PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
ESCOLA POLITÉCNICA E DE ARTES
CURSO DE CIÊNCIA DA COMPUTAÇÃO

CLASSIFICAÇÃO DE ATIVOS NO MERCADO DE AÇÕES UTILIZANDO MINERAÇÃO DE DADOS

DANIEL BUENO DE OLIVEIRA

GOIÂNIA
2023

DANIEL BUENO DE OLIVEIRA

CLASSIFICAÇÃO DE ATIVOS NO MERCADO DE AÇÕES UTILIZANDO MINERAÇÃO DE DADOS

Trabalho de Conclusão de Curso apresentado à Escola Politécnica e de Artes, da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para a obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Sibelius Lellis Vieira
Banca examinadora: Prof. Me. Fernando Gonçalves Abadia,
Prof. Me. Gustavo Siqueira Vinhal

GOIÂNIA
2023

DANIEL BUENO DE OLIVEIRA

CLASSIFICAÇÃO DE ATIVOS NO MERCADO DE AÇÕES UTILIZANDO MINERAÇÃO DE DADOS

Trabalho de Conclusão de Curso aprovado em sua forma parcial pela Escola Politécnica e de Artes, da Pontifícia Universidade Católica de Goiás, para obtenção do título de Bacharel em Ciência da Computação, em _____/_____/_____.

_____

Orientador: Prof. Dr. Sibelius Lellis Vieira

_____

Prof. Me. Fernando Gonçalves Abadia

_____

Prof. Me. Gustavo Siqueira Vinhal

GOIÂNIA

2023

RESUMO

Visando auxiliar investidores do mercado acionário em suas tomadas de decisões, este trabalho tem como objetivo a aplicação de técnicas de mineração de dados na bolsa de valores para predizer tendências futuras de ações. Inicialmente, foi realizada uma pesquisa bibliográfica sobre o mercado acionário e mineração de dados, com o propósito de compreender e conceituar o mercado de ações e as técnicas de mineração de dados. Em seguida, os dados históricos da ação ITSA4 foram levantados e extraídos referente ao ano de 2019 por meio do site InfoMoney. A seleção, limpeza e estruturação desses dados foram conduzidas tanto por um programa como manualmente. Utilizando o software WEKA, um software para mineração de dados, foi realizada a aplicação da mineração de dados e os resultados obtidos foram apresentados e analisados, indicando a possibilidade de prever uma tomada de decisão sobre a ação, sendo elas uma compra, venda ou neutralização, ou seja, não tomar nenhuma decisão. Toda essa avaliação é feita dentro de um intervalo de trinta dias. Os resultados indicam que a mineração de dados pode ser aplicada para prever a tendência do comportamento das ações.

Palavras-chave: precificação de ação, classificação, mineração de dados, descoberta de conhecimentos

ABSTRACT

Aiming to help investors in the stock market in their decision-making, this work aims to apply data mining techniques in the stock market to predict future stock trends. Initially, a bibliographic research on the stock market and data mining was carried out, with the purpose of understanding and conceptualizing the stock market and data mining techniques. Then, the historical data of the ITSA4 share were collected and extracted for the year 2019 through the InfoMoney website. The selection, cleaning and structuring of these data were conducted both by a program and manually. Using the WEKA software, a software for data mining, the application of data mining was carried out and the results obtained were presented and analyzed, indicating the possibility of predicting a decision-making on the action, being a purchase, sale or neutralization, that is, do not make any decision. All of this assessment is done within a thirty-day period. The results indicate that data mining can be applied to predict the trend of stock behavior.

Keywords: stock pricing, ranking, data mining, knowledge discovery

LISTA DE ABREVIATURAS

ARFF	*Attribute-Relation File Format*
B3	Brasil, Bolsa, Balcão
BM&FBovespa	Bolsa de Mercadorias e Futuros de São Paulo
CSV	*Comma-separated Values*
CVM	Comissão de Valores Mobiliários
DCBD	Descoberta de Conhecimento em Base de Dados
DM	*Data Mining*
FF	*Feed Forward*
IA	Inteligência Artificial
IBOVESPA	Índice da Bolsa de Valores de São Paulo
KDD	*Knowledge Dicovery in Databases*
MLP	*Multilayer Perceptron*
NYSE	*New York Stock Exchange*
RNA	Rede Neural Artificial
SLP	*Single Layer Perceptron*
SVM	*Support Vector Machine*
WEKA	*Waikato Environment for Knowledge Analysis*

LISTA DE FIGURAS

SUMÁRIO

1. INTRODUÇÃO

1.1 Contextualização

Conforme a CVM, Comissão de Valores Mobiliários (2019), o mercado acionário desempenha um papel significativo na economia de uma nação, uma vez que representa uma fonte de engajamento de recursos de investidores por meio da bolsa de valores, o que viabiliza o financiamento de atividades empresariais, impulsionando a economia. De acordo com a B3 (2022), o número de investidores tem experimentado um crescimento no Brasil, registrando um aumento de 35% no total de investidores individuais na bolsa de valores brasileira, passando de 3,3 milhões para 4,6 milhões em relação ao terceiro trimestre de 2021.

Para obter ganhos significativos no mercado de ações, é essencial utilizar ferramentas de previsões. No entanto, prever o comportamento do mercado acionário é uma tarefa complexa. As variáveis que impactam esse mercado são numerosas e suas interconexões dificultam a obtenção de previsões precisas. Eventos políticos, informações assimétricas e expectativas dos investidores podem influenciar os preços e as tendências. Identificar e mensurar essas alterações é uma tarefa desafiadora, e fazer previsões com absoluta precisão é praticamente impossível. Por outro lado, mesmo previsões com um certo grau de imprecisão possuem valor.

Existem diversas técnicas computacionais disponíveis que podem auxiliar na extração de conhecimento a partir de grandes volumes de dados. Um exemplo são as técnicas baseadas em inteligência artificial, aprendizado de máquina e ciência de dados, que são amplamente utilizadas por empresas com o intuito de identificar padrões ou informações relevantes para suas atividades comerciais. Essas abordagens permitem explorar os dados de maneira eficiente e eficaz, possibilitando a tomada de decisões embasadas em evidências e a obtenção de novas possibilidades que podem ser valiosas na compreensão de problemas.

Segundo Russell e Norvig (2013) desde 1943 pesquisadores de inteligência artificial (IA) e estatísticos têm demonstrado interesse nas propriedades abstratas das redes neurais, como sua capacidade de computação distribuída, tolerância a ruídos e habilidade de aprendizado. Embora temos conhecimento de outros tipos de sistemas, como redes bayesianas, que possuem essas mesmas propriedades, as redes neurais

continuam sendo uma das abordagens mais populares e eficazes para o aprendizado de sistemas, merecendo assim um estudo aprofundado.

As redes neurais apresentam um desempenho satisfatório devido à sua capacidade de lidar eficientemente com dados que possuem flutuações e oscilações periódicas significativas, ou seja, isso significa que os valores dos dados estão constantemente se alterando, seguindo um padrão previsível e repetitivo. Essas características fazem das RNAs uma ferramenta poderosa para previsões de séries temporais, não apenas no mercado financeiro de ações, mas em diversas áreas como diagnósticos médicos, avaliação de crédito, processamento de imagens, sistemas de segurança, dentre outras.

Além disso, uma técnica amplamente empregada na mineração de dados é a árvore de decisão, que lida com a classificação de dados. Sua capacidade de ser facilmente explicada e compreendida a torna uma das melhores opções para tarefas nesse campo. Além disso, sua construção é simples e resumida, resultando em uma representação mais direta (CASTRO; FERRARI, 2016).

O objetivo deste trabalho é realizar previsões utilizando técnicas de mineração de dados aplicadas a séries temporais do mercado de ações. A análise realizada constituiu-se no preço de fechamento da ação ITSA4, seu índice de liquidez, médias móveis e outros indicadores de tendências, que são detalhados na metodologia. Esses dados são utilizados como entrada para a rede neural e árvore de decisão em um período de trinta dias. Após o treinamento e teste desses modelos com a análise dos resultados, foram realizadas previsões do comportamento futuro da ação ITSA4.

1.2 Justificativa

O aumento do uso da mineração de dados para apoiar a tomada de decisões é claramente observado não apenas na economia, mas também em diversas outras áreas de conhecimento. Em setores como a medicina, por exemplo, essas técnicas têm sido empregadas no cruzamento de dados de pacientes com quadros clínicos semelhantes, permitindo diagnósticos mais precisos. Da mesma forma, no sistema financeiro, técnicas de mineração de dados são utilizadas para combater fraudes no sistema de cartão de crédito e irregularidades em processos licitatórios. Dessa forma, os resultados positivos alcançados pela mineração de dados justificam plenamente sua utilização neste trabalho.

1.3 Objetivos

1.3.1 Geral

O objetivo geral deste trabalho é principalmente aplicar e analisar técnicas de mineração de dados para prever a tendência da ação da Itausa (ITSA4), utilizando o software WEKA. O objetivo é determinar uma tomada de decisão adequada, que pode ser comprar, vender ou se manter em neutralidade, ou seja, não realizar nenhuma tomada de decisão.

1.3.2 Específicos

Para alcançar o objetivo geral, são propostos os seguintes objetivos específicos:

- Conceituar o mercado de ações.
- Conceituar a mineração de dados.
- Utilizar o processo de descoberta de conhecimento em bases de dados.
- Avaliar os resultados obtidos por meio da aplicação da mineração de dados utilizando o software WEKA.
- Avaliar a aplicabilidade da mineração de dados no contexto do mercado acionário.

1.4 Estrutura da pesquisa

O trabalho se encontra organizada em cinco capítulos. O Capítulo 1 corresponde à introdução, onde é fornecida uma visão geral do estudo. No Capítulo 2, é apresentada a revisão bibliográfica, que abrange conceitos relacionados ao mercado financeiro e ao processo de descoberta de conhecimento em bases de dados. O Capítulo 3 aborda os materiais e métodos utilizados nesta pesquisa. No Capítulo 4, são apresentados os resultados obtidos e, em seguida, são discutidos. Por fim, o Capítulo 5 traz a conclusão da pesquisa.

2. REFERENCIAL TEÓRICO

Neste tópico são apresentados os conceitos abordados no decorrer do projeto, tais como a fundamentação sobre bolsa de valores e suas tendências, principais conceitos sobre mercado financeiro e tratados os conceitos da teoria de Dow, e a base para previsão de padrões baseada em inteligência artificial.

2.1 Bolsa de valores

Em geral, as bolsas de valores representam o estado de saúde da economia e das empresas de um país, determinando e fornecendo as bases para a criação de tendências e previsões econômicas. As bolsas promovem o desenvolvimento e a capitalização de empresas, que podem usar parte de seus ativos para levantar recursos no mercado, garantindo assim a possibilidade de lucro para os investidores que acreditam em seu desempenho e resultado.

Conhecida ainda como "Bovespa", a principal bolsa de valores do Brasil é a BM&FBovespa, criada em 2008 a partir da fusão da Bolsa de Mercadorias & Futuros (BM&F) e Bolsa de Valores de São Paulo (Bovespa). (FOGAÇA, 2015)

A Bovespa comercializava basicamente ações e títulos de empresas abertas no mercado, antes de sua fusão, enquanto a BM&F negociava papéis de *commodities* e também seus derivados, contratos e outros valores mobiliários. Desde 2008 na BM&FBovespa, todos os tipos de valores mobiliários, títulos e contratos estão concentrados na mesma plataforma. O Índice Bovespa – IBOV é formado a partir de uma carteira teórica de ativos, conforme determina a metodologia do indicador e é o principal índice da bolsa de valores brasileira pois seu rendimento indica o desempenho médio do mercado de ações brasileiro. Da mesma forma que o Brasil possui o Bovespa, outros países também têm bolsas de valores próprias, contudo, seus indicadores de desempenho diferem. (FOGAÇA, 2015).

Existem outros índices que também são usados para representar o mercado de ações como um todo, um deles é o índice Brasil – IBRX. O índice Brasil envolve as 100 ações mais negociadas dos últimos 12 meses e cada uma delas deve ter sido negociado em pelo menos 70% do horário de funcionamento nesse período. (FOGAÇA, 2015).

2.2 Mercado financeiro

Define-se mercado financeiro por uma entidade que permeia na economia e reúne pessoas ou empresas interessadas em levantar ou emprestar recursos financeiros por diferentes razões e propósitos. Desta forma, por um lado, existem os depositantes cuja renda pode atender às necessidades imediatas de consumo, e eles podem optar por investir parte de seus recursos no mercado financeiro. Por outro lado, há aqueles que necessitam de recursos adicionais para atender suas necessidades imediatas, seja para consumo ou para investimento produtivo.

Segundo a Comissão de Valores Mobiliários (2019) De modo geral, o sistema financeiro é subdividido em quatro mercados principais: mercado monetário, mercado de crédito, mercado de câmbio e mercado de capitais. Esta classificação ajuda a compreender melhor cada um desses mercados, incluindo suas características, riscos e vantagens.

Em suma, o mercado financeiro é composto por um grupo organizado de intermediários e instituições de apoio – o Sistema Financeiro Nacional – que reúne os interesses de tomadores e credores, de modo que o capital a flua pela economia.

2.3 Ações

Ações são títulos que representam a menor proporção do capital social da empresa (seja ela uma instituição, sociedade anônima ou companhia por ações). Acionistas não são credores, mas sim coproprietários que possuem o direito de compartilhar seus resultados da empresa. Quando ações são compradas, seus proprietários se tornam parceiros da empresa, se dispondo a arriscar lucros e perdas como qualquer outro empresário. (MARANGONI, 2010)

As ações não possuem um período de resgate, o que significa que elas são convertidas em dinheiro no momento em que ocorrem as transações no mercado, ou seja, quando é feito a compra ou venda de uma ação, a transação é concluída instantaneamente e o valor correspondente em dinheiro é transferido. Não há necessidade de esperar ou aguardar um prazo específico para resgatar o valor das ações. Os investidores ou proprietários podem alterar suas ações e desfazer seu número de títulos detidos ou mesmo vendidos em uma empresa e adquirir mais títulos em outras empresas. (MARANGONI, 2010)

2.3.1 Tipos de ações

As ações que são encontradas no mercado de capitais brasileiro são limitadas a em dois tipos principais: ações preferenciais e ações ordinárias.

- Ações ordinárias: define-se em direitos básicos concedidos aos titulares e acionistas, especialmente em participação no desempenho da empresa e nos direitos de voto na Assembleia geral de acionistas. Corresponde também a votos nas deliberações da Assembleia Geral, e são nominativas fazendo com que tenham a marcação ON (BERENSTEIN, 2010).

- Ações preferenciais: concede ao proprietário uma certa vantagem patrimonial (prioridade em distribuição de dividendos e no reembolso de capital) relacionados às ações ordinárias em troca de renunciar a outros direitos, como direitos de voto na assembleia de acionistas da empresa, e são nominativas fazendo com que tenham a marcação PN (BERENSTEIN, 2010).

As iniciais ON e PN no final de cada nome em companhias negociadas publicamente em notícias, programas de TV e sites relacionados a bolsa de valores pode causar algumas dúvidas, ainda mais porque, dependendo da situação, essas ações, mesmo que pertençam à mesma empresa, podem apresentar um preço completamente diferente ou até mesmo uma tendência reversa. Nesse sentido, esses dois conceitos devem ser detalhados. (FOGAÇA, 2015)

Os dois tipos devem ser nominativos, e desta forma seu proprietário é identificado nos livros de registro das ações nominativas e a empresa também pode criar o número necessário de classes dentro de cada tipo para assim emiti-los (BERENSTEIN, 2010).

As ações são emitidas através de medidas preventivas da empresa e devem ser registradas no citado livro de registro de ações nominativas para refletir formalmente a posse do título. No entanto, algumas operações anônimas não têm suas próprias prevenções e são chamadas de escriturais. Essas ações são

controladas por uma instituição fiel depositária da empresa na qual a mesma abre uma conta de depósito em nome de seus donos. (MARANGONI, 2010)

2.3.2 Preços de ações

Ao final de um período de negociação, em geral o horário comercial diário, assim que todas as negociações são concluídas, a maneira mais comum de análise do desempenho da ação é observando seu preço final. Este indicador é usado pela mídia e pelo público comum, mas geralmente da abertura para o final do pregão, pode haver um grande número de transações para compra e venda dessa ação, fazendo com que o preço flutue para cima e para baixo ao longo do dia até que o mesmo alcance seu valor final. Sendo assim olhar apenas para o preço final não demonstra com clareza como a ação se manteve no decorrer do dia (MARANGONI, 2010).

Algumas métricas são usadas por investidores mais profissionais para avaliação de preços das ações, como a métrica de abertura, fechamento, máximo, mínimo e volume:

A abertura é o preço da ação no início da sessão de negociação. Não necessariamente igual ao preço de fechamento do dia anterior, pois geralmente há uma operação chamada de *pre-market* que inicia o dia em alta ou em queda devido suas transações realizadas antes do horário de abertura, e *after-market* após o término do último pregão.

Fechamento mostra o montante que a ação alcançou no final do período. É o índice mais selecionado e usada na análise de mercado.

A métrica de máximo mostra o valor máximo atingido durante o período. Esse valor ajuda a explicar o preço, pois quando o valor máximo da ação é muito maior do que o preço de fechamento, o que se pressupõe que o preço da ação houve queda neste período e a tendência é que continue diminuindo no próximo período.

A métrica de mínimo exibe o menor valor atingido por uma cotação em um dia. Sua explicação é inversamente proporcional ao valor máximo.

Volume é a soma do valor envolvido em todas as operações de compra e venda de uma determinada ação em um dia. Para empresas menores, tem uma variação entre vários milhares de reais, já para as reservas de caixa de grandes empresas são ampliadas em centenas de milhões de reais além de alguns bancos e construtoras. É impossível inferir o comportamento do mercado com base apenas no volume de

transações, mas uma análise no histórico de volume de transações pode determinar o início ou o fim de uma tendência. Para exemplificar, quando o volume de transações é alto por um determinado período de tempo, e então começa a diminuir, isso significa um mercado esgotado devido à tendência de aquisição de ações, o que pode significar mudanças na mesma.

2.3.3 Representação gráfica das ações

Existem 3 (três) formas de representar graficamente o preço de uma ação na bolsa de valores: linha, barras e *candlestick*.

A maneira mais fácil de apresentar um gráfico de preços é através do gráfico de linha. O gráfico de linha, como ilustra a figura 1, é caracterizado pelo preço de fechamento diário: é um gráfico simples e fácil de visualizar para a identificação de alguns padrões gráficos, porém é pouco usado na prática. (Comissão de Valores Mobiliários, 2017).

Figura 1: BOVA11 representado pelo gráfico de linha

Fonte: B3

No entanto, outros valores também são importantes, além do valor de fechamento: o valor de abertura, o valor máximo e o valor mínimo do dia. Uma das maneiras de representar os 4 (quatro) valores é o gráfico de barras, e é identificado em uma barra e a abertura é indicada por um simples contorno horizontal à esquerda

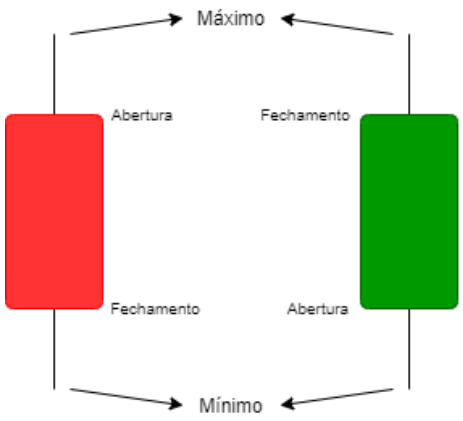
da barra e o fechamento à direita da barra (Comissão de Valores Mobiliários, 2017), como ilustra a figura 2.

Figura 2: Representação do gráfico de barras

Fonte: Elaborada pelo autor

Similar ao gráfico de barras, preços de abertura, fechamento, máximo e mínimo de uma ação podem ser vistos em uma *candlestick*, que possui esse nome porque seu formato se assemelha o de uma vela, como ilustra a figura 3. Um *candle* pode ser usado para exibir o comportamento de uma ação para um determinado período. Os valores de abertura e fechamento são demonstrados por uma barra na horizontal. Um retângulo é pintado de verde quando o fechamento é maior que a abertura (indica que ocorreu um aumento de preço) e pintado em vermelho quando o preço for menor que a abertura (sugere que houve queda no preço) (MARANGONI, 2010).

Figura 3: Representação de um candlesticks nos cenários de alta e queda



Fonte: Elaborado pelo autor

Os limites superior e inferior, ilustrados por uma linha, indicam os valores máximo e mínimo obtidos durante o período selecionado. Todos os valores oscilados pela ação ao longo deste período estão incluídos na linha vertical, que são delimitados pelas quantidades máximas e mínimas (MARANGONI, 2010).

Um *candle* pode retratar o preço a qualquer intervalo de tempo: um dia, uma semana, um mês ou em um período intradiário, como 15 minutos. Como exemplifica a figura 4, o gráfico diário é utilizado com mais frequência, pois os preços de abertura, o máximo, o mínimo e de fechamento têm significados mais importantes, devido ao planejamento diário do mercado para suas operações (Comissão de Valores Mobiliários, 2017).

Figura 4: Representação de vários *candlesticks* em um gráfico diário

Fonte: B3.

2.4 Tendência

Na condução de preços, quando visualizada em um gráfico, nota-se claramente um padrão flexível e cheio de variações. Essa direção pode ser crescente, indeterminada ou descendente. A prolongação de um preço em uma determinada direção por um tempo, refere-se ao conceito de tendência (ROQUE, 2009).

Uma tendência aparenta-se de forma clara no gráfico de preços, mas para a detecção correta, é essencial explorar definições mais técnicas. A partir desse

parâmetro, pode-se analisar padrões gráficos mais desenvolvidos (Comissão de Valores Mobiliários, 2017).

2.4.1 Tendência de alta

É determinado por fundos ascendentes de modo que os fundos são pontos de apoio em que a força dos compradores supera os vendedores. Os fundos crescentes significam que os compradores estão dispostos a comprar a um valor cada vez mais alto, o que dá suporte e continuidade à tendência alta (Comissão de Valores Mobiliários, 2017).

Uma linha de tendência alta é puxada conectando as extremidades crescentes de um movimento de alta. O desenho de uma linha de tendência parte conectando 2 (dois) pontos (fundos). No entanto, a linha é confirmada apenas com o toque de um terceiro ponto. A figura 5 ilustra um exemplo de como é feito a análise de tendência de alta:

Figura 5: Representação de uma tendência de alta.

Fonte: Comissão de Valores Mobiliários, 2017.

2.4.2 Tendência de baixa

É determinado por topos descendentes de modo que os topos são pontos de resistência, em que a força dos vendedores excede os compradores. Os topos

significam que os vendedores estão dispostos a vender a um preço bem baixo, o que dá apoio e continuidade à tendência de baixa (Comissão de Valores Mobiliários, 2017).

Assim como a tendência de alta, porém de forma invertida, como ilustra a figura 6, uma linha de tendência baixa é realizada conectando topos cada vez mais baixos de um movimento de baixa.

Figura 6: Representação de uma tendência de baixa

Fonte: Comissão de Valores Mobiliários, 2017.

2.4.3 Tendência lateral

É caracterizada pela formação de topos e fundos no mesmo plano horizontal. Caracteriza a estabilidade entre pressão compradora e vendedora. A figura 7 ilustra como é feita a tendência lateral. Os preços são negociados dentro de uma linha horizontal restringindo os mesmos. Técnicas de rastreamento de tendências não se aplicam na tendência lateral.

Figura 7: Representação de uma tendência lateral

Fonte: Comissão de Valores Mobiliários, 2017.

2.4.4 Características e rompimentos da linha de tendência

O grau de atividade de uma linha de tendência é equivalente a quantidade de vezes que a linha toca nos *candlesticks*. Pode-se usar o preço de fechamento, máximo ou mínimo. O preço de fechamento tem maior relevância no gráfico diário. O foco deve ser na frequência de vezes que o mesmo nível de preço se repete, independentemente do tipo de preço. A linha de tendência pode sofrer pequenas penetrações, ela não é totalmente precisa, mas deve ser usada como critério para um bom parâmetro de análise.

O rompimento de uma linha de tendência alta (baixa) não indica necessariamente o início de uma tendência baixa (alta); O mercado pode começar a trabalhar em uma estabilização (tendência lateral). A perda de uma linha de tendência indica apenas uma permutação entre elas, ou seja, em fase de encerramento de uma tendência para que outra se inicie; A quebra da linha de alta tendência deve ser confirmada com um fechamento abaixo da mesma.

2.5 Teoria de Dow

Entre 1900 e 1902, foi anunciado por Charles H. Dow em uma série de artigos no *Wall Street Journal*, a teoria do Dow. A teoria tem como foco a identificação de tendências do mercado e é considerada a mais antiga e concreta das declarações

teóricas sobre a permanência de grandes tendências no mercado de capitais (Comissão de Valores Mobiliários, 2017).

Charles Dow criou o índice médio de cotação, como ferramenta de avaliação dos preços das ações da Bolsa de Valores de Nova York (em inglês *New York Stock Exchange – NYSE*) em 1884. Ele criou índices que são usados constantemente por investidores sendo o Dow Jones Industrial o principal desses índices (Comissão de Valores Mobiliários, 2017).

Os escritos acadêmicos de Charles Dow estabeleceram seis princípios que juntos, determinam o que se entende sobre teoria de Dow. Serão apresentados a seguir.

O primeiro princípio é o que os índices de preço já descontam tudo**.** De acordo com esse princípio, os índices de mercado (IBOVESPA por exemplo) já representam a ação conjunta de vários investidores, desde os mais instruídos (os que possuem mais entendimento do mercado e informações privilegiadas) até os principiantes. As variações diárias dos preços de um índice acompanham os eventos que acontecem ou que irão acontecer e conseguem conciliar todos estes movimentos. Consequentemente, todo fator que afeta a relação de oferta/demanda está refletido no preço do índice;

O segundo princípio estabelece que o mercado se desenvolve em 3 (três) tendências. O mercado possui três tipos de movimento: primário, secundário e terciário.

A tendência primária é classificada como a principal tendência do mercado. É um longo movimento que pode ser alto ou baixo e leva a uma grande avaliação ou desvalorização de ativos. Não há regras precisas para definir a duração das tendências, mas as tendências primárias se permanecem cerca de 1 a 2 anos. Na Figura 8 as linhas verticais fazem uma separação entre as três tendências primária no índice Bovespa.

Figura 8: Tendências primárias

Fonte: Adaptado no site da B3.

O conjunto de várias correções e impulsos de alta e baixa dentro de uma tendência primária são chamados de tendências secundárias. Uma tendência secundária dura algumas semanas ou meses e pode corrigir até dois terços da tendência primária que ela pertence. Tendências terciárias são parte do secundária. Existem pequenos movimentos de, em média, até 3 (três) semanas que se comportam em relação às tendências secundárias da mesma forma que as secundárias são relacionadas às primárias.

O terceiro princípio estabelece as três fases dos movimentos. Cada tendência é formada e conduzida em três etapas: acumulação, movimento e distribuição.

- Acumulação é o momento no qual investidores mais experientes começam a ocupar posições em seus investimentos alinhados nas tendências formais. É nesse instante que o mercado já assimilou todas as informações ruins que mantiveram a tendência de baixa, e começaram a dar sinais de retorno para uma alta tendência;
- Movimento ou Participação Pública é o ponto no qual os investidores que seguem tendências começam a investir em ativos. A tendência é confirmada assimilando as boas notícias e a movimentação flui;

- Distribuição é quando os canais de notícias começam a escrever sobre o forte aumento nos lucros do ativo no mercado de ações, o volume de operações começa a aumentar e a participação pública se tornar maior. Neste momento, o investidor que entrou no ativo na fase de Acumulação começa a desfazer suas posições, resgatando seus lucros.

O quarto princípio é o da confirmação, significa que uma reversão de tendências deve ser válida, o fato deve ocorrer em dois índices de diferentes composições. Os dois índices são usados para interpretar um ponto de vista diferente entre si para validar os acontecimentos ou especificar uma falsa tendência. No caso brasileiro, esses dois índices podem ser, por exemplo, o índice Bovespa e índice Brasil. Desta forma, um índice confirma o outro demonstrando que não houve variação temporária do movimento.

O volume deve confirmar a tendência é o quinto princípio**.** A teoria de Dow admite o volume como um termo secundário, mas ainda assim importante para confirmar novas tendências de preço. Desta forma o volume está relacionado com as tendencias da seguinte maneira:

- Tendência de alta: em uma tendência principal de alta, espera-se que o volume aumente com a valorização dos ativos e decaia nas reações de desvalorização.
- Tendência de baixa: em uma tendência principal de baixa, espera-se que o volume aumente com a desvalorização dos ativos e decaia nas reações de valorização.

O sexto princípio afirma que uma tendência ocorre enquanto não houver sinais de reversão. De acordo com este princípio, o mercado não cairá apenas porque atingiu um nível elevado de alta ou subir porque atingiu uma baixa expressiva. Portanto, para mudar a posição em relação a uma tendência, é necessário provas concretas que a mesma foi finalizada. Existem muitas técnicas desenvolvidas por investidores para confirmar se a tendência ainda se mantem ou não como indicadores técnicos, padrões gráficos e até mesmo os padrões de *candlesticks*.

2.6 Descoberta de conhecimento em base de dados

Com o passar dos anos, a capacidade de gerar e armazenar dados vem crescendo gradativamente. Extensas bases de dados são geradas nos mais diversos ramos de atividade. Com esse aumento na quantidade de dados armazenados, nasce também a necessidade de novas tecnologias e ferramentas para ajudar a transformar esses dados em Informações e conhecimentos úteis.

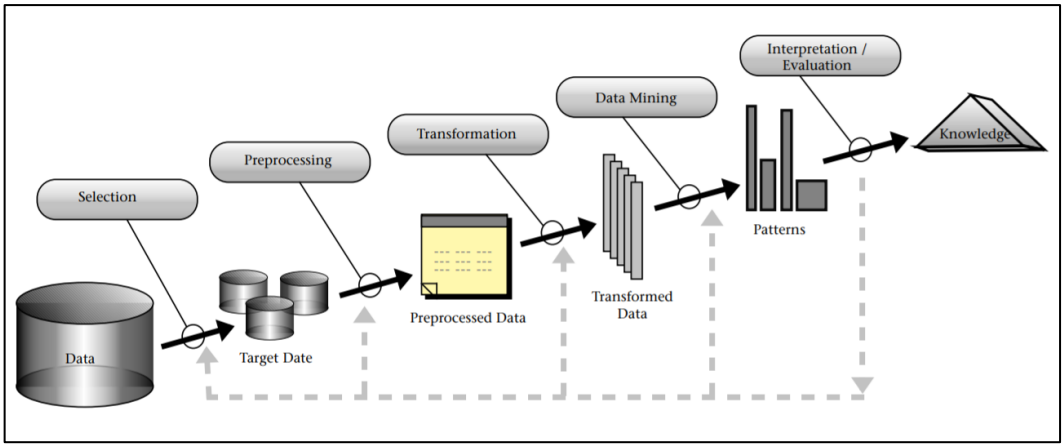
Visto isso, o processo de Descoberta de Conhecimento em Banco de Dados (DCBD), ou *Knowledge Dicovery in Databases* (KDD) surge para encontrar uma maneira explorar esses bancos de dados e encontrar os padrões existentes utilizando modelagem. A Mineração de Dados ou *Data Mining* (DM) é classificado como um dos passos que integra o processo de DCBD ao qual explora e analisa um grande número de dados para encontrar padrões, regras e relacionamentos interessantes e importantes para algum fim (BERRY, 2000).

A entrada para este processo consiste em dados organizacionais. O *enterprise data warehouse*, definido como um armazenamento de dados em grande escala que é usado em empresa para apoio à decisão, permite que o KDD seja implementado de forma eficiente porque fornece uma única fonte de dados a serem extraídos (SHARDA; DELEN; TURBAN, 2019).

2.7 Etapas da DCBD

As etapas da DCBD consistem em uma série de passos que auxilia a tomar as mais variadas decisões. Cada estágio tem uma interseção com outros estágios, fazendo com que o resultado seja melhorado a cada fase (RELICH; MUSZYNSKI, 2014). Dunham (2003) resumiu o processo de DCBD consistindo em cinco etapas que são elas: seleção de dados, pré-processamento de dados, transformação de dados, mineração de dados e interpretação/avaliação. Estas etapas estão ilustradas na figura 9 a seguir, representando suas fases.

Figura 9: Etapas da Descoberta de Conhecimento em Base de Dados.



Fonte: Fayyad et al. (1996).

Segundo Rosa (2017) a etapa da seleção de dados é também conhecida como “Redução de Dados”. É a primeira etapa no processo de descoberta de informação e possui papel indispensável no resultado final, uma vez que nesta etapa é definida o conjunto de dados contendo todas as possíveis variáveis (atributos) e registros (instâncias ou casos ou observações ou padrões) que se pretende analisar. Em grande parte dos casos, esta seleção é feita por um especialista da área, ou seja, alguém que realmente tenha conhecimento sobre assunto em questão.

A segunda etapa é a de pré-processamento e Limpeza dos Dados, onde é feito os processos que excluem dados redundantes e inconsistentes, recuperam dados incompletos e avaliam possíveis divergências nos dados. Ter um especialista que domina as técnicas é fundamental, pois é o mesmo que definirá se os atributos adquiridos são realmente relevantes, se o conhecimento é válido, novo e útil, ou se será necessário retornar a alguma etapa anterior (ANUMALLA, 2007). Nesta etapa, também é analisado a possível chance de diminuir o número de variáveis envolvidas no processo, procura entender a identificação de quais informações, dentre as bases de dados existentes, devem ser analisadas durante o processo KDD, intencionando melhorar o desempenho dos algoritmos de análise (GOLDSCHMIDT; PASSOS, 2005).

De acordo com Goldschmidt e Passos (2005) a etapa de transformação dos dados ou codificação dos dados tem a principal finalidade de transformar o conjunto bruto de dados em uma forma padrão de uso. A fase três é elaborada através de um

processamento dos dados, com o objetivo de organizar os dados para facilitar o trabalho feito pelas etapas posteriores do processo KDD. Essa transformação dos dados se refere a uma transformação que seja destinada a todos os valores de um certo atributo para todos os atributos. No entanto, não há um único critério de transformação dos dados e várias técnicas podem ser realizadas de acordo com os objetivos que foram planejados (ROSA, 2017).

A quarta etapa tem como finalidade a realização da mineração de dados, onde Witten et al, (2011) definiram que as técnicas de DM abrangem a extração automatizada de padrões, representando conhecimento implicitamente armazenado em grandes bases de dados, armazéns de dados, e outros repositórios de informação de alta escala. As técnicas e algoritmos de mineração de dados que são aplicadas, constatam a hipótese e extraem padrões de forma autônoma a partir dos dados definidos na etapa de transformação dos dados. Assim como na fase anterior, os modelos podem ser aplicados várias vezes, ou até mesmo refeitos, dependendo dos objetivos a serem alcançados (CASTRO; FERRARI, 2016).

A quinta e última etapa é também conhecida como pós-processamento e engloba todos os participantes que avaliam de forma cuidadosa os resultados favorecendo então uma interpretação para o modelo, de onde se absorve o conhecimento (ROSA, 2017). Caso o resultado não seja satisfatório, o processo pode retornar a qualquer uma das fases anteriores. Essa interpretação deve ser incluída no algoritmo minerador, porem determinadas vezes é conveniente a implementação isolada. Desta forma, o principal objetivo dessa etapa é assegurar um grau bom de compreensão do conhecimento descoberto pelo algoritmo minerador, realizando validações através de percepção de um analista de dados e principalmente por meio de métricas de qualidade da solução (NGAI et al., 2011).

2.8 Mineração de dados

A mineração de dados pode ser vista como um conjunto de técnicas operacionais automáticas para grandes volumes de dados com o intuito de descobrir novos padrões e relacionamentos que, devido a quantidade elevada de dados, não seria facilmente descoberto analisando manualmente pelo ser humano. Na verdade, são várias as técnicas usadas, porem a mineração de dados é considerada mais uma arte que uma ciência (AMORIM, 2006).

Definido então de forma simples, para Sharda, Delen e Turban (2019) a mineração de dados é um termo usado para descrever a descoberta ou “mineração” de conhecimento de grandes quantidades de dados. Quando visto por analogia, é facilmente entendido que o termo mineração de dados é de certa forma inapropriado. Neste caso, mineração de dados talvez devesse ser chamado de "mineração de conhecimento" ou "descoberta de conhecimento". Apesar da incompatibilidade entre o termo e seu significado, a mineração de dados se tornou a escolha de muitas comunidades. Muitos outros nomes associados à mineração de dados incluem conhecimento extração, análise de padrões, arqueologia de dados, coleta de informações, pesquisa de padrões, e escavação de dados.

Mineração de dados não é um termo novo, mas sim uma boa definição para o uso de muitas disciplinas. Como ilustra a figura 10, DM está bem posicionada na interseção de muitas disciplinas, incluindo estatística, inteligência artificial, aprendizado de máquina, ciência da gestão, sistemas da informação (SI) e bancos de dados. Usando avanços em todas essas disciplinas, os dados mineração se esforça para progredir na extração de informações úteis e conhecimento de grandes bancos de dados (SHARDA; DELEN; TURBAN, 2019).

Figura 10: Disciplinas envolvidas na mineração de dados.

Fonte: Sharda, Delen e Turban (2019).

2.8.1 Tarefas e técnicas da mineração de dados

Conforme mencionado por Castro e Ferrari (2016), as tarefas de extração de mineração de dados são separadas em duas categorias principais. As atividades preditivas objetivam prever valores desconhecidos ou futuros com base em informações já conhecidas. Em contrapartida, nas atividades descritivas, o objetivo é identificar padrões nos dados, ou seja, detectar características intrínsecas, de forma que seja possível interpretá-las. As atividades de classificação e regressão são categorizadas como atividades preditivas, enquanto as atividades de clusterização, modelagem de dependências, sumarização e detecção de anomalias são atividades descritivas. Este tópico apresentará algumas das principais atividades de extração de informações dos dados.

A atividade de classificação consiste em analisar um novo registro e determinar a classe correspondente com base em um histórico de registros já classificados. Por exemplo, em um banco que possui um histórico de seus clientes, é possível usar essas informações para prever se um novo cliente, com determinadas características, será ou não capaz de pagar um empréstimo. Essa atividade é realizada em duas etapas: a primeira é o treinamento, em que um modelo é criado a partir de um conjunto de dados já classificados. Em seguida, um novo conjunto de dados, que não foi utilizado na etapa anterior, é usado para realizar os testes, de forma que seja possível verificar a eficiência do modelo em responder corretamente aos dados. Essa atividade é considerada do tipo aprendizagem supervisionada (CASTRO; FERRARI, 2016).

De forma semelhante, a atividade de regressão também é uma atividade que segue o modelo de aprendizagem supervisionada, separando os dados para o treinamento e os testes. No entanto, a forma como seus resultados são avaliados é diferente, já que a regressão visa prever um valor contínuo, e não leva em consideração a quantidade de acertos e erros, como na classificação. Em vez disso, a precisão da previsão é determinada pelo cálculo da distância entre a saída esperada e a saída estimada (CASTRO; FERRARI, 2016).

A atividade de clusterização, como o nome sugere, visa agrupar objetos em clusters, de acordo com as relações existentes entre eles. São realizadas buscas por semelhanças e diferenças entre os objetos analisados, e a distância entre eles é usada para determinar em qual grupo cada um se encaixa. Ou seja, objetos semelhantes têm uma distância menor, sendo agrupados em um mesmo cluster.

Apesar de parecer semelhante à classificação, na clusterização não existem classes previamente definidas, já que os objetos são agrupados com base em suas semelhanças (SILVA; PERES; BOSCARIOLI, 2016).

Na atividade de associação, o objetivo é encontrar relações entre os atributos que ocorrem em uma base de dados transacionais. A frequência de ocorrência dos atributos em transações indica que há uma relação forte entre eles. Um exemplo clássico é o carrinho de supermercado, em que as pessoas que compram leite também compram pão, o que indica que existe uma relação entre esses dois produtos (SILVA; PERES; BOSCARIOLI, 2016).

2.9 Árvore de decisão

De acordo com Russell e Norvig (2013) um modelo de árvore de decisão simboliza uma função que recebe um conjunto de dados com atributos e fornece uma "resposta" - um resultado único. Os dados podem ter valores discretos ou contínuos. No momento, é evidenciado questões em que os dados possuem valores discrepantes e o resultado possua exatamente dois desfechos possíveis. Essa é uma classificação binária, em que cada caso é separado como verdadeiro (positivo) ou falso (negativo).

Para chegar a uma decisão, a árvore de decisão realiza uma sequência de avaliações. Cada nó interno da árvore conduz uma avaliação no valor de um dos atributos de entrada, A_i, e as ramificações dos nós são categorizadas com os valores possíveis do atributo, $A_i = v_{ik}$. Cada nó folha da árvore determina o valor a ser devolvido pela função. A representação de árvores de decisão parece ser muito intuitiva para os seres humanos; de fato, muitos manuais do tipo "como fazer" (como, por exemplo, manuais de conserto de automóveis) são redigidos inteiramente como uma única árvore de decisão que se estende por centenas de páginas.

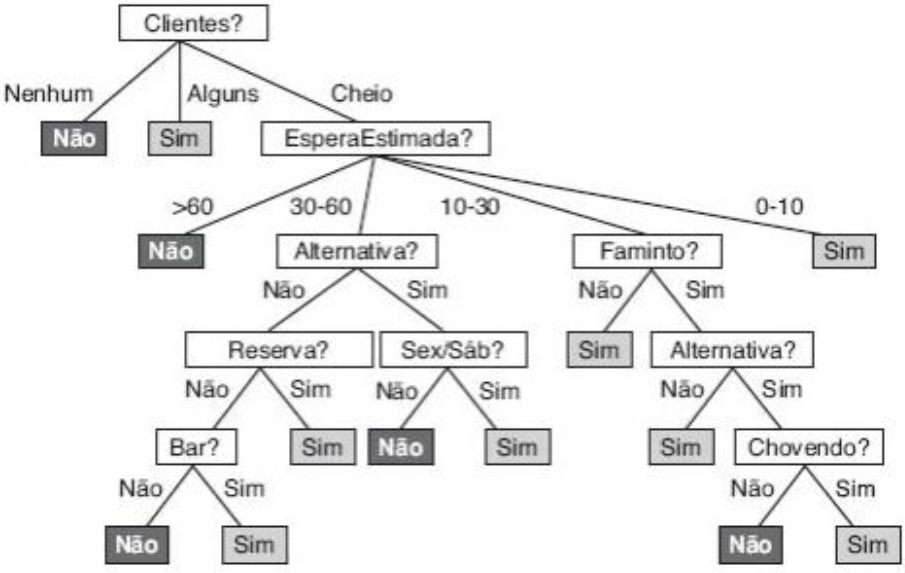
Para ilustrar, foi criado uma árvore de decisão com o propósito de decidir se uma pessoa pode esperar ou não por uma mesa em um restaurante. O objetivo é aprender uma definição para o predicado de objetivo *VaiEsperar*. Primeiramente, enumera-se os atributos em que é considerado como parte da entrada:

- Opção: Se há outro restaurante apropriado nas proximidades.
- Bar: Se o restaurante possui um bar confortável onde possamos esperar.
- Fim de semana: Verdadeiro nas sextas e sábados.

- Fome: Se está com fome.
- Clientes: Quantidade de pessoas no restaurante (os valores são: Nenhum, Alguns e Cheio).
- Preço: A faixa de preços do restaurante ($, $$, $$$).
- Clima: Se está chovendo do lado de fora.
- Reserva: Se realizou uma reserva.
- Tipo: O tipo de restaurante (francês, italiano, tailandês ou hamburgueria).
- Tempo de espera estimado: O tempo de espera estimado pelo gerente (0-10 minutos, 10-30, 30-60, >60).

Observe que cada variável tem um conjunto limitado de valores possíveis. O valor de *EsperaEstimada*, por exemplo, não é um número inteiro, mas sim um dos quatro valores discretos 0-10, 10-30, 30-60 ou >60. Na figura 11 é apresentada a árvore de decisão de acordo com o domínio descrito. Observe que a árvore desconsidera o preço e o tipo dos atributos. Os exemplos são processados pela árvore, partindo da raiz e seguindo a ramificação apropriada até alcançar uma folha. Por exemplo, uma situação em que Clientes = Cheio e *EsperaEstimada* = 0-10 será classificada como positiva (isto é, sim, vamos esperar por uma mesa).

Figura 11: Árvore de decisão sobre a espera por uma mesa.



Fonte: Russell e Norvig (2013).

A busca gulosa empregada na aprendizagem de árvore de decisão foi concebida para reduzir aproximadamente a profundidade da árvore resultante. O conceito consiste em selecionar o atributo que alcance a maior extensão possível com o intuito de proporcionar uma classificação precisa dos exemplos. Um atributo ideal separa os exemplos em conjuntos, cada um dos quais será totalmente positivo ou negativo e, posteriormente, se converterão nas folhas da árvore. Alguns atributos podem gerar conjuntos de exemplos com proporções praticamente iguais de exemplos positivos e negativos em relação ao conjunto original (RUSSELL; NORVIG, 2013).

Sendo assim, é essencial dispor de uma medida formal de "muito bom" e "completamente inútil" para implementar uma função de importância com relação aos atributos. Para isso, é empregado o conceito de ganho de informação, que se caracteriza como termo de entropia, a grandeza fundamental na teoria da informação segundo Shannon e Weaver (1949).

A incerteza de uma variável aleatória é medida pela entropia, e quando se obtém informações, a entropia é reduzida. Quando a variável aleatória apresenta sempre o mesmo resultado, como uma moeda que sempre cai cara, sua entropia é zero. Se uma moeda honesta é lançada, a chance de cair cara ou coroa é igual, representando uma entropia de "1 bit". Quando um dado de quatro lados é lançado, há 2 bits de entropia, já que são necessários dois bits para descrever cada uma das quatro escolhas igualmente prováveis. Agora, considere uma moeda viciada que caia cara em 99% das vezes. Essa moeda tem menos incerteza do que a moeda honesta, e sua entropia deve estar próxima de zero, mas positiva. Em geral, a entropia de uma variável aleatória V, com valores v_k cada um com probabilidade $P(v_k)$, é definida como

$$Entropia\colon\ H(V) = \sum_k P(v_k)\, log_2 \frac{1}{P(v_k)} = -\sum_k P(v_k) log_2 P(v_k)$$

2.10 Redes neurais

Redes neurais são modelos computacionais inspirados no funcionamento do cérebro humano, que têm como objetivo aprender a partir de exemplos e realizar tarefas complexas. Essas redes são compostas por neurônios artificiais, que realizam operações matemáticas para processar e transmitir informações (GIACOMEL, 2016).

Uma rede neural típica é composta por várias camadas de neurônios interconectados, cada uma das quais realiza uma transformação nos dados de entrada. As camadas iniciais da rede geralmente extraem características básicas dos dados de entrada, enquanto as camadas posteriores combinam essas características para realizar tarefas mais complexas, como classificação ou regressão. as Redes Neurais são um poderoso modelo de aprendizado de máquina que têm sido amplamente utilizados em diversas áreas. Com sua capacidade de processar e extrair informações úteis a partir de dados abstratos, esses modelos têm mostrado um grande potencial para resolver problemas difíceis (MARANGONI, 2010).

De acordo com Giacomel (2016), existem diversos tipos de Redes Neurais, cada uma com suas características e aplicações específicas. Entre as mais utilizadas, podemos citar as Redes Neurais *Multilayer Perceptron*, *Feed Forward* e *Backpropagation*. Estes conceitos serão detalhados em seguida.

2.10.1 *Multilayer Perceptron*

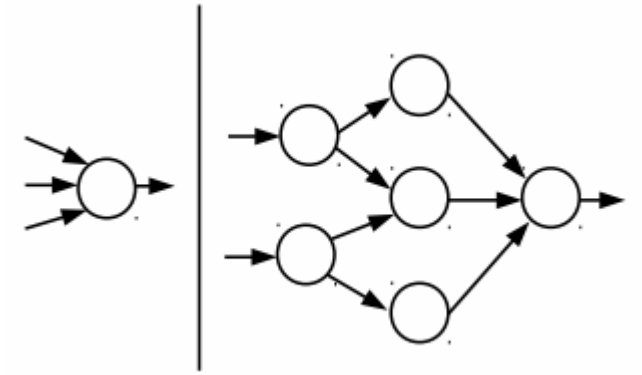
A arquitetura de rede neural denominada *Multilayer Perceptron* (MLP), é uma modalidade avançada que possibilita solucionar problemas de maior complexidade. Na estrutura de uma MLP, os neurônios são organizados em múltiplas camadas, interligadas por sinapses: a primeira camada recebe as entradas, enquanto a última camada emite as saídas.

Opcionalmente, uma MLP pode apresentar uma ou mais camadas intermediárias, também chamadas de camadas ocultas, que são responsáveis por potencializar a capacidade de aprendizado e processamento da rede neural. O mecanismo de funcionamento de uma MLP é semelhante ao de um Perceptron Simples (SLP): as entradas são inseridas na camada inicial e são transmitidas entre as camadas intermediárias, se existirem, até alcançarem a camada de saída. A cada passagem de um valor de uma camada para outra, o valor que é enviado do neurônio de origem é multiplicado pelo peso que foi estabelecido na sinapse entre os dois neurônios e filtrado pela função de ativação do neurônio de destino. O resultado desses cálculos é enviado para a próxima camada (GIACOMEL, 2016).

Para exemplificar a organização das duas redes neurais, a figura 12 apresenta uma comparação entre a estrutura de um Perceptron Simples (no lado esquerdo) e

um Perceptron Multicamadas (no lado direito). Os neurônios são representados pelos nós e as sinapses entre os neurônios são representadas pelas arestas.

Figura 12: Ilustração de uma rede neural do tipo *Single Layer Perceptron* e *Multilayer Perceptron*.

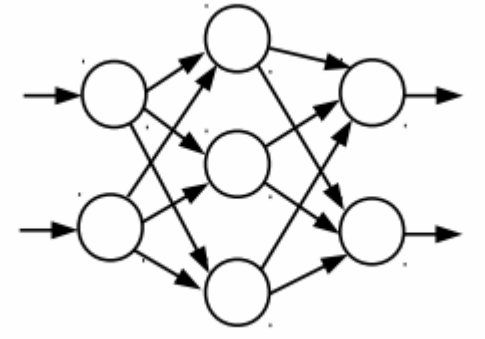


Fonte: Giacomel (2016).

2.10.2 *Feed Forward*

Uma MLP do tipo *Feed Forward* (FF) é caracterizada pelo atendimento de duas condições: a primeira consiste em direcionar todas as sinapses para a frente, em direção à saída; a segunda envolve a replicação da saída de cada nó para todos os outros nós na camada seguinte (GIACOMEL, 2016). A arquitetura de uma rede FF com duas saídas é apresentada na figura 13, na qual é possível observar que ambas as condições foram satisfeitas, diferentemente da rede MLP mostrada na figura 12.

Figura 13: Ilustração de uma rede neural do tipo *Feed Forward*.



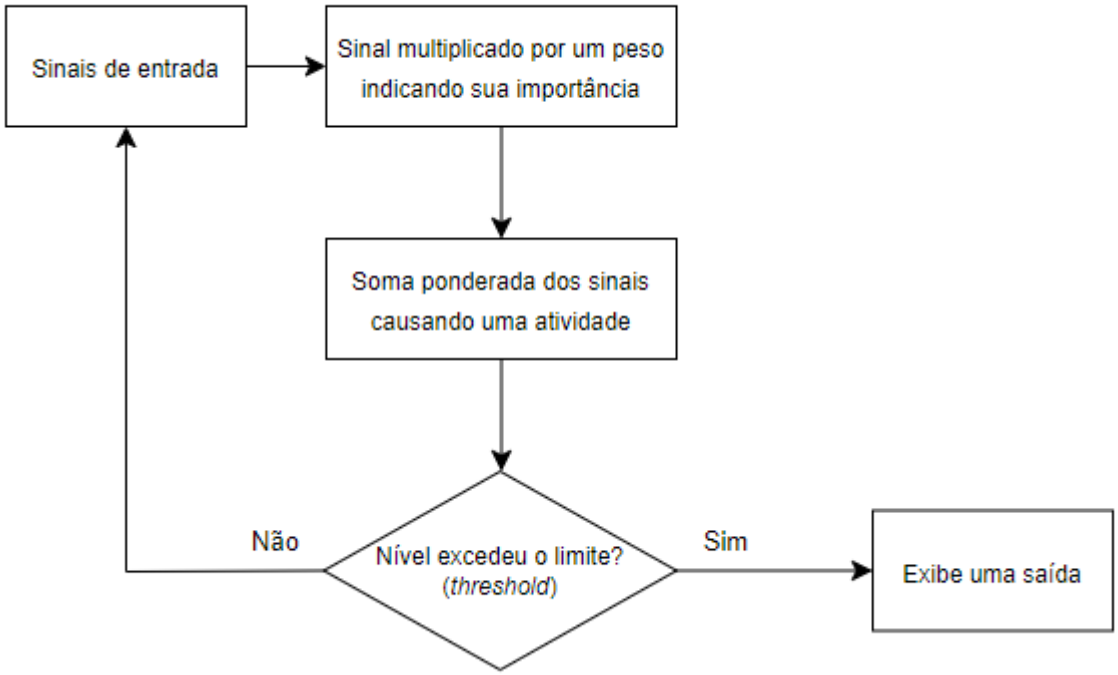
Fonte: Giacomel (2016).

2.10.3 *Backpropagation*

As redes que utilizam o algoritmo de *Backpropagation* têm como meta diminuir o termo de erro entre a saída da rede neural e o valor de saída desejado. O cálculo do termo de erro é feito pela comparação entre a saída e o resultado almejado, e posteriormente é alimentado de volta na rede, ocasionando uma alteração nos pesos sinápticos, com o intuito de minimizar o erro. Esse procedimento é reiterado até que o erro alcance um valor mínimo (MARANGONI, 2010).

De maneira geral, a demonstração intuitiva da operação de cada nó da rede é evidenciada na figura 14. A noção fundamental é que os valores são renovados conforme cada nova entrada de cálculo em cada nó, agregados e, então, reavaliados. Se o erro não tiver alcançado um número máximo de épocas estabelecido, os valores serão reavaliados (MARANGONI, 2010).

Figura 14: Esquema funcional de uma RNA com *Backpropagation*.



Fonte: Elaborado pelo autor.

Um sistema de rede neural que utilize a técnica de Backpropagation é indicado para situações que apresentem os aspectos de grande quantidade de dados de entrada, cuja conexão com a saída esperada não é completamente conhecida; que seja possível construir um conjunto de exemplos que represente o comportamento

esperado; situações em que a solução pode sofrer alterações ao longo do tempo, mesmo quando as entradas são as mesmas (GIACOMEL, 2016).

No entanto, é importante mencionar que esse tipo de rede não deve ser utilizado em problemas que possam ser representados por meio de um fluxograma (nesse caso, é mais adequado utilizar a programação convencional) ou problemas que exijam muita precisão ou uma solução analítica nos valores numéricos de saída (já que a rede neural sempre fornece um resultado aproximado, ainda que com um erro insignificante) (GIACOMEL, 2016).

Problemas como processamento de imagens, reconhecimento de fala e previsão de séries temporais compartilham das características mencionadas acima. Portanto, a aplicação de uma rede neural com Backpropagation pode ser bem-sucedida na resolução desses tipos de problemas (GIACOMEL, 2016).

2.11 Ferramenta Weka

Conforme afirmado por Silva (2008), o Waikato Environment for Knowledge Analysis (WEKA) representa um programa de código aberto dotado de diversos algoritmos e instrumentos de pré-processamento e exploração de dados empregados no processo KDD. A ferramenta, que tem sua gênese na Nova Zelândia, mais especificamente na Universidade de Waikato, dispõe de uma interface gráfica simples, de simples manejo, além de disponibilizar relatórios e histogramas dos dados.

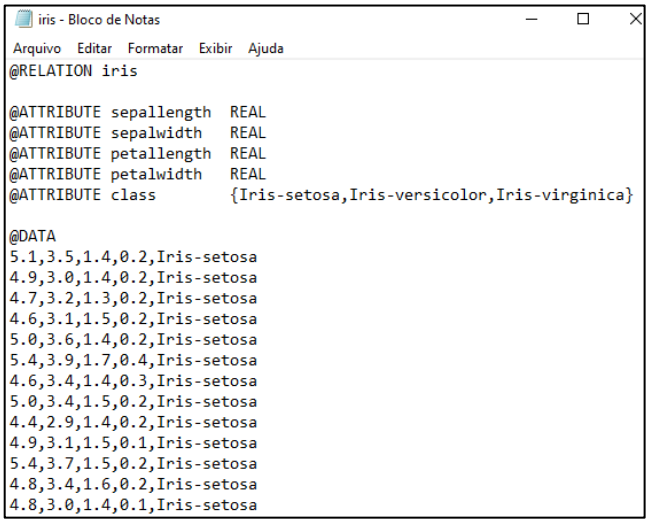
Na tela principal do software, como exemplifica a figura 15, são disponibilizadas algumas escolhas de aplicativos. De acordo com Agostini (2017), a alternativa Explorer, como o próprio nome sugere, é uma ferramenta para descobrir os dados, na qual é viável manusear os registros de forma simplificada, para isso são empregadas diversas opções de processamento de dados e mineração de dados providenciadas pelo WEKA. A aplicação Experimenter viabiliza a execução de experimentos e testes estatísticos envolvendo os sistemas de aprendizagem. A função KnowlwdgeFlow possui aplicações análogas às do Explorer, contudo, funciona no sistema clique e arraste. O Workbench é como uma estação de trabalho com todas as opções anteriores de aplicativos em um único lugar. E por último, a escolha Simple CLI, distinta das outras escolhas, traz uma interface com linhas de comando para trabalhar.

Figura 15: Página inicial do WEKA.

Fonte: Elaborada pelo autor.

Para que seja possível manipular os dados pelo WEKA, é imprescindível que os registros sejam reconhecidos pelo software e estejam organizados em um arquivo no formato Attribute-Relation File Format (ARFF). O arquivo ARFF é um arquivo de texto com a estrutura de lista, em que os registros descritos compartilham dos mesmos atributos. A composição básica de um arquivo ARFF é constituída por duas seções, conforme pode ser observado na figura 16, em que na primeira seção são expostas informações do cabeçalho, na qual são informados os atributos, que têm a capacidade de ser do tipo NUMERIC, DATE ou STRING, e na segunda seção do arquivo são encontrados os dados, que são organizados de acordo com as declarações da primeira seção e são delimitados por vírgula (AGOSTINI, 2017).

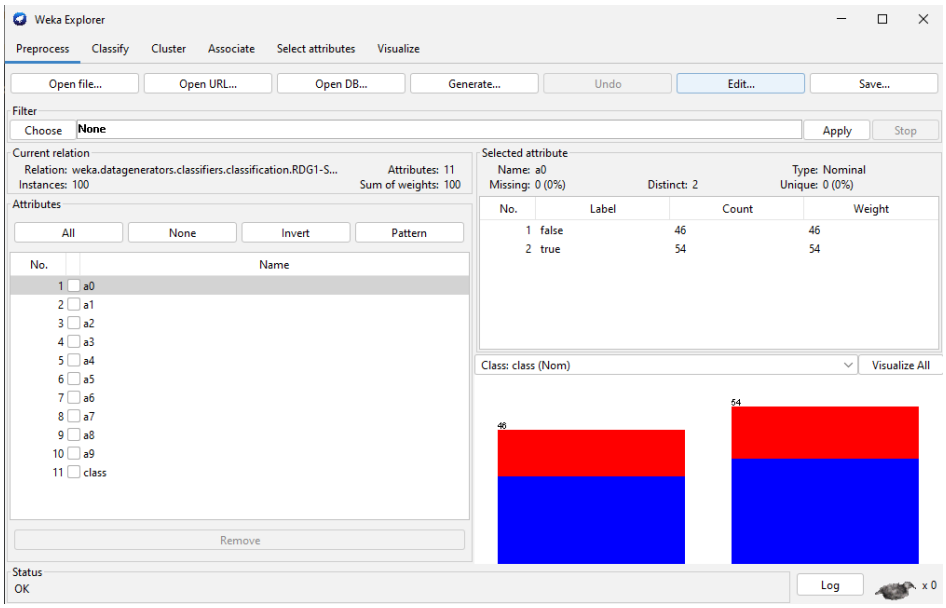
Figura 16: Exemplo de um arquivo ARFF.



Fonte: Elaborado pelo autor.

Com um arquivo ARFF devidamente configurado em mãos, é viável observar e manipular os dados, além de realizar experimentos com WEKA. A função Explorer (figura 17) provê várias atividades para lidar com mineração de dados, incluindo tarefas de classificação, agrupamento e associação. Para cada atividade de mineração de dados, existem diversas metodologias e fórmulas que podem ser utilizadas na execução dos testes no WEKA. Na classificação, por exemplo, pode-se mencionar árvore de decisão, multilayer perceptron e Naive Bayes. Em tarefas de agrupamento, o WEKA fornece algoritmos como Canopy, Cobweb e EM. Já na atividade de associação é permitido empregar os algoritmos Apriori, FilterAssociator e FP Growth.

Figura 17: Página Explorer do WEKA.



Fonte: Elaborado pelo autor.

De acordo com Brownlee (2016), quando o assunto é aprendizado de máquina aplicado, há muito o que aprender, como por exemplo, os algoritmos, os dados, o problema em específico em que está trabalhando, a matemática por trás de tudo, ferramenta que você planeja usar. Muitas vezes é dito que aprender uma nova linguagem de programação antes de iniciar no aprendizado de máquina aplicado, como Python ou linguagens mais esotéricas como Matlab ou R é o principal caminho. Porém, o fator essencial quando se fala em aprendizado de máquina é como entregar um resultado. Isto é, dado um problema, como trabalhá-lo e entregar um conjunto de previsões ou como entregar um modelo que pode gerar previsões. Não apenas

previsões, mas previsões precisas que pode ser entregue de forma robusta e confiável. Esta é a habilidade mais importante e isto geralmente envolve etapas como:

- Definir o problema
- Preparar os dados.
- Ter a avaliação de um conjunto de algoritmos.
- Melhorar os resultados com afinações e conjuntos.
- Finalizar o modelo e apresentar os resultados.

Este é o processo de aprendizado de máquina aplicado. Selecionar conjuntos de dados padrão de uma variedade de problemas domínios, como biologia, física e publicidade, e uma variedade de tipos de problemas, como classificação binária e multiclasse, regressão, conjuntos de dados não balanceados e muito mais. No aprendizado de máquina aplicado, a recuperação rápida, confiável e sistemática dos resultados é mais importante do que a maioria das outras coisas. Para isso e muito mais, o Weka é o caminho mais confiável a se seguir (BROWNLEE, 2016).

2.12 Estudos relacionados

Há diversos estudos relevantes que se relacionam ao emprego de inteligência artificial utilizando técnicas de mineração de dados para prever ativos na bolsa de valores. Silva (2015) explica sobre importância da previsão do mercado de ações para o desenvolvimento de algoritmos de negociação mais eficazes. O autor desenvolve modelos baseados em redes neurais MultiLayer Perceptron e em uma abordagem ensemble, que combina duas MLPs, para prever a direção do preço das ações em um curto intervalo de tempo.

O algoritmo de negociação proposto utiliza a saída do modelo para tomar decisões e maximizar o retorno. Conduziu simulações em um simulador realístico da Bolsa de Valores de São Paulo, mostrando que as técnicas de aprendizado de máquina melhoram a eficácia no processo de tomada de decisão. A abordagem ensemble mostrou-se mais precisa na previsão e gerou melhores resultados na simulação realística.

Giacomel (2016) propõe modelos de conjuntos baseados em redes neurais para a previsão de séries temporais no contexto do mercado de ações. Há dois

conjuntos distintos, que são adaptados de acordo com o perfil do investidor: o primeiro é destinado a investidores com perfil moderado e o segundo é para aqueles que desejam arriscar mais, com o perfil mais agressivo. Ambos objetivam classificar saídas por meio da aplicação de redes neurais em séries temporais, prevendo tendências de aumento e diminuição dos preços dos ativos, para que assim, com esses resultados, auxiliem o investidor a decidir entre comprar ou vender. O autor utiliza o framework Encong para a criação das redes neurais, o qual pode ser encontrado nas linguagens de programação Java, C++ e .Net.

Na pesquisa de Souza (2021) foi baseada em utilizar técnicas de mineração de dados para prever preços futuros de ações, ajudando os investidores na tomada de decisões no mercado acionário. Os dados históricos da ação PETR4 foram extraídos do Yahoo Finance para a realização do experimento. A limpeza e estruturação dos dados foram realizadas com um programa e manualmente. Utilizando o software WEKA, a mineração de dados foi realizada e os resultados indicaram a aplicação de prever se a ação tem comportamento de aumento ou diminuição de preço em um intervalo definido de dias.

3. MATERIAIS E MÉTODOS

O propósito desta seção consiste em explanar os materiais e métodos empregados na elaboração deste projeto, bem como as fases e procedimentos seguidos no decorrer da execução das tarefas.

3.1 Materiais

Para a execução deste projeto, foram empregados o software WEKA, informações adquiridas no site InfoMoney e um notebook fabricado pela Samsung, o qual atende às seguintes especificações:

- Modelo: Samsung 550XBE/350XBE;
- Sistema operacional: Windows 10 Pro;
- Processador: Intel® Core™ i5-8265U CPU 1.60 GHz – 1.80 GHz;
- Capacidade de armazenamento: SSD de 446 GB;
- Memória RAM: 12 GB.

3.2 Métodos

A condução deste projeto foi segmentada em quatro fases, que se inicia com a revisão bibliográfica, seguindo com a seleção dos dados, pré-processamento e, por fim, a análise preditiva. Na primeira fase, a revisão bibliográfica, buscou-se por literaturas semelhantes ao tema selecionado, bem como obras relacionadas à ciência de dados, mercado de ações e técnicas de mineração de dados. Estudos foram realizados com o intuito de adquirir informações relacionadas ao tema, a fim de definir a melhor estratégia para solucionar o problema proposto.

Seguindo para a segunda fase, foram realizadas pesquisas por dados da bolsa de valores e escolhidos dados históricos diários da ação da Itausa (ITSA4). Utilizando a o site do InfoMoney, foram coletados dados históricos da ITSA4, sendo que a plataforma forneceu um arquivo no formato de valores separados por vírgula (CSV) com os dados do período entre 02/01/2019 e 30/12/2019.

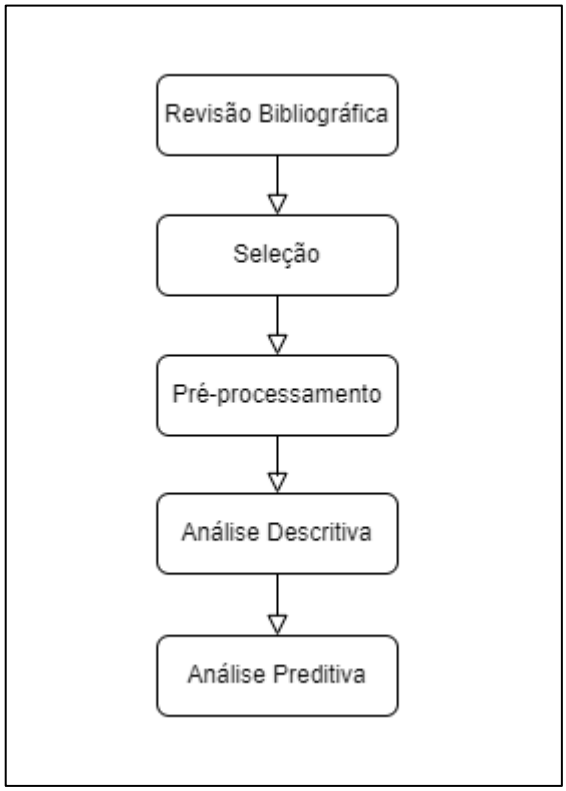
A fase de pré-processamento consiste na depuração e transformação dos dados. Na terceira fase, foram realizados procedimentos para estruturar, organizar e

depurar os dados. Em seguida, na quarta fase, foi realizada a análise preditiva, em que os dados foram processados e aplicaram-se técnicas de mineração de dados para prever os preços da ação PETR4. Nessa etapa, empregaram-se técnicas de árvore de decisão e redes neurais por meio do software WEKA. Em cada técnica, foram utilizados três arquivos com estruturas diferentes e analisados índices que demonstram o nível de acurácia obtido na classificação usando cada técnica.

De acordo com Prodanov e Freitas (2013), há várias maneiras de classificar uma pesquisa científica, sendo as principais por sua finalidade, objetivos, abordagem e procedimentos técnicos. Em termos de sua finalidade, este trabalho busca gerar conhecimento para solucionar problemas específicos na prática, caracterizando-se como uma pesquisa aplicada.

De acordo com Nascimento e Sousa (2015), na pesquisa experimental, estabelece-se um objeto de estudo, selecionam-se as variáveis que podem influenciá-lo e definem-se mecanismos e formas de controle e observação dos efeitos causados pelas variáveis selecionadas sobre o objeto pesquisado, conforme ilustrado na figura 18.

Figura 18: Método para análise dos dados.



Fonte: Elaborado pelo autor.

Em termos de seus objetivos, segundo Gil (2017) essa pesquisa pode ser categorizada como exploratória e descritiva, pois inicialmente foi conduzida uma revisão bibliográfica para adquirir uma melhor compreensão sobre a bolsa de valores e a mineração de dados. Em seguida, dados da bolsa de valores foram coletados e experimentos de mineração de dados foram realizados para estabelecer conexões entre as variáveis obtidas, e, posteriormente, os resultados obtidos foram descritos. A abordagem utilizada nesta pesquisa é quantitativa, uma vez que a análise dos resultados foi baseada em técnicas estatísticas.

Em termos de procedimentos técnicos, segundo Prodanov e Freitas (2013) podemos definir essa pesquisa como bibliográfica e experimental, já que foi elaborado a partir de material bibliográfico já publicado, como livros, teses e monografias, com o objetivo de obter conhecimento a partir de material escrito sobre mineração de dados e bolsa de valores. A manipulação de variáveis para observar os resultados é uma característica de pesquisas experimentais. Neste trabalho, as variáveis da bolsa de valores são manipuladas com o propósito de analisar os resultados obtidos na mineração de dados.

4. RESULTADOS E DISCUSSÃO

Esta seção tem como propósito expor os resultados alcançados ao longo deste estudo. Foram efetuados experimentos empregando algoritmos de árvore de decisão e redes neurais, e nos experimentos realizados com a árvore de decisão, utilizou-se o algoritmo J48, que representa uma variante do algoritmo C.45. O J48 gera árvores de decisão, em que cada nó analisa a presença e a importância dos atributos individualmente. Essas árvores são construídas de maneira ascendente, escolhendo o atributo mais apropriado em cada caso à medida que se avança da parte superior até a base. Nos experimentos com redes neurais empregou-se o *multilayer perceptron*, mencionado no capítulo 2.

As avaliações foram realizadas utilizando duas das quatro opções de testes oferecidas pelo software WEKA. A primeira opção de teste é o *Percentage split*, que consiste em dividir o conjunto de dados fornecido previamente em duas bases de dados, reservando um percentual pré-estabelecido dos dados para o treinamento e o restante para os testes. Na segunda opção de teste, utiliza-se toda a base de dados tanto para o treinamento quanto para os testes. Esse tipo de avaliação é ativado pela opção *Use training set*.

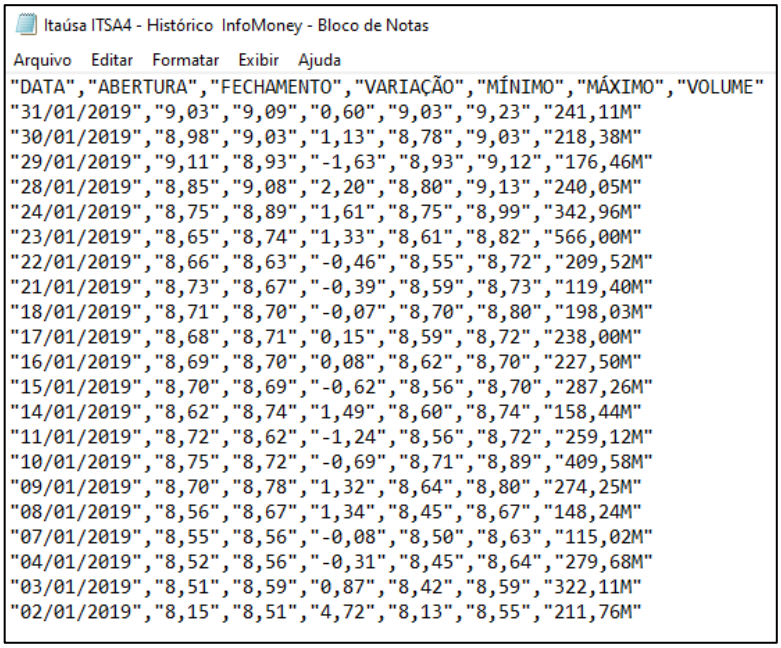
4.1 Escolha e pré-processamento dos dados

Há várias origens para esse gênero de dados de valores de compra e venda de ações e muitas plataformas digitais provêm informações acerca da bolsa de valores. Uma delas é a plataforma do InfoMoney, que disponibiliza informações de todas as ações comercializadas na bolsa brasileira, bem como indicadores de ações transacionadas em bolsas internacionais. O InfoMoney fornece os seguintes dados históricos de ações: data, abertura, máximo, mínimo, fechamento, fechamento adequado e volume. É permitido também selecionar o período em que se quer visualizar os dados históricos, além da plataforma possibilitar o download de arquivos no formato CSV contendo os dados.

Para este experimento, foram adquiridos no InfoMoney informações históricas diárias referentes à ação ITSA4 da Itaúsa no período entre 02/01/2019 e 30/12/2019, por intermédio da transferência de um arquivo CSV. O arquivo obtido possui o formato

ilustrado na figura 19, em que cada coluna é segregada por vírgula e representa respectivamente os valores de cada um dos atributos. Posteriormente foi realizado uma adaptação manualmente para organizar e extrais apenas os dados de interesse para a realização dos testes.

Figura 19: Arquivo com dados da ação ITSA4



Itaúsa ITSA4 - Histórico InfoMoney - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
"DATA","ABERTURA","FECHAMENTO","VARIAÇÃO","MÍNIMO","MÁXIMO","VOLUME"
"31/01/2019","9,03","9,09","0,60","9,03","9,23","241,11M"
"30/01/2019","8,98","9,03","1,13","8,78","9,03","218,38M"
"29/01/2019","9,11","8,93","-1,63","8,93","9,12","176,46M"
"28/01/2019","8,85","9,08","2,20","8,80","9,13","240,05M"
"24/01/2019","8,75","8,89","1,61","8,75","8,99","342,96M"
"23/01/2019","8,65","8,74","1,33","8,61","8,82","566,00M"
"22/01/2019","8,66","8,63","-0,46","8,55","8,72","209,52M"
"21/01/2019","8,73","8,67","-0,39","8,59","8,73","119,40M"
"18/01/2019","8,71","8,70","-0,07","8,70","8,80","198,03M"
"17/01/2019","8,68","8,71","0,15","8,59","8,72","238,00M"
"16/01/2019","8,69","8,70","0,08","8,62","8,70","227,50M"
"15/01/2019","8,70","8,69","-0,62","8,56","8,70","287,26M"
"14/01/2019","8,62","8,74","1,49","8,60","8,74","158,44M"
"11/01/2019","8,72","8,62","-1,24","8,56","8,72","259,12M"
"10/01/2019","8,75","8,72","-0,69","8,71","8,89","409,58M"
"09/01/2019","8,70","8,78","1,32","8,64","8,80","274,25M"
"08/01/2019","8,56","8,67","1,34","8,45","8,67","148,24M"
"07/01/2019","8,55","8,56","-0,08","8,50","8,63","115,02M"
"04/01/2019","8,52","8,56","-0,31","8,45","8,64","279,68M"
"03/01/2019","8,51","8,59","0,87","8,42","8,59","322,11M"
"02/01/2019","8,15","8,51","4,72","8,13","8,55","211,76M"
```

Fonte: Elaborado pelo autor

Com os dados coletadas, optou-se por trabalhar somente com a variável de encerramento da ação ITSA4. Foram conduzidos dois testes, nos quais os valores de fechamento foram organizados de forma a ficarem agrupados em uma linha, sequenciados em blocos dos últimos trinta dias, de acordo com o experimento efetuado. A equação a seguir mostra a estrutura dos últimos cinco dias. Esse arranjo foi submetido a uma categorização, na qual foi considerado o dia D_n (dia atual) e o dia D_{n-1} (dia anterior).

$$D_{n-4}, D_{n-3}, D_{n-2}, D_{n-1}, D_n$$

Em seguida, foi criado um atributo adicional, capaz de assumir três situações: a classe "c" indica que o preço do dia anterior é inferior ao preço atual, isto é, haverá uma valorização do preço de um dia para o outro, sendo recomendável efetuar a compra a fim de obter lucro; a classe "v", por sua vez, indica o oposto, pois o preço do dia anterior é maior que o do dia atual, de modo que é aconselhável realizar a

venda para evitar prejuízos; a classe “n” indica que o preço não houve variação o suficiente para uma tomada de decisão de compra ou venda, neste caso a recomendação é não realizar nenhuma ação, estado neutro. Por fim, a partir do resultado desse pré-processamento, foi gerado um arquivo no formato ARFF para cada teste. A figura 20 ilustra o formato final do arquivo gerado com 5 dias.

Figura 20: Configuração definitiva do arquivo ARFF com fechamento de 5 dias.

```
@RELATION ITSA42019_30dias

@ATTRIBUTE data DATE "dd/mm/yyyy"
@ATTRIBUTE fechamento-1 REAL
@ATTRIBUTE fechamento-2 REAL
@ATTRIBUTE fechamento-3 REAL
@ATTRIBUTE fechamento-4 REAL
@ATTRIBUTE fechamento-5 REAL
@ATTRIBUTE acao {c,n,v}

@DATA
08/01/2019,11.29,11.44,11.36,11.22,11.39,c
07/01/2019,11.14,11.29,11.44,11.36,11.22,v
04/01/2019,11.15,11.14,11.29,11.44,11.36,n
03/01/2019,11.19,11.15,11.14,11.29,11.44,v
02/01/2019,11.09,11.19,11.15,11.14,11.29,v
```

Fonte: Elaborado pelo autor

Foi utilizado algumas métricas para definir se os resultados foram bem-sucedidos ou não:

O primeiro e mais simples de ser compreendido é o indicador de instâncias classificadas corretamente, que informa a taxa de acerto do algoritmo.

O segundo é o coeficiente kappa que pode ser definido como uma medida do grau de concordância entre dois conjuntos de dados categorizados. O resultado do Kappa varia entre os intervalos de 0 a 1. Quanto maior o valor de Kappa, mais forte é a concordância/vínculo. Se Kappa = 1, então há uma concordância perfeita. Se Kappa = 0, então não há concordância. Se os valores do coeficiente Kappa variarem na faixa de 0,40 a 0,59, são considerados moderados; de 0,60 a 0,79, são considerados substanciais; e acima de 0,80 são considerados excelentes.

O terceiro indicador é a matriz de confusão. A matriz de confusão se resume em uma tabela que exibe a quantidade de previsões para cada classe em comparação

com o número de instâncias que realmente pertencem a cada classe. Essa tabela é extremamente útil para obter uma visão geral dos diferentes tipos de erros cometidos pelo algoritmo.

4.2 Testes empregando técnica de árvore de decisão e redes neurais no período de 30 dias.

4.2.1 Redes neurais

Foram executados testes com redes neurais durante um período de trinta dias, utilizando um arquivo ARFF contendo 248 registros. Dentre esses registros, 102 foram definidos como compra (c), 120 como venda (v) e 26 como neutro (n). No teste inicial foi empregado o modelo de teste *percentage split*, que particionou o conjunto de dados em 66% para treinamento e 34% para testes.

Figura 21: Processamento com 30 dias usando redes neurais - *percentage split*.

```
Time taken to build model: 0.65 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances          74               88.0952 %
Incorrectly Classified Instances        10               11.9048 %
Kappa statistic                          0.7948
Mean absolute error                      0.1066
Root mean squared error                  0.2673
Relative absolute error                 27.1361 %
Root relative squared error             60.0591 %
Total Number of Instances               84

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,909    0,137    0,811      0,909    0,857      0,759    0,934     0,865     c
                 0,400    0,027    0,667      0,400    0,500      0,469    0,782     0,489     n
                 0,976    0,023    0,976      0,976    0,976      0,952    0,989     0,986     v
Weighted Avg.    0,881    0,068    0,874      0,881    0,872      0,819    0,943     0,879

=== Confusion Matrix ===

  a  b  c   <-- classified as
 30  2  1 |  a = c
  6  4  0 |  b = n
  1  0 40 |  c = v
```

Fonte: Elaborado pelo autor

Ao analisar a figura 21, observamos que o experimento apresentou uma acurácia considerável acertando 74 dos 84 registros destinados para testes. A matriz de confusão corrobora o resultado, mostrando que o modelo classificou corretamente 30 registros como compra (c) e 3 registros foram classificados de forma equivocada nessa categoria. Foram classificados corretamente 4 registros como neutro (n) e 6 de forma errônea nessa categoria. Além disso, foram classificados corretamente 40 registros como venda (v) e apenas um registro como erro nas classificações dessa categoria. O resultado foi considerado bom, pois esse modelo de teste apresentou uma porcentagem de 88,09% de acurácia para as instâncias classificadas corretamente e o índice kappa ficou bem próximo de 0.80.

O segundo teste com redes neurais, utilizando o mesmo conjunto de dados do treinamento, gerou resultados satisfatórios como mostrado na figura 22. O modelo conseguiu uma taxa de acerto de 98,38%, errando apenas 4 entre os 248 registros. A matriz de confusão apresentou que 100 registros foram classificados corretamente como compra e apenas 2 registros classificações incorretamente, 24 registros foram classificados como neutro com 2 registros incorretos, e na classe de venda 40 registros tiveram classificações corretas e apenas 1 registro com classificação incorreta. O índice kappa apresentou um valor de 0.9723.

Figura 22: Processamento com 30 dias usando redes neurais – treinamento.

```
Time taken to build model: 0.56 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         244               98.3871 %
Incorrectly Classified Instances         4                1.6129 %
Kappa statistic                          0.9723
Mean absolute error                      0.0232
Root mean squared error                  0.1051
Relative absolute error                  5.9217 %
Root relative squared error             23.7744 %
Total Number of Instances              248

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,980    0,007    0,990      0,980    0,985      0,975    0,988     0,981     c
                 0,923    0,000    1,000      0,923    0,960      0,956    0,936     0,933     n
                 1,000    0,023    0,976      1,000    0,988      0,976    0,989     0,973     v
Weighted Avg.    0,984    0,014    0,984      0,984    0,984      0,974    0,983     0,972

=== Confusion Matrix ===

   a   b   c   <-- classified as
 100   0   2 |   a = c
   1  24   1 |   b = n
   0   0 120 |   c = v
```

Fonte: Elaborado pelo autor

O experimento com treinamento apresentou resultados mais precisos, entretanto, não é possível afirmar que o modelo seja adequado, uma vez que o teste foi realizado no mesmo conjunto de dados utilizado para o treinamento. Tal abordagem pode levar a *overfitting*, já que o modelo treinado já está familiarizado com os dados de teste. Logo, um modelo que emprega diferentes conjuntos de dados para treinamento e teste pode ser considerado mais aplicável ao mundo real.

4.2.2 Árvore de decisão

Os experimentos realizados com trinta dias também empregaram árvore de decisão. O conjunto de dados utilizado nesses experimentos é o mesmo empregado nos experimentos anteriores. Neste primeiro processamento, o conjunto de dados foi dividido em grupos, com 66% dos registros destinados para treinamento e o restante separado para testes.

Figura 23: Processamento com 30 dias usando árvore de decisão - *percentage split*.

```
Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          58               69.0476 %
Incorrectly Classified Instances        26               30.9524 %
Kappa statistic                          0.4773
Mean absolute error                      0.214
Root mean squared error                  0.4396
Relative absolute error                 54.4586 %
Root relative squared error             98.771  %
Total Number of Instances               84

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,636    0,216    0,656      0,636    0,646      0,423    0,697     0,653     c
                 0,200    0,108    0,200      0,200    0,200      0,092    0,582     0,152     n
                 0,854    0,163    0,833      0,854    0,843      0,691    0,840     0,776     v
Weighted Avg.    0,690    0,177    0,688      0,690    0,689      0,514    0,753     0,654

=== Confusion Matrix ===

  a  b  c   <-- classified as
 21  6  6 |  a = c
  7  2  1 |  b = n
  4  2 35 |  c = v
```

Fonte: Elaborado pelo autor

O experimento resultou em uma árvore com 45 nós, sendo 23 deles nós-folha. Pode-se notar que o primeiro experimento, utilizando redes neurais com trinta dias, obteve resultados superiores aos experimentos com árvore de decisão. A figura 23 exibe os resultados obtidos e a matriz de confusão revela que o modelo conseguiu efetuar 21 classificações como compra (c), apenas 2 como neutro (n) e 35 como venda (v). O índice kappa se mostrou abaixo dos valores encontrados nos experimentos anteriores retornando um valor de 0.4773. O resultado foi considerado bom, pois esse modelo de teste apresentou o indicador de classificações corretas com uma acurácia de 69,04%.

No segundo experimento, foi empregado o modelo de teste que usa o conjunto de dados completo tanto para treinamento quanto para testes. Nesse modelo de teste, a árvore gerada pelo experimento possui a mesma do experimento anterior com 45 nós, sendo 23 nós-folha. O modelo conseguiu uma acurácia de 97,98%, acertando 243 das 248 classificações possíveis, como ilustrado na figura 24. A matriz de confusão revela que dos 102 registros definidos como compra, 97 foram classificados corretamente resultando em apenas 5 classificações de forma errônea. Nas classes de venda e neutro o modelo foi capaz de classificar 100% dos registros corretamente. O índice kappa foi considerado excelente retornando o valor de 0.9657.

Figura 24: Processamento com 30 dias usando árvore de decisão – treinamento.

```
Time taken to build model: 0.04 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         243               97.9839 %
Incorrectly Classified Instances         5                2.0161 %
Kappa statistic                          0.9657
Mean absolute error                      0.023
Root mean squared error                  0.1072
Relative absolute error                  5.8747 %
Root relative squared error             24.2576 %
Total Number of Instances              248

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,951    0,000    1,000      0,951    0,975      0,959    0,995     0,993     c
                 1,000    0,009    0,929      1,000    0,963      0,959    0,999     0,982     n
                 1,000    0,023    0,976      1,000    0,988      0,976    0,998     0,996     v
Weighted Avg.    0,980    0,012    0,981      0,980    0,980      0,967    0,997     0,993

=== Confusion Matrix ===

   a   b   c   <-- classified as
  97   2   3 |   a = c
   0  26   0 |   b = n
   0   0 120 |   c = v
```

Fonte: Elaborado pelo autor

Nesses experimentos com trinta dias utilizando árvore de decisão, os resultados foram inferiores em relação aos experimentos com redes neurais, também feito com trinta dias. Tanto no teste utilizando a opção *Percentage split* quanto nos testes utilizando a opção *Use training set*, a diferença dos resultados entre de redes neurais e árvore de decisão foi razoavelmente considerável, demonstrando que, para o modelo com trinta dias, o uso de redes neurais apresenta um potencial pouco maior do que o uso de árvore de decisão.

4.3 DISCUSSÕES

Os resultados apresentados na seção 4.2 revelaram uma melhoria modesta quando realizados com redes neurais, com índices de precisão acima de 88%. De forma semelhante, os níveis de precisão alcançados nos experimentos com árvore de decisão se mantiveram entre de 69% e 97%.

A tabela 1 mostra comparações entre os algoritmos classificadores de redes neurais e árvore de decisão usando duas opções de teste, ou seja, o método de treinamento e o método de *percentage split*. Neste artigo, dois parâmetros são utilizados para analisar o desempenho do classificador e determinar qual método é melhor. A partir das estatísticas da tabela 1, é evidente que o método de treinamento apresenta um desempenho superior em relação ao método de *percentage split*.

Tabela 1: Comparação dos classificadores usando os métodos de teste.

Classificador	**Conjunto de treinamento**		***Percentage Split***	
	Índice Kappa	Instâncias Classificadas Corretamente (%)	Índice Kappa	Instâncias Classificadas Corretamente (%)
Redes Neurais	0.9723	98,38	0.7948	88,09
Árvore de decisão	0.9657	97,98	0.4773	69,04

Fonte: Elaborado pelo autor

Ao analisar o desempenho dos algoritmos, observamos que o modelo de rede neural multicamadas perceptron apresenta um melhor desempenho utilizando o método *percentage split*. O valor máximo da estatística Kappa foi de 0.7948 e o indicador de classificações corretas alcançou 88,09%. Por outro lado, o desempenho da árvore de decisão J45 foi inferior nos resultados, mas ainda pode ser considerado. O valor máximo da estatística Kappa foi de 0.4773 e o indicador de classificações corretas foi de 69,04%.

5. CONSIDERAÇÕES FINAIS

A pesquisa relacionada à ciência de dados, especialmente à mineração de dados, desempenha um papel crucial devido ao crescente volume de informações disponíveis. É essencial ter a capacidade de extrair conhecimento valioso desses dados em constante expansão. Portanto, é fundamental contar com uma disciplina científica dedicada a essa tarefa, que se torna cada vez mais indispensável.

Com o objetivo de alcançar o propósito geral deste trabalho, foram atendidos os elementos dos objetivos específicos, os quais foram estabelecidos no primeiro capítulo desta pesquisa. A definição dos conceitos relacionados ao mercado de ações e mineração de dados foi resultado de uma pesquisa exploratória, que envolveu a busca por renomados autores e trabalhos relevantes sobre esses temas.

No que diz respeito à aplicação do processo de descoberta de conhecimento, foi realizada uma investigação de dados históricos, sendo decidido que os dados seriam obtidos do site InfoMoney, ao qual oferecem uma variedade de informações relacionadas ao mercado financeiro. As etapas do processo foram aplicadas de acordo com o que foi apresentado no terceiro capítulo. Na etapa de mineração de dados, foram utilizadas técnicas de redes neurais e árvore de decisão.

Com base em todas as pesquisas realizadas neste trabalho, foi concluído que é possível aplicar a mineração de dados no contexto dos investimentos, desde que a tarefa seja selecionada de maneira adequada. Nesse estudo, verificou-se que o uso da técnica de árvore de decisão se mostrou tão eficiente quanto a técnica de redes neurais. Comparando os resultados entre as técnicas, tiveram pouca diferença de precisão em contraste com os resultados obtidos em outros estudos relacionados. Em contrapartida, ao analisar os modelos que utilizaram redes neurais, todos os experimentos apresentaram resultados consistentes e satisfatórios.

5.1 Recomendações para trabalhos futuros

Dentro do contexto de previsões por meio da mineração de dados, reconhece-se que seu potencial adaptativo é vasto, permitindo sua aplicação em diversas áreas do conhecimento. Com base nisso, recomenda-se a realização de estudos envolvendo essa metodologia em conjunto com os seguintes parâmetros:

- Desenvolver múltiplas redes neurais e realizar previsões conjuntas, como relacionar previsões de tendências com previsões de taxas de câmbio.
- Utilizar bases de dados com intervalo de dias acima de 30 dias para prever oscilações maiores.
- Realizar previsões dos preços no mercado de derivativos.
- Realizar testes com outras variáveis de entrada, como câmbio, índices de bolsas de outros países, taxas de juros internas e externas, entre outras.
- Utilizar outras técnicas de mineração de dados com mais variações de algoritmos.
- Utilizar outras variações de indicadores fornecidos pelo WEKA para análise.

REFERÊNCIAS

AGOSTINI, Michel**. Estudo comparativo entre as ferramentas weka e sas no processo de descoberta de informações**. 2017. 55 f. Monografia (Especialização) - Curso de Especialização em Banco de Dados, Universidade Federal de Mato Grosso, Cuiabá, 2017.

AMORIM, Thiago. **Conceitos, técnicas, ferramentas e aplicações de mineração de dados para gerar conhecimento a partir de bases de dados**. 2006. Trabalho de Conclusão de Curso (Graduação em ciência da computação) - Centro de Informática, Universidade Federal de Pernambuco, Recife, 2006. Disponível em: <http://www.cin.ufpe.br/~tg/2006-2/tmas.pdf>. Acesso em: 20 abr. 2021.

ANUMALLA, K. **Sistema de Gestão da Pré-Processamento de Dados** (Dissertação de Mestrado). USA: Universidade de Akron, 2007.

BERENSTEIN, Marcelo. **Uso de mineração de dados na bolsa de valores.** 2010. 95 f. TCC (Graduação) - Curso de Ciência da Computação, Universidade do Vale do Itajaí, Itajaí, 2010.

BERRY, M., LINOFF, G. **Mastering Data Mining**: the art and science of customer relationship management. John Wiley & Sons, 2000.

BROWNLEE, J. **Machine Learning Mastery With Weka**. Machine Learning Mastery, 2016

CASTRO, Leandro Nunes de; FERRARI, Daniel Gomes. **Introdução à mineração de dados:** conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

COMISSÃO DE VALORES MOBILIÁRIOS (org.). **Mercado de valores mobiliários brasileiro.** 4. ed. Rio de Janeiro: Comissão de Valores Mobiliários, 2019.

COMISSÃO DE VALORES MOBILIÁRIOS (org.). **Análise de Investimentos: histórico, principais ferramentas e mudanças conceituais para o futuro.** 1. ed. Rio de Janeiro: Comissão de Valores Mobiliários, 2017.

DUNHAM, M. **Data mining: Introductory and advanced topics**. Upper Saddle River, NJ: Prentice Hall, 2003.

FAYYAD, U.; PIATESKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R**. Advances in knowledge discovery and data mining.** 1996. Cambridge: MIT Press, 1996. 560p.

FOGAÇA, André. **Bolsa de valores para leigos.** São Paulo: Guiainvest, 2015.

GIACOMEL, Felipe dos Santos. **Um Método Algorítmico para Operações na Bolsa de Valores Baseado em Ensembles de Redes Neurais para Modelar e Prever os Movimentos dos Mercados de Ações.** 2016. 92 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 6. ed. São Paulo: Atlas, 2017.

GOLDSCHMIDT, R., PASSOS, E. **Data Mining um Guia Prático**. Conceitos, Técnicas, Ferramentas, Orientações e Aplicações. Ed. Campus, Rio de Janeiro, 2005.

NASCIMENTO, Francisco Paulo do; SOUSA, Flávio Luís Leite. **Metodologia da Pesquisa Científica: Teoria e Prática.** Brasília: Thesaurus, 2015.

NGAI, E.W.T., HU, Y., WONG, Y.H., CHEN. Y., SUN, X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. **Decision Support Systems**, v. 50, 2011.

MARANGONI, Pedro Henrique. **Redes Neurais Artificiais para Previsão de Séries Temporais no Mercado Acionário.** 2010. 80 f. TCC (Graduação) - Curso de Ciências Econômicas, Universidade Federal de Santa Catarina, Florianópolis, 2010.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico.** 2. ed. Novo Amburgo: Universidade Feevale, 2013. Disponível em: https://www.feevale.br/institucional/editora-feevale/metodologia-do-trabalhocientifico---2-edicao. Acesso em: 103 maio 2023.

RELICH, M., MUSZYNSKI, W. The use of intelligent systems for planning and scheduling of product development projects. **Procedia Computer Science**, v. 35, p. 1586-1595, 2014.

ROQUE, Reginaldo do Carmo. **Estudo sobre a empregabilidade da previsão do índice BOVESPA usando redes neurais artificiais**. 2009. 102 f. TCC (Graduação) - Curso de Engenharia Eletrônica e de Computação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2009.

ROSA, C.R.M. **Uma metodologia para a descoberta de conhecimento em bases de dados visando a classificação de padrões**. 2017. 158 p. Tese (Doutorado) – Programa de Pós-Graduação em Engenharia de Produção e Sistemas, Pontifícia Universidade Católica do Paraná, Curitiba, 2017.

ROSS, S. A. et alli. **Administração Financeira: Corporate Finance.** São Paulo: Atlas, 2002. SZEWCZYK, S. H. et alli. Do Dividend Omissions Signal Future Earnings of Past Earnings? In: Journal of Investing, Spring, 1997.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência artificial.** 3. ed. Rio de Janeiro: Elsevier, 2013.

SHANNON, C. E. & WEAVER, W. **The Mathematical Theory of Communication. University of Illinois Press,** 1949.

SHARDA, DELEN & TURBAN, **Business Intelligence, Analytics, and Data Science**: a managerial perspective, New York, NY: Pearson, 2018.

SILVA, Marcelino Pereira dos Santos; **Mineração de dados: conceitos, aplicações e experimentos com weka**. In: Simpósio de informática do CEFET-PI, 6, 2008, Teresina. Minicurso, Teresina: CEFET-PI. P. 1-20.

SILVA, E.J. **Modelagem e aplicação de técnicas de aprendizado de máquina para negociação em alta frequência em bolsa de valores,** Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, 2015.

SOUZA, W. B. C. **Mineração de Dados Aplicada a Previsão de Preços de Ações Utilizando WEKA**. Trabalho de Conclusão de Curso apresentado à Escola de Ciências Exatas e da Computação, da Pontifícia Universidade Católica de Goiás. 2021.

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
GABINETE DO REITOR

Av. Universitária, 1069 • Setor Universitário
Caixa Postal 86 • CEP 74605-010
Goiânia • Goiás • Brasil
Fone: (62) 3946.1000
www.pucgoias.edu.br • reitoria@pucgoias.edu.br

RESOLUÇÃO nº 038/2020 – CEPE

ANEXO I

APÊNDICE ao TCC

Termo de autorização de publicação de produção acadêmica

O(A) estudante Daniel Bueno de Oliveira do Curso de Ciência da Computação, matrícula 2016100280377-9, telefone: 62991401468 e-mail bueno98daniel@gmail.com, na qualidade de titular dos direitos autorais, em consonância com a Lei nº 9.610/98 (Lei dos Direitos do Autor), autoriza a Pontifícia Universidade Católica de Goiás (PUC Goiás) a disponibilizar o Trabalho de Conclusão de Curso intitulado Classificação de ativos no mercado de ações utilizando mineração de dados, gratuitamente, sem ressarcimento dos direitos autorais, por 5 (cinco) anos, conforme permissões do documento, em meio eletrônico, na rede mundial de computadores, no formato especificado (Texto(PDF); Imagem (GIF ou JPEG); Som (WAVE, MPEG, AIFF, SND); Vídeo (MPEG, MWV, AVI, QT); outros, específicos da área; para fins de leitura e/ou impressão pela internet, a título de divulgação da produção científica gerada nos cursos de graduação da PUC Goiás.

Goiânia, 26 de junho de 2023.

Assinatura do autor: Daniel Bueno de Oliveira

Nome completo do autor: Daniel Bueno de Oliveira

Assinatura do professor–orientador: Sibelius Lellis Vieira

Nome completo do professor-orientador: Sibelius Lellis Vieira