito conhecimento sobre o modelo de negócio da organização e conceitos básicos da ciência de dados. Deve ser pensado cuidadosamente o cenário de uso e o problema a ser resolvido, de forma que um ou mais subproblemas envolvam a construção de modelos de classificação, regressão, estimativa de probabilidade, assim por diante.

- Entendimento dos Dados - Os dados são a matéria prima disponível a partir da qual a solução do problema de negócio será construída. A compreensão dos dados é muito importante, visto que nem sempre há uma correspondência exata com o problema. Essa etapa do processo envolve não só o entendimento dos dados disponíveis, mas também a análise das fontes de dados, e estimação dos custos e benefícios de cada uma, para decidir se são necessários mais investimentos para obtê-las.

- Preparação dos Dados - As ferramentas de mineração de dados são poderosas, mas impõem determinadas exigências sobre os dados que usam. Frequentemente, o formato dos dados disponível não é compatível com o requisitado pelos algoritmos, portanto são necessárias uma série de técnicas de limpeza e transformação, que exigem boa parte do tempo do processo de extrair conhecimento a partir dos dados. Muitas vezes, uma fase de preparação dos dados é realizada juntamente com a compreensão dos mesmos. É comum que a etapa de pré-processamento consuma mais tempo do que as etapas de aplicação dos algoritmos de mineração propriamente ditos (CASTRO; FERRARI, 2016), dado que ela deve ser feita de maneira estruturada e cuidadosa.

- Modelagem - Esta etapa é o principal local onde as técnicas de mineração de dados são aplicadas aos dados. O resultado da modelagem é algum tipo de modelo ou padrão que captura regularidades nos dados. Para Tan, Kumar e Steinbach (2005), as tarefas de mineração de dados são divididas em duas categorias principais:

  - Descritivas: O objetivo é encontrar padrões, correlações, grupos e anomalias que resumem as relações subjacentes nos dados. Os principais tipos de algoritmos descritivos são os de agrupamento, associação e detecção de anomalias;
  - Preditivas: Buscam prever o valor de um atributo específico, com base em outros atributos. Os principais tipos de modelos preditivos são os de classificação e de regressão. No presente estudo, como o objetivo é classificar ações da bolsa de valores, serão utilizados algoritmos de classificação. Na subseção 2.3, serão explicados mais a fundo o que são modelos de classificação, e cada um dos algoritmos que serão utilizados.

- Avaliação - A fase de avaliação tem como objetivo estimar os resultados de mineração de dados e obter a confiança de que são válidos e confiáveis antes de implementar

o modelo criado. São necessários uma série de testes para atestar se o os padrões extraídos dos dados são regularidades verdadeiras e não apenas idiossincrasias ou anomalias de amostra. Ademais, a avaliação é importante para garantir que o modelo satisfaça os objetivos de negócios originais. Afinal, o projeto foi criado para apoiar a tomada de decisão, logo, deve-se analisar a qualidade das decisões tomadas pelo modelo.

- Implantação - Nesta fase, os resultados da mineração de dados são implantadas, a fim de contribuir com a tomada de decisão da organização, e trazer um retorno sobre o investimento. Além disso, as próprias técnicas de mineração de dados vêm sendo implantadas cada vez mais, para criar e testar automaticamente modelos em produção.

2.3 Algoritmos de Classificação

Uma das principais técnicas de mineração de dados, a classificação, é utilizada para determinar a qual classe cada indivíduo de uma população pertence, com base em seus atributos. O rótulo dado a cada instância é um valor discreto, e é chamado de atributo alvo, pois é o valor que busca-se prever neste tipo de tarefa. De acordo com CASTRO e FERRARI (2016), quando cada registro histórico está rotulado, o objetivo da análise é construir um modelo que possa ser usado para predizer qual seria essa saída para novos registros, ou seja, registros cuja classe ou valor de saída são desconhecidos.

A classificação é uma tarefa de aprender uma função alvo f que mapeia cada exemplo X do conjunto de dados para rótulos de classe específicos y (TAN; KUMAR; STEINBACH, 2005). Conhecida como modelo de classificação, esta função é útil para classificar registros de dados que não foram vistos ainda, por isso, Tan, Kumar e Steinbach (2005) destacam que ela pode ser tratada como uma "*black box*" (caixa preta), que automaticamente atribui um rótulo de classe quando é apresentado a um conjunto de atributos de um registro desconhecido.

Há uma grande variedade de algoritmos de classificação na literatura, e através deles é possível criar modelos preditivos que conseguem atribuir um rótulo a um objeto, de acordo com a categoria à qual ele pertence. Para que isso seja possível, um algoritmo de classificação é usado na construção de um modelo de classificação, também chamado de classificador, o qual é construído com base em um conjunto de treinamento com dados rotulados, ou seja, um conjunto de pares entrada-saída, $(x_i, d_i)_i = 1, ..., n$, onde x_i representa os objetos e d_i as respectivas classes que foram previamente conhecidas (CASTRO; FERRARI, 2016).

As subseções 2.3.1, 2.3.2, 2.3.3, 2.3.4, 2.3.5 e 2.3.6 apresentam os conceitos básicos de alguns dos principais métodos de classificação, e que serão utilizados no presente estudo.

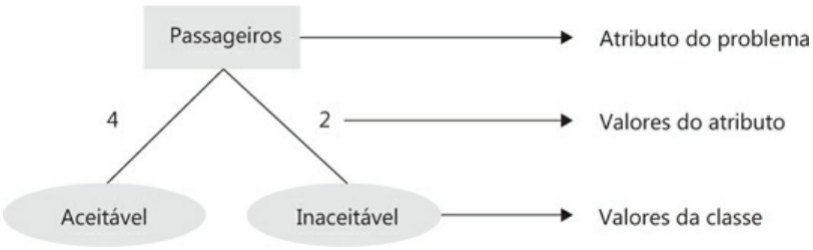
2.3.1 *Decision Tree* (Árvore de Decisão)

Uma árvore de decisão (*decision tree*) é um algoritmo que assume a estrutura de uma árvore, onde cada nó interno denota um teste em um atributo, cada ramificação representa um resultado do teste e cada nó folha (nó terminal) contém um rótulo de classe. Tan, Kumar e Steinbach (2005) salientam que existem três tipos de nós em uma árvore de decisão:

- um nó raiz não tem arestas de entrada e zero ou mais arestas de saída;
- os nós internos possuem uma aresta de entrada e uma ou mais arestas de saída;
- nós folhas possuem uma aresta de entrada e nenhuma de saída.

Uma vez construída a árvore, ela pode ser usada para classificar um objeto cuja classe é desconhecida. Para isso, é necessário apenas testar os atributos na árvore até chegar no nó folha, que corresponde à classe a qual aquele objeto pertence. A Figura 3 mostra um exemplo simples de uma árvore de decisão, usando o objeto carro, o atributo número de passageiros, e as classes“Aceitável” e “Inaceitável”. Neste exemplo, se um carro tiver a capacidade de quatro passageiros é rotulado como "Aceitável", entretanto, se tiver a capacidade de dois passegeiros, é rotulado como “Inaceitável”.

Figura 3 – Exemplo de árvore de decisão



Fonte: CASTRO e FERRARI (2016, p. 343)

As árvores de decisão são fáceis de compreender e visualizar. Além disso, elas possuem facilidade para explicar a classificação, pois basta percorrer a árvore para se identificar por que um objeto foi classificado naquela categoria (CASTRO; FERRARI, 2016).

2.3.2 *Logistic Regression*

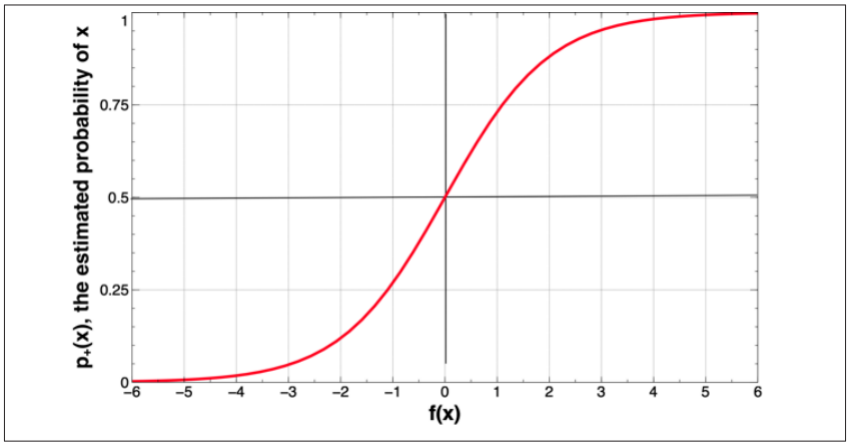
A regressão logística (*Logistic Regression*) é um processo de modelagem da probabilidade de um resultado discreto em função de uma ou mais variáveis de entrada. A

melhor maneira de pensar sobre regressão logística é que ela é uma regressão linear, mas para problemas de classificação. A partir dela é possível calcular ou prever a probabilidade de um evento ocorrer, dada uma observação aleatória. Nela, usa-se essencialmente uma função logística definida abaixo para modelar uma variável de saída binária:

$$p_+(x) = \frac{1}{1+e^{-f(x)}}$$

Na equação da função logística, x é a variável de entrada. Conforme ilustrado na Figura 4, essa função pode ser representada graficamente, e mostra a estimativa de probabilidade de classe de regressão logística como função de f(x), ou seja, a distância a partir do limite de separação (PROVOST; FAWCETT, 2013). Esta curva é chamada de curva "sigmoide" por causa de seu formato em "S", que comprime as probabilidades para sua extensão correta (entre zero e um).

Figura 4 – Sigmoide (Função Logística)



Fonte: Provost e Fawcett (2013, p. 101)

De acordo com Provost e Fawcett (2013), o eixo vertical é a estimativa de probabilidade e o eixo horizontal é o limite de decisão. A Figura 4 mostra que, no limite de decisão (na distância $x = 0$), a probabilidade é $0,5$. A probabilidade varia quase de forma linear próximo ao limite de decisão, logo, a certeza é menor. Provost e Fawcett (2013) evidenciam que a rapidez com que a certeza aumenta ao se afastar do limite é parte do "ajuste" do modelo aos dados.

Pode-se dizer que a interpretabilidade da regressão logística não é tão fácil quanto a interpretação de kNN ou regressão linear, mas ainda muito mais fácil do que mais modelos de "caixa preta" como Redes Neurais. O principal obstáculo é a natureza multiplicativa das "probabilidades".

2.3.3 *k-Nearest Neighbors*

O *k-Nearest Neighbors* (KNN), também conhecido como k-vizinhos mais próximos, é um classificador baseado em distância. Ele funciona da seguinte forma: dado um objeto x_0, cuja classe se deseja inferir, encontram-se os k objetos $x_i, x = 1, ..., k$ da base que estejam mais próximos a x_0, e, depois, se classifica o objeto x_0 como pertencente à classe da maioria dos k vizinhos. Caso haja um empate, a classe é escolhida de forma aleatória.

CASTRO e FERRARI (2016) ressaltam que qualquer medida de distância pode ser usada para determinar os k vizinhos mais próximos, porém, caso os objetos da base de dados sejam representados numericamente, a medida mais comum é a Euclidiana. Essa escolha deve considerar as características do conjunto de dados, e o contexto da aplicação.

A estratégia de *voto majoritário* para determinar a classe do objeto de análise pode trazer um problema se a distribuição de objetos nas classes não for equilibrada. As classes mais frequentes podem dominar a predição dos objetos desconhecidos. Isso pode ser resolvido ao utilizar o voto ponderado dos vizinhos mais próximos, ou seja, considerar a distância entre o objeto x_0 e os k vizinhos, ao definir a classe a qual ele pertence. A *votação ponderada* ou *voto de similaridade moderada* escalona a contribuição dos vizinhos pela sua similaridade (PROVOST; FAWCETT, 2013).

O KNN, apesar da simplicidade, apresenta resultados satisfatórios em vários cenários. Além disso, ele é um algoritmo conhecido como "preguiçoso" (*lazy learning*), pois sua saída é calculada apenas quando se deseja saber a classe de um novo objeto, logo não é treinado a priori.

2.3.4 *Naive Bayes*

O *Naive Bayes* (NB) é um classificador baseado no Teorema de Bayes, e é usado para predizer a probabilidade de pertinência de um objeto a determinada classe. De acordo com CASTRO e FERRARI (2016), os classificadores Naive Bayes possuem alta acurácia e velocidade quando aplicados a grandes bases de dados.

Com o objetivo de simplificar os cálculos, esse algoritmo possui uma premissa denominada "independência condicional da classe" (*class conditional independence*), o qual assume que o efeito do valor de um atributo em uma dada classe é independente dos valores dos outros atributos. Por causa dessa premissa que o algoritmo é denomidado *naive* (ingênuo).

O classificador NB, apesar de muito simples, leva em conta todas as evidências características. Além disso, ele é muito eficiente em termos de espaço e tempo, e possui um desempenho muito bom para várias tarefas do mundo real. Todavia, Provost e Fawcett (2013) destacam que se o foco do projeto for as estimativas de probabilidade, o NB deve ser usado com cautela para a tomada de decisão, e deve-se avaliar os custos e benefícios.

Portanto, este classificador é utilizado com mais frequência em situações onde os valores reais das probabilidades não são relevantes, apenas os valores relativos de exemplos nas diferentes classes.

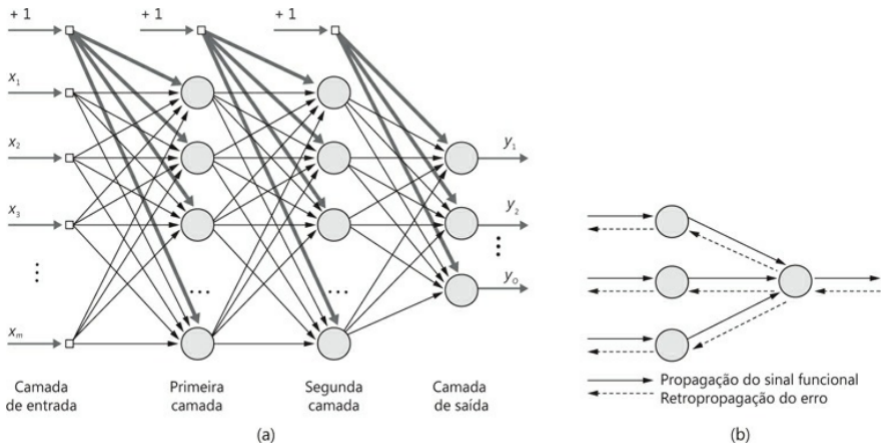
Outra vantagem do NB é que ele é um "aprendiz incremental", ou seja, ele não precisa reprocessar todos os exemplos antigos de treinamento quando novos dados de treinamento surgirem, logo, atualiza o modelo um exemplo por vez (PROVOST; FAWCETT, 2013).

2.3.5 *Multilayer Perceptron*

O *Multilayer Perceptron* (MLP) é uma rede *Perceptron* com múltiplas camadas, que foi desenvolvida com o intuito de conseguir ser aplicado a dados não lineares, e superar o limitação da rede *Perceptron* simples. Ele possui camadas de entrada e saída e uma ou mais camadas ocultas com muitos neurônios empilhados. Diferentemente do *Perceptron*, onde é utilizado apenas uma função de ativação que impõe um limite, nele é possível utilizar qualquer função de ativação arbitrária.

A rede MLP se enquadra na categoria de algoritmos *feedforward*, pois as entradas são combinadas com os pesos iniciais em uma soma ponderada e sujeitas à função de ativação. Cada camada alimenta a próxima com o resultado de sua computação, e dessa forma, todo o caminho é percorrido através das camadas ocultas até a camada de saída.

Figura 5 – Rede neural de múltiplas camadas (a), sentido de propagação do sinal de entrada e retropropagação do erro (b)



Fonte: CASTRO e FERRARI (2016, p. 421)

Segundo CASTRO e FERRARI (2016), o treinamento da rede MLP é feita utilizando-se um algoritmo chamado *backpropagation* (retropropagção do erro), que consiste de dois passos:

1. Propagação positiva do sinal funcional, durante a qual todos os pesos da rede são mantidos fixos.

2. Retropropagação do erro, durante a qual os pesos da rede são ajustados com base no erro.

O *backpropagation*, ilustrado na Figura 5(b), é o mecanismo de aprendizado que permite ao *Multilayer Perceptron* ajustar iterativamente os pesos na rede, com o objetivo de minimizar a função custo. Caso contrário, a rede MLP iria apenas calcular as somas ponderadas em cada neurônio, e propagar os resultados para a camada de saída.

2.3.6 *Random Forest*

A *Random Forest* (Floresta Aleatória) é um algoritmo de aprendizado de máquina supervisionado amplamente utilizado em problemas de classificação e regressão. Ele constrói aleatoriamente várias árvores de decisão, formando uma floresta, e cada árvore é utilizada na escolha do resultado final.

Este algoritmo utiliza um método chamado *ensemble*, que combina diferentes modelos para obter um único resultado. Assim, uma coleção de modelos é usada para fazer previsões em vez de um modelo individual. Em problemas de regressão são utilizadas a média dos valores dados pelos modelos para obtenção do resultado final, e em problemas de classificação o resultado que mais se repete é o escolhido. Essa característica torna as florestas aleatórias mais robustas e complexas, com maior custo computacional, embora com melhores resultados.

De acordo com o Sruthi (2021), estas são as etapas envolvidas no algoritmo de *Random Forest*:

1. São selecionadas aleatoriamente algumas amostras dos dados de treino.

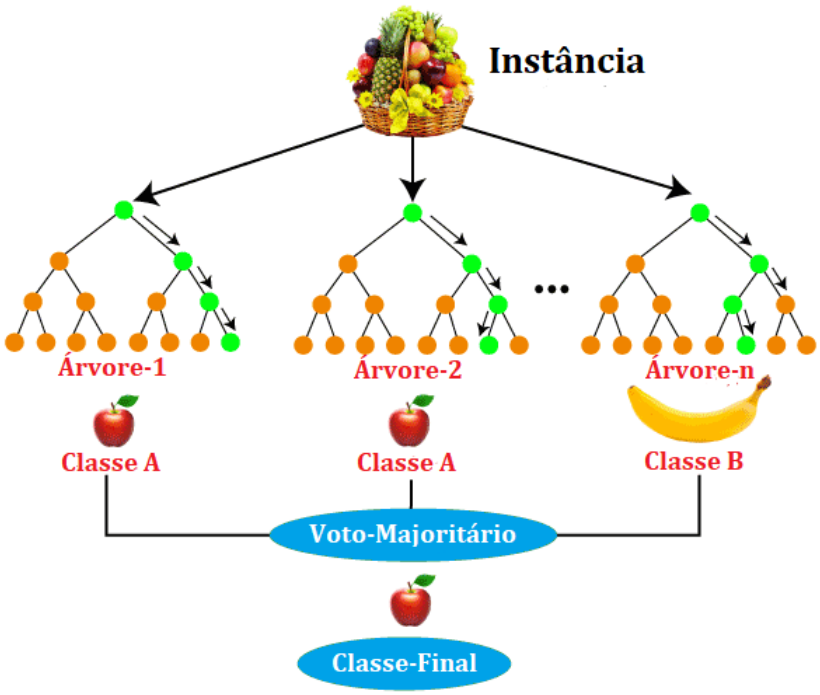
2. As árvores de decisão individuais são construídas para cada amostra.

3. Cada árvore de decisão irá gerar uma saída.

4. O resultado final é considerado com base na votação majoritária ou média para classificação e regressão, respectivamente.

A Figura 6 tem uma ilustração do funcionamento das árvores aleatórias. Neste exemplo, o resultado final, classe A, é escolhido com voto majoritário, pois a maioria das árvores de decisão retornam este resultado. É importante ressaltar que nem todos as variáveis são consideradas ao fazer uma árvore individual, cada árvore usa diferentes conjuntos de variáveis. Logo, segundo Sruthi (2021), há diversidade no uso dos atributos, e

Figura 6 – Representação do funcionamento de uma Floresta Aleatória



Fonte: Adaptado de Sruthi (2021)

o algoritmo não sofre com a alta dimensionalidade dos dados, visto que o espaço de feição é reduzido. Ademais, a *Random Forest* resolve o problema de *overfitting* (sobreajuste dos dados), dado que a saída é baseada no voto majoritário ou na média.

Em contrapartida, a floresta aleatória é altamente complexa quando comparada às árvores de decisão, onde as decisões podem ser tomadas seguindo o caminho da árvore. Além disso, o tempo de treinamento é relativamente maior em relação a outros modelos e sempre que tiver que fazer uma previsão, cada árvore de decisão deve gerar uma saída para os dados de entrada fornecidos.

2.4 Trabalhos Relacionados

O estudo de Rodrigues (2016) apresentou a aplicação da mineração de dados em uma base de dados contendo indicadores fundamentalistas de empresas da bolsa de valores BM&FBovespa. O objetivo era classificar as empresas como investimento "Fraco", "Bom" ou "Muito Bom" comparando seus rendimentos com a taxa Selic no período pesquisado. Como resultado, foi obtido uma alta taxa de acertos para a classe "Fraco", porém poucas empresas das classes "Bom" e "Muito Bom" foram classificadas corretamente. Similar ao presente estudo, Rodrigues (2016) pretendia, por meio da análise fundamentalista e mineração de dados, rotular as empresas da bolsa de valores e atestar a sua confiabilidade

para investimento. No entanto, neste trabalho, a proposta foi não apenas realizar análises preditivas nos dados das empresas, mas também implementar a solução encontrada em uma ferramenta que fosse acessível aos investidores.

Souza (2021) desenvolveu um trabalho que utilizou técnicas de mineração de dados para predizer o preço futuro de ações, com o intuito de auxiliar a tomada de decisão de investidores no mercado acionário. Para isso, ele utilizou os algoritmos Árvore de Decisão e Redes Neurais, e realizou um levantamento de dados históricos da ação PETR4, referentes ao ano de 2019. Com o auxílio da ferramenta *WEKA*, constatou-se que as redes neurais obtiveram os melhores resultados, sendo superior à Árvore de Decisão em 5 dos 6 testes realizados. Analogamente, o presente estudo também busca contribuir com as pessoas para que possam realizar melhores investimentos em ativos financeiros, contudo, a variável alvo não era o preço e sim se as ações iriam render ou não.

Lima (2016) buscou desenvolver um modelo para predizer o comportamento da bolsa de valores, baseado na mineração de opinião. Com a aplicação de técnicas de inteligência artificial, como processamento de linguagem natural e Máquinas de Vetor de Suporte, buscou-se averiguar como a análise de sentimentos pode contribuir para antecipar o comportamento de um ativo financeiro. Diferentemente deste trabalho, onde os dados analisados foram das empresas de capital aberto. Para tal fim, foram utilizados dados da rede social *Twitter* com o objetivo de analisar as publicações de usuários referentes as ações da Petrobras, e comparar a tendência de alta ou baixo do ativo com a variação de humor expressos pelos usuários em suas postagens. Evidenciou-se que postagens do *Twitter* representam uma rica fonte de informação, no entanto, exigem de um grande esforço na etapa de pré-processamento dos dados. Além disso, constatou-se que o humor coletivo não reflete o sentimento do mercado para um ativo específico, visto que comentários de usuários podem ter conotações políticas, distorcendo o foco da análise. Logo, no âmbito da bolsa de valores, a análise de sentimentos deve ser vista de forma complementar no processo de tomada de decisão.

Na pesquisa de Souza (2015) foi realizado um estudo sobre séries temporais financeiras. Em métodos tradicionais, considera-se que o ambiente é estático, ou seja, o mecanismo gerador da série financeira é o mesmo durante todo o intervalo de tempo de interesse. No entanto, no caso de séries temporais financeiras, isso pode não ocorrer. Logo, foi proposto a abordagem Aproximação por Valor Agregado de Segmento - Otimização por Enxame de Partículas Auto Adaptativa com detecção de mudança de conceito (PAA-IDPSO-CD) para descoberta de padrões em séries temporais financeiras. O objetivo é lidar explicitamente com mudanças de conceito na série e descobrir os melhores padrões representativos dos dados das séries temporais que serão utilizados junto a uma estratégia de investimento formulada para automatizar as operações a serem feitas no mercado de ações utilizando técnicas de mineração de dados. Os experimentos foram comparados com os resultados

das versões do método proposto entre si e com os resultados obtidos pelas abordagens *Buy and Hold* (BH) e Aproximação por Valor Agregado Simbólico - Algoritmos Genéticos (SAX-GA). Concluiu-se que PAA-IDPSO-CD apresentou resultados estatisticamente melhores que o BH e o SAX-GA para todas as vinte ações em que os testes foram executados. Além disso, a estratégia que opera nas posições comprado e vendido é melhor quando comparada àquela que opera apenas na posição comprado. Ao contrário do presente estudo, Souza (2015) utilizou séries temporais financeiras para tentar antecipar o comportamento futuro das ações. Todavia, é interessante notar que há várias técnicas diferentes que podem ser usadas para este mesmo objetivo.

Raminelli e Santos (2019) fizeram uma revisão sistemática sobre a aplicação da mineração de dados e aprendizagem de máquina no mercado de ações. Utilizou-se as bases de dados *Scopus*, *Web of Science* e *ScienceDirect* até o ano de 2018. Em relação às técnicas de mineração de dados a apredizagem de máquina, a que mais se destacou foi a predição, evidenciando-se que prever os preços futuros de ações é a tarefa mais almejada pelos investidores. Além disso, constatou-se que a técnica de associação também foi bastante utilizada, indicando a busca por correlações existentes entre o preço de ações e fatores externos. Quanto às técnicas e ferramentas utilizadas para manipular os dados, houve um relevante destaque da aplicação de *Neural Networks* e *Deep Learning*. No presente estudo, a proposta foi criar uma ferramenta que utilize a mineração de dados no mercado de ações, enquanto o trabalho de Raminelli e Santos (2019) foi revisar os últimos trabalhos publicados a respeito deste tema.

Por fim, Almeida e Rocha (2019) utilizaram técnicas de mineração de dados para encontrar e analisar padrões no conjunto de dados de empresas da bolsa de valores BM&FBovespa utilizando dados de indicadores fundamentalistas, cotações e indicadores socioeconômicos. Os resultados obtidos mostram que os algoritmos de classificação conseguiram encontrar padrões relacionando indicadores fundamentalistas e as cotações futuras. Na clusterização hierárquica, descobriu-se agrupamentos de empresas com situação financeira semelhantes. Esse trabalho assemelha-se ao presente estudo, pois busca testar o poder preditivo de modelos de aprendizagem de máquina acima dos dados de balanço patrimonial e de outros indicadores de análise fundamentalista. Outrossim, mostraram que é possível, com o auxílio da mineração de dados, encontrar padrões e tendências que contribuam para a tomada de decisão dos investidores. Por outro lado, Almeida e Rocha (2019) focaram em encontrar padrões utilizando diferentes técnicas de mineração de dados, enquanto este trabalho buscou resolver um problema específico do mercado de ações.

3 Desenvolvimento

Neste capítulo serão descritas as etapas de entendimento e preparação dos dados. A execução do trabalho se deu a partir do método proposto no manual CRISP-DM.

Nas próximas seções serão apresentados em detalhes o desenvolvimento, desafios e aprendizados de cada etapa. Todo o código do projeto foi desenvolvido utilizando a linguagem de programação *Python*[1] e está disponível no GitHub [2].

3.1 Entendimento do Problema

Esta fase é vital para entender o problema a ser resolvido. Provost e Fawcett (2013) ressaltam que a criatividade desempenha um grande papel em projetos de ciência de dados, pois formular um problema com engenhosidade permite tecer soluções inteligentes que tiram proveito das ferramentas e técnicas de mineração de dados.

Esta etapa começou quando os objetivos foram definidos no início do trabalho. Todas as tarefas subsequentes foram realizadas com base no problema e objetivos definidos.

3.2 Entendimento dos Dados

A etapa de compreensão dos dados teve início a partir do levantamento dos dados necessários para a realização da análise. Foram coletados os dados de balanço patrimonial, demonstrativos de resultados e dados das cotações das ações.

3.2.1 Balanços e Demonstrativos das Empresas

Os dados com os balanços e demonstrativos das empresas foram obtidos através do site *Fundamentus* [3]. Trata-se de um sistema online que tem como principal objetivo a disponibilização de informações financeiras e fundamentalistas de empresas que têm suas ações listadas na bolsa de valores. Nele é possível realizar o *download* de planilhas Excel de balanços históricos das empresas. Foram coletados dados das empresas que listaram ações no período de junho de 2011 a setembro de 2021.

Algumas observações acerca dos arquivos coletados é que estavam, por padrão, divididos em duas abas, balanço patrimonial e demonstrativo de resultados, e as informações eram disponibilizadas trimestralmente.

[1] <https://www.python.org>
[2] <https://github.com/endhel/InvestBot>
[3] <https://fundamentus.com.br>

3.2.2 Cotações

As cotações das empresas utilizadas foram obtidas através do site *Yahoo! Finances*[4]. Este site tem como objetivo a disponibilização de informações acerca do mercado financeiro (foco principal nas cotações das ações de empresas). Os dados foram coletados dentro do mesmo período mencionado na subseção anterior.

Ao contrário do *Fundamentus*, onde a coleta foi realizada manualmente, no *Yahoo! Finances* foi possível coletar as cotações através de uma biblioteca chamada *Pandas DataReader*[5], e os dados de cada empresa foram inseridos em um dicionário (uma estrutura de dados do *Python* do tipo coleção), onde as chaves eram os códigos das ações, e os valores eram os arquivos de cotações.

3.2.3 Tratamento dos Dados

Após a coleta dos dados, foram realizadas uma série de modificações com o intuito de unir tudo em um único *dataframe* (uma estrutura de dados da biblioteca *Pandas*, utilizada para análise e tratamento de dados). A seguir serão citadas cada uma das alterações realizadas:

1. União dos dados de balanço e demonstrativos de cada empresa em *dataframes*. Os dados de cada uma das empresas foram armazenados em um *dataframe* separado e todos foram inseridos em um dicionário.

2. Remoção das empresas que não tinham dados de cotações no período analisado. Os dados das empresas foram removidos de ambos dicionários de cotações e de balanço.

3. União dos dados de balanços com as cotações de cada uma das empresas. Para este trabalho, foram utilizados apenas os dados de fechamento ajustado das cotações e o restante foram descartados.

4. Remoção das empresas que não tinham todos os dados de balanço patrimonial e demonstrativo de resultados disponíveis.

5. Tratamento das variáveis com nomes repetidos. Como algumas variáveis do balanço patrimonial e demonstrativo de resultados possuiam nomes iguais, foi concatenado a string "_1" ao final de algumas delas.

6. Remoção de colunas com muitos valores faltantes.

[4] <https://br.financas.yahoo.com>
[5] <https://pandas-datareader.readthedocs.io>

3.2.4 Criação dos Rótulos para as Ações

Para alcançar o objetivo de se predizer se uma ação é confiável ou não para investir até o próximo trimestre, foi necessário rotular os dados históricos, para que eles possam ser utilizados nas fases de treinamento e teste do modelo. As ações foram rotuladas com base no seguinte critério: se a ação render 3% ou mais até o próximo trimestre, ela será classificada como "confiável", caso contrário, será classificada como "não confiável".

Historicamente, o Índice Ibovespa, o mais importante indicador de desempenho das ações negociadas na B3, rendeu em média 11,74% ao ano (ECONOMATICA, 2017). Ao arrendondar este número para 12%, pode-se dizer que o rendimento médio trimestral é 3%. Assim, ao utilizar este valor como critério de classificação, certificou-se que as ações rotuladas como "confiáveis" tivessem uma margem de segurança, além de garantir que a rentabilidade delas seja igual ou superior ao de vários outros tipos de investimentos.

Em suma, para calcular a rentabilidade de uma ação é necessário ter o valor da cotação atual e do próximo trimestre. Logo, foi possível realizar o cálculo para as ações de todas as empresas, exceto do último trimestre coletado, correspondente à data de setembro de 2021, pois no momento da coleta ainda não haviam os dados do próximo trimestre. Como detalhado em XPEED (2021), o cálculo da rentabilidade deve ser realizado da seguinte forma:

$$(\text{cotação do próximo trimestre} \;/\; \text{cotação atual}) - 1$$

Ao fim desta etapa, com o propósito de facilitar o processo de análise, todos os dados foram fundidos em um único *dataframe*, uma vez que as datas dos trimestres já não eram mais necessárias. Com isso, a base de dados ficou pronta, sendo possível passar para a próxima etapa do CRISP-DM, a preparação dos dados, onde foram realizados vários tratamentos e transformações nos dados, para que possam ser submetidos a algoritmos de classificação.

3.3 Preparação dos Dados

Como foi citado na descrição do CRISP-DM, a preparação dos dados é um dos aspectos mais importantes e mais demorados da mineração de dados. Aqui, os dados foram tratados, selecionados e transformados para as próximas fases do trabalho. Tais processos serão descritos a seguir.

3.3.1 Descrição dos Dados

O primeiro passo da etapa de preparação dos dados foi a descrição dos dados. Em síntese, este passo é responsável por fornecer uma visão geral sobre os dados e do quão

desafiador é o problema através de uma série tarefas.

- Dimensão dos Dados: Consiste em saber a quantidade de dados que irão ser utilizadas no trabalho, ou seja, número de linhas e de colunas, e o número de empresas que serão analisadas. Esta tarefa é importante para que se tenha conhecimento se os recursos disponíveis para o projeto de mineração de dados são suficientes, no que diz respeito ao armazenamento e processamento dos dados. O cientista de dados deve sempre estar preparado para processar dados massivamente.

- Tipos de Variáveis: Nesta tarefa, são conhecidos os tipos das variáveis dos dados, numéricas, categóricas, temporais, etc. Isto pode ser útil, por exemplo, se uma variável "idade" estiver com o tipo categórico, e neste caso, basta alterar o seu tipo para numérico. Além disso, ter uma noção dos tipos dos dados logo no começo do projeto pode ser uma guia inicial para utilizar as técnicas de transformação dos dados posteriormente.

- Quantidade de Dados Faltantes: Também é necessário saber o volume de dados Not Applicable (NA) - dados faltantes. Ao descobrir o número de NAs, pode-se tomar duas decisões diferentes: a primeira é simplesmente não fazer o projeto, pois não há a quantidade de dados suficientes; a segunda decisão é preencher os campos vazios. Neste último caso, é fundamental compreender a razão dos dados faltantes, pois será relevante na hora de inserir novos dados.

- Estatística Descritiva: Esta tarefa consiste em fornecer um resumo geral dos dados, a fim de ter uma noção da grandeza dos dados, isto é, valores mínimos e máximos das variáveis, qual é o intervalo, qual é a média e mediana, entre outros. São estatísticas básicas para descrever os dados de uma maneira macro.

No presente estudo, foram analisadas 415 empresas, totalizando 15254 linhas e 75 colunas. Além disso, todos as variáveis do trabalho são do tipo numérico, exceto pela variável que corresponde ao código da ação. Logo, não foi necessário alterar nenhum tipo de variável. Outrossim, como na etapa de entendimento dos dados foram excluídas várias variáveis do conjunto de dados, pois continham muitos NAs, restaram apenas 3 linhas com dados faltantes, as quais foram descartadas visto que não haveria muita perda de informação.

Por fim, foram calculadas as seguintes medidas estatísticas sobre os dados: média e mediana para medir a tendência central; valor máximo, valor mínimo, intervalo e desvio padrão para medir a dispersão; assimetria e curtose para medir a simetria. Os resultados mostraram que a variável do valor da cotação, possuia valores negativos e valores discrepantes, ao observar os seus valores máximo, mínimo, intervalo e a média. Após uma pesquisa na base de dados, constatou-se que os papéis "MMAQ3", "MMAQ4" e

"VSPT3" estavam com valores discrepantes, e provavelmente houve um erro no momento de preenchimento dos dados. Além disso, verificou-se que praticamente todas as variáveis não possuiam uma distribuição normal, ao considerar os seus valores de assimetria e curtose, que estavam distantes do ideal. Esta foi uma informação relevante para decidir qual técnica utilizar para normalizar a escala dos dados em tarefas posteriores.

3.3.2 Engenharia de Variáveis

A engenharia de variáveis é uma das tarefas mais importantes dos projetos de mineração de dados, e fundamenta-se na derivação de novas variáveis a partir do conjunto de dados original. Este passo necessita de criatividade, e muito conhecimento sobre o problema e os dados. Criar novas variáveis muitas vezes é essencial para explicar o fenômeno que se está modelando e pode melhorar significativamente a performance do modelo.

Dado que o objetivo do presente estudo era tentar modelar o fenômeno de valorização das ações, verificou-se então que calcular indicadores fundamentalistas a partir das variáveis existentes, como o P/L ou **ROE!** (**ROE!**) poderiam ser vitais para prever se uma ação iria render mais que 3% até o próximo trimestre.

A Tabela 1 expõe todos os indicadores que foram calculados e serão utilizados no decorrer do trabalho:

Tabela 1 – Indicadores Fundamentalistas Calculados.

Liquidez Corrente	LPA	P/L	VPA
P/VP	EBIT	EV/EBIT	NOPLAT
ROIC	ROE	Pay-Out	Dividend Yield
Margem Ebit	Margem Líquida	P/Ebit	PSR
P/Ativos	Giro do Ativo Total	ROA	EBITDA

3.3.3 Filtragem de Variáveis

Antes de realizar a análise exploratória dos dados, é muito importante filtrar os dados. Mais precisamente, deseja-se nesta tarefa descartar os dados que não serão mais necessárias para o restante do projeto, que estão corrompidos, ou que infringem alguma restrição de negócio.

1. Filtragem de Linhas: foram excluídas as linhas dos papéis "MMAQ3", "MMAQ4" e "VSPT3", em virtude de seus dados discrepantes, conforme foi visto na tarefa de estatística descritiva.

2. Filtragem de Variáveis:

   - "Perc", que se refere a variação percentual das cotações até o próximo trimestre, foi utilizada para rotular as ações, não sendo mais necessária.

- "quantity", variável que continha a quantidade de ações ofertadas e usada na etapa de engenharia de variáveis, para realizar o cálculo dos indicadores fundamentalistas, não sendo mais necessária.
- "Adj Close", a variável de cotação, imprescindível para que se pudesse classificar as ações, não sendo mais necessária.

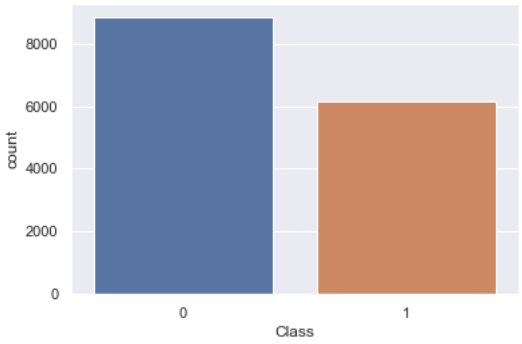
3.3.4 Análise Exploratória de Dados

A análise exploratória de dados busca descobrir quais variáveis impactam o fenômeno e qual é a força do impacto. Em outras palavras, ela serve para quantificar o efeito das variáveis em relação à variável resposta. Esta etapa expõe quais variáveis da base de dados são mais importantes para explicar para o modelo o comportamento do fenômeno.

3.3.4.1 Variável Resposta

Primeiramente, foi realizada a análise da variável alvo. Ações classificadas como "não confiáveis" receberam o valor 0, enquanto as "confiáveis" receberam o valor 1. Após esta análise, concluiu-se que há um número significativamente maior de ações classificadas como "não confiáveis", conforme é ilustrado na Figura 7. Neste conjunto de dados há uma clara desproporção entre o número de exemplos de cada classe. Contudo, não é o bastante a ponto de caracterizar um problema para o processo de classificação.

Figura 7 – Análise da Variável Resposta

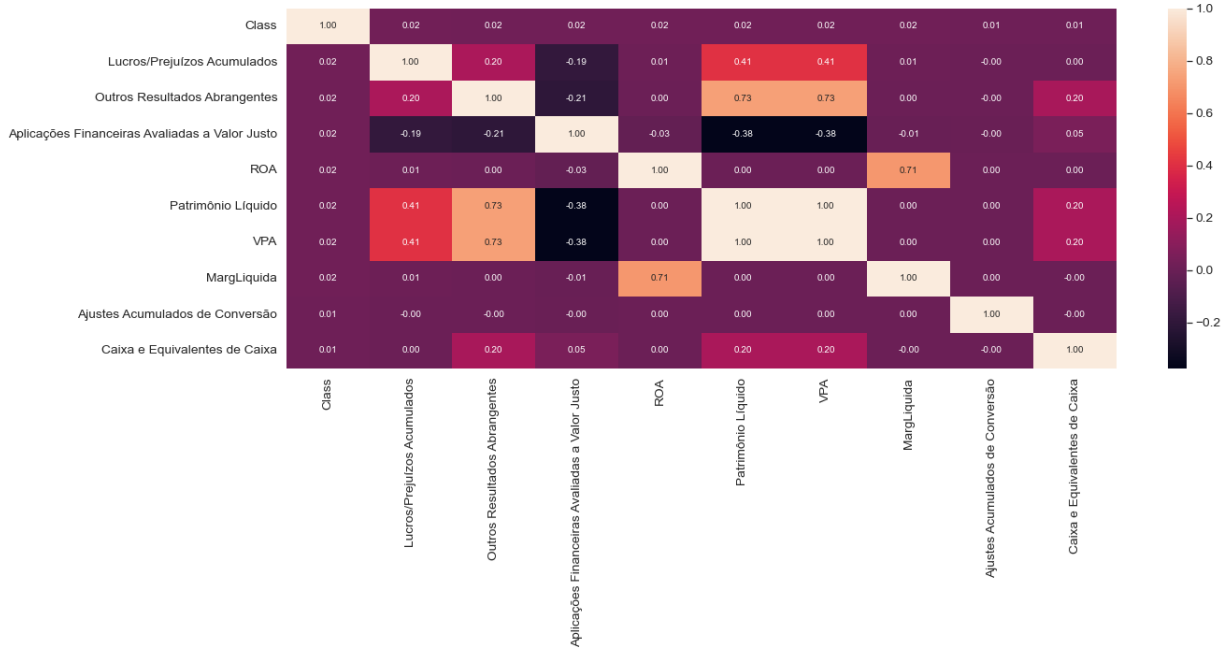


3.3.4.2 Análise Multivariada

A análise multivariada é definida como o estudo estatístico de dados onde são feitas medições múltiplas em cada unidade experimental e onde as relações entre medições multivariadas e sua estrutura são importantes (VIGHNESH, 2021). Ela pode ser usada para encontrar variáveis que carregam o mesmo conteúdo de informação, ou seja, variáveis que são linearmente dependentes. Neste caso, não é necessário que ambas as variáveis

sejam utilizadas para treinar o modelo, e caso uma seja tirada da base de dados, não há perda de informação. Uma das técnicas mais utilizadas na análise multivariada é a correlação, e diz respeito à relação de interdependência de duas ou mais variáveis.

Figura 8 – Análise de Correlações



No presente estudo, foi utilizado um gráfico de correlações, com o objetivo de expor não só quais variáveis são mais correlacionadas entre si, como também as que possuem mais influência sobre a variável alvo. A fim de permitir uma melhor visualização, a Figura 8 apresenta uma matriz de correlações apenas com as 10 variáveis mais correlacionadas com a variável alvo. Nota-se que nenhuma variável possui uma alta correlação com a variável resposta, revelando a alta complexidade do fenômeno no qual buscou-se modelar. Ademais, observa-se que as variáveis "VPA" e "Patrimônio Líquido" estavam totalmente correlacionadas, uma vez que o patrimônio líquido das empresas é utilizado para calcular o indicador fundamentalista VPA. Logo, como estas duas variáveis carregam a mesma informação, não foi necessário utilizar ambas para construir o modelo de classificação.

3.3.5 Preparação dos Dados

Também conhecido como modelagem dos dados, este passo tem por objetivo facilitar o aprendizado dos algoritmos de mineração de dados, ao garantir que todos os dados sejam numéricos e estejam na mesma escala. Além disso, nesta etapa também pode ser realizada a transformação na natureza dos dados, que de alguma forma beneficie o desempenho do modelo de classificação.

3.3.5.1 Transformações dos Dados para Variações Percentuais

A primeira ação realizada para preparar os dados foi transformá-los em variações percentuais. Isto é, com base em um relatório trimestral de qualquer ação escolhida, observou-se o quanto os valores mudaram do relatório anterior para o selecionado atualmente. Por exemplo, o ativo total da ação "PETR4" valorizou 2% do trimestre anterior para o atual, enquanto o passivo total desvalorizou 1,5%. O propósito desta transformação é dada por dois motivos: a primeira é em virtude dos dados variarem bastante de uma ação para a outra. Um patrimônio líquido de 60.000.000, por exemplo, pode ser excelente para uma empresa, e nem tanto para outra. Como o propósito deste trabalho é encontrar ações que irão se valorizar até o próximo trimestre, o tamanho dos dados não importa de fato, e sim o quanto influenciam na variável alvo; já o segundo motivo é em razão da própria variável alvo, e da forma como foi criada. Conforme foi explicado na subseção 3.2.4, as ações foram rotuladas com base na variação da cotação até o próximo trimestre. Isso permite que as variações percentuais dos dados possam ser usadas para explicar uma possível valorização ou desvalorização de uma ação. Essencialmente, com esta transformação, detectou-se se mudanças fundamentais do trimestre anterior para o atual afetaram os preços futuros, ou seja, os modelos foram treinados para analisar como a variação dos valores de cada um dos indicadores afetam a valorização das ações até o próximo trimestre. É importante ressaltar que só foi possível rotular as ações com base em trimestres futuros, pois foram coletados dados históricos das cotações dos anos anteriores e com isso o modelo criado foi treinado para ser capaz de predizer o rótulo das ações de um trimestre que nunca viu antes. A Figura 9 mostra um diagrama que esclarece melhor esta ideia.

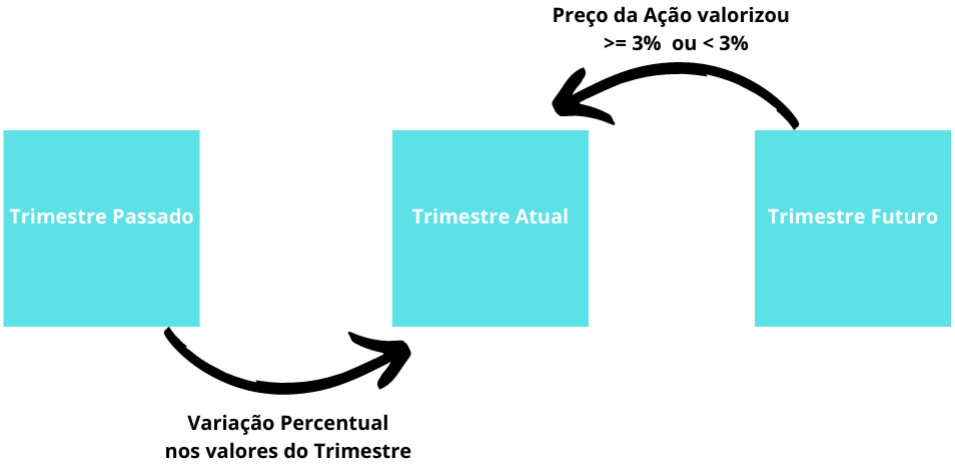
Esse método julgou se vale a pena investir em uma ação no trimestre. Evidentemente, não foi possível calcular a variação percentual dos dados do primeiro trimestre porque a natureza da análise exige um relatório trimestral passado.

3.3.5.2 Divisão dos Dados em Treino e Teste

Antes de normalizar os dados, é importante dividir o conjunto de dados em treino e teste. A ideia por trás deste passo é simples: evitar o *overfitting*. De acordo com Nikolaiev (2021), esse fenômeno ocorre quando um modelo funciona muito bem nos dados de treinamento, mas não consegue generalizar bem para dados não vistos antes. Dentre outras razões pelas quais isso pode acontecer, pode ser devido ao ruído nos dados ou pode ser que o modelo tenha aprendido a prever entradas específicas em vez dos parâmetros preditivos que poderiam ajudá-lo a fazer previsões corretas. Normalmente, quanto maior a complexidade de um modelo, maior a chance de sofrer um *overfitting*.

Por outro lado, o *underfitting* ocorre quando o modelo apresenta desempenho ruim mesmo nos dados que foram usados para treiná-lo (NIKOLAIEV, 2021). Na maioria dos casos, o *underfitting* ocorre porque o modelo não é adequado para o problema que se está

Figura 9 – Método para classificar as ações



As ações foram rotuladas com base em trimestres futuros, e seus indicadores foram transformados com base em trimestres passados

tentando resolver. Normalmente, isso significa que o modelo é menos complexo do que o necessário para aprender as variáveis que podem ser comprovadamente preditivos.

Portanto, criar diferentes amostras de dados é a abordagem mais comum que pode ser usada para identificar esses tipos de problemas. Dessa forma, pode-se utilizar o conjunto de dados de treinamento para treinar os modelos, e o conjunto de teste para avaliar qual modelo apresenta o melhor desempenho. Esta forma de divisão é chamada de *hold-out*.

No presente estudo, foram selecionados, aleatoriamente, 70% do conjunto de dados para treinar os modelos, com o intuito de garantir que os modelos sejam treinados com dados que contenham padrões de todo o período escolhido para a realização deste trabalho. Os 30% restantes foram utilizados para avaliar a performance dos modelos.

3.3.5.3 *Rescaling*

Uma das características desta base de dados é a diferença entre os *ranges* das variáveis, visto que cada uma delas possuem escalas totalmente diferentes umas das outras. Os algoritmos de mineração de dados tendem a dar mais importância para as variáveis que possuem um *range* maior. Por exemplo, o atributo “Ativo Circulante” possui um intervalo muito maior do que o indicador P/L, mas este não é um motivo para julgar qual é mais importante. Assim, com o objetivo de equalizar esta importância, deve-se colocar todas as variáveis na mesma escala, para que não haja um enviesamento do modelo.

Uma das técnicas utilizadas para isto é a normalização, que reescala o centro da variável para 0, com o desvio-padrão igual a 1. Todavia, ela é ideal para variáveis que possuem uma distribuição normal, e como foi visto nas estatísticas descritivas, na

subseção 3.3.1, os dados estão longe da normalidade.

Por isso, neste estudo foi utilizado o *Robust Scaler*, uma técnica eficaz em dados que não possuem uma distribuição normal. Ela normaliza os dados utilizando a seguinte fórmula:

$$X_{norm} = \frac{\text{Valor a ser Normalizado} - \text{Q}_1(x)}{\text{Q}_{3(x)} - \text{Q}_{1(x)}}$$

Outro motivo para utilizar esta técnica foi devido ao grande número de *outliers* (valores discrepantes) presentes na base de dados, dado a expressiva diferença de valores entre as empresas. A técnica *MinMax Scaler*, por exemplo, utiliza o intervalo entre o seu valor mínimo e máximo em sua fórmula, e por isso, é afetado por valores fora do padrão. Por outro lado, como o *Robust Scaler* é calculado utilizando a intervalo interquartil, como pode ser visto em sua fórmula, ele é imune a este tipo de problema.

3.3.6 *Feature Selection* (Seleção de Variáveis)

A seleção de variáveis vai de encontro a um dos princípios da aprendizagem estatística, que é uma teoria que garante o aprendizado dos modelos. Se trata do princípio de *Occam's Razor* (A navalha de Occam), que diz que a explicação mais simples sobre um fenômeno observado deveria prevalecer sobre explicações mais complexas. Isto significa que deve-se dar preferência por modelos mais simples, pois provavelmente ele irá generalizar melhor.

Para garantir a simplicidade dos modelos é necessário diminuir o número de variáveis do conjunto de dados, e manter apenas aquelas que melhor explicam o fenômeno. Adicionalmente, existem variáveis que carregam o mesmo conteúdo de informação para explicar o fenômeno, chamadas de variáveis colineares. Em suma, a seleção de variáveis é responsável por tornar os modelos mais simples e remover colinearidade dos dados.

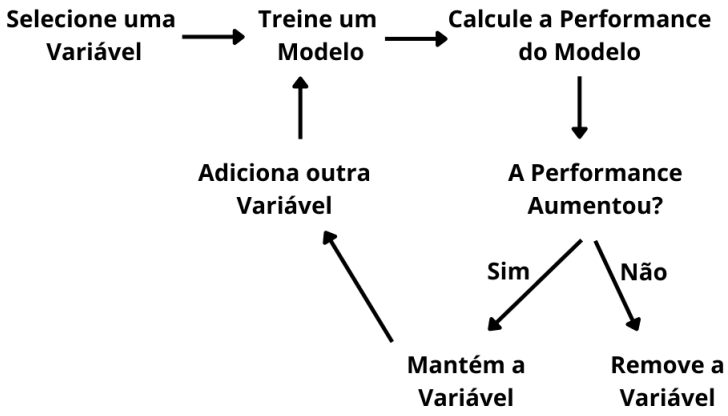
Neste trabalho foi utilizado um método chamado Seleção por *Subset* (*Wrapper Method*). O seu funcionamento é simples e consiste em manter variáveis que agregam para melhorar a performance do modelo, e descartar as que não agregam. O diagrama ilustrado na Figura 10 mostra o passo a passo utilizado neste método:

É importante ressaltar que as variáveis são escolhidas de forma aleatória, e a performance do modelo é calculado a partir de sua acurácia, isto é, dentre todas as classificações, quantas o modelo classificou corretamente.

3.3.6.1 Boruta

O Boruta é um algoritmo de seleção de recursos robusto, estatisticamente fundamentado, e se baseia na seleção por *subset*. Segundo Mazzanti (2020), este algoritmo

Figura 10 – Seleção por *Subset*



possui duas ideias brilhantes que lhe permite ser tão performático para vários problemas de mineração de dados. A primeira é que as variáveis não competem entre si. Em vez disso, elas competem com uma versão aleatória delas, chamadas *Shadow Features*, como pode ser observado na Figura 11. Neste exemplo dado por Mazzanti (2020) é possível notar que as *shadow features* são versões das próprias variáveis existentes, idade (*age*), altura (*height*) e peso (*weight*), com os seus valores embaralhados. Basicamente, o Boruta treina um modelo que é passado para ele, em várias iterações, e em cada uma das vezes as variáveis originais são comparadas com a *Shadow Feature* mais importante, segundo a estimação do modelo. Conforme pode ser visto na Figura 12, a *shadow feature* com maior importância é a "shadow_height", cujo valor é 14. Observa-se que as variáveis originais que possuem uma importância menor do que ela recebem o valor 0, caso contrário, 1, e isso é chamado de *hit* (acerto). Este passo é repetido várias vezes, com um embaralhamento diferente das linhas das *shadow features*, com a finalidade de garantir uma confiabilidade maior sobre a importância das variáveis originais. Na Figura 13, é dado um exemplo da tabela de *hits* após 20 repetições.

Figura 11 – Novo Conjunto de Dados Criado pelo Boruta

	age	height	weight	shadow_age	shadow_height	shadow_weight
0	25	182	75	51	176	75
1	32	176	71	32	182	71
2	47	174	78	47	168	78
3	51	168	72	25	181	72
4	62	181	86	62	174	86

Fonte: Mazzanti (2020)

Figura 12 – Tabela de Acertos

	age	height	weight	shadow_age	shadow_height	shadow_weight
feature importance %	39	19	8	11	14	9
hits	1	1	0	-	-	-

Fonte: Mazzanti (2020)

Figura 13 – Resultado após 20 Repetições

	age	height	weight
hits (in 20 trials)	20	4	0

Fonte: Mazzanti (2020)

A segunda ideia engenhosa do Boruta, de acordo com Mazzanti (2020), é utilizar a distribuição binomial para decidir se uma variável é útil. Para isso, ele utiliza o *p-valor* (um conceito estatístico utilizado para aceitar ou rejeitar uma hipótese) para determinar a importância das variáveis. Então é calculado o *p-valor* de cada uma das variáveis originais, e é construído uma distribuição binomial. Para decidir quais variávies serão aceitas e quais serão rejeitadas, é escolhido um *threshold* (limite), para comparar com o *p-valor* de cada um dos recursos calculados. As variáveis que tiverem o *p-valor* maior do que o *threshold* serão aceitas, e as que forem menores serão rejeitadas. O resultado final é o conjunto de variáveis que o Boruta considerou relevantes para o modelo.

Ao utilizar o Boruta neste trabalho, obteve-se 11 variáveis consideradas importantes para o modelo, utilizando um *threshold* fornecido pelo próprio algoritmo. Contudo, dentre elas haviam as variáveis “VPA” e “Patrimônio Líquido”, e como foi analisado na subseção 3.3.4.2, elas são totalmente correlacionadas entre si, indicando que carregam o mesmo conteúdo de informação. Logo, para não aumentar desnecessariamente a dimensionalidade dos dados, optou-se por descartar a variável “VPA”. O critério de escolha foi a variável com menor correlação com a variável alvo.

A Tabela 2 apresenta quais foram as variáveis utilizadas para treinar os modelos de classificação:

Tabela 2 – Variáveis Escolhidas pelo Boruta.

Ativo Total	P/Ativos
Caixa e Equivalentes de Caixa	P/L
Imobilizado	MargEbit
Passivo Total	Financeiras
Patrimônio Líquido	P/VP
Lucros/Prejuízos Acumulados	-

4 Resultados

Neste capítulo estão descritas as etapas de modelagem, avaliação e implantação, do manual CRISP-DM.

4.1 Modelagem e Avaliação

O processo de Modelagem é onde, de fato, a mineração de dados ocorreu. Os dados que foram preparados nas etapas anteriores puderam ser submetidos aos algoritmos de classificação. Os métodos utilizados foram os seguintes: *Dummy Classifier*; *Logistic Regression*; *Decision Tree*; *K-Nearest Neighbors*; *Naive Bayes*; *Multilayer Perceptron*; *Random Forest*.

O processo de Avaliação é onde os resultados obtidos são avaliados. Neste passo também é importante definir quais métricas serão utilizadas para avaliar a performance dos modelos. Neste caso, deve-se levar em consideração o contexto do problema, reforçando a importância do entendimento do problema na seção 3.1.

4.1.1 Métricas de Avaliação

Antes de começar a etapa de modelagem, é fundamental definir quais serão as métricas utilizadas para determinar a performance dos modelos. Esta escolha deve ser feita com precaução, pois irá afetar diretamente o resultado final do projeto de mineração de dados. Em problemas de classificação, existem várias métricas diferentes para avaliar modelos. Contudo, para entendê-las, é necessário primeiro entender o que é uma matriz de confusão.

4.1.1.1 Matriz de Confusão

De acordo com Rodrigues (2019), uma matriz de confusão é uma tabela que indica os erros e acertos do modelo, comparando-os com o resultado esperado. A Figura 14 demonstra um exemplo de uma matriz de confusão:

- Verdadeiro Positivo: número de vezes que o modelo classificou como Positivo, e estavam corretos;
- Verdadeiro Negativo: número de vezes que o modelo classificou como Negativo, e estavam corretos;

Figura 14 – Matriz de Confusão

		Prevista	
		Não	**Sim**
Real	**Não**	Verdadeiro Negativo	Falso Positivo
	Sim	Falso Negativo	Verdadeiro Positivo

- Falso Positivo: número de vezes que o modelo previu a classe Positivo quando o valor real era classe Negativo;
- Falso Negativo: número de vezes que o modelo previu a classe Negativo quando o valor real era classe Positivo.

4.1.1.2 Principais Métricas de Avaliação

Com base na matriz de confusão, é possível calcular as métricas de avaliação para a classificação. Estas são as principais métricas, segundo Rodrigues (2019):

- Acurácia: indica uma performance geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente. É calculada da seguinte forma:

$$\frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}$$

- Precisão: dentre todas as classificações de classe Positivo que o modelo fez, quantas estão corretas. É calculada da seguinte forma:

$$\frac{\text{VP}}{\text{VP} + \text{FP}}$$

- *Recall*: dentre todas as situações de classe Positivo como valor esperado, quantas estão corretas. É calculada da seguinte forma:

$$\frac{\text{VP}}{\text{VP} + \text{FN}}$$

- *F1-Score*: média harmônica entre precisão e recall. É calculada da seguinte forma:

$$\frac{2 \times \text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}$$

4.1.1.3 Métricas Utilizadas

Neste trabalho, o grande objetivo é classificar corretamente ações como "confiáveis" em um dado trimestre. Logo, como o conjunto de dados estava desbalanceado, optou-se por não utilizar a acurácia, visto que podia levar a falsas conclusões sobre a performance dos modelos. Por exemplo, um modelo podia ter uma boa acurácia, no entanto, previa muitas ações "não confiáveis" corretamente, mas poucas ações "confiáveis" corretamente. Isto significa que apesar do modelo ter classificado erroneamente muitas ações "não confiáveis" como "confiáveis" (Falsos Positivos), é possível que ele tenha uma boa acurácia se tiver uma alta taxa de acertos com as ações "não confiáveis", dado que elas representam a grande maioria dos dados.

Em contrapartida, a precisão pode ser usada em uma situação em que os Falsos Positivos são considerados mais prejudiciais que os Falsos Negativos. Por exemplo, ao classificar uma ação como "confiável" em um certo trimestre, é necessário que o modelo esteja correto, mesmo que acabe classificando ações "confiáveis" como ações "não confiáveis" (Falso Negativo) no processo. Ou seja, o modelo deve ser preciso em suas classificações, pois a partir do momento que consideramos uma ação como "confiável" quando na verdade ela não é, uma grande perda de dinheiro pode acontecer. Por este motivo, a precisão foi a principal métrica para avaliar os classificadores.

Outrossim, o *recall* também foi uma importante métrica de avaliação, pois não pretendia-se ter um modelo muito conservador, que se arriscava pouco. Por exemplo, um modelo podia ter uma alta precisão, mas classificava apenas 4 ações no total. Entretanto, o *F1-Score* não foi considerado, pois havia uma clara ordem de prioridade entre a precisão e a recall, e as duas não tinham o mesmo grau de importância.

4.1.2 Modelagem

A seguir serão relatados os resultados e considerações de cada um dos modelos criados. Ao final, será escolhido um modelo para ser utilizado nas etapas finais do projeto.

4.1.2.1 *Dummy Classifier*

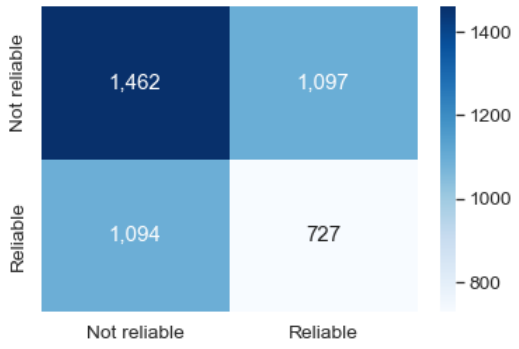
O *Dummy Classifier* foi utilizado neste trabalho como um *baseline*, a fim de compará-lo com modelos mais complexos. Ele simplesmente ignora os dados de entrada e depende exclusivamente dos valores da variável alvo para fazer as previsões. O principal parâmetro deste algoritmo é a estratégia (*strategy*) e permite definir o comportamento específico desta *baseline*. No presente estudo, utilizou-se a estratégia estratificada (*stratified*), ou seja, as previsões realizadas por este classificador serão feitas na mesma proporção da variável resposta, vista na subseção 3.3.4.1. Portanto, a ideia de se utilizar o *Dummy Classifier* foi com o objetivo de verificar se os modelos utilizados não são melhores que um simples

palpite.

Tabela 3 – Métricas do *Dummy Classifier*.

Classe	**Precisão**	**Recall**
Não Confiável	57%	57%
Confiável	40%	40%

Figura 15 – Matriz de Confusão do *Dummy Classifier*



Os resultados deste modelo podem ser vistos na Tabela 3 e Figura 15. Como o foco era classificar corretamente as ações "confiáveis", foram analisadas apenas a *precision* e *recall* das ações com o rótulo "*Reliable*". Observa-se que o *Dummy Classifier* teve uma precisão de 40%, e este valor foi a linha de comparação para o restante dos modelos. Em relação a *recall*, percebe-se que a porcentagem de "chutes" é igual a porcentagem de classes rotuladas como "confiáveis", em conformidade com a estratégia definida anteriormente.

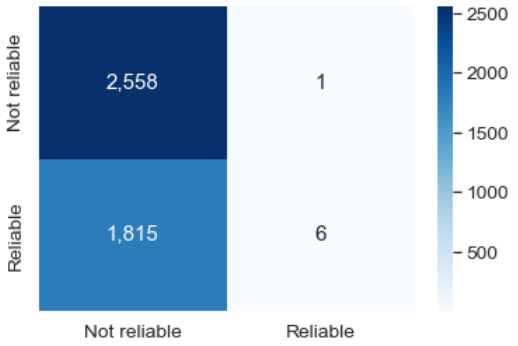
4.1.2.2 *Logistic Regression*

O *Logistic Reression* obteve uma precisão de 86% para classificar ações confiáveis, uma performance muito melhor que o modelo *baseline*. No entanto, o seu *recall* foi muito baixo, conforme é mostrado na Tabela 4, e como foi mencionado anteriormente, um modelo que rotula poucas ações "*Reliable*" não é o ideal para este tipo de problema. A Figura 16 mostra que o *Logistic Reression* preveu apenas 6 ações confiáveis, de um total de 1821 presentes nos dados de teste.

Tabela 4 – Métricas do *Logistic Reression*.

Classe	**Precisão**	**Recall**
Não Confiável	58%	99.9%
Confiável	86%	0.01%

Figura 16 – Matriz de Confusão do *Logistic Reression*



4.1.2.3 *Decision Tree*

Os resultados do modelo de *Decision Tree* foram melhores que os dois anteriores, ao analisar conjuntamente a precisão e *recall*. Observa-se na Tabela 5 e Figura 17 que ele obteve 49% de precisão para a classe "*Reliable*", e conseguiu prever mais da metade das ações confiáveis presentes na base de dados. Todavia, ainda não era o desempenho desejado.

Tabela 5 – Métricas do *Decision Tree*.

Classe	**Precisão**	**Recall**
Não Confiável	64%	63%
Confiável	49%	51%

Figura 17 – Matriz de Confusão do *Decision Tree*

4.1.2.4 *K-Nearest Neighbors*, *Naive Bayes* e *Multilayer Perceptron*

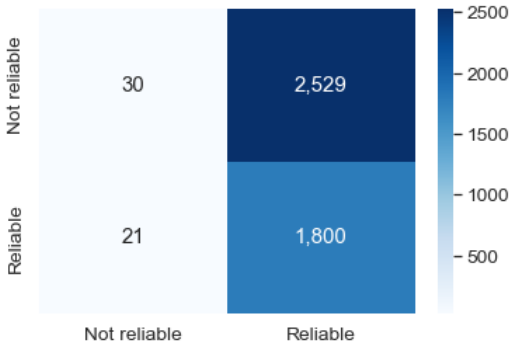
Os modelos de *K-Nearest Neighbors*, *Naive Bayes* e *Multilayer Perceptron* também não tiveram uma boa performance, tendo obtido precisões de 50%, 42% e 47%, respectivamente, ao prever ações "confiáveis". Contudo, foi interessante analisar os resultados do

Naive Bayes, pois o seu *recall* foram 99% (Tabela 6), ou seja, das 1821 ações rotuladas como "Reliable", o modelo conseguiu prever 1800 delas, como pode ser vista na Figura 18. Apesar disso, o modelo classificou equivocadamente 2529 ações como "confiáveis", esclarecendo a razão de a sua precisão ter sido baixa.

Tabela 6 – Métricas do *Naive Bayes*.

Classe	**Precisão**	**Recall**
Não Confiável	59%	01%
Confiável	42%	99%

Figura 18 – Matriz de Confusão do *Naive Bayes*



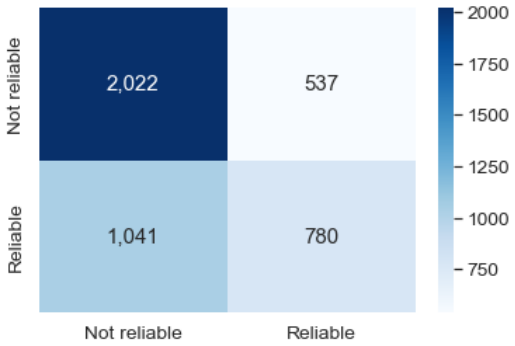
4.1.2.5 *Random Forest*

O modelo de *Random Forest* obteve a melhor performance dentre todos criados, como pode ser observado na Tabela 7 e Figura 19. A sua precisão para classificar ações "confiáveis" foi de 59%, com um *recall* de 43%, considerados bons resultados. Em vista de todos os resultados encontrados até esta etapa do projeto, a escolha mais sensata do modelo a ser utilizado nas próximas tarefas seria o *Random Forest*. Entretanto, ainda era necessário utilizar uma técnica imprescindível para avaliar a real performance dos modelos: o *Cross-Validation*.

Tabela 7 – Métricas do *Random Forest*.

Classe	**Precisão**	**Recall**
Não Confiável	66%	79%
Confiável	59%	43%

Figura 19 – Matriz de Confusão do *Random Forest*



4.1.2.6 *Cross Validation*

Também conhecida como validação cruzada, o *Cross-Validation* é um método estatístico comumente usado em projetos de mineração de dados para estimar a habilidade de modelos em realizar predições (BROWNLEE, 2018). Ele diminui o viés em relação ao *hold-out*, pois possibilita pegar toda a variabilidade do conjunto de dados. Em outras palavras, é possível que os dados de teste sejam propícios para que um modelo tenha mais acertabilidade nas predições, porém, ao utilizar o *Cross-Validation*, todos os dados serão utilizados para treiná-lo e testá-lo. Logo, obtêm-se uma resposta do comportamento do modelo em todos os cenários possíveis. A validação cruzada funciona da seguinte forma:

1. Embaralha o conjunto de dados.
2. Divide o conjunto de dados em k grupos.
3. O algoritmo se repete k vezes, e a cada iteração um grupo é escolhido para teste e o restante dos grupos é utilizado para treinamento, até que todos os grupos sejam usados para validação.
4. A cada iteração é calculada a performance do modelo.
5. No final, a real performance do modelo é a média dos valores da métrica escolhida.

Neste trabalho, para avaliar a performance dos modelos com o *Cross-Validation*, foram utilizadas as mesmas métricas da etapa anterior e o valor do k igual a 5. Contudo, foram calculadas com base apenas na classe “Reliable”, com excessão da acurácia que se refere à taxa geral de acertos. Conforme pode ser visto na Tabela 8, o *Random Forest* obteve o melhor resultado novamente, com precisão de 60% e um desvio padrão de 2%. Em razão da complexidade do fenômeno que se estava tentando modelar, considerou-se que esta é uma boa precisão. Portanto, o *Random Forest* foi o algoritmo escolhido para ajustar os seus hiperparâmetros, e criar o modelo final.

Tabela 8 – Performance dos Modelos - *Cross Validation*.

Modelo	**Precisão**	**Recall**
Random Forest	60%	42%
Multilayer Perceptron	50%	24%
Decision Tree	49%	50%
K-Nearest Neighbors	49%	43%
Dummy	42%	41%
Logistic Regression	37%	0.01%
Naive Bayes	37%	59%

4.1.3 Ajuste de Hiperparâmetros

No contexto de mineração de dados, hiperparâmetros são todos os parâmetros que um modelo possui para aprender um comportamento. Por exemplo, redes neurais tem o número de neurônios, número de camadas escondidas, a função de ativação, etc. A etapa de ajuste de hiperparâmetros é necessária para descobrir quais os melhores valores para os parâmetros do modelo, a fim de fazê-lo ter um melhor desempenho, ou seja, encontrar o conjunto de parâmetros que maximiza o resultado do modelo.

Segundo Johnson (2016), existem 3 estratégias diferentes para ajustar os hiperparâmetros:

- *Random Search*: Define valores para cada um dos hiperparâmetros aleatoriamente.
- *Grid Search*: Define todas as combinações possíveis de valores que os hiperparâmetros podem assumir.
- *Bayesian Search*: Define valores para os hiperparâmetros seguindo a teoria de Bayes.

A estratégia escolhida no presente estudo foi o *Random Search*. Este método escolhe, aleatoriamente, um valor diferente para cada parâmetro do algoritmo que esteja sendo utilizado, e avalia a performance do modelo com esta configuração de hiperparâmetros. O conjunto de valores de cada parâmetro, que podem ser selecionados pelo *Random Search*, é escolhido arbitrariamente. Portanto, é importante entender o funcionamento do algoritmo de classificação, e quais os seus parâmetros mais importantes. Este passo é repetido várias vezes, a depender do número de iterações escolhido. No final, é salvo a combinação de parâmetros que forneceu o melhor resultado.

Neste trabalho, optou-se por testar 100 combinações diferentes de hiperparâmetros para o modelo de *Random Forest*. A Figura 20 mostra o conjunto de valores para cada um dos parâmetros. A cada iteração, foi escolhido um valor aleatório de cada parâmetro, dentro dessas opções disponíveis. É importante ressaltar que o restante dos parâmetros que não apareceram na Figura 20 receberam o valor padrão.

Figura 20 – Conjuntos de Valores Escolhidos para os Parâmetros do *Random Search*

```
params = {'n_estimators':      [500, 1000, 1600, 2000, 2500, 3000, 3500],
          'max_features':      ['auto', 'sqrt'],
          'max_depth':         [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None],
          'min_samples_split': [2, 5, 10],
          'min_samples_leaf':  [1, 2, 4],
          'bootstrap':         [True, False],
          'random_state':      [13]
          }
```

De acordo com a documentação do Scikit-Learn[1] (biblioteca utilizada para implementar o *Random Forest*), estas são as definições de cada um dos parâmetros usados:

- *n_estimators*: O número de árvores na floresta.
- *max_features*: O número de recursos a serem considerados ao procurar a melhor divisão.
- *max_depth*: A profundidade máxima da árvore.
- *min_samples_split*: O número mínimo de amostras necessárias para dividir um nó interno.
- *min_samples_leaf*: O número mínimo de amostras necessárias para estar em um nó folha.
- *bootstrap*: Se as amostras de bootstrap serão usadas ao construir as árvores.

De forma análoga à etapa de modelagem, a performance do modelo foi avaliada com base na sua precisão para prever ações "confiáveis". A Tabela 9 mostra qual foi a configuração de hiperparâmetros ideal encontrada em comparação com os valores padrões utilizados na etapa anterior do projeto:

Tabela 9 – Combinação de Hiperparâmetros.

Hiperparâmetro	**Padrão**	**Ajustado**
n_estimators	100	3000
min_samples_split	2	10
max_features	sqrt	auto
max_depth	None	80
bootstrap	True	True

Com os hiperparâmetros definidos, foi criado um novo modelo de *Random Forest*, e os mesmos dados da etapa de modelagem foram utilizados para treiná-lo e testá-lo. A

[1] <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

Tabela 10 revela que a sua precisão para ações "confiáveis" melhorou 1% em relação ao modelo não ajustado, com a técnica *hold-out*. Além disso, é importante ressaltar que o *recall* para as ações "não confiáveis" aumentou para 81%, evidenciando que o modelo detectou grande parte das ações com este rótulo na base de dados, como pode ser visto na Figura 21.

Ao realizar a validação cruzada, o modelo ajustado de *Random Forest* também melhorou em 1% a sua precisão, com um desvio padrão de 2%, como é mostrado na Tabela 11.

Tabela 10 – Métricas do Modelo Ajustado.

Classe	**Precisão**	**Recall**
Não Confiável	66%	81%
Confiável	60%	40%

Figura 21 – Matriz de Confusão do Modelo Ajustado

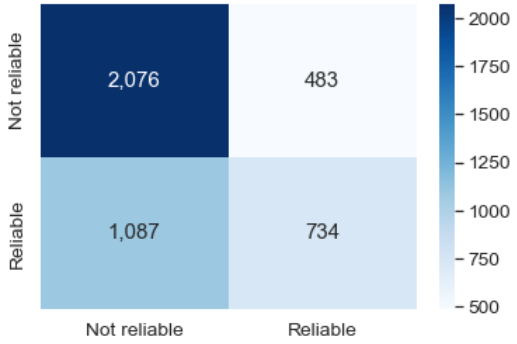


Tabela 11 – Métricas do Modelo Ajustado - *Cross Validation*.

Precisão	**Recall**
61%	39%

A Figura 22 mostra os valores do modelo *Random Forest* antes e depois do ajuste de hiperparâmetros:

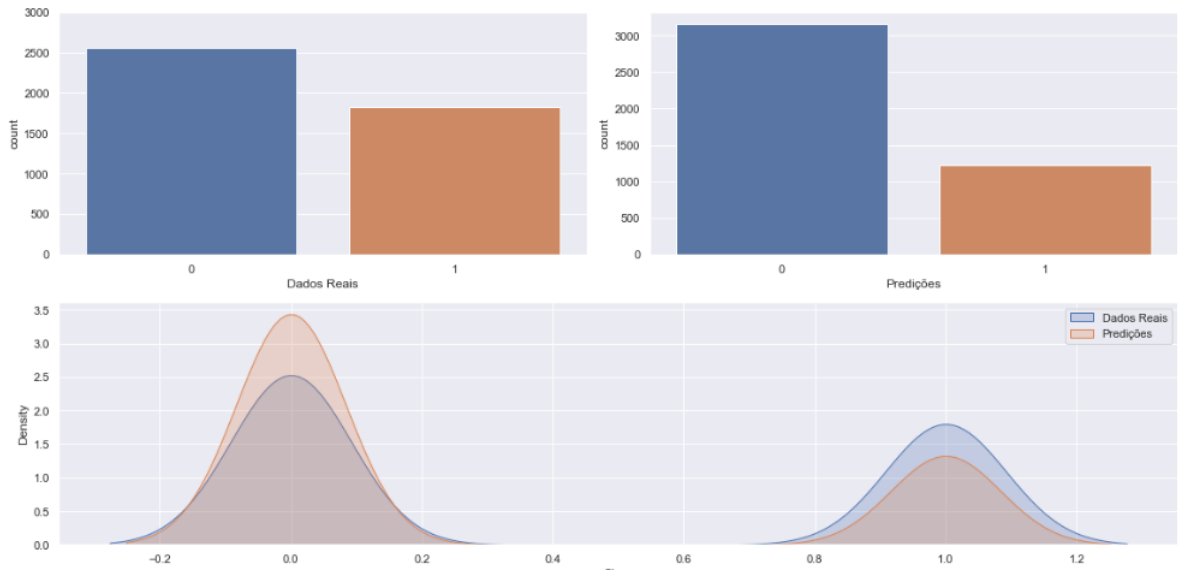
Figura 22 – Comparação do Modelo de *Random Forest* Antes e Depois do Ajuste de Hiperparâmetros

		Modelo Inicial	Modelo Ajustado
Hold-Out	**Precisão**	59%	60%
	Recall	43%	40%
Cross-Validation	**Precisão**	60%	61%
	Recall	42%	39%

4.1.4 Avaliação do Modelo

Para avaliar a performance do modelo, foram criados três gráficos, ilustrados na Figura 23. Os dois primeiros são gráficos de barras que mostram a quantidade de dados reais e preditos para as classes "não confiável" e "confiável". De forma análoga, o terceiro gráfico trata-se de um gráfico de densidade para visualizar a distribuição de observações dos dados reais e preditos, para ambas as classes do conjunto de dados. Ao analisar estes gráficos, concluiu-se que o modelo criado tende a prever muito mais ações "não confiáveis" do que de fato são. No entanto, como foi exposto no subseção 4.1.1.3, não há problema em o modelo prever erroneamente muitas ações "não confiáveis", contanto que preveja, com boa precisão, um número razoável de ações "confiáveis".

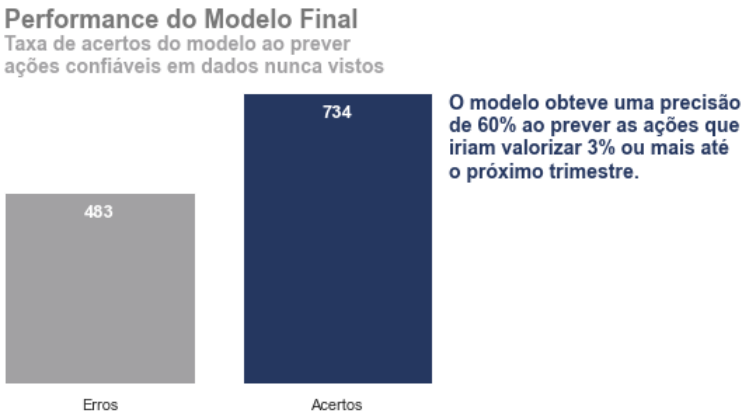
Figura 23 – Gráficos de Avaliação do Modelo Final



A Figura 24 mostra o desempenho do modelo final para prever ações confiáveis. Ele obteve uma precisão de 60% ao prever 734 ações "confiáveis" corretamente e 483 incorretamente. Apesar de não ser o resultado ideal, esta precisão correspondeu às expectativas

do trabalho. É importante ressaltar que vários fatores importantes para o mercado de ações não foram considerados, como a análise setorial, macroeconômica, ou até análise de sentimentos, onde poderia-se considerar a opinião pública. Desse modo, visto que foram utilizados apenas os dados de balanço patrimonial, demonstrativos de resultados e alguns indicadores fundamentalistas, entende-se que a solução encontrada foi aceitável. No Capítulo 5, foram propostas possíveis melhorias, e algumas recomendações que podem vir a melhorar o resultado do presente estudo.

Figura 24 – Performance do Modelo Final



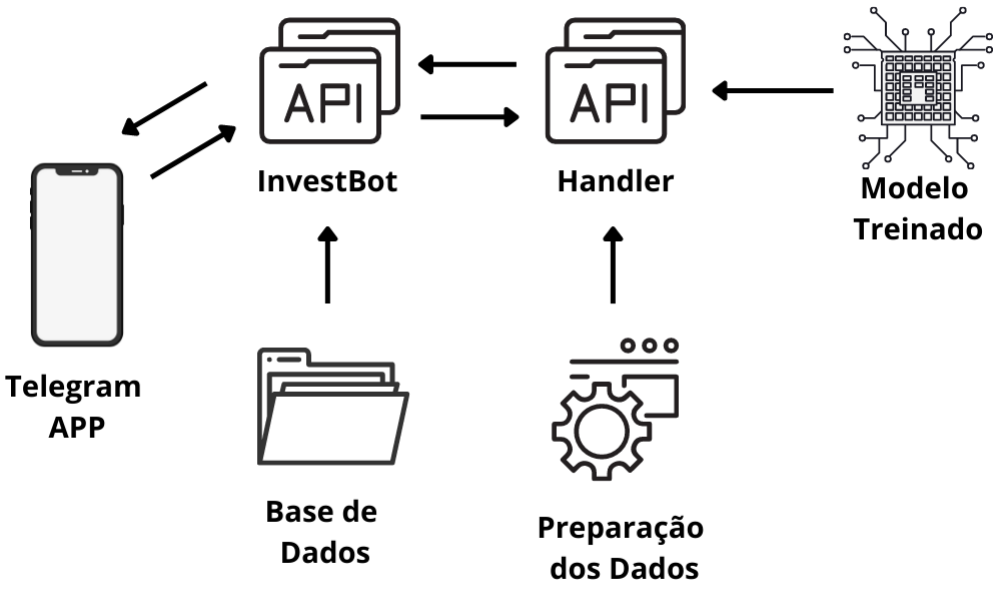
4.2 *Deployment* (Implementação)

O último passo deste trabalho foi colocar o modelo criado em produção para que outras pessoas também possam utilizá-lo. Ou seja, tirar todo o trabalho desenvolvido na máquina local e colocá-lo em um servidor em nuvem, que possibilite acessar as predições de qualquer máquina remota. Este procedimento se chama *deploy* e se trata de uma publicação de um projeto em um ambiente de produção, com o intuito de tornar a solução acessível a qualquer consumidor.

Após a implementação, é necessário criar uma ferramenta que utilize o modelo em nuvem para fornecer as predições aos usuários finais. Visto que este trabalho foi desenvolvido para resolver o problema dos investidores, optou-se por criar um bot no Telegram, pois é uma ferramenta acessível e fácil de usar.

A arquitetura do ambiente de produção criada para este trabalho é dada na Figura 25:

Figura 25 – Arquitetura do Ambiente de Produção



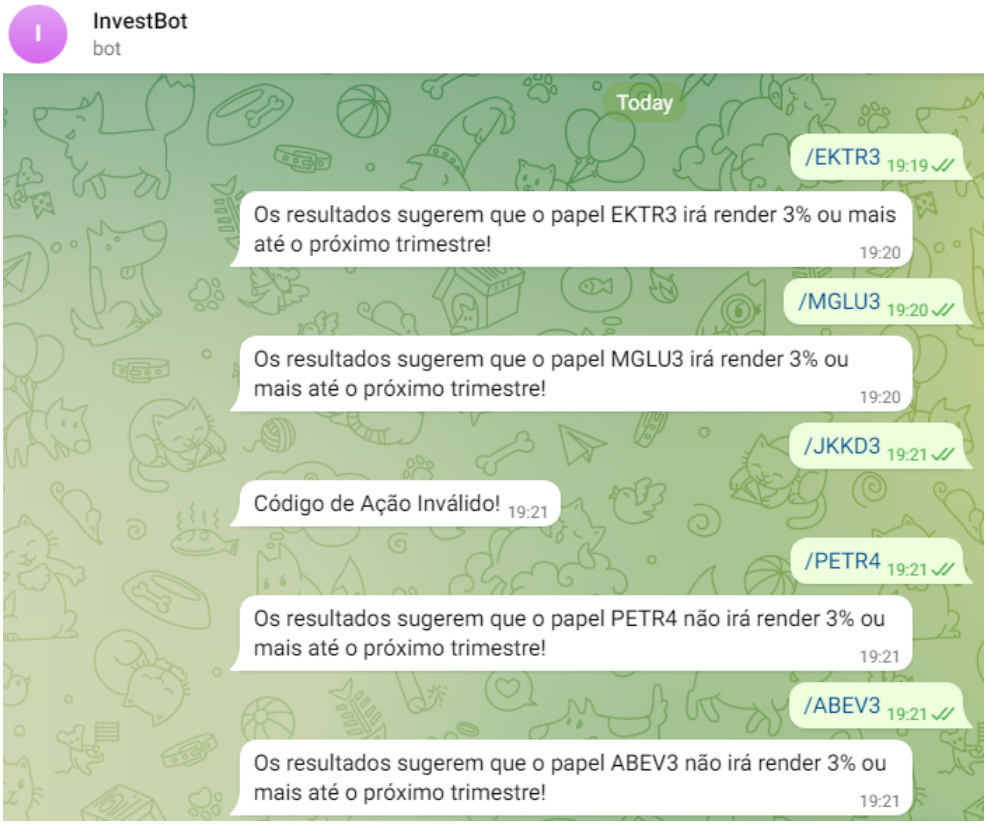
O primeiro componente desenvolvido foi uma *Application Programming Interface* (API) chamada *Handler*, que foi implementada utilizando um *framework* chamado *Flask* [2] (uma biblioteca de desenvolvimento *web* do *Python*). Esta API espera uma requisição e recebe como entrada os dados de um papel. Assim que o *Handler* recebe a requisição, o modelo é carregado, e ele realiza a preparação dos dados utilizando uma classe com exatamente todas as limpezas e transformações utilizadas para criar o modelo. Esta é a única forma de garantir que o modelo tenha o mesmo desempenho do treinamento no ambiente de produção. Após consultar o modelo e a classe de preparação dos dados, o *Handler* retorna a previsão para a ação, ou seja, se ela irá ou não render 3% ou mais até o próximo trimestre.

Além disso, também foi criada uma API chamada *InvestBot*. Se trata de um bot no Telegram, que recebe como entrada o código de uma ação, e carrega os dados referentes aos dois últimos trimestres da ação presentes na base de dados. O motivo de ter coletado os dados dos dois últimos trimestres é para que fosse possível transformar os dados para variações percentuais, similar ao que foi feito nas etapas anteriores do trabalho, desta forma, a predição buscada é apenas a do último trimestre da ação. Assim que os dados são carregados, o *InvestBot* faz a requisição ao *Handler*, envia para ele os dados da ação e recebe a predição.

[2] <https://flask.palletsprojects.com>

É importante ressaltar que tanto *Handler* quanto o *InvestBot* foram armazenados em um servidor na nuvem através de uma ferramenta chamada *Heroku*[3]. Com isso, foi possível criar o bot no Telegram, que recebe do usuário o código da ação e faz a requisição às duas APIs mencionadas anteriormente para receber a predição. Como é possível observar na Figura 26, o bot retorna uma mensagem ao usuário com a previsão se a ação irá ou não render 3% ou mais até o próximo trimestre. Entretanto, caso o usuário envie um código errado ou de uma ação que não está presente na base de dados, uma mensagem de erro é enviada.

Figura 26 – InvestBot - Bot no Telegram



Ao analisar a solução criada, constatou-se que é uma ferramenta útil e muito acessível, pois está na palma da mão do investidor. O próximo passo para finalizá-la é coletar os dados atuais das ações, para que as previsões sejam de um trimestre futuro. Apesar disso, considerou-se que o *InvestBot* realmente pode apoiar a tomada de decisão dos investidores, e ajudá-los a realizarem melhores investimentos. Entretanto, é importante deixar claro que este trabalho é um projeto de pesquisa e não deve ser utilizado como única fonte de informação para realizar investimentos na bolsa de valores.

[3] <https://heroku.com>

5 Conclusão

Neste trabalho foram coletados dados de balanço patrimonial e demonstrativos de resultados das empresas, com o intuito de criar um modelo de mineração de dados que seja capaz de classificar ações da bolsa de valores B3, como confiáveis ou não para investir. Inicialmente, os dados foram obtidos através dos *sites Fundamentus* e *Yahoo Finances*, no período de junho de 2011 a setembro de 2021. Em seguida, foram realizadas uma série de limpezas e transformações nos dados, a fim de prepará-los para serem submetidos aos algoritmos de mineração de dados. Além disso, nesta fase, as ações foram rotuladas como "confiáveis" se a ação rendeu 3% ou mais até o próximo trimestre, e caso contrário, foram rotuladas como "não confiáveis".

Com os dados preparados, foram criados os modelos *Dummy Classifier*, *Logistic Regression*, *Decision Tree*, *K-Nearest Neighbors*, *Naive Bayes*, *Multilayer Perceptron* e *Random Forest*, com o propósito de escolher o algoritmo que obtivesse a melhor performance. Utilizando as métricas de precisão e *recall*, constatou-se que o *Random Forest* obteve o melhor desempenho, com 59% de precisão para rotular ações "confiáveis".

Com o modelo escolhido, foi utilizado o algoritmo *Random Search* para ajustar os hiperparâmetros do modelo e melhorar a sua performance. Após o ajuste, o *Random Forest* obteve 60% de precisão ao classificar ações que renderam 3% ou mais até o próximo trimestre. Ao avaliar o modelo criado, concluiu-se que a sua performance para classificar ações "confiáveis" é aceitável, dada a complexidade e o período de previsão.

Ao final do trabalho, foram armazenados em um servidor em nuvem, o modelo e uma API que realiza a limpeza e preparação dos dados, para que a solução pudesse ser acessada por qualquer pessoa. Foi criado um bot no Telegram que recebe do usuário um código de uma ação, faz uma requisição à API, recebe a predição, e retorna uma mensagem ao usuário se a ação irá ou não render 3% ou mais até o próximo trimestre.

O objetivo principal do trabalho foi alcançado, uma vez que através do modelo criado, utilizando o algoritmo *Random Forest*, foi possível classificar ações como confiáveis ou não para se investir com 60% de precisão. Além disso, os objetivos específicos também foram alcançados, sendo criado um bot no Telegram, que com o uso do modelo, apoiasse a tomada de decisão. Como conclusão, observou-se que a ferramenta construída é simples, acessível, fácil de usar, e fica como sugestão para auxiliar os investidores e apoiar a tomada de decisão.

Para trabalhos futuros, sugere-se encontrar mais variáveis dos dados, que possuem um maior impacto na variável resposta, a fim de aumentar a precisão do modelo para classificar ações "confiáveis". Ademais, também pode ser interessante aumentar o período

de previsão do modelo para seis meses ou até um ano, pois assim poderá contribuir com os investidores que visam o longo prazo. Por fim, pode-se aperfeiçoar o bot no Telegram, para que ele faça uma coleta automática dos dados atuais das empresas assim que o usuário enviar o código da ação.

Referências

ALMEIDA, L. G.; ROCHA, R. da S. Mineração de dados de ações com indicadores fundamentalistas. 2019. Citado na página 35.

ALVARENGA, D. 2021 já soma 13 ipos e outras 31 empresas estão na fila para entrar na bolsa. 2021. Disponível em: <https://g1.globo.com/economia/noticia/2021/02/19/2021-ja-soma-13-ipos-e-outras-31-empresas-estao-na-fila-para-entrar-na-bolsa-veja-lista.ghtml>. Citado na página 18.

B3. Uma das principais empresas de infraestrutura de mercado financeiro do mundo. 2021. Disponível em: <https://www.b3.com.br/pt_br/b3/institucional/quem-somos/>. Citado na página 18.

BROWNLEE, J. A gentle introduction to k-fold cross-validation. 2018. Disponível em: <https://machinelearningmastery.com/k-fold-cross-validation/>. Citado na página 55.

CASTRO, L. N. D.; FERRARI, D. G. *Introdução a mineração de dados*. [S.l.]: Saraiva, 2016. ISBN 9788547200992. Citado 5 vezes nas páginas 26, 27, 28, 30 e 31.

DEBASTIANI, C. A.; RUSSO, F. A. *Avaliando Empresas, Investindo em Ações*: A aplicação prática da análise fundamentalista na avaliação de empresas. [S.l.]: Novatec, 2008. ISBN 9788575221792. Citado 2 vezes nas páginas 19 e 21.

D'ÁVILA, M. Z. Bolsa conquista 1,5 milhão de novos investidores em 2020, um aumento de 92% no ano. 2021. Disponível em: <https://www.infomoney.com.br/onde-investir/bolsa-conquista-15-milhao-de-novos-investidores-em-2020-um-aumento-de-92-no-ano/>. Citado na página 13.

ECONOMATICA. Desempenho do ibovespa: 50 anos de história. 2017. Disponível em: <https://insight.economatica.com/desempenho-do-ibovespa-50-anos-de-historia/>. Citado na página 38.

ELDER, A. *Aprenda a operar no mercado de ações: um guia completo para trading*. [S.l.]: Alta Books, 2016. ISBN 9788550800998. Citado na página 13.

INFOMONEY. Análise fundamentalista de ações: como identificar empresas sólidas e rentáveis a longo prazo. 2021. Disponível em: <https://www.infomoney.com.br/guias/analise-fundamentalista/#guia-analise-fundamentalista-criticas>. Citado na página 19.

JOHNSON, A. Evaluating hyperparameter optimization strategies. 2016. Disponível em: <https://sigopt.com/blog/evaluating-hyperparameter-optimization-strategies/>. Citado na página 56.

LIMA, M. L. Um modelo para predição de bolsa de valores baseado em mineração de opinião. 2016. Citado na página 34.

MAZZANTI, S. Boruta explained exactly how you wished someone explained to you. 2020. Disponível em: <https://towardsdatascience.com/boruta-explained-the-way-i-wish-someone-explained-it-to-me-4489d70e154a>. Citado 3 vezes nas páginas 45, 46 e 47.

NIKOLAIEV, D. Overfitting and underfitting principles: Understand basic principles of underfitting and overfitting and why you should use particular techniques to deal with them. 2021. Disponível em: <https://towardsdatascience.com/overfitting-and-underfitting-principles-ea8964d9c45c#:~:text=underfitting>. Citado na página 43.

PINHEIRO, J. L. *Mercado de Capitais*. 9th. ed. [S.l.]: Atlas, 2019. ISBN 9788597021745. Citado 4 vezes nas páginas 13, 17, 18 e 20.

PROVOST, F.; FAWCETT, T. *Data Science for Business*: What you need to know about data mining and data-analytic thinking. [S.l.]: O'Reilly Media, 2013. ISBN 9781449361327. Citado 5 vezes nas páginas 25, 29, 30, 31 e 36.

RAMINELLI, D. G. de T. L.; SANTOS, B. S. dos. Aplicação de técnicas de mineração de dados e aprendizagem de máquina no mercado de ações: Uma revisão sistemática. 2019. Citado na página 35.

RASSIER, L. H.; HILGERT, S. P. *Aprenda a Investir na Bolsa de Valores*. [S.l.]: IESDE, 2009. ISBN 9788538707998. Citado na página 19.

RODRIGUES, T. Uso de técnicas de mineração de dados para encontrar tendências em mercados financeiros. 2016. Citado na página 33.

RODRIGUES, V. Métricas de avaliação: acurácia, precisão, recall... quais as diferenças? 2019. Disponível em: <https://vitorborbarodrigues.medium.com/métricas-de-avaliaç~ao-acurácia-precis~ao-recall-quais-as-diferenças-c8f05e0a513c>. Citado 2 vezes nas páginas 49 e 50.

SILVA, C. A. T. da; RODRIGUES, F. F. *Curso Prático de Contabilidade*. [S.l.]: Atlas, 2018. ISBN 9788597017946. Citado na página 20.

SINATORA, J. R. P. *Mercado de Capitais*. [S.l.]: Editora e Distribuidora Educacional S.A, 2016. ISBN 9788584824366. Citado na página 17.

SOUZA, V. L. de F. Sistema automático para negociação de ações usando técnica de mineração de dados com detecção de mudança de conceito. 2015. Citado 2 vezes nas páginas 34 e 35.

SOUZA, W. B. C. de. Mineração de dados aplicada a previsão de preços de ações utilizando weka. 2021. Citado na página 34.

SRUTHI, E. R. Understanding random forest. 2021. Citado 2 vezes nas páginas 32 e 33.

TAN, P.-N.; KUMAR, V.; STEINBACH, M. *Introduction to Data Mining*. [S.l.]: Addison-Wesley Professional, 2005. ISBN 9780321321367. Citado 4 vezes nas páginas 24, 26, 27 e 28.

VIGHNESH, D. Multivariate analysis: an overview. 2021. Disponível em: <https://s4be.cochrane.org/blog/2021/09/09/multivariate-analysis-an-overview/>. Citado na página 41.

XPEED, R. Como calcular a rentabilidade de ações? 10 passos práticos! 2021. Disponível em: <https://xpeedschool.com.br/blog/como-calcular-rentabilidade-de-acoes/>. Citado na página 38.