UNIÃO DE ENSINO DO SUDOESTE DO PARANÁ – UNISEP
FACULDADE DE EDUCAÇÃO DE FRANCISCO BELTRÃO - FEFB
COORDENADORIA DE TRABALHO DE CONCLUSÃO DE CURSO
SISTEMAS DE INFORMAÇÃO

ANÁLISE DE EMPRESAS PARA INVESTIDORES A LONGO PRAZO COMO SÓCIO DENTRO DA BM&FBOVESPA UTILIZANDO MINERAÇÃO DE DADOS

FRANCISCO BELTRÃO
(2017)

EDUARDO ALEXANDRE FRANCISCON

ANÁLISE DE EMPRESAS PARA INVESTIDORES A LONGO PRAZO COMO SÓCIO DENTRO DA BM&FBOVESPA UTILIZANDO MINERAÇÃO DE DADOS

Orientador: Profª. Roberto Cesar da Silva Padilha, Esp.

FRANCISCO BELTRÃO
(2017)

TERMO DE APROVAÇÃO

EDUARDO ALEXANDRE FRANCISCON

ANÁLISE DE EMPRESAS PARA INVESTIDORES A LONGO PRAZO COMO SÓCIO DENTRO DA BM&FBOVESPA UTILIZANDO MINERAÇÃO DE DADOS

Trabalho de Conclusão de Curso apresentado ao Curso de Sistemas de Informação da União de Ensino do Sudoeste do Paraná – UNISEP, como requisito parcial para a obtenção do título de Bacharel em Sistemas de Informação.

BANCA EXAMINADORA

Titulação. Nome Professor(a)
Orientador(a)

Titulação. Nome Professor(a)
Membro da Banca

Titulação. Nome Professor(a)
Membro da Banca

Francisco Beltrão, 12 de dezembro de 2017

AGRADECIMENTOS

Gostaria de agradecer a todos que de alguma forma participaram, da elaboração deste presente projeto. Gostaria de agradecer ao professor Dr. Jones Granatyr, por ter feito o trabalho de orientação do projeto durante a primeira fase do trabalho de conclusão de curso, e ter acompanhado o desenvolvimento até o seu final, e por ter possibilitado que este mesmo viesse a se tornar um artigo científico mais tarde.

Também agradeço ao professor Esp. Roberto Cesar da Silva Padilha, por ter acolhido o trabalho na segunda fase de desenvolvimento e ter orientado até o seu término, sua ajuda foi muito importante, pois a sua orientação possibilitou que este trabalho viesse a ser concluído com êxito e tornar-se o artigo dito anteriormente.

Agradeço ao professor Me. Fábio Alexandre Taffe por ajudar na continuidade dos estudos que este trabalho de conclusão de curso disponibilizou e por toda a ajuda no decorrer do ano letivo prestada.

Agradeço também ao professor Gilson Reis, pela sua ajuda e esclarecimento no momento em que precisei para a elaboração do trabalho.

Agradeço a minha família pelo apoio e pela paciência nas horas difíceis que passei para a elaboração do trabalho, sem a ajuda e compreensão deles teria sido muito mais difícil.

SUMÁRIO

LISTA DE FIGURAS

pag.

LISTA DE TABELAS

pag.

LISTA DE ABREVIAÇÕES

WEKA..... *Waikato Environment for Knowledge Analysis*

IPO..... *Initial Public Offering*

IA..... Inteligência Artificial

PCA.....*Principal Components Analysis*

PC.....Computador

GPL..... *General Public License*

TDM..... *Text Data Minning*

PUCPR.....Universidade Católica De Pontifica – Paraná

BOVESPA..... Bolsa de Valores de Oficial de São Paulo

BM&FBOVESPA.....Bolsa de Valores, Mercadorias e Futuros de São Paulo

API.....*Application Programming Interface*

LISTA DE CÓDIGOS-FONTE

RESUMO

FRANCISCON, E. A. **Análise de empresas para investidores a longo prazo como sócio dentro da BM&FBOVESPA utilizando mineração de dados.** 2017. Francisco Beltrão. Trabalho de Conclusão do Curso de Sistemas de Informação da Faculdade de Educação de Francisco Beltrão - FEFB

Neste presente trabalho será apresentado um modelo de detecção de boas empresas para investimento a longo prazo dentro da bolsa de valores. Este modelo de detecção, funciona com a utilização de mineração de dados, onde através de técnicas de seleção, será treinado e testado uma base de conhecimento que retorna uma posição de boa, média ou má empresa para investimento. O *software* utilizado para a elaboração dos testes e resultado deste projeto foi o Weka *(Waikato Environment for Knowledge Analysis)*. Esta posição sobre as empresas foram alcançadas com o teste de todos os algoritmos disponíveis no *software* do Weka, *software*, que mais tarde passaram por filtros, restando apenas cinco algoritmos, os que apresentaram melhores resultados. A base de conhecimento utilizada aqui foi construida com dados reais de variáveis financeiras das empresas cadastradas na BM&FBOVESPA. As validações dos testes realizados para encontrar o resultado final do projeto, foram elaboradas com a utilização do software do Weka, que disponibilizou os algoritmos e os métodos necessários para a construção da base de conhecimento. A base foi testada com a utilização dos métodos de seleção de atributos, Wrapper e Discretize para a fase de treinamento da base de conhecimento, e para a realização dos testes da base, foi utilizado o método de validaçao cruzada *(cross-validation)*.

Palavras chaves: Mineração de dados, bolsa de valores, BM&FBOVESPA, identificação de boas empresas, investimento como sócio.

1.0 - INTRODUÇÃO

A bolsa de valores é alvo de muitos investidores hoje no Brasil, investidores estes que vem crescendo consideravelmente de número nos últimos anos. Pode-se ressaltar a respeito disso:

> Investidores individuais tiveram participação de 30% do volume negociado na BOVESPA nos três primeiros meses de 2010. De fevereiro de 2008 a fevereiro de 2010, o aumento de contas de investidores pessoa física com posição de custódia foi de 19%. (MARANGONI, 2010)

Com o aumento crescente destes investidores no mercado acionário é muito comum que haja muitas dúvidas de como investir, ou em qual empresa negociar ações. Sendo que o investimento por conta do acionista pode muitas vezes lhe causar prejuízos, se não estudar com atenção sobre as ações desejadas.

A mineração de dados vem sendo utilizada para o auxílio de tomada de decisões para investidores que desejam apostar em ações com as mais variadas situações. É comum encontrar acionistas que apostam em resultados derivados da mineração de dados para suas negociações na bolsa de valores, tanto a curto quanto a longo prazo. Este projeto mostra uma forma de detectar empresas com ações que são visadas para a aplicação a longo prazo, ou seja, para investimentos em interesses com o passar dos anos. O objetivo é detectar boas ações para este modelo de investimento.

Segundo os consultores da ToroRadar.com, um dos grandes motivos pelos quais as pessoas não entram neste modelo de negócio é a dificuldade de interpretação das ações, ou do próprio negócio em si. Em meio a tantas variáveis que podem moldar um bom ou mal negócio, encontram-se diversas dificuldades, que acabam resultando em prejuízo para investidores sem um bom conhecimento do mercado acionista. Este trabalho tem a finalidade de ajudar um investidor a perceber boas ações de empresas para investimentos a longo prazo, como investidor sócio, ou seja, que tem interesse em colher os lucros gerados pelas empresas em um período de longo prazo. Isto traz mais tranquilidade e segurança para investidores sem muita experiência no ramo.

Para a elaboração do presente trabalho foi necessário estudar o funcionamento deste modelo de negócio, assim como acontecem as mudanças que podem ocorrer e por que motivos. Por ser um mercado repleto de particularidades é necessário também o entendimento aprofundado das variáveis que o pertencem. É

obrigatório o trabalho de extração destas informações de alguma fonte de dados de confiança, pois é crucial para mais tarde utilizar na mineração de dados. Neste projeto foi utilizado os dados do site Bastter, um site que contém as informações das ações das empresas listadas na BM&FBOVESPA. A continuidade do trabalho de extração dos dados anteriormente elencado se dá em montar um banco de dados que mais tarde será utilizado para testes e também para obter um resultado que seja satisfatório para a aplicação dos determinados objetivos do projeto.

O resultado deste projeto é a aplicação de regras de associação, derivadas da mineração de dados. Para a execução do banco de dados e a aplicação da técnica é utilizado o WEKA, um *software* gratuito que trabalha com estes dados de maneira satisfatória.

2.0 - OBJETIVOS

2.1 - Geral

- Construir e avaliar uma solução de mineração de dados para escolher as melhores empresas para investimento a longo prazo na bolsa de valores.

2.2 – Específicos

- Compreender os indicadores da análise fundamentalista para construir a base de dados.
- Criar a base de conhecimento contendo todas as variáveis necessárias.
- Avaliar e testar algoritmos de classificação e métodos de estatísticos.
- Executar e coletar os resultados dos algoritmos de classificação e métodos estatísticos mais relevantes para verificar a eficácia da abordagem proposta para classificar perfis das empresas.

3.0 - JUSTIFICATIVA

Nos dias atuais a mineração de dados vem se tornando cada vez mais utilizada para a elaboração de trabalhos que visam prever ou salientar algumas informações que são precisas ou indispensáveis para a tomada de decisões sobre os mais diversos temas ou aplicações. Técnicas de *data mining* são utilizadas em diversas áreas com várias finalidades. Como por exemplo: Na medicina, no marketing, nas telecomunicações, na educação ou nas finanças, como é o caso da aplicação deste trabalho. (GOLDSCHMIDT; BEZERRA, 2016)

A aplicação de mineração de dados no mercado acionista pode trazer inúmeros benefícios, a longo ou a curto prazo. Benefícios estes que podem trazer resultados muito satisfatórios. Pois através de uma análise pode ser percebido que algumas variáveis podem ter grande participação no resultado de uma empresa ou ação. E estes resultados podem ser positivos ou negativos. (GOLDSCHMIDT; BEZERRA, 2016)

Na área de finanças a aplicação de mineração de dados podem trazer informações que nas mãos de profissionais capacitados podem gerar lucros muito bons. Pois estas informações permitem que o profissional se antecipe de seus concorrentes na tomada de decisões e assim leve uma vantagem para com os demais. (VELOSO; MOREIRA; SILVA; SILVA, 2011)

Outro benefício do *data mining* é a avaliação e resultado de um banco de dados enorme. Uma vez que seja necessário a avaliação de grandes quantidades de dados para o estudo de algum fator importante para uma decisão ou aplicação, torna-se muito difícil e demorado o resultado final. Pois a análise é na maioria das vezes complicada, pois se trata de inúmeras variáveis capazes de alterar resultados ou interferir diretamente neles. E como é sabido, bolsa de valores está longe de ser simples e fácil de ser avaliada. (VELOSO; MOREIRA; SILVA; SILVA, 2011)

Com a utilização da mineração de dados dentro da bolsa de valores, será possível identificar pontos ou informações relevantes que propiciam um bom investimento nas ações desejadas. Uma vez que o relatório adquirido através do *data mining* venha a ser estudado e analisado. Será capaz de fazer comparações sobre os dados da empresa para que se obtenham informações capazes de nortear o usuário na decisão de aplicação ou não nas ações da empresa. Isto é possível graças ao histórico de dados que é alimentado no banco de informações processados depois da aplicação do *data mining*.

Será possível também avaliar o desempenho financeiro dentro da bolsa de valores das empresas que estiverem sob a análise da mineração de dados. Uma vez que o histórico de uma empresa conter todos os dados necessários sobre seus relativos anos. A análise trará um resultado positivo ou negativo a respeito da empresa que foi submetida a avaliação. Ajudando assim, o usuário a ter mais noção do histórico financeiro de determinada empresa. Histórico este muito importante para a avaliação, pois como o foco do trabalho é investimento na bolsa a longo prazo, como sócio.

Outra possibilidade será a possível avaliação da legitimidade das previsões dos especialistas a respeito da bolsa de valores a respeito dos anos já passados. Será possível a comparação das previsões feitas pelos especialistas em finanças com os dados adquiridos através da análise *data mining*.

A conclusão deste projeto trará uma ferramenta capaz de auxiliar pessoas leigas ou com um conhecimento que ainda não é o suficiente para um investimento no mercado de ações com segurança a tomarem uma decisão mais segura e que propiciará um bom negócio a longo prazo.

4.0 - REVISÃO DE LITERATURA

Neste capítulo são apresentados conceitos fundamentais para o desenvolvimento do projeto, é explicado a funcionalidade da mineração de dados, as técnicas de algoritmos para o funcionamento da mesma.

4.1 MINERAÇÃO DE DADOS

Atualmente o conceito de mineração de dados vem sendo utilizado com eficiência no mercado atual. E o resultado dessas aplicações são informações valiosíssimas que podem levar uma empresa a conseguir novos resultados satisfatórios em seu mercado trabalhado ou até mesmo descobrir falhas que podem explicar ou até mesmo resolver problemas recorrentes. (CORTES; PORCARO; LIFSCHITS, 2002)

A mineração de dados pode ser muito útil em situações em que o volume de dados para se avaliar seja muito grande. A partir da aplicação de técnicas de mineração de dados, de forma correta, pode-se obter como resultado informações importantíssimas que podem serem utilizadas estrategicamente pela empresa, ou no caso, também pode descobrir ou solucionar desde problemas de desvios de informações e fluxos de informações errôneos. (CORTES; PORCARO; LIFSCHITS, 2002)

Data Mining como também é conhecido a mineração de dados é um processo de exploração de uma grande quantidade de dados a procura de padrões consistentes e relevantes, podendo ser regras de associação ou sequências temporais, assim detectando relacionamentos entre variáveis alimentadas no banco de dados, e por fim detectando novos subconjuntos de dados. (BERENSTEIN, 2010)

A mineração de dados é um conjunto de regras e técnicas, baseado em redes neurais, que são aplicadas em um algoritmo de aprendizagem. Sendo assim, *data mining* estuda e analisa um conjunto de dados, aprende com eles e evidencia novos conhecimentos através de novos padrões encontrados. Por fim este novo conhecimento é apresentado, podendo ser através de tabelas, regras, árvore de decisões, hipóteses ou grafos. (BERENSTEIN, 2010)

Atualmente a mineração de dados é utilizada tanto no mercado empresarial podendo impulsionar suas áreas, como em pesquisas científicas, podendo auxiliar em diversos estudos. *Data mining* é muito utilizada em áreas como estatísticas,

matemática e a computação, devido sua demanda de análise de grandes quantidades de dados ou informações.

Normalmente mineração de dados se aplica na necessidade de encontrar padrões sobre grandes quantidades de dados brutos, ou seja, o mercado atual é alimentado todos os dias com dados que dizem respeito sobre compras, transações ou quantidades por exemplo. Estes dados analisados a olho nu, são extremamente difíceis de se chegar a alguma conclusão sólida para detectar alguma tendência ou afinidade entre outros dados, devido a sua enorme quantidade analisada.

4.2 TÉCNICAS DE MINERAÇÃO DE DADOS

Nesta seção serão abordadas as sete técnicas de mineração de dados utilizadas para o treinamento de uma base de conhecimento.

4.2.1 CLASSIFICAÇÃO

O método de classificação consiste em examinar um conjunto de dados e atribuir a eles ou a cada um deles uma classe, previamente definida. Estes dados são associados a um conceito ou a uma classe utilizando um processo de discriminação ou caracterização. (CORTES; PORCARO; LIFSCHITZ, 2002)

O processo de discriminação caracteriza-se pelo seu resultado obtido através da atribuição de um valor para um atributo no registro, em função de mais atributos do mesmo. Por exemplo, em uma loja de materiais de construção, pode-se separar os produtos da prateleira por um tipo, como elétrico, de acabamento ou construção bruta. Por sua vez, a sumarização é o atributo de estudo para uma característica de um ou mais atributos. Por exemplo, pode-se caracterizar um lojista pelo seu desenvolvimento anual de vendas, identificando sua receita mensal de vendas por faixas, como baixa, média e alta. (CORTES; PORCARO; LIFSCHITZ, 2002)

Na figura 1 é possível observar uma árvore de classificação, onde cada opção é discriminada em outras opções que por sua vez sofrem a mesma discriminação até não ter mais opções a serem discriminadas, e assim finalizando uma árvore de classificação.

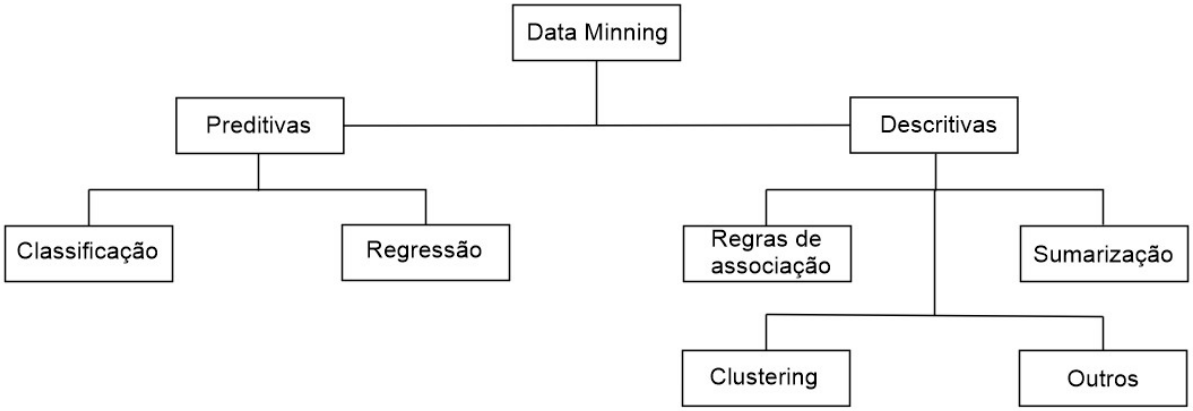

Figura 1: Árvore de Classificação.
FONTE: Faria, 2009.

4.2.2 REGRESSÃO

Segundo Granatyr (2017), o modelo de regressão funciona de maneira muito parecida com a de classificação, porém são analisados números em seu processo de discriminação. Geralmente este método é muito utilizado por bancos, pois utilizam em funções para controle de limites de contas e cartões de clientes.

O método de regressão é muito usado para definir um valor para alguma variável contínua desconhecida, como um saldo de cartão de crédito por exemplo. Ela lida com resultados contínuos, diferente da classificação que lida com resultados discretos. Este método pode ser usado para executar uma tarefa de classificação, convencionando-se que diferentes faixas (intervalos) de valores contínuos correspondem a diferentes classes. (MARANGONI, 2010)

4.2.3 DETECÇÃO DE DESVIOS

Este método visa encontrar dados ou conjunto de dados que não obedecem a um comportamento ou a um modelo de dados previamente estabelecido, ou seja, o método encontra dados que se diferenciam da maioria dos outros dados. Uma vez que encontrados ou separados estes dados, os mesmos podem ser tratados para sua utilização ou até mesmo descartados. (CORTES; PORCARO; LIFSCHITZ, 2002)

O método de detecção de desvios é importante, pois é utilizado para descobrir probabilidades de desvios ou riscos que podem ser tratados conforme as regras traçadas inicialmente na mineração de dados. E é claro, também muito utilizada para o descobrimento de possíveis falhas sobre estes dados. Esta técnica vem sendo

utilizada por empresas com o intuito de descobrir e avaliar métodos de vendas para determinados tipos de produtos, regiões onde são vendidas ou até mesmo a época do ano que são vendidos. Uma vez que aplicado a mineração de dados utilizando esta técnica, é possível detectar um tipo de comportamento dos clientes, e assim traçar um plano mais eficiente de vendas. Observe na figura 2, um padrão de detecção de desvios. (CORTES; PORCARO; LIFSCHITZ, 2002)

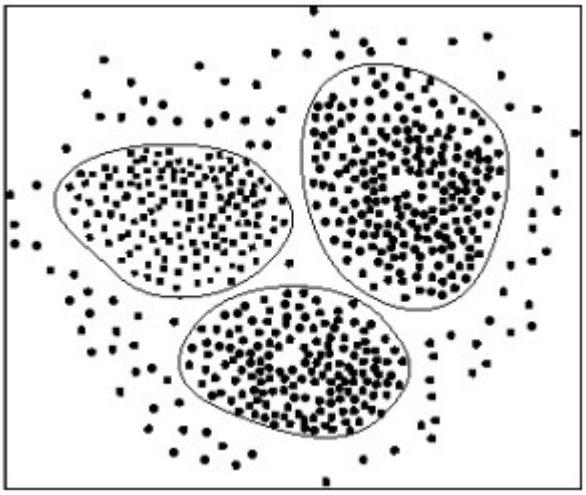

Figura 2: Detecção de Desvios
Fonte: Cortes, Porcaro, Lifschitz, 2002.

4.2.4 REGRAS DE ASSOCIAÇÃO

O padrão de associação é uma das técnicas mais utilizadas no mercado atual. Esta técnica tem por objetivo encontrar relações ou padrões entre dois ou mais dados analisados. A técnica determina que é possível prever um movimento que envolve mais de um dado, observando a movimentação apenas de um. Um exemplo clássico desta técnica é a prateleira do supermercado, onde por exemplo, foi descoberto que no final de semana a probabilidade de vender carne e cerveja era muito alta. Assim, os mercados se preocupavam com estes itens e elaboravam estratégias de vendas envolvendo estes produtos. Os resultados como esperados, foram muito significativos. (AMO, 2004)

Conforme Berenstein (2010) “Há diversos tipos de algoritmos que podem ser utilizados na tarefa de associação, com estruturas e características diversas, mas os utilizados com mais frequência são: Regras de Associação, Teoria dos Conjuntos, Estatísticas e Apriori”.

4.2.5 PADRÕES SEQUÊNCIAIS

Este método funciona de maneira muito parecida com regras de associação, porém, uma vez que que o método de associação procura relações entre dois ou mais itens na mesma transação de tempo, os padrões sequenciais procuram relações entre dois ou mais itens em transações diferentes, ou seja, em mais de uma transação. Isto é, através de uma transação em um determinado período de tempo, é possível prever outra transação em outro período de tempo. Este método aumenta empresas que trabalham com vendas a se planejarem de acordo com seus itens contidos em estoque, pois uma vez que se vende um item que tem associação com outros itens, a porcentagem de vender os outros itens aumenta. (GRANATYR, 2017)

Por exemplo, pessoas que compram o livro 1 de uma determinada saga literária, tendem a comprar o livro 2 e assim por diante. É importante perceber a transação de cada item não ocorre no mesmo tempo, pois é necessário que a pessoa primeiro leia o livro 1 para mais tarde pensar em possuir o livro 2.

Conforme pode-se observar na figura 3, alguns círculos nomeados com uma sequência alfabética de A à E, também observe que à 5 espaços e cada um deles com uma sequência diferente. As letras nos círculos significam uma sequência que vem sendo formada em diferentes espaços de tempos representadas pelos espaços quadrados. Observe como no último espaço de tempo as sequências já se relacionam.

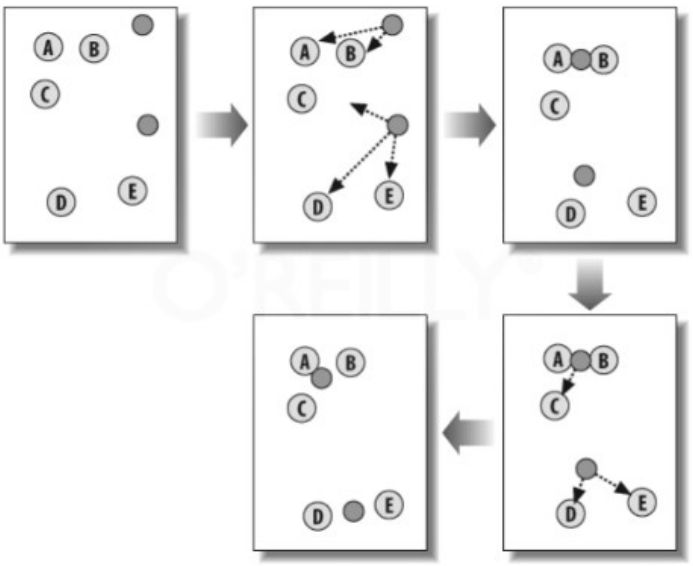

Figura 3: Padrões Sequenciais.

Fonte: Researchgate, 2012

4.2.6 AGRUPAMENTO

Este método tem a finalidade de segmentar um conjunto de dados em grupos homogêneos. O método visa formar grupos baseados no princípio de que os grupos dos dados devem ser mais homogêneos em si e heterogêneos entre si. A principal diferença entre a formação dos grupos é a classificação, que no agrupamento não existem classes predefinidas para a classificação dos dados estudados. Os grupos são formados de acordo com a similaridade de seus dados, ou seja, o método agrupa os dados de acordo com algumas variáveis iguais entre si. Separando os dados em dois ou mais grupos, dependendo do banco de dados avaliado. (CORTES; PORCARO; LIFSCHITZ, 2002)

> A aplicação mais comum do agrupamento é o marketing, e o objetivo é analisar todos os clientes de uma base de dados e agrupar aqueles que possuem características semelhantes, ou seja, que pertençam ao mesmo grupo. Com isso, uma mala direta pode ser enviada somente para as pessoas certas e que tem maiores chances de comprar um determinado produto. O banco Itaú foi um dos pioneiros no Brasil a utilizar essas análises, conseguindo aumentar a taxa de respostas da mala direta de 3% para 30%, fora a enorme diminuição da conta no correio! (GRANATYR, 2017)

Na figura 4 pode-se observar um aglomerado de dados que foram submetidos a esta técnica. Os mesmos passaram pelos processos do método e se agruparam em três grupos como pode ser observado na imagem “c”.

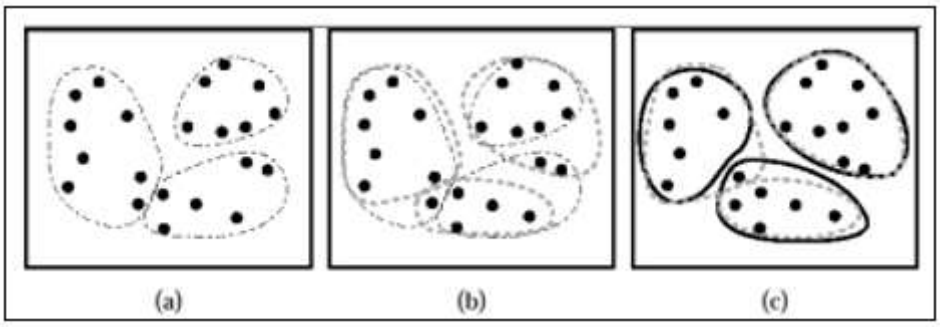

Figura 4: Agrupamento

Fonte: Cortes, Porcaro, Lifschitz, 2002.

4.2.7 SUMARIZAÇÃO

Hosokawa (2011) diz que a descrição compacta de um subconjunto é determinada pela sumarização, exemplos simples são as medidas de posição e variabilidade. Funções mais elaboradas utilizam técnicas de visualização e a determinação de relações funcionais entre variáveis. Granatyr (2017) diz que a sumarização consiste em extrair as características de um grupo já formado.

Existem técnicas de sumarização mais sofisticadas que são chamadas de visualização, nas quais é possível através de diagramas baseados em proporções, dispersão, histogramas, entre outros, obter um entendimento sobre um conjunto de dados que muitas vezes é intuitivo. (HOSOKAWA, 2011)

Um exemplo prático de sumarização se dá por uma aplicação de uma loja que trabalhou seu banco de dados com a técnica. Assim, foi possível considerar dois grupos de clientes que frequentam esta loja. A dos clientes que gastam muito e a dos clientes que gastam pouco. Desta forma, pode-se extrair as informações mais relevantes sobre estes clientes classificados nos dois grupos e mandar propagandas de mala direta mais direcionadas para estas pessoas, por exemplo, contendo produtos mais caros ou mais baratos dependendo das condições do cliente.

4.3 BOLSA DE VALORES

A bolsa de valores é um espaço para negociação entre investidores que podem comprar ou vender seus títulos emitidos por empresas, conhecidos por ações, títulos estes que podem ser capitais públicos, mistos ou privados. Esse processo de negociação de ações é acompanhado e intermediado por corretoras através de correspondentes de negociações. O objetivo principal deste ambiente é organizar negociações confiáveis entre acionistas ou investidores. Assim, a bolsa garante que os investidores recebam as ações compradas de maneira eficiente e segura e que as transações sejam rápidas e práticas. No Brasil as negociações da bolsa acontecem através da BM&FBOVESPA. (BTG, 2017)

O funcionamento da bolsa de valores se dá pela decisão de uma empresa abrir seu capital e disponibilizar ações em troca de verba, ou seja, vender uma porcentagem da empresa para acionistas dispostos a obterem uma porcentagem do capital desta empresa. Desta forma os interessados podem se tornar sócios da empresa comprando uma porcentagem, geralmente não muito alta do negócio. A negociação

de ações da empresa que abre seu capital na bolsa começa pelo processo do IPO (*Initial Public Offering*, em inglês; Oferta Pública Inicial, em português) que permite que as ações começam a ser negociadas, e assim, investidores fazem ofertas de compra e venda. Essa negociação faz parte do mercado primário, que é a primeira negociação da ação, é a primeira vez que uma ação é comercializada, sempre entre empresa titular e acionista. Este processo ajuda a delimitar uma relação de oferta e demanda pelos títulos. (BTG, 2017)

É normal que um acionista queira vender suas ações, pelo simples motivo de precisar de dinheiro ou por achar que o valor da ação desta empresa possa cair e deixá-lo no prejuízo. Assim, o investidor primário vai lançar uma ordem de venda para suas ações, estipulando o valor que pretende obter em troca delas, através de uma corretora. Feito isso, pode haver outro investidor interessado nas ações a serem negociadas que queira comprá-las. Este, através de sua corretora envia uma ordem de compra para a bolsa, no valor que está disposto a investir, e quando a bolsa recebe o mesmo valor em ambas as ordens o negócio é realizado e assim se concretiza uma negociação de ações. Este modelo de negócio configura um mercado secundário, onde as ações são negociadas pela segunda ou mais vezes entre acionistas. (TORORADAR, 2016)

4.3.1 AÇÕES

As ações são papéis que representam uma parte do capital social da empresa. Ao comprá-las, um acionista se torna um sócio, de uma margem da empresa. Desta forma, como sócio, vai correr os riscos do empreendimento em si, ou seja, perder dinheiro caso o valor da ação deprecie, porém também pode lucrar caso o valor da ação valorize. (BTG, 2017)

O valor da ação se dá pelo interesse de compra destas ações. Quando mais investidores interessados na ação, mais valorizada ela se torna para negociação, elevando assim seu valor. Por outro lado, quanto menos interessados, menor o valor da ação. A valorização e a depreciação de ações de uma empresa podem acontecer por vários motivos. Uma empresa em dia com suas contas pode estar com o valor de ação baixo, por motivos muito variados, como por exemplo, a dificuldade de a empresa crescer e se valorizar no mercado. Também, uma empresa que pode estar com uma dívida grande e que até então tinha um valor baixo de ação, pode fechar um contrato de aquisição com uma companhia maior e bem difundida no mercado, fazendo assim

o valor da ação valorizar e chamar atenção de vários investidores, aumentando assim seu valor para a aquisição. (BTG, 2017)

4.4 FORMAS DE INVESTIMENTO EM AÇÕES

Existem formas de fazer investimento em ações negociadas dentro da bolsa de valores. Cada forma determina um modelo de administração e rentabilidade para o acionista que fazer os movimentos das ações. São três os princípios de investimento para se fazer na bolsa de valores. Para iniciantes na bolsa de valores, existe o método de renda fixa, um modelo de negócio com menos riscos. Para acionistas que desejam participar dos dividendos da empresa a longo prazo, aconselha-se que se torne sócio, este é o outro método. E para pessoas que desejam recorrer a lucros a curto prazo, tem-se a opção de *trade*, que é a mais ariscada e consiste em compra e revenda de ações.

4.4.1 INVESTIDOR SÓCIO

O acionista que investe dinheiro em ações para um período de tempo de longo prazo, tem a intenção de fazer parte da empresa como sócio, mesmo que de uma parte ínfima da empresa. Este modelo de investidor, não se preocupa com os *déficits* e *superávits* da empresa em curto períodos de tempo, pois ele aposta na lucratividade a longo prazo. Sua estratégia é aguardar o desenvolvimento e a valorização da empresa, o que geralmente leva um período de tempo mais longo. Assim, ele participa dos lucros anuais da empresa, ou seja, ele recebe o pagamento dos dividendos da ou das sociedades em que ele investiu. O acionista pode optar por aplicar este recebimento dos lucros na empresa novamente, adquirindo novas ações, como também pode sacar seu valor e fazer o uso conforme sua necessidade. (BTGPACTUAL, 2017)

Para fazer um investimento com características de investidor sócio é importante esclarecer alguns pontos. Sempre que se efetua uma compra de ações, indiferente das ações das empresas em que foi aplicado o dinheiro, isto caracteriza o investidor como um sócio, pois o mesmo agora detém de uma parte desta empresa. O que diferencia um investidor sócio de longo prazo de um investidor *trader*, é o fato de que o investidor *trader* compra uma carteira de ações para revendê-la em um curto período de tempo, isto caracteriza um *trade*. Desta forma, investidores que procuram uma

carteira de ações para aplicação a longo prazo, procuram fazer um investimento em empresas bem consolidadas e com um histórico bom de lucros sobre suas receitas, ou seja, um modelo de negócio mais seguro, que caracteriza resultados positivos no decorrer dos anos, e que à tendência de continuar com bons resultados. Empresas assim são o alvo de acionistas que procuram investimentos para acumular capital ou usufruir dos dividendos gerados pelas participações dos lucros anuais de tais empresas. (BTGPACTUAL ,2017)

4.4.2 *TRADE*

O investidor que faz dinheiro com um modelo de investimento *trade* tem como objetivo comprar ações com preços baratos para vender mais caro, ou vende uma ação com um bom preço, esperando que o valor da ação caia para recomprar novamente por um preço mais barato no futuro, se achar que irá valorizar novamente. O modelo de negócio *trade*, tira proveito das oscilações do mercado acionário, ou seja, o investidor faz uma compra ou venda de ações espiculando que em um momento futuro ele possa renegociar estas ações e gerar lucro mais uma vez. Geralmente estas oscilações no preço de ações acontecem em curtos períodos de tempo, dando oportunidade assim para acionistas trabalharem como investidores *traders*. O benefício deste modelo de negócio se dá na liquidez do mercado acionário, isto por que com um volume de investidores pensando em fazer negócio a períodos de tempo curtos, a facilidade de venda ou de compra de ações é maior, pois é provável que haja um investidor afim de fazer negócio. (APRENDAINVESTIMENTO ,2016)

4.4.3 RENDA FIXA

Investimentos de renda fixa acontecem através de títulos que geram um fluxo fixo de ativos, ou que tem seu fluxo configurado por uma fórmula, desde que o emissor deste título não esteja listado como falido. Ou seja, o fluxo da rentabilidade é igual ou muito parecido. (ALMEIDA, 2010)

Para Almeida (2010) ”Existem títulos de renda fixa variando desde riscos de inadimplência quase nulos, como títulos do tesouro, até aqueles com riscos moderados (obrigações de rendimento alto) ”. Em outras palavras, renda fixa funciona de maneira parecida com um empréstimo, ou seja, o valor que será recolhido dos dividendos pelo investidor é definido no momento da aplicação ou no momento do

resgate, no final da aplicação. Basicamente, quando um investidor compra um título de renda fixa, é como se ele estivesse fazendo um empréstimo para o emissor do título, que pode ser o governo, uma empresa ou um banco. E o recolhimento dos dividendos que o investidor recebe, são nada mais que os juros cobrados pelo empréstimo realizado. (INFOMONEY, 2005)

4.5 CARACTERÍSTICAS DE AÇÕES PARA INVESTIMENTO A LONGO PRAZO.

Segundo especialistas na área de ações dos serviços de consultoria do Tororadar.com (2017), empresas com boas oportunidades para investimento a longo prazo carregam características que não podem ser ignoradas em sua avaliação. Uma empresa com ações propícias a valorização em um período longo é muito procurada para compra, assim, seu valor por ação sobe naturalmente, pois tem muitos compradores dispostos a negociar, e quanto mais acionistas procurando por uma ação, maior é o valor de venda.

Por sua vez, o que faz um investidor a procurar ações de uma determinada empresa são suas características administráveis e financeiras. Uma empresa que contém uma receita líquida muito grande indica um grande volume de movimentação de seus produtos, ou seja, tem volume de venda e compra. Se o resultado desta receita for positivo, acaba gerando um caixa disponível, que será usado para pagamentos de contas e mais tarde o que sobrar se torna o lucro líquido. Este lucro líquido é o valor que é dividido com os acionistas nos períodos determinados, é a participação dos lucros. Assim, quanto mais alto o valor de lucro líquido, melhor para os acionistas que recebem os dividendos das ações desta empresa.

Uma empresa também é composta por valores de caixa e dívidas, sendo que o valor de caixa é o montante de dinheiro disponível para uso dos mais variados tipos, sendo para compra, quitação ou aplicação de várias espécies. A dívida bruta é encontrada na soma de todas as dívidas de todas as espécies que a empresa deve prestar contas, ou seja, é a reunião de todos os valores devidos. É muito natural que a empresa utilize do valor em caixa para a quitação destes valores devidos, se não for capaz de quitar todo o valor, pelo menos uma parte é direcionada para extinguir uma porcentagem da dívida. A empresa pode realizar um cálculo de liquidez de suas dívidas, que são nada mais que a dívida bruta menos o caixa, resultando assim em uma dívida líquida. (TORORADAR, 2016)

É interessante para o investidor, que esta dívida líquida esteja abaixo de zero, isto quer dizer que a empresa tem mais dinheiro em caixa do que dívidas, e isto é uma característica muito boa para a prosperidade da empresa, que pode aplicar seus lucro e recursos para a prosperidade da mesma, isto por sua vez chama a atenção de investidores que procuram fazer investimentos para longo prazo, e alavancando o valor das ações como consequência. Um valor próximo de zero também pode ser bom, isto indica que a empresa tem pouca dívida, e que tem o controle do seu setor financeiro, o que também é um bom sinal para os investidores capazes de fazerem investimentos com grandes valores. (BASTTER, 2017)

Investidores que apostam alto em empresas, costumam avaliar algumas variáveis dentro da bolsa de valores das respectivas empresas que deseja fazer investimento. Uma destas variáveis mede a liquidez corrente das dívidas da empresa, ou seja, a capacidade que a empresa tem de quitar suas dívidas. Isto é importante pois gera uma posição que pode vir a ser útil em uma possível negociação da empresa, uma vez que a empresa não está em boas condições de investimento e acaba por gerar outra dívida de valores consideráveis, significa que, pode levar um tempo para a quitação da mesma. (BASTTER, 2017)

Estes indicadores acabam por depreciar o valor da ação se alguns investidores resolverem se desfazer de suas ações. Um fator importante a se avaliar em uma empresa é sua margem de lucro. Uma vez que a empresa tem um grande volume de vendas e compras, consequentemente sua receita líquida é grande. Porém, é viável para a empresa que a margem de lucro desta receita seja positiva. Ou seja, deve haver lucros consistentes para o volume de negociações da empresa. Uma vez que a empresa gere uma receita muito grande e seus lucros não são altos, o custo benefício de investimento em suas ações não se torna boa. O que não chama a atenção de bons investidores. (TORORADAR, 2016)

4.6 MINERAÇÃO DE DADOS APLICADA NA BOLSA DE VALORES

Através de técnicas de mineração de dados foi-se desenvolvido sistemas, pesquisas ou métodos que prestam o devido auxílio a investidores da bolsa de diversificadas formas, como, auxílio em predições de ações, de empresas ou do próprio mercado, como veremos no conteúdo a seguir. Com grandes quantidades de dados e a evolução dos sistemas de informação, hoje é possível chegar a conclusões

mais precisas e com mais facilidade no mercado acionário graças a utilização da mineração de dados.

4.6.1 USO DE REDES NEURAIS ARTIFICIAIS PARA A PREDIÇÃO DA BOLSA DE VALORES

Krieger (2012), desenvolveu uma ferramenta que estabelece previsões de uma ação, em seu projeto ele trabalhou com a predição da ação BVMF3 (BMFBOVESPA) e também testou a ferramenta nas ações GETI4 (AES Tietê) e na PETR4 (Petrobrás) para avaliar a redundância da ferramenta em outros casos. Krieger criou uma rede neural utilizando a ferramenta *Neuroph Studio*, utilizando o algoritmo *BackPropagation* e constatou que a utilização deste método pode ser vantajosa, pois teve como resultado uma margem de acerto de 83% para predição da ação BVMF3.

Em relação á analise da ação GETI4, os resultados não foram satisfatórios, pois a ferramenta não foi capaz de prever uma queda no valor das ações que decorreram de problemas internos da ação, que derivou de uma mudança tributária em relação a empresas de energia elétrica como é o caso da AES Tietê. Na análise da ação PETR4 da Petrobras, constatou-se uma variação muito forte, o que não significou um bom resultado, não sendo satisfatório, porém foram executadas algumas mudanças de testes com a mudança do modelo da tipologia para a *momentum* e se obteve uma melhora significativa, aproximando os resultados do mercado com as previsões do método. (KRIEGER, 2012)

4.6.2 USO DE MINERAÇÃO DE DADOS NA BOLSA DE VALORES

Em 2010, Berenstein realizou um trabalho que trabalha de forma parecida com o citado anteriormente, porém este visa a descoberta ou a predição do comportamento das cotações em seus diversos setores. Para a realização dos métodos o mesmo trabalhou com um banco de dados com mais de 5 mil itens e os aplicou em regras de classificação, associação e agrupamento. Com isto se obteve um resultado não satisfatório com as técnicas de agrupamento e associação. Porém a técnica de classificação trouxe resultados importantes na predição do rumo das cotações ou na descoberta de pontos importantes que podem ser utilizados para a análise e aplicabilidade de renda nos setores indicados.

Berenstein utilizou o algoritmo J48 para a análise de seu banco de dados e em seu resultado final constatou que foi o modelo de análise mais satisfatório, pois obteve mais confiabilidade e significância.

4.6.3 MINERAÇÃO DE DADOS E ANÁLISE DE SENTIMENTOS

Barroso e Branco em 2014 utilizaram técnicas modernas de classificação e predição, e mostraram que é possível chegar a bons resultados, que se bem analisados e utilizados podem resultar em bons lucros no mercado acionário. Estas duas técnicas utilizadas juntamente com a mineração de dados e a análise de sentimento, podem definir qual é a tendência ou reação de uma determinada ação no momento da negociação e entender a influência que uma notícia pode gerar no preço da ação. O *software Stock MiningTec*, utilizado neste projeto, auxilia o investidor no momento de tomar as cabíveis decisões, ajudando-o assim a vender ou comprar determinadas ações que podem ser melhores administradas posteriormente. Isto tudo com o uso do modelo gerado através da mineração de dados e informações da análise de sentimento. (BRANCO; BARROSO, 2014)

Barroso e Branco (2014) diz que “A análise de sentimentos visa entender qual o significado da informação que está sendo utilizada e seu impacto no meio em que está inserida”.

4.6.4 MINERAÇÃO DE TEXTO

A *TheStreet*, um site americano de notícias e serviços financeiros, lançou em 2012 uma notícia dizendo que uma equipe de profissionais vinha utilizando uma técnica de mineração de texto, através de análises textuais. Onde sua principal fonte de dados se dá por jornais, revistas ou o próprio jornalismo tradicional. Enquanto a imprensa ativa lança notícias em seus meios de noticiário todos os dias, existem pessoas aproveitando destas informações e utilizando em mineração de dados para a obtenção de informações que podem se tornar boas oportunidades de negócio no memento de sua aplicação.

Mineração de texto é a análise de dados de obras de linguagem natural (artigos, livros, jornais, entre outros), usando o texto como uma forma de dados. Muitas vezes associado com a mineração de dados. A análise numérica de trabalhos de dados (como arquivamentos e relatórios) é referido como "texto e mineração de

dados". TDM, como também é conhecido envolve o uso de *software* avançado que permite aos computadores lêrem e digerirem informações digitais muito mais rapidamente do que um ser humano pode. O *software* TDM divide informações digitais em dados brutos e texto, analisá-o e apresentando novas conexões (*THESTREET*, 2012)

Em 2008, um relatório do grupo Aite, com sede em Boston, descobriu que a porcentagem de jogadores financeiros extraindo dados não-estruturados, incluindo conteúdo de empresas como Dow Jones e Thomson Reuters, subiu de 2% para 35%, e as despesas foram projetadas para quase dobrar nos próximos dois anos. Isto aconteceu graças a mineração de textos. (*THESTREET*, 2012)

Na tabela 1 é possivel observar um súmario onde é apresentado os trabalhos comentados anteriormente. Pode-se observar na primeira coluna os autores dos respectivos projetos. Na segunda coluna apresenta-se o ano em que foi elaborado os trabalhos. Na terceira coluna observa-se o título do trabalho. E na quarta coluna é observado a abordagem de IA que foi utilizada no decorrer de cada projeto.

Tabela 1: Trabalhos de IA na área de bolsa de valores.

Autor	**Ano**	**Trabalho Realizado**	**Abordagem em IA**
Paulo Eduardo Krieger	2012	Uso de redes neurais artificiais para a predição da bolsa de valores	Aprendizagem com Algoritmo *BackPropagation*
Marcelo Berenstein	2010	Uso de mineração de dados na bolsa de valores	Algoritmo J48 na utilização das técnicas classificação, associação e agrupamento
Gustavo Mendonça do Rio Branco, Marcos André Rosendo Barroso	2014	*Mining StockTec*: Predição de preço de ações através de mineração de dados e análise de sentimentos.	Uso do *software Stock MiningTec* para análise de sentimentos
TheStreet	2012	*How Traders Are Using Text and Data Mining to Beat the Market*	Uso do software TDM para mineração de dados textuais

Fonte: Do autor.

Ao analisar a tabela 1, é notável que a mineração de dados alia-se através de diversas pesquisas de campo com a área do mercado de ações, é possível perceber isso com os títulos dos autores Krieger, Berenstein e o site *TheStreet*. Outro fator

importante a respeito do site *TheStreet* é o uso de mineração de texto, um conceito recente em mineração de dados.

5.0 - MATERIAL E MÉTODOS

No capítulo a seguir, é mostrado o desenvolvimento do projeto na parte prática, como foi realizada a pesquisa para obtenção do conhecimento necessário, para a continuação do presente trabalho, métodos utilizados para a aprendizagem, como foi capturado os dados necessários para montar o banco de dados.

Também é apresentado as ferramentas utilizadas tanto para a o conhecimento que foi agregado como também para a elaboração dos materiais utilizados para a conclusão no término deste trabalho.

5.1 WEKA

O WEKA é um *software* que teve sua origem na Universidade de Waikato (Nova Zelândia) e foi implementado pela primeira vez em sua forma moderna em 1997. O mesmo utiliza a GNU *General Public License* (GPL). O *software* foi escrito em Java e contém uma interface para interagir com arquivos de dados e produzir resultados visuais, como imagens de pontos e curvas. O WEKA dispõe de uma API para a incorporação do *software* com outras bibliotecas. O *software* tem seus próprios aplicativos, que podem ser utilizados para a elaboração de trabalhos ao lado de servidores. Na figura 5, a interface de início do Weka. (IBM, 2010)

Figura 5: Janela de abertura do software Weka.

Fonte: Do autor.

A ferramenta é amplamente utilizada por profissionais que desejam aprender os conceitos básicos sobre mineração de dados, pois atende a diversos níveis de

exigência e sofisticação. Através de sua interface gráfica, conhecida como *WEKA Explorer* é possível elaborar processos de mineração de dados de forma fácil e simples, realizando a avaliação dos resultados obtidos e a comparação de algoritmos. A ferramenta também oferece recursos para a execução de tarefas relacionadas ao pré-processamento de dados como seleção e a transformação de atributos. (PACHIAROTTI, 2012)

5.2 LEVANTAMENTO DE ESTUDOS

Para um aprendizado mais aprofundado sobre o funcionamento e comportamento da bolsa de valores, foi realizado uma pesquisa do tipo exploratória. Estas pesquisas contribuíram diretamente com a elaboração da metodologia necessárias para o levantamento de informações que auxiliam na detecção de pontos importantes para a escolha de empresas com boas ações para fazer investimentos a longo prazo, como sócio. Estas pesquisas aconteceram através de estudos realizados em sites especialistas na área de finanças, tanto em renda variável como em renda fixa. (GIL, 2008)

Os levantamentos das informações aconteceram principalmente nos sites do Bastter.com, no ToroRadar.com.br e no Euqueroinvestir.com, que são sites qualificados que prestam consultoria profissional a respeito da bolsa de valores. Estes conteúdos deram-se através de cursos disponíveis nos mesmos através de *e-books*, vídeos e artigos. Nos sites do ToroRadar.com.br e do Euqueroinvestir.com é possível o contato direto através de telefone ou outros meios de comunicação com um especialista da área. Desta forma, utilizou-se deste meio para o aprendizado de um conhecimento que veio a ser necessário na aplicação prática do projeto.

5.2.1 SITES ESTUDADOS

O site Euqueroinvestir.com trabalha com uma assessoria aos investidores que procuram por uma ajuda ou que não contém o conhecimento necessário para fazer negócios com segurança dentro do mercado de ações. O modelo de negócio deste site é focado de acordo com um perfil de investidor, ou seja, de acordo com a realidade da pessoa afim de investir no mercado de ações. Assim, é recebido uma acessória qualificada de acordo com suas características, que podem ser limitações financeiras, conhecimento do mercado acionista, tipos de investimentos desejados, quantidade ou

tempo que se deseja obter um lucro, e assim por diante. Pensando nisso, o acionista interessado preenche um formulário de acordo com suas características, assim o especialista do Euqueroinvestir.com analisa e oferece a acessória correta para o investidor. (EUQUEROINVESTIR, 2017)

O Bastter.com é um site de assessoria de investimento para pessoas que necessitam de ajuda para a administração de sua carteira de ações ou até mesmo que precise de ajuda para montar uma carteira. O site conta com o conhecimento de alguns especialistas na área de ações. Através destes especialistas um investidor pode obter o auxílio necessário para a melhor administração de seus recursos. O site tem disponível uma série de informações que vem a ser muito útil a investidores que querem acompanhar de perto o desenvolvimento da empresa. No site está disponível uma lista de todas as empresas cadastradas na BM&FBOVESPA, e nestas empresas um histórico das variáveis financeiras desde 2001. É importante deixar claro que nem todas as empresas cadastradas tem dados registrados desde 2001 dentro do Bastter.com. De acordo com a disponibilidade das informações na BOVESPA sobre as empresas, foi-se registrado os dados. O Basttter.com também conta com uma variedade de cursos, vídeos, artigos, livros e aulas particulares para o aprendizado no mercado de ações. (BASTTER, 2017)

O site Tororadar.com.br reúne o cadastro das principais empresas de investimento dentro da BM&FBOVESPA. O site é especializado em assessoria de curto, médio e longo prazo. No Tororadar.com.br é possível assistir a uma série de vídeos que todos juntos formam um curso sobre bolsa de valores, voltado para quem está iniciando na área. Os serviços prestados pelo Tororadar.com.br se dividem em três categorias, sendo elas: *Day-Trade*, curto prazo e longo prazo. (TORORADAR, 2017)

A categoria de *Day-Trade* é possível observar informações relevantes para negociações do tipo que acontecem em menos de um dia, ou seja, a ação é comprada ou vendida no mesmo dia. Na categoria de Curto prazo é possível observar informações que dizem respeito a negócios que acontecem em um período de tempo de um dia a duas semanas. Por fim tem-se a categoria de longo prazo, para negociações entre duas semanas e dois anos, que se divide em outras três categorias, sendo elas: Dividendos, Moderada e Agressiva.

Os serviços disponibilizados pela equipe da Tororadar.com.br são através de análises técnicas, para investimentos de curto e médio prazo, e análises fundamentalistas, para investimentos de longo prazo. Uma vez que se cadastrado

dentro do site, é possível manter contato, via chat, e-mail ou telefone com um especialista da área que lhe presta auxílio sempre que necessário e lhe informa sobre possíveis tendências de sua carteira de ações. (TORORADAR, 2017)

5.3 ANÁLISE FUNDAMENTALISTA

A análise fundamentalista procura observar a saúde financeira em que uma empresa se encontra, com o intuito de projetar um valor ou significância aos seus parâmetros para ajudar a estabelecer a confiança ou valor das ações da empresa. (TORORADAR, 2017)

Para o levantamento de tais valores, analistas levam em consideração os fundamentos da empresa, que são fatores macro e microeconômicos que impactam no desempenho das empresas dentro da bolsa de valores. Levando em consideração análises minuciosas de vários parâmetros da companhia, pode-se estabelecer uma situação para a mesma, com possíveis resultados a longo prazo, geralmente em um período de tempo de cinzo a dez anos. A análise fundamentalista é uma observação detalhada sobre o momento da empresa que pode ajudar analistas ou estatísticos a projetar uma opinião sobre os resultados da companhia no futuro. (EXAME, 2013)

Uma análise fundamentalista avalia fatores quantitativos e qualitativos. Ao se avaliar aspectos quantitativos o analista leva em consideração valores dos tipos numerais, como por exemplo inflação, taxas de juros, valores de caixa, câmbio, entre outros. E ao se avaliar aspectos qualitativos, é observado fatores administrativos, como gerência, seus controladores, composição do conselho administrativo e decisões governamentais por exemplo. (EXAME, 2013)

Com isto, percebe-se como a análise fundamentalista pode impactar e ajudar um acionista no mercado de ações a estipular o valor de suas ações como também administrá-las com um grau de segurança, procurando sempre o melhor momento para vender sua carteira de ações como também aumentá-la através da percepção de bons momentos para fazer negócio.

5.4 VARIÁVEIS EXTRAIDAS PARA AVALIAÇÃO

Para a melhor compreensão do funcionamento dos valores de uma empresa, avaliou-se o significado de cada variável que constitui as informações de cadastro, os valores por ação, o lucro e a geração de caixa da empresa, o caixa e a dívida da

organização, a liquidez e a solvência da empresa, e o fluxo de caixa da mesma. Todas estas informações foram retiradas de uma conta *premium* do site de assistência de renda variável, Bastter.com. (BASTTER, 2017)

Todas estas variáveis representam informações para análise de uma empresa, desde sua geração de receita até o pagamento dos dividendos, utilizadas então para meios que justifiquem uma análise fundamentalista, sendo esclarecedor e mais fácil de entender o fluxo administrativo dos valores trabalhados no decorrer das etapas que empresa passa para manter-se no mercado. As variáveis representadas nas tabelas 2 a 7, são retiradas do site do Bastter.com.

5.4.1 DADOS DE CADASTRO

Com as variáveis levantadas foi possível entender melhor o fluxo das informações e da administração do dinheiro que a empresa controla. Na tabela 2, as informações sobre os cadastros das empresas, dados que ditam informações concretas sobre as mesmas.

Tabela 2: Variáveis de dados de cadastro.

DADOS DE CADASTRO	
Variável	**Descrição**
Empresa	Nome da empresa.
Segmento	Área de atuação no mercado da empresa.
Valor de Mercado	Valor da empresa no mercado.
EV	Enterprise Value = Valor de mercado + Dívida bruta – Caixa.
Majoritário	Proprietário da maior porcentagem da empresa.
Pessoas físicas	Total de acionistas das ações da empresa.
Categoria	Classificação evolutiva que a empresa se encontra atualmente no mercado.

Fonte: Do autor

5.4.2 VARIÁVEIS DA AÇÃO

As variáveis de ação caracterizam-se por mostrar os movimentos financeiros que acontecem com as ações das empresas. Estes movimentos estão relacionados a valor, oscilação, lucro e significância em relação ao patrimônio da empresa como um todo. Verifique na tabela 3.

Tabela 3: Variáveis de ação

VARIÁVEIS DA AÇÃO	
Variável	**Descrição**
LPA	Lucro líquido por ação.
LPA Descontado	Lucro líquido (descontado dos não recorrentes) por ação.
VPA	Valor Patrimonial por ação.
P/L	Preço da ação dividido pelo lucro líquido por ação.
P/L Descontado	Preço da ação dividido pelo lucro líquido (descontado dos não recorrentes) por ação.
EV/EBITDA	Utilizado para trade de valor assim como o P/L. Teoricamente quanto menor o valor, mais atraente estaria a compra da ação.
P/VPA	Preço da ação dividido pelo valor patrimonial por ação.
DPA	Dividendos por ação distribuídos aos acionistas nos últimos 12 meses.
Dividend Yield	Dividendos distribuídos nos últimos 12 meses dividido pelo preço da ação.
Payout	Taxa de distribuição do lucro da empresa para os acionistas na forma. De dividendos ou juros sobre o capital próprio. DPA dividido pelo LPA
Margem de Segurança	Diferença positiva entre o potencial de ganho na ação e a taxa de juros praticada pelo mercado ou a perspectiva de ganho investido em títulos do governo.

Fonte: Do autor.

5.4.3 LUCRO E GERAÇÃO DE CAIXA

Na tabela 4, pode-se observar as variáveis que juntas caracterizam os movimentos de geração de caixa e seus respectivos lucros. Estes valores configuram as finanças relacionadas aos valores de caixa, como liquidez, receitas e margens brutas que impactam no valor patrimonial ou nos movimentos financeiros da empresa.

Tabela 4: Variáveis de lucro e geração de caixa

LUCRO E GERAÇÃO DE CAIXA	
Variável	**Descrição**
Receita Líquida	São as vendas da empresa diminuída de: Devoluções e vendas canceladas, descontos concedidos incondicionalmente e impostos e contribuições sobre as vendas.
Lucro Líquido	A partir da receita líquida se diminui os custos e as despesas das vendas para se chegar ao lucro líquido.
Lucro Líquido x NR	A partir da receita líquida se diminui os custos e as despesas das vendas para se chegar ao lucro líquido e calcula os “não recorrentes”.
Resultado Bruto	É a receita líquida menos o custo dos bens e serviços vendidos.
Margem Bruta	A margem bruta é o resultado bruto dividido pela receita líquida.
EBIT	É o lucro operacional, obtido nas demonstrações de resultados das empresas.
D&A	Depreciação e amortização.

EBITDA	Indica propriamente o quanto a empresa gera de caixa das suas atividades operacionais.
Margem EBITDA	É igual a EBITDA dividida pela receita líquida. Significa o indicador da margem operacional de uma empresa.
Resultado Financeiro	É igual a receitas financeiras – despesas financeiras. Assim, o resultado final das operações financeiras da empresa, o quanto ela lucrou com aplicações financeiras menos o quanto ela pagou de juros de dividas e/ou perdeu em aplicações financeiras.
Margem Líquida	É o lucro líquido dividido pela receita líquida. Mede a fração de cada real de vendas que resultou em lucro.
ROE	É o lucro líquido dividido pelo patrimônio líquido. Mede a rentabilidade de uma empresa.
ROA	É o retorno sobre o ativo (EBIT/Ativo total). Mede a eficiência operacional em gerar lucros a partir dos ativos da empresa, antes do efeito financeiro.
SSS	Significa "*Same Store Sales*" e são as vendas realizadas na mesma base de lojas do ano anterior.
Patrimônio Líquido	Representa os valores que os sócios ou acionistas tem na empresa em um determinado momento. É a diferença entre o total dos ativos e dos passivos.
RIF	Significa "Receita de Intermediação financeira" e tem toda a receita obtida com as operações de intermediação financeira do banco no período.
Margem Bancaria	É o lucro líquido / Receita de intermediação financeira. Quanto maior o resultado teoricamente melhor para o banco.
Índice de Eficiência	É o indicador das despesas operacionais sobre as receitas.
Índice de Basileia	Quanto mais sobra, mais o banco pode emprestar com segurança. Um índice ruim é menor que 11, ou índice bom fica entre 11 e 14 e um índice ótimo é mais que 14.
PDD	Significa "Provisão para devedores Duvidosos" que é feita pelo banco para se proteger de casos de inadimplência.
PDD/LL	Mede a qualidade da carteira de crédito. Quanto menor melhor. Abaixo de 0,75 é bom, acima de 1 é ruim.
Equity Multiplier	É o (Ativo total / Patrimônio líquido). Quanto maior, mais alavancada está a empresa. Demonstra quantos reais a empresa está operando para cada real de dinheiro do acionista.

Fonte: Do autor.

5.4.4 CAIXA E DÍVIDA

Na tabela 5 estão dispostas as variáveis que caracterizam as dívidas da empresa e os valores em caixa da mesma. Estas variáveis estão relacionadas aos

valores que controlam ou impactam de alguma forma na geração de caixa ou dívidas que a empresa possui.

Tabela 5: Variáveis de Caixa de Dívida.

CAIXA E DÍVIDA	
Variável	**Descrição**
Patrimônio Líquido	Representa os valores que os sócios ou acionistas tem na empresa em um determinado momento. É a diferença entre o total dos ativos e dos passivos.
Caixa	Representado pelas reservas financeiras disponíveis para a empresa que podem ser acessadas imediatamente. Compostas de: Dinheiro em caixa, aplicações financeiras de curto prazo, títulos e valores imobiliários de curto prazo.
Dívida Bruta	Representada pelos empréstimos e financiamentos bancários e debêntures emitidas ainda válidas.
Dívida Líquida	Dívida Líquida é igual a dívida bruta – caixa. Demonstra a dívida da empresa diminuindo o caixa.
Dívida Bruta/PL	É a dívida bruta dividida pelo patrimônio líquido. Uma das formas de avaliar o endividamento de uma empresa.
Dívida Líquida/EBITDA	Como padrão se tolera até 3, mas isso tem de ser visto de empresa para empresa. P = prejuízo.
Índice de Cobertura	Índice de cobertura representa: (EBIT / Despesas com juros). É uma forma de analisar quanto os juros estão pressionando o operacional da empresa.
Despesas com Juros	Quanto a empresa gasta com o pagamento de juros de empréstimos, financiamentos e debêntures.
EF	Significa “Endividamento financeiro” e se dá pela fórmula: (Dívida Bruta / (Dívida Bruta + Patrimônio líquido)).
ECP	Significa “Endividamento de curto prazo” e funciona através do cálculo (dívidas de CP / dívida bruta).
Custo % da dívida	É igual a (Despesas com juros / dívida bruta). São os juros pagos no período divididos pela dívida bruta.
EM	Significa “*Equity Multiplier*” e é igual a (Ativo total / Patrimônio líquido). Quanto maior, mais alavancada está a empresa. Demonstra quantos reais a empresa está operando para cada real do dinheiro do acionista.
IPL	Significa “Imobilização do Patrimônio líquido” ((Imob+Inv+Intang)/PL). Quanto maior o índice, mais a empresa está investindo no ativo permanente e menos sobra para o Ativo Circulante, aumentando a dependência de capitais de terceiros.

Fonte: Do autor.

5.4.5 LIQUIDEZ E SOLVÊNCIA

Na tabela 6 é possível observar as variáveis que dão destaque aos movimentos que geram a liquidez da empresa em seus negócios. Estas variáveis impactam nos

valores financeiros relacionados a atribuições de recebimento ou dívidas de uma empresa, com um determinado tempo para serem cumpridas, como depósitos bancários e financiamentos.

Tabela 6: Variáveis de Liquidez e Solvência.

LIQUIDEZ E SOLVÊNCIA	
Variável	**Descrição**
Ativo Circulante	São considerados ativos circulantes: dinheiro em caixa, conta movimento em banco, aplicações financeiras, contas a receber, estoques, despesas antecipadas, numerário em caixa, depósito bancário, mercadorias, matérias-primas e títulos.
Passivo Circulante	São as obrigações que normalmente são pagas dentro de um ano: contas a pagar, dívidas com fornecedores de mercadorias ou matéria-prima, impostos a recolher, empréstimos bancários com vencimento nos próximos 360 dias.
Liquidez Corrente	Liquidez corrente é igual ao Ativo circulante dividido pelo passivo circulante.
Liquidez Imediata	Liquidez imediata é igual a caixa dividido pelo passivo circulante. Que é a capacidade da empresa de honrar seus compromissos de curto prazo.
Capital de Giro	O capital de giro é igual ao Ativo – Passivo Circulante.

Fonte: Do autor.

5.4.6 FLUXO DE CAIXA

Aqui ficam destacadas as variáveis que caracterizam o fluxo do caixa da empresa. Basicamente, são as variáveis que explicam como o caixa recebe movimentos e como estes funcionam, o fluxo de caixa são as operações que vem a movimentar valores dentro do caixa da empresa. Observe na tabela 7, as descrições.

Tabela 7: Variáveis de Fluxo de caixa.

FLUXO DE CAIXA	
Variável	**Descrição**
FCO	Significa Fluxo de Caixa Operacional. É o caixa gerado nas operações da empresa menos as despesas e gastos decorrentes da industrialização, comercialização ou prestação de serviços da empresa.
FCI	Significa Fluxo de Caixa de Investimentos. São os gastos em investimentos, no imobilizado e no intangível, e aplicações financeiras. Entradas por venda de ativos e resgate de aplicações financeiras.
FCF	Fluxo de Caixa de Financiamentos. Entradas com empréstimos e financiamentos de curto prazo. Saídas com pagamentos destas dívidas e pagamentos aos acionistas de dividendos e distribuição de lucros.

FCT	Fluxo de Caixa Total. Representa a efetiva entrada e saída de recursos do caixa da empresa no período determinado. FCT = FCO + FCI + FCF.
FCL	Significa Fluxo de Caixa Livre. Saúde financeira da empresa. De forma bastante prática, se sobra dinheiro das atividades operacionais, descontados os investimentos. FCL = FCO + FCI.
FCI/LL	É o Fluxo de Caixa de Investimentos dividido pelo Lucro Líquido.
CAPEX	Investimentos realizados em equipamentos e instalações de forma a manter a produção de um produto ou serviço ou manter em funcionamento o negócio. CAPEX = Inv no Intangível + Inv no mobilizado.
FCL CAPEX	Quantidade de dinheiro gerado pela empresa após descontos de investimentos utilizados para expandir seus ativos. FCL CAPEX = FCO + Inv Intangível + Inv Imobilizado.
CAPEX/LL	É o CAPEX dividido pelo Lucro Líquido.
CAPEX/FCO	Esse indicador informa quanto a empresa investiu em sua manutenção e expansão (Capex) se comparado com a sai geração de caixa de através das atividades operacionais.

Fonte: Do autor.

5.5 MONTAGEM DA BASE DE DADOS

Para a montagem do banco de dados, primeiramente procurou-se uma fonte de onde se pudesse extrair a maior quantidade de informações possíveis a respeito das empresas cadastradas na BM&FBOVESPA. Para isto foi adquirida uma conta no site do Bastter.com, para que se tivesse acesso total as informações dispostas do grande volume de empresas que o site contém, empresas essas, todas cadastradas na bolsa de valores.

O próximo passo foi copiar o histórico de valores da empresa em um período de oito anos, para que mais tarde fosse gerado uma média deste histórico. Salvo as empresas que não tinham os oito anos de informações disponíveis, desta forma, foi-se usado as informações que estavam disponíveis, vale informar, que foram poucas as empresas com estas condições. Todas estas informações foram compiladas em um arquivo Excel. Dentro do Excel as informações foram separadas de acordo com a letra inicial dos nomes de suas respectivas empresas, desta forma era mais fácil a identificação e gerenciamento. Observe o exemplo na tabela 8.

Tabela 8: Fragmento de tabela de Excel com dados recolhidos.

Empresa	Segmento	Valor de M.
aes elpa s.a	energia eletrica	1334627184
aiet aes tiete	energia eletrica	15190603555
aflu afluente	energia eletrica	94653192
aflu afluente t	energia eletrica	281988609

Fonte: Do autor.

É importante ressaltar que estavam separadas as variáveis que juntas formavam o conjunto de informações de cada empresa.

O site do Bastter.com disponibiliza informações do tipo cadastrais, que continham informações de características das empresas, uma tabela com o nome de “Múltiplos” que continham as variáveis responsáveis por mostrar as ações da empresa. Gerou-se uma tabela para cada tipo de variável, que foi apresentada nos tópicos 5.3.1 a 5.3.6. As variáveis foram separadas dentro do Excel de acordo com suas tabelas no momento de retirá-las do site do Bastter.com. Na figura 6, é possível observar como estavam organizadas.

	2017	2016	2015	2014	2013	2012	2011	2010
FCO	36.987	21.635	15.720	27.791	32.235	33.082	42.062	35.376
FCI	-13.326	-16.639	-20.117	-22.359	-23.150	-30.094	-24.703	-31.585
FCF	-18.363	-2.919	5.250	-8.634	-9.397	2.240	-23.605	-3.474
FCT	5.298	2.077	853	-3.202	-312	5.228	-6.246	317
FCL	23.661	4.996	-4.397	5.432	9.085	2.988	17.359	3.791
FCO/EBITDA	78%	58%	-107%	104%	76%	142%	71%	77%
CAPEX	-14.004	-17.191	-27.784	-26.916	-29.332	-31.962	-29.008	-23.666
FCL CAPEX	22.983	4.444	-12.064	875	2.903	1.120	13.054	11.710
CAPEX/FCO	38%	79%	177%	97%	91%	97%	69%	67%

Figura 6: Tabela de Fluxo de Caixa do site do Bastter.com

Fonte: Do autor.

A partir da compilação dos dados no Excel, realizou-se uma média a respeito das mesmas, que continha em analisar os 8 anos de valores de cada variável da empresa e retirar uma média dos respectivos valores, pois assim foi possível construir os atributos necessários para mais tarde se usar na mineração de dados.

Cada um dos valores foi retirado anualmente, logo a média era subtraída de 8 valores respectivos dos 8 anos analisados, pois desta forma conseguiu-se volume e consistência de dados para ser analisados. Este processo se repetiu em cada uma

das variáveis, exceto para os valores da tabela de Liquidez e Solvência que contem valores únicos. Na tabela 9, um exemplo dos dados já trabalhados após a compilação no Excel.

Tabela 9: Fragmento de tabela com dados com médias prontas.

LPA	LPA desconctado	VPA	P/L
0,57	0,56	36,03	24,71
0,18	0,18	0,8	14,92
0,16	0,15	2,07	31,55
0,21	0,2	1,23	21,69
0,6	0,6	10,42	5,65

Fonte: Do autor.

Com as médias retiradas de todas as variáveis de todas as empresas construiu-se uma única tabela, em um formato que mostrava várias colunas, cada uma para sua respectiva variável e uma linha para cada respectiva empresa e, foram separadas as empresas de acordo com a letra inicial do seu nome. Foram criadas tabelas iguais para cada letra do alfabeto em diferentes arquivos do Excel. Assim foram preenchidas as tabelas com as empresas e suas respectivas variáveis em seus devidos arquivos, que continham suas iniciais. O próximo passo foi classificar as empresas com posições entre boas, médias ou ruins para investimento a longo prazo.

5.5.1 CLASSIFICAÇÃO DAS EMPRESAS PARA INVESTIMENTO A LONGO PRAZO

Todas as empresas depois de tabeladas receberam uma classificação que deriva entre boas, médias ou ruins para investimento a longo prazo. Estas posições foram retiradas das análises dos especialistas em assessoria do site do ToroRadar.com e no aplicativo Dinheiro do Windows. Esta classificação serve para o treinamento da base de conhecimento, para que mais tarde seja utilizado nos testes de mineração de dados.

Entende-se que uma empresa com a classificação “boa”, é uma empresa que demonstra valores ascendentes ou contínuos em relação aos seus lucros e retornos para os acionistas. Uma empresa com a classificação “médio” tem uma ligeira situação positiva de seus lucros e retornos patrimoniais, ou uma certa consistência em relação bons resultados mantidos no decorrer dos anos, porém não tão bons os resultados como a classificação “boa”. E a classificação “ruim” se dá pela decadência aos valores

patrimoniais, lucros ou retornos a seus investimentos, assim como também resultados que estão consistentes sem muita variância, porém resultados ruins.

Com isto obteve-se um total de 99 empresas com a qualificação “boa”, 89 empresas com a qualificação “média” e um total de 189 empresas com a qualificação “ruim”

Uma parte destas classificações foram retiradas do site da ToroRadar.com, onde especialistas em finanças determinam a sua classificação, que é realizada através de cálculos estatísticos, é importante esclarecer que apenas as empresas mais relevantes para a ToroRadar.com estão dispostas no site, sendo assim, não são todas as empresas da BM&FBOVESPA que estão presentes.

Outra parte foi retirada do aplicativo “Dinheiro”, existente no Windows 8 e no Windows 10. Este aplicativo, mostra através de um gráfico a evolução das cotações da empresa. Este gráfico pode ser configurado para períodos, de um dia, uma semana, um mês, um ano, cinco anos ou para todo o período que o aplicativo disponibiliza. Desta forma é simples observar o quadro da empresa e chegar uma conclusão para a classificação das empresas. O aplicativo “Dinheiro” também disponibiliza outras informações a respeito das empresas e suas ações, como rendas, variáveis de suas finanças e notícias relevantes sobre a empresa.

5.6 ARQUIVO ARFF

O *software* do WEKA, para realização da leitura dos dados submetidos aos seus algoritmos, não aceita qualquer extensão de arquivo para o recebimento destes mesmos dados. Desta forma, é necessário que o arquivo que é executo pelo WEKA, tenha um formato de arquivo .arff. Desta forma, é importante entender que, para a elaboração de um arquivo .arff, é necessário primeiramente organizar os dados que farão parte do arquivo, em outro local, sendo muito comum e usado uma tabela de Excel.

Uma vez que se tenha os dados todos organizados em algum local, é preciso entender o formato de um arquivo com extensão .arff. Basicamente o arquivo é composto por três partes, sendo elas: *@relation*, que é onde se nomeia a relação dos arquivos que serão construídos em seu corpo. A seguir encontra-se a etapa de nomear e definir as extensões dos atributos, a etapa *@attribute*. Aqui um atributo que é composto por dados numéricos variáveis, é configurado com a extensão “REAL”.

Atributos que são compostos por dados alfabéticos ou dados numéricos pré-definidos, ganham em sua configuração a extensão “NOMINAL”.

Para cada atributo novo definido no corpo do arquivo .arff, deve-se criar um novo indicador *“@attribute”,* diferente do indicador *@relation,* que aparece apenas uma vez. Por fim, é preciso definir as instâncias que farão parte do arquivo. Para indicar o início das instâncias, é necessário primeiramente definir o indicador *@data,* feito isso, o processo segue de maneira simples, colocando os valores dos atributos na mesma ordem em que os mesmos foram definidos anteriormente.

Logo se existe cinco atributos definidos, deverá de haver em cada linha de instância, cinco valores separados por vírgulas. Se alguma instância não tiver valor a ser preenchido em algum atributo, este mesmo deve ser preenchido com o caractere “?”. Para ficar claro o que é uma instância, é preciso entender que uma tabela por exemplo, contém cinco colunas, onde cada uma delas corresponde a um atributo, e contém 10 linhas, onde cada linha representa uma instância contendo todos os atributos. Imagina-se que em uma farmácia, todo cliente que realiza compra pela primeira vez, deve fazer seu cadastro, e neste cadastro se tem alguns dados que são obrigatórios o preenchimento. Assim, com o tempo a farmácia irá ter uma tabela com vários clientes “instâncias”, e cada cliente com seus dados “atributos” preenchidos. Visualize no texto a seguir um exemplo de arquivo com extensão .arff. (WEKA, 2017)

```
@relation pima_diabetes

@attribute preg REAL
@attribute plas REAL
@attribute pres REAL
@attribute skin REAL
@attribute insu REAL
@attribute mass REAL
@attribute pedi REAL
@attribute age REAL
@attribute class{tested_negative, tested_positive}

@data
6,148,72,35,0,33.6,0.627,50,tested_positive
1,85,66,29,0,26.6,0.351,31,tested_negative
8,183,64,0,0,23.3,0.672,32,tested_positive
1,89,66,23,94,28.1,0.167,21,tested_negative
0,137,40,35,168,43.1,2.288,33,tested_positive
```

Código-Fonte 1: Exemplo de texto de arquivo com extensão .arff.
Fonte: Do autor.

Uma vez que é elaborada e organizada a base de dados, é necessário processá-la em técnicas de mineração de dados, que são processadas no WEKA.

No processo para a elaboração do arquivo .arff, escolheu-se os atributos que teriam impacto no momento de avaliação dos algoritmos de mineração, ficando fora do arquivo .arff apenas os atributos de (empresa e majoritário), pois os mesmos não são relevantes a o treinamento da base de dados devido aos seus dados, que são informações únicas.

Sendo assim, os atributos foram divididos em dois tipos de extensões, atributos nominais e reais. Os atributos nominais são: segmento, categoria e situação. O atributo de situação é o que corresponde aos resultados, (boa, médio e ruim), que foram as classificações retiradas dos sites anteriormente elencados para a obtenção da informação para o treinamento da base. Os atributos reais por sua vez são: Valor_de_M, EV, Pes._Fisicas, LPA, LPA_desconctado, VPA, P/L, P/L_descontado, EV/EBITDA, P/VPA, DPA, Dividend_Yield, Payout, Margem_Segurança, Rec._Liquida, Luc._Liquido, Luc._Liq_*_NR, Resultado_Bruto, Margem_Bruta, EBIT, D&A, EBITDA, Margem_EBITDA, Res._Financeiro, Margem_Liquida, ROE, ROA, SSS, RIF, Margem_Bancaria, Indc._Eficiencia, Indc._Basileia, PDD, PDD/LL, Equity_Multi, Patri._Liquido, Caixa, Divida_bruta, Divida_Liquida, Divida_Bruta/PL, Div_Liquida/EBITDA, Indice_de_Cobertura, Despesas_com_juros, EF, ECP, Custo_da_divida, EM, IPL, At._Circulante, Pas._Circulante, CaixaLS, Liq._Corrente, Liq._Imediata, Capital_de_giro, FCO, FCI, FCF, FCT, FCL, FCI/LL, CAPEX, FLC_CAPEX, CAPEX/LL, CAPEX/FCO. Estes valores são reais devido as suas características da informação, por serem dados numéricos. Todas as variáveis aqui citadas serão utilizadas nos métodos de classificação dentro do *software* do WEKA, para fins de uma análise fundamentalista que retornará os resultados desejados.

O arquivo .arff foi criado a partir de um aplicativo em Java (ConversorXls2Arff) que transforma uma tabela do Excel com a extensão .xls em um arquivo .arff, depois de processada pelo mesmo. A definição das extensões dos atributos são feitas no aplicativo, desta forma apenas se escolheu a extensão para cada atributo e executou-se para a criação do arquivo .arff.

5.7 TREINAMENTO DA BASE DE DADOS

Para as próximas etapas do projeto, é necessário entender como o *software* do WEKA, através de técnicas de mineração de dados, executa seus algoritmos e retorna os valores em porcentagem sobre os acertos em relação ao treinamento da base de dados. Primeiramente, escolhe-se o algoritmo a ser executado e os métodos de

classificação para a avaliação da base de dados. Em seguida, é preciso configurar os métodos de testes, neste trabalho foi utilizado o método de teste *cross-valitation*, devido ao seu aceite na comunidade científica.

Assim, uma vez que selecionado o algoritmo desejado e configurado o método de teste, que neste caso é o *cross-valitation* com configuração de *10 folds*, que por sua vez é a quantidade de vezes que as instâncias contidas no arquivo .arff serão repartidas para o trabalho de treinamento e testes da base de dados.

5.7.1 *CROSS-VALIDATION (10 FOLDS)*

Para a obtenção dos resultados das execuções realizadas pelo WEKA, com os algoritmos definidos, é necessário separar a fase de treinamento da base de dados em duas partes, a parte de “aprendizagem” e a de “teste”. Entende-se que quando uma base de dados, independente do algoritmo, é submetida a um processo de mineração de dados com o método de *cross-valitation*, inicialmente a base é fracionada em dez partes, já que anteriormente foi dito *10 folds*, sendo que nove delas serão usadas no processo de aprendizagem e uma será usada no processo de teste.

Desta forma, ao se iniciar o processo de treinamento da base de dados, primeiramente serão processadas as partes do 1 ao 9, e isto servirá para o algoritmo que está sendo executado encontre e aprenda como os atributos podem interagir entre si e destacar informações que possam ser utilizadas para se combinarem e aumentar a taxa de acerto de novas instâncias que podem ser usadas mais tarde novamente. Terminado este processo de aprendizagem, inicia-se a execução da décima parte da base de dados. Que será utilizada para teste do aprendizado adquirido da primeira à nona parte.

Assim, levando em consideração a quantidade de vezes que houve acertos da base testada em relação ás outras nove treinadas, levanta-se uma porcentagem, a qual é utilizada para a análise da eficácia do algoritmo em relação e aquela base de dados com suas configurações. Cada instância da base de dados contém um atributo que está preenchido com seu resultado final, por exemplo, se uma quantidade de instâncias pode retornar os resultados *dia* e *noite.*

É necessário que todas as instâncias estejam preenchidas com informações dos atributos, caso não haja valor para um atributo, o mesmo é substituído pelo caractere “?”. Mais tarde, com a execução da base de dados à algum algoritmo, estes

resultados serão comparados entre as partes de aprendizagem e a parte de teste, elaborando assim uma média através dos acertos. (DAMASCENO, 2017)

5.8 AVALIAÇÃO DE ALGORITMOS COM ZEROR

Para Sayad (2010) ZeroR é um método de classificação muito simples que não utiliza nenhum preditor, depéndendo apenas do alvo trabalhado. O classificador ZeroR simplesmente prediz a categoria maioritária, ou seja, o algoritmo faz a soma de cada possivel resultado e produz uma média de acordo com o valor do resultado mais presente. Por exemplo, se o uma base de dados tem 100 instâncias, com 3 possiveis resultados, sendo que o primeiro resultado seja a classificação *manhã*, o segundo seja *tarde* e o terceiro seja *noite*, e que das 100 instâncias, 25 delas correspondem ao resultado *manhã*, 30 delas correspondem ao resultado *tarde*, e 45 correspondem ao resultado *noite*. Desta forma, o resultado final da execução do algoritmo ZeroR será de 45%, pois corresponde ao valor que mais somou resultado, neste caso o resultado *noite*.

Embora não exista um poder de previsibilidade no ZeroR, é muito útil e utilizado para determinar um desempenho média mínima aceitavel como referência para a utilização de outros métodos de classificação. (SAYAD, 2010)

Após a finalização do arquivo .arff, o mesmo foi submetido ao algoritmo ZeroR para a obtenção da média mínima aceitável. Com isto, obteve-se o resultado de 50,13%, que é a menor média que se tem de ser acançada e melhorada. Que se deu pelo maior número de empresas com o atributo situação caracterizada como ruim. Sendo que existem na base de dados 99 empresas com o atributo situação como boa, 89 empresas com o atributo situação médio e 189 empresas com o atributo situação ruim.

5.8.1 SELEÇÃO DE ALGORITMOS

Com a obtenção da média mínima necessária para o treinamento da base de dados iniciou-se a fase de submissão aos algoritmos, primeiramente testou-se todos os algoritmos disponíveis no WEKA. A base de dados foi submetida a cada um dos algoritmos com testes realizados de 1 a 5 *Random Seeds*, que por sua vez é a quantidade de vezes que se randomiza as partes da base de dados para treinamento. Todos os testes foram realizados com a técnica de *cross-validation* com 10 *folds*. Foi

realizado 5 testes para a elaboração de uma média que ajudaria a verificar a consistência dos resultados para posteriormente escolher-se os melhores algoritmos para colher os resultados. Na tabela 10 é possível observar os algoritmos e suas respectivas médias.

Tabela 10: Algoritmos com suas médias de 5 testes.

Algoritmo	Média	Algoritmo	Média
NaiveBayes	41,11	CVParameterSelection	50,13
J48	59,99	FilteredClassifier	59,84
OneR	62,7	IterativeClassifierOptimizer	63,55
JRip	62,06	LogitBoost	65,09
IBK3	29,06	MultiClassClassifier	45,46
LibSVM	50,13	MultiClassClassifierUpdateable	56,54
MultilayerPerceptron	58,35	MultiScheme	50,13
BayesNet	59,62	RandomCommittee	62,01
BayesNetMultinominalText	50,13	RandomizadeFilteredClassifier	54,85
NaiveBayesUpdateable	41,11	RandomSubSpace	58,35
Logistic	45,3	Stacking	50,13
MLPClassifier	54,9	Vote	50,13
SimpleLogistic	57,61	WeightedInstancesHandlerWrapper	50,13
SMO	53,05	ImputMappedClassifier	50,13
KStar	25,19	DecisionTable	59,89
LWL	62,28	PART	52,67
AdaBoostM1	60,79	ZeroR	50,13
AttributeSelectedClassifier	59,73	DecisionStump	60,79
Bagging	49,6	HoeffdingTree	52,03
ClassificationViaRegression	57,87	LMT	58,4
CostSensitiveClassifier	0	RandomForest	63,71
REPTree	48,16	RandomTree	56,07

Fonte: Do autor.

Após os testes realizados sobre todos os algoritmos encontrados no WEKA, filtrou-se os 11 algoritmos com os maiores e mais consistentes resultados. Na tabela 11 é possível observá-los seguidos de suas médias.

Tabela 11: Algoritmos com melhores médias.

Algoritmo	Média
AdaBoostM1	60,79
DecisionTable	59,89
DecisionStump	60,79
LWL	62,28
FilteredClassifier	59,84
RandomSubSpace	58,35
OneR	62,7
RandomForest	63,71
JRip	62,06
IterativeClassifierOptimized	63,55
LogitBoost	65,09

Fonte: Do autor.

5.9 MÉTODOS DE TESTES NA BASE DE DADOS

Durante a fase de testes na base de dados foram utilizadas diferentes técnicas com o intuito de aumentar a porcentagem de acerto sobre a base testada. Nesta fase ainda são utilizadas 5 *Random Seeds* para observação da consistência dos resultados. Todos os testes realizados, foram feitos usando a técnica de *cross-validation,* com *10 folds*. Escolheu-se esta técnica devido a sua eficácia em questão as bases de dados e sua aceitação na comunidade científica.

As técnicas de testes na base de dados foram realizadas com suas configurações mais relevantes, que de alguma maneira aumentasse a porcentagem de acertos da base de conhecimento.

É de suma importância os testes na base de dados serem executados com diversos métodos e configurações. Cada método colabora de uma maneira diferente, pois utilizam de técnicas e configurações que podem fazer uma diferença drástica em relação aos resultados obtidos ou alcançados sem a utilização dos mesmos. É muito interessante e muito útil também mesclar os métodos que forem possíveis, pois, uma vez que trabalhado com as melhorias dos dois juntos, existe uma chance de se obter um resultado ainda melhor nos testes realizados, desta forma, a seguir serão explicados e mostrados os testes efetuados com os métodos que mostraram os resultados mais relevantes em relação as melhorias apresentadas.

5.9.1 MÉTODO DE SELEÇÃO DE ATRIBUTOS

Técnicas de seleção de atributos do tipo filtro selecionam um conjunto de atributos na fase de pré-processamento de forma independente para com o classificador a ser utilizado. O filtro funciona de forma que os atributos com menor relevância são eliminados e apresentados apenas os atributos com chances melhores de formularem melhores resultados com sua aplicação. Desta forma, no relatório final da aplicação do método de seleção de atributos, são mostrados apenas os atributos mais relevantes na base de dados, ou os que podem apresentar melhores resultados em sua aplicação. (PEREIRA, 2009)

É importante a utilização do método de seleção de atributos, pois a base de dados a ser explorada neste projeto contém um número considerável de empresas cadastradas, gerando assim 377 instâncias no arquivo .arff, tendo isto em mente é importante lembrar que cada empresa carrega consigo 67 atributos, e que uma parte considerável destes atributos em certas empresas não estão preenchidos, então, não tem valor atribuído, ou seja, uma quantidade relevante destes atributos carregam consigo um valor nulo, e assim não apresentam significância na execução dos métodos e algoritmos nos momentos de testes.

É nesta situação em que o método de seleção de atributo faz uma diferença positiva e importante, pois como se trata de um método de filtragem, o mesmo seleciona os atributos com maior relevância na base de dados e os seleciona. Com a utilização apenas dos atributos mais importantes, os testes realizados apresentam uma melhoria significativa, pois reduz drasticamente o número de processos entre atributos e permite que estes mesmos processos sejam mais eficientes com a aplicação apenas dos atributos com maior significância.

Os resultados obtidos com o método, foram os atributos Segmento, Valor de mercado, Pessoas Físicas, Categoria, LPA, VPA, P/VPA, Lucro Líquido, Margem Líquida, Patrimônio Líquido e CaixaLS. O resultado destes atributos selecionados se dá pela relevância de seus valores. Os mesmos contém valores preenchidos na maioria das instâncias do arquivo .arff, o que indica uma significância importante para meios da análise fundamentalista, que através dos métodos e algoritmos de seleção fazem uso destes atributos. É comum que seja preenchido nos cadastros das empresas dentro da BM&FBOVESPA, as variáveis mais impactantes em uma análise fundamentalista, pois é necessária esta análise para períodos de longo prazo. Assim, como o método de seleção de atributos utiliza das variáveis mais usadas para

detecção de valores, é normal que as variáveis mais selecionadas são as mesmas utilizadas por especialistas em bolsa de valores para a utilização de uma análise fundamentalista.

Desta forma os atributos que recebem maior quantidade de informação guardada, é mais importante que os atributos que não contém informação alguma, representados pelo caractere “?”, ou seja, com valores nulos. Também é importante deixar claro, que estes atributos selecionados pelo método de seleção de atributos apresentam um papel importante devido ao seu impacto na execução dos algoritmos, uma vez que podem ser trabalhados para a obtenção de resultados mais satisfatórios, devido a significância de seus valores. É possível observar na figura 7 o resultado da técnica de seleção de atributos.

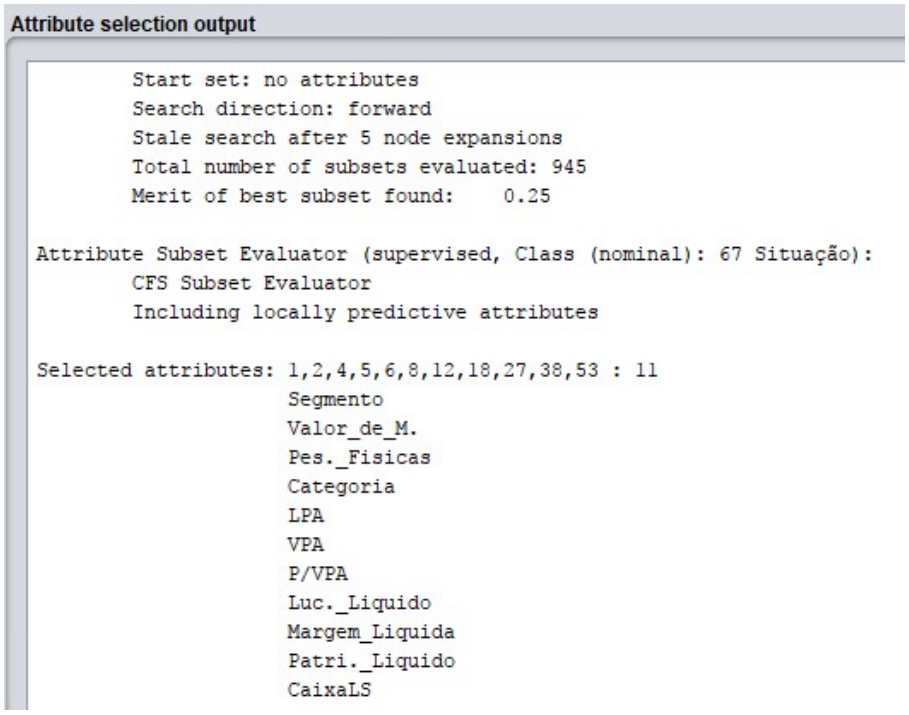

Figura 7: Retorno de resultados de método de seleção de atributos no Weka.

Fonte: Do autor.

Os resultados das médias obtidas nas aplicações dos algoritmos com a utilização da seleção de atributos foram positivos para os algoritmos DecisionTable, FilteredClassifier, RandomSubSpace, IterativeClassifierOptimizer e LogitBoost. Nos algoritmos AdaBoostM1, DecisionStump e OneR não se obteve resultados, as médias continuaram as mesmas. Com os algoritmos LWL, RandomForest, JRip os resultados

foram negativos, ou seja, após a aplicação da seleção de atributos as médias ficaram mais baixas. A tabela 12 apresenta os resultados.

Tabela 12: Médias dos algoritmos com o método de seleção de atributos.

Algoritmo	Média
AdaBoostM1	60,79
DecisionTable	60,52
DecisionStump	60,79
LWL	59,89
FilteredClassifier	62,06
RandomSubSpace	58,67
OneR	63,44
RandomForest	59,62
JRip	61,69
IterativeClassifierOptimized	65,14
LogitBoost	65,94

Fonte: Do autor.

5.9.2 MÉTODO *WRAPPER* (EMBRULHO)

A abordagen do método *Wrapper* utiliza um esquema de aprendizagem para avaliar a significância de um conjunto de atributos determinados, que são usados também em uma análise fundamentalista, como as variáveis são selecionadas por sua significância e os dados retirados da Bastter.com estão preenchidos principalmente pela sua importância em uma análise fundamentalista, é comum que se utilize destes dados para o funcionamento deste método.

Dentro do *software* WEKA, isto é possivel através das configurações da opção "*WrapperSubseEval*" que implementa a abordagem repetida de validação cruzada, *"cross-validation".* Na abordagem do *Wrapper*, o algoritmo de seleção do subconjunto engloba o algoritmo de indução. O algoritmo de seleção de subconjuntos de recursos conduz uma busca por um bom subconjunto usando o próprio algoritmo de indução como parte da função de avaliação de subconjuntos de recursos. (PEREIRA, 2009)

Para esclarecimento da funcionalidade e dos resultados com a utilização da técnica de *Wrapper*, primeiramente pode-se observar o impacto da técnica sobre os diferentes algoritmos. É importante esclarecer que esta técnica foi aplicada em

conjunto com o método de seleção de atributos, pois seus resultados mostraram grande melhoria nas médias dos algoritmos com resultados mais altos.

Os resultados com a técnica de *Wrapper* foram positivos para os algoritmos DecisionTable, LWL, FilteredClassifier, RandomSubSpace, RandomForest e JRip. Para os algoritmos AdaBoostM1, DecisionStump, OneR, IterativeClassifierOptimizer e LogitBoost foram negativos, pois as médias diminuíram. Pode-se observar estes resultados na tabela 13.

Tabela 13: Médias dos algoritmos com o método *Wrapper.*

Algoritmo	**Média**
AdaBoostM1	59,73
DecisionTable	61,37
DecisionStump	59,31
LWL	60,21
FilteredClassifier	62,75
RandomSubSpace	62,38
OneR	63,02
RandomForest	64,35
JRip	64,08
IterativeClassifierOptimized	63,66
LogitBoost	63,13

Fonte: Do autor.

5.9.3 MÉTODO PCA (*Principal Components Analysis*)

Neste capítulo é abordado os resultados obtidos com a utilização do método PCA e as configurações para a funcionalidade do mesmo. É importante deixar esclarecido, que as configurações utilizadas para a submissão da base de dados aos algoritmos, foram as configurações propostas pelo WEKA, sem exclusões de regras geradas pelo método, esta escolha se deu pela grande quantidade de atributos que são comparados para a geração de regras, sendo assim muito difícil classificá-las e compará-las. Neste método não foi utilizado a seleção de atributos anteriormente explicada, pois como o método PCA é uma combinação dos principais componentes a serem analisados, não é aconselhável que se exclua métodos, pois isto pode afetar nos resultados de forma negativa.

O método do PCA executa uma análise dos princípais componentes e a transformação dos dados. Em um conjunto com uma pesquisa de classificados, a redução de dimensão é realizada através da escolha de autovetores suficientes para

explicar alguns percentuais da variância nos dados originais, utilizando automaticamente um filtro de análise fundamentalista, detectando os dados mais importantes e trabalhando com eles, pois como já dito, os principais dados obtidos, foram os que são necessários para uma análise fundamentalista. O ruído do atributo pode ser filtrado transformando-se para o espaço do PC, eliminando alguns dos piores vetores próprios e depois transformando de volta ao espaço original. (WEKA, 2017)

Os resultados obtidos com a utilização do método PCA, mostraram uma perca na porcentagem dos algoritmos FilteredClassifier, OneR, RandomForest, JRip, IterativeClassifierOptimized, LogitBoost executados em relação a base de dados com a seleção de dados aplicada. Já os algoritmos AdaBoostM1, DecisionTable, DecisionStump, LWL e RandomSubSpace mostraram melhorias nos resultados. Acompanhe na tabela 14 os resultados.

Tabela 14: Médias dos algoritmos com o método PCA.

Algoritmo	**Média**
AdaBoostM1	62,06
DecisionTable	62,12
DecisionStump	62,06
LWL	62,22
FilteredClassifier	59,41
RandomSubSpace	61,53
OneR	57,71
RandomForest	58,09
JRip	60,37
IterativeClassifierOptimized	62,06
LogitBoost	59,68

Fonte: Do autor.

5.9.4 MÉTODO *DISCRETIZE*

Aqui será abordado os testes realizados com a utilização do método *Discretize*, é importante salientar que na configuração de "*binRangePrecision*" o número selecionado foi 6, a determinação deste número aconteceu por seus melhores resultados em comparação de outros testados. Também é importante esclarecer que o mesmo foi testado com a opção de "*makeBinary = true*" quanto com a opção de "*makeBinary = false*". Assim como no método *Wrapper*, esta técnica também foi testada com a seleção de atributos configurada na base de dados, também pelo motivo de melhoria nos resultados finais.

O filtro *Discretize* funciona de maneira símples, de forma que um filtro de instância discretiza, ou seja, divide uma variedade de atributos numéricos no conjunto de dados em atributos nomínais. Ou seja, o algoritmo divide em grupos dados que são muito exparços, criando assim grupos que podem ir de dado A á M, e de N á Z, por exemplo, e os transforma em nomínais, para que se agregue os valores novos. A discretização é por *binning* símples. O mesmo ignora o atributo de classe se estiver configurado. (WEKA, 2017)

Os resultados que este método retornou foram separados por suas configurações de "*makeBinary = true*" e "*makeBinary = false*". Na configuração de "*makeBinary = true*" as médias obtiveram melhoras nos resultados dos algoritmos AdaBoostM1, DecisionTable, DecisionStump, LWL, FilteredClassifier, RandomSubSpace e JRip. Já os algoritmos OneR, RandomForest, IterativeClassifierOptimizer e LogitBoost apresentaram perca de valores nas médias. Observe a tabela 15.

Tabela 15: Médias dos algoritmos com a utilização do método *Discretize.*

Algoritmo	**Média com makeBinary = true**	**Média com makeBinary = false**
AdaBoostM1	61,32	61,32
DecisionTable	63,34	63,45
DecisionStump	61,32	61,32
LWL	62,59	63,5
FilteredClassifier	63,23	64,4
RandomSubSpace	60,47	61,21
OneR	45,03	63,12
RandomForest	64,93	65,25
JRip	66,89	65,62
IterativeClassifierOptimized	65,99	67,38
LogitBoost	67,16	67,79

Fonte: Do autor.

5.9.5 MÉTODOS *DISCRETIZE* E *WRAPPER* COMBINADOS

Neste capítulo é abordado a forma como foi elaborada os devidos testes com a técnica de combinação entre os métodos *Wrapper* e *Discretize*, lembrando que as configurações sobre os dois métodos são as mesmas explicadas anteriormente. A configuração de "*binRangePrecision*" tem o valor 6, devido aos seus bons resultados.

Como no método de *Discretize* alguns algoritmos apresentam melhores resultados com a configuração de "*makeBinary = true*" e outros com a configuração

de “*makeBinary = false*”, para cada algoritmo testado, foi-se realizada as configurações que lhe retornasse o melhor resultado. Os testes realizados com esta técnica também foram feitos com a seleção de atributos aplicada, por conta de suas melhorias, assim como dito anteriormente.

Os resultados obtidos com está técnica foram positivas para os algoritmos DecisionTable, LWL, FilteredClassifier, RandomSubSpace, OneR, RandomForest e JRip, pois suas médias tiveram melhorias muito significativas. Assim, os algoritmos AdaBoostM1, DecisionStump, IterativeClassifierOptimizer e LogitBoost não mostraram melhorias, pois obtiveram queda nos valores de suas médias. Acompanhe na tabela 16 os resultados.

Tabela 16: Médias dos algoritmos com a utilização do método *Wrapper* e *Discretize.*

Algoritmo	**Média**	**Configuração do makeBinary**
AdaBoostM1	58,78	Indiferente
DecisionTable	64,61	False
DecisionStump	60,79	Indiferente
LWL	63,92	False
FilteredClassifier	65,78	False
RandomSubSpace	64,08	False
OneR	66,57	False
RandomForest	64,08	False
JRip	66,62	True
IterativeClassifierOptimized	66,84	False
LogitBoost	66,47	False

Fonte: Do autor.

5.10 SELEÇÃO DOS RESULTADOS MAIS ALTOS

Nesta etapa do projeto é apresentado os algoritmos que obtiveram os maiores resultados, independente da média, pois como dito anteriormente, a média que até agora veio acontecendo através da execução de cinco testes para cada algoritmo selecionado, era somente para verificar a consistência dos resultados dos algoritmos. Uma vez, que detectado os resultados com maior valor de porcentagem de acerto sobre o treinamento da base de dados, e verificado que a consistência dos resultados uma vez que testado com várias configurações de “*seeds*” é positiva, então, seleciona-se os cinco melhores, de forma que deste ponto em diante do projeto os algoritmos escolhidos irão ser submetidos a outros testes para a comprovação e a finalização da

competência ao verificar as devidas empresas e classificá-las se são boas, médias ou ruins para investimento a longo prazo.

Os testes a seguir serão realizados com os algoritmos LogitBoost, IterativeClassifierOptimizer, JRip, RandomForest e OneR, pois obtiveram os maiores resultados, independente dos métodos de testes e das médias. Observe na tabela 17.

Tabela 17: Cinco algoritmos com os resultados mais altos e suas configurações.

Algoritmo	Maior Resultado	Configuração
LogitBoost	68,7	Método Discretize com makeBinary = true e seed = 4
IterativeClassifierOptimizer	68,43	Método Discretize e Wrapper combinados com makeBinary = false e seed = 3
JRip	68,43	Método Discretize com makeBinary = true e seed = 3
RandomForest	66,84	Método Discretize com makeBinary = true e seed = 2
OneR	66,57	Método Discretize e Wrapper combinados com makeBinary = false e seed = 1~5

Fonte: Do autor.

Todos os testes elencados na tabela estão configurados juntamente com o método de seleção de atributos. Também é importante observar como os métodos *Discretize* e *Wrapper* estão presentes nos resultados. Todos os algoritmos precisaram do *Discretize* para alcançar seu maior resultado nos testes realizados, é valido observar como a configuração do "*makeBinary*" pode apresentar variância nos resultados quando estão com seus valores "*true/false*". Os algoritmos LogitBoost, JRip e RandomForest obtiveram melhores resultados com o "*makeBinary = true*", e os algoritmos IterativeClassifierOptimizer e o OneR obtiveram os melhores resultados com a configuração de "*makeBinary = false*". É importante explicitar que o algoritmo OneR obteve o mesmo resultado em todas as configurações de "*seeds*".

A partir da seleção dos 5 melhores algoritmos para a avaliação da base de dados, o próximo passo é realizar os testes de 30 *seeds* para cada um dos algoritmos, utilizando técnicas com o método de seleção de atributos, com a técnica do método de seleção de atributos combinada com o método *Discretize*, também a técnica do

método de seleção de atributos combinada com o método *Wrapper*, e por fim, com a técnica do método de seleção de atributos mesclada com o método *Discretize* e o método *Wrapper*. A escolha destes métodos para serem realizados os testes finais para a obtenção dos resultados finais se dá pela importância dos mesmos na obtenção dos maiores resultados, como visto na tabela anterior.

5.11 TESTE DE 30 *SEEDS*

Nesta seção é abordado o teste de 30 *seeds*, que é o último teste a ser realizado, também é o teste que trará os resultados finais mais relevantes em relação ao treinamento da base de dados. Para este teste de conclusão, é importante salientar que ele foi elaborado com a seleção dos 5 algoritmos que apresentaram os resultados mais altos, sendo eles o LogitBoost, IterativeClassifierOptimizer, JRip, RandomForest e OneR. A escolha dos algoritmos com os testes mais altos, se deu pela possibilidade de aumentar o resultado ou manter o resultado mais alto, uma vez que a configuração do resultado maior para cada algoritmo é importante para o treinamento da base de dados. Pois quanto maior a média de acerto do algoritmo, mais confiável é a aplicação de uma instância nova na base de dados treinada.

Também foi determinado para o treinamento dos 30 *seeds*, os métodos mais eficazes, sendo eles, o método de seleção de atributos, o método de *Wrapper*, o método de *Discretize* e o método de *Wrapper* e *Discretize* juntos, como é possível observar nas tabelas 13, 14, 16 e 17, anteriormente exibidas.

Estes métodos com a aplicação destes algoritmos foram realizados de acordo com as configurações anteriormente elencados nas seções *5.8.1 Método de Seleção de atributos*, *5.8.2 Método Wrapper (Embrulho), 5.8.4 Método Discretize* e *5.8.5 Método Discretize e Wrapper combinados.* O método de validação de teste para a seleção de cada *seed* na base de dados, foi por sua vez a *cross-validation com 10 folds,* devido a competência do método e a sua aceitação na comunidade científica, que é atualmente a mais aceita.

As tabelas que seguem com os valores dos testes a seguir, correspondem uma para cada método. As tabelas são compostas com os algoritmos em colunas, totalizando cinco colunas, e os *seeds* em linhas totalizando trinta linhas e uma última linha no final da tabela que corresponde à média do teste no final dos 30 resultados. Esta média tem a função de exibir um resultado que mede a consistência da base de dados testada com algoritmo escolhido no método determinado.

5.11.1 MÉTODO DE SELEÇÃO DE ATRIBUTOS

A tabela 18, exibe os resultados da base de dados testada com a utilização do método de seleção de atributos, explicada na seção 5.8.1. As configurações aqui utilizadas também são as mesmas realizadas nos testes com 5 *seeds*.

Tabela 18: Teste de30 *seeds* com o método de seleção de atributos.

Método de Seleção de Atributos					
	LogitBoost	**Iterative Classifier Optimizer**	**JRip**	**RandomForest**	**OneR**
Seed 1	66,57	66,04	62,86	60,47	64,72
Seed 2	67,1	65,51	60,74	58,88	63,66
Seed 3	65,25	64,98	61,27	59,41	62,86
Seed 4	66,84	66,31	62,33	59,68	63,92
Seed 5	63,92	62,86	61,27	59,68	62,06
Seed 6	63,39	63,92	61	60,47	63,66
Seed 7	64,98	64,19	62,06	58,09	62,86
Seed 8	63,39	62,86	60,47	59,15	62,33
Seed 9	66,84	63,92	58,88	57,82	62,06
Seed 10	65,78	64,72	60,74	58,09	63,66
Seed 11	66,84	64,72	62,33	57,55	62,59
Seed 12	64,19	63,66	60,21	59,15	60,74
Seed 13	67,9	64,19	58,62	61,27	63,66
Seed 14	67,1	65,51	62,06	60,47	65,25
Seed 15	65,25	64,19	63,39	58,09	63,66
Seed 16	66,04	64,19	64,45	59,68	63,92
Seed 17	62,86	63,66	62,06	58,35	64,45
Seed 18	65,78	62,59	59,68	57,02	63,39
Seed 19	65,78	64,19	59,41	59,41	62,59
Seed 20	63,39	61,27	59,94	57,55	62,06
Seed 21	64,72	63,66	63,39	58,35	64,45
Seed 22	63,92	62,59	60,47	58,35	63,66
Seed 23	65,51	65,25	61	57,29	63,39
Seed 24	64,98	65,78	62,59	59,68	63,39
Seed 25	64,98	64,72	61,53	59,41	62,06
Seed 26	63,13	62,06	63,13	58,09	58,88
Seed 27	64,19	63,92	62,59	59,94	63,92
Seed 28	64,98	61,53	63,39	58,62	62,33
Seed 29	64,98	65,78	63,39	58,09	62,59
Seed 30	64,45	63,66	59,68	58,88	62,33
Média	**65,17**	**64,08**	**61,50**	**58,90**	**63,04**

Fonte: Do autor.

5.11.2 MÉTODO *WRAPPER*

A tabela 19, mostra os resultados da base de dados testada com a utilização do método de *Wrapper*, explicada na seção 5.8.2. As configurações utilizadas neste método também são as mesmas realizadas nos testes com 5 *seeds*.

Tabela 19: Teste de 30 *seeds* com o método *Wrapper.*

Base com método Wrapper					
	LogitBoost	**IterativeClassifierOptimizer**	**JRip**	**RandomForest**	**OneR**
Seed 1	64,72	64,19	64,45	63,92	64,72
Seed 2	63,39	63,66	64,72	64,19	63,66
Seed 3	62,59	63,92	65,51	65,51	62,86
Seed 4	62,86	63,66	62,33	64,72	61,8
Seed 5	62,09	62,86	63,39	63,39	62,06
Seed 6	62,06	63,13	66,04	66,84	62,06
Seed 7	63,13	63,39	63,92	66,04	62,86
Seed 8	62,33	64,45	63,66	62,33	62,33
Seed 9	61,27	63,39	61,8	65,51	62,06
Seed 10	63,39	65,25	64,45	64,98	63,66
Seed 11	61,27	63,39	64,45	65,72	62,59
Seed 12	63,66	64,19	63,13	65,78	60,74
Seed 13	61,53	64,72	61,53	63,13	63,66
Seed 14	64,72	63,39	65,78	65,51	65,25
Seed 15	64,45	64,19	64,19	64,98	63,66
Seed 16	62,59	64,72	63,92	65,51	62,06
Seed 17	63,13	64,98	63,13	65,51	64,45
Seed 18	64,98	63,66	63,66	63,13	63,39
Seed 19	61,27	63,13	62,59	65,51	62,59
Seed 20	63,39	63,13	62,33	63,92	62,06
Seed 21	61,8	64,45	63,66	64,19	64,45
Seed 22	59,68	63,39	65,78	63,92	63,66
Seed 23	62,86	64,19	65,78	64,45	63,39
Seed 24	63,92	63,66	62,33	63,39	63,39
Seed 25	63,13	63,13	63,92	66,31	62,06
Seed 26	63,92	64,98	64,19	66,57	62,86
Seed 27	60,74	62,86	64,19	65,25	63,92
Seed 28	64,72	63,66	63,92	66,04	62,33
Seed 29	64,72	65,25	65,51	65,25	62,59
Seed 30	62,86	63,92	64,72	65,51	62,33
Média	**62,91**	**63,90**	**63,97**	**64,90**	**62,98**

Fonte: Do autor.

5.11.3 MÉTODO *DISCRETIZE*

A tabela 20 a seguir, exibe os resultados da base de dados testada com a utilização do método de *Discretize*, explicada na seção 5.8.4. As configurações usadas aqui também são as mesmas usadas nos testes com 5 *seeds*.

Tabela 20: Teste de 30 *seeds* com o método de *Discretize*.

Base com método Discretize					
	LogitBoost	**IterativeClassifierOptimizer**	**JRip**	**RandomForest**	**OneR**
Seed 1	66,04	66,04	65,78	64,45	62,59
Seed 2	68,43	64,72	66,57	66,84	66,57
Seed 3	66,04	66,57	66,04	63,39	62,33
Seed 4	67,37	67,37	66,31	63,92	63,66
Seed 5	66,31	64,19	66,31	66,04	60,47
Seed 6	65,78	64,19	66,31	64,98	60,47
Seed 7	66,31	65,78	68,96	65,78	66,57
Seed 8	66,31	66,84	67,63	63,39	63,13
Seed 9	68,16	65,78	**70,02**	63,92	56,76
Seed 10	65,25	66,31	68,43	64,19	59,94
Seed 11	66,31	66,57	67,9	64,19	66,57
Seed 12	63,92	63,66	67,1	64,19	59,15
Seed 13	67,63	66,84	64,45	65,78	58,62
Seed 14	66,04	65,78	67,37	66,31	60,47
Seed 15	67,78	66,04	67,63	66,31	63,13
Seed 16	64,72	63,39	64,98	63,66	66,57
Seed 17	63,92	64,72	66,31	64,45	63,13
Seed 18	66,04	65,25	66,84	65,25	62,86
Seed 19	66,04	64,72	68,16	62,06	63,66
Seed 20	66,04	65,51	66,04	61,8	57,82
Seed 21	66,84	65,25	67,37	64,98	62,33
Seed 22	68,43	64,98	67,1	64,98	63,92
Seed 23	64,72	65,25	64,72	63,92	66,57
Seed 24	66,84	65,25	67,9	66,84	66,57
Seed 25	65,78	66,84	68,16	66,04	62,59
Seed 26	65,25	64,45	68,7	63,66	62,86
Seed 27	66,04	65,25	67,37	66,04	62,86
Seed 28	65,51	66,04	65,51	65,51	66,57
Seed 29	64,45	65,51	68,96	64,45	66,57
Seed 30	67,1	65,51	67,1	63,92	62,59
Média	**66,18**	**65,49**	**67,07**	**64,71**	**62,93**

Fonte: Do autor.

5.11.4 MÉTODO *DISCRETIZE* E *WRAPPER* COMBINADOS

A tabela 21, mostra os resultados da base de dados testada com a utilização dos métodos de *Wrapper* e *Discretize* combinados, explicada na seção 5.8.5. As configurações utilizadas aqui também são as mesmas realizadas nos testes com 5 *seeds*.

Tabela 21: Teste de 30 *seeds* com o método de *Discretize* e *Wrapper.*

Base com métodos Discretize e Wrapper					
	LogitBoost	**IterativeClassifierOptimizer**	**JRip**	**RandomForest**	**OneR**
Seed 1	66,04	66,31	67,9	64,98	66,57
Seed 2	66,31	64,72	65,78	63,92	66,57
Seed 3	66,04	66,04	67,1	64,98	66,57
Seed 4	65,78	65,78	66,84	62,59	66,57
Seed 5	66,84	66,31	65,51	62,33	66,57
Seed 6	64,98	65,78	66,04	63,92	66,57
Seed 7	66,84	64,98	66,04	63,66	66,57
Seed 8	67,1	64,98	66,84	66,31	66,57
Seed 9	66,31	67,63	66,57	65,25	66,57
Seed 10	64,72	67,37	64,98	60,21	66,57
Seed 11	66,04	65,78	65,78	68,43	66,57
Seed 12	65,25	63,92	65,25	66,31	66,57
Seed 13	64,72	65,78	65,78	65,25	66,57
Seed 14	67,63	66,57	65,25	64,45	66,57
Seed 15	66,31	67,37	66,04	66,57	66,57
Seed 16	64,19	65,51	66,84	66,57	66,57
Seed 17	67,63	64,19	65,51	66,31	66,57
Seed 18	67,9	67,1	65,25	64,98	66,57
Seed 19	66,31	65,25	66,31	62,06	66,57
Seed 20	63,39	66,31	64,45	64,98	66,57
Seed 21	67,37	67,37	65,25	65,51	66,57
Seed 22	64,19	64,19	64,19	64,72	66,57
Seed 23	65,78	67,63	64,45	63,92	66,57
Seed 24	65,51	64,72	65,25	64,45	66,57
Seed 25	65,51	67,1	66,84	67,37	66,57
Seed 26	63,92	65,25	66,84	63,39	66,57
Seed 27	67,37	62,86	65,78	64,45	66,57
Seed 28	66,57	63,66	65,25	65,51	66,57
Seed 29	65,25	65,25	64,19	63,92	66,57
Seed 30	65,51	63,92	64,98	64,72	66,57
Média	**65,91**	**65,65**	**65,77**	**64,73**	**66,57**

Fonte: Do autor.

6.0 – RESULTADOS E DISCUSSÃO

Neste capítulo são abordados os resultados obtidos com a elaboração do presente projeto. Serão explicados e detalhados os resultados extraídos com a elaboração da parte prática do trabalho. São estes, o algoritmo que apresentou o resultado mais alto para o treinamento da base de dados, e o algoritmo que mostrou o menor resultado, levando em consideração que para os testes finais com 30 *seeds* foram escolhidos apenas cinco algoritmos.

A menor média obtida com o algoritmo *ZeroR,* que é a média de mínima que o algoritmo pode alcançar, e a média a ser melhorada e trabalhada, neste caso, 50,13%. Também é apresentado neste capítulo os resultados obtidos com cada método utilizado, como também os algoritmos usados para a elaboração do treinamento da base de dados.

6.1 RESULTADOS COM O MÉTODO DE SELEÇÃO DE ATRIBUTOS.

Todos os algoritmos executados com este método passaram pela quantidade de 30 testes. Sendo eles, respectivamente os testes do *seed* 1 ao 30, como observado nas tabelas referentes aos métodos.

Os resultados com a utilização apenas do método de seleção de atributos, aplicada com algoritmo *LogitBoost* demonstrou uma variância de 4,77%, o menor resultado obtido com este algoritmo foi de 63,13% de acerto, o que não é um resultado confiável para o treinamento da base de dados, pois sua porcentagem não é alta o suficiente para demonstrar confiança. O maior resultado retornou um valor de 67,90% de acerto no treinamento da base de dados, resultado este que já mostra uma melhora de porcentagem, devido do aumento de valor. A média geral dos resultados dos testes com os 30 *seeds* apresentou um valor de 65,17% como é possível observar na tabela 22, mostrando que os resultados dos testes alcançaram um valor intermediário levando em consideração a menor média e a maior média.

Com a utilização do algoritmo *IterativeClassifierOptimizer,* com este mesmo método, se obteve o resultado de 61,27%, como o menor valor adquiro no teste. Por sua vez se obteve o valor de 66,31% como sendo o valor mais alto alcançado, desta forma se obtém uma variância de 5,04%. A média total dos 30 *seeds*, ficou com um valor de 64,08%. Que é também um valor intermediário levando em consideração o menor e o maior valor retornado com este algoritmo.

Utilizando o algoritmo *JRip,* obteve-se uma média geral de 61,50% nos acertos, sendo que a maior porcentagem de acerto foi de 64,45%. O menor resultado foi de 58,62%. A variância entre o menor e o maior valor adquiridos com os testes foi de 5,83%, o que é um valor intermediário, se levar em consideração que foram realizados 30 testes no total.

O algoritmo *RandomForest*, mostrou uma variância de 4,25%, tendo em vista que o seu maior valor de acerto foi de 61,27%. Por sua vez, o menor resultado dos testes retornou o resultado de 57,02%. A média final da execução dos 30 *seeds* retornou o valor de 58,90%.

Por fim, o algoritmo *OneR* mostrou uma média geral de 63,04%. Uma variância de 6,37%. O menor valor ocorrido deu-se com um valor de 58,88%, e o maior valor resultante foi equivalente a 65,25%, assim como nos algoritmos *JRip* e *RandomForest*. Na tabela 22 é possível observar os resultados dos algoritmos.

Tabela 22: Valores obtidos com o método de Seleção de Atributos.

	Menor Valor	Maior Valor	Variância	Média Geral
LogitBoost	63,13%	67,90%	4,77%	65,17%
IterativeClassifierOptimizer	61,27%	66,31%	5,04%	64,08%
JRip	58,62%	64,45%	6,37%	63,04%
RandonForest	57,02%	61,27%	4,25%	58,90%
OneR	58,88%	65,25%	6,37%	63,04%

Fonte: Do autor.

Com a tabela 22, é possível observar que o maior resultado, se obteve com o algoritmo *LogitBoost*, com um valor de 67,90%. Sua variância alcançou um valor de 4,77%, o que é bom, pois é uma porcentagem baixa, se comparada com as outras. Por outro lado, o algoritmo *JRip* obteve uma variância de 6,37%, o que não é bom, levando em consideração que foram calculados 30 valores para a retirada destes valores. Seu maior valor alcançou apenas 64,45%, o que não é uma média boa para o treinamento de uma base de dados, pois é menos de 2/3 de acertos, o que pode ser arriscado no momento da escolha da empresa. Neste teste, com o método de seleção de atributos, o algoritmo *IterativeClassifierOptimizer*, mostrou um desempenho mediano se levado em consideração as variáveis de menor valor, maior valor, variância e média geral. Pois o mesmo obteve um valor de maior média, com 66,31%, e um valor de 61,27% como menor valor. A sua variância foi de 5,04%, o que é um valor intermediário para 30 testes. A média geral deste algoritmo apresentou o melhor

resultado, com 64,08%, o que aponta um valor bom e considerável de testes com boas médias.

O algoritmo *OneR* mostrou um desempenho não satisfatório, pois obteve a menor média de todos os algoritmos com 58,88% e uma média geral de 63,04%, o que também não é um resultado satisfatório. Por fim, o algoritmo *RandomForest* mostrou o seu menor resultado com 57,02% e seu maior resultado com 65,25%, resultados considerados baixos para treinamento de uma base de dados.

6.2 RESULTADOS COM O MÉTODO *WRAPPER*.

Como nos testes com o método de seleção de atributos, os testes com o método de *Wrapper* também foram utilizadas 30 *seeds*, sendo os testes também do seed 1 ao 30.

Tabela 23: Valores obtidos com o método de *Wrapper.*

	Menor Valor	Maior Valor	Variância	Média Geral
LogitBoost	59,68%	64,98%	5,3%	62,91%
IterativeClassifierOptimizer	62,86%	65,25%	2,39%	63,9%
JRip	61,53%	66,04%	4,51%	63,97%
RandonForest	62,33%	66,84%	4,53%	64,9%
OneR	60,74%	65,25%	4,51%	62,98%

Fonte: Do autor.

Com a utilização do algoritmo *LogitBoost* é possível observar na tabela 23 que, seus resultados não são muito satisfatórios, pois, seu maior valor apresentou uma porcentagem de acertos de 64,98%. Também é importante observar que a média geral deste algoritmo no final dos testes com 30 *seeds* ficou em 62,91%, o que também não é um valor recomendado para o treinamento de uma base de dados, devido ao valor baixo. Desta forma, este algoritmo não demonstrou bons resultados para a sua utilização.

O algoritmo *IterativeClassifierOptimizer* demonstrou um resultado no seu menor valor de 62,86%, o que veio a ser o resultado de menor valor mais alto com este método, por outro lado, seu valor máximo foi de apenas 65,25%, o que não é um bom resultado, pois a porcentagem não é alta o suficiente. Porém, é importante observar que a sua variância foi de apenas 2,39%, o que é muito bom. Mas como seu maior resultado foi de 65.25%, não é viável a utilização para o treinamento da base de dados.

Nos testes com a utilização do algoritmo *JRip*, é visível uma ligeira melhora do maior resultado que alcançou um valor de 66,04%, mas ainda não é um bom valor para o treinamento da base de dados, pois não é uma porcentagem alta o suficiente. Sua média geral apresentou um valor de 63,97%, como é possível observar na tabela 24, o que também apresenta uma melhora.

O algoritmo *RandomForest,* por sua vez, apresentou os melhores valores, com um resultado de 66,84%, o que é o resultado mais alto nos testes com o método *Wrapper*. Obteve também uma boa média geral ao final dos 30 *seeds* testados, uma média de 64,9%. Com o menor resultado de 62,33%, que é a segunda maior média neste método, obteve-se uma variância de 4,53%, que não é um valor ruim, pois é baixa a porcentagem, se levar em consideração a boa média dos 30 *seeds* alcançada pelo algoritmo.

Para finalizar os testes o algoritmo *OneR* teve no seu maior valor, 65,25%, valor este que não é significativo o suficiente para o treinamento da base de dados. Sua média ao final dos 30 testes realizados retornou um valor de 62,98%, o que também não é relevante o suficiente para destacar o algoritmo como bom.

Com isto, é possível observar que o algoritmo com o resultado mais alto alcançado foi o *RandomForest* com 66,84%, com uma variância de 4,53%. Já o algoritmo *IterativeClassifierOptimizer* apresentou a menor variância dentre todos os algoritmos, um valor de 2,39%, porém como observado anteriormente, o maior resultado foi de 65,25%, e isto não o faz um bom algoritmo para uso destas configurações. O algoritmo *LogitBoost* apresentou os piores resultados com o menor valor de 59,68% e o maior valor de 64,98%, o que não demonstra confiança para um treinamento de base de dados. Os algoritmos *OneR* e *JRip* também não apresentaram resultados relevantes a serem analisados, com o maior resultado considerado baixo e também com as médias gerais sem valores altos.

6.3 RESULTADOS COM O MÉTODO *DISCRETIZE*

Neste capítulo é verificado os testes realizados com a utilização do método *Discretize*, assim como nos testes anteriores, os resultados aqui obtidos também são referentes aos resultados obtidos ao final do teste com 30 *seeds*.

Tabela 24: Valores obtidos com o método *Discretize*.

	Menor Valor	Maior Valor	Variância	Média Geral
LogitBoost	63,92%	68,43%	4,51%	66,18%
IterativeClassifierOptimizer	63,39%	67,37%	3,98%	65,49%
JRip	64,45%	70,02%	5,57%	67,07%
RandonForest	61,08%	66,84%	5,76%	64,71%
OneR	56,76%	66,57%	9,81%	62,93%

Fonte: Do autor.

Com a utilização do algoritmo *LogitBoost,* obteve-se um resultado de 68,43% como maior valor. É possível observar na tabela 24 também que a média geral deste algoritmo resultou em 66,18%. O que é considerado uma boa média, levando em consideração que sua variância foi de 4,51%.

O algoritmo *IterativeClassifierOptimizer* mostrou uma boa variância, de 3,98%, o que mostra que ao final dos 30 testes, se obteve resultados bem consistentes. Seu menor valor resultou em de 63,39%, e seu maior valor retornou um resultado de 67,37%. A média final dos 30 testes retornou uma média geral de 65,49%, um valor intermediário entre o menor e maior resultado, comprovando sua consistência de resultados.

O algoritmo *JRip* mostrou o maior valor mais alto encontrado sobre todos os testes, com 70,02%, o que é um valor muito bom para se levar em consideração o treinamento de uma base de dados, já que este valor representa, mesmo que ligeiramente, mais de 2/3 dos resultados para acerto, representando assim uma maioria. Este algoritmo também mostrou seu menor valor de 64,45%, o que também não é um valor ruim, se comparados com os outros testes, já que sua média geral ao final dos 30 *seeds* mostrou um valor de 67,07%, o que é a maior média geral já obtida com os testes. Sua variância mostrou um valor de 5,57%, o que é um valor mediano, levando em consideração as outras médias gerais obtidas com os testes até então realizados.

Com o algoritmo *RandonForest* obteve-se valores medianos, se comparados com os outros algoritmos, seu resultado de menor média ficou com 61,08%, e seu valor maior retornou uma porcentagem de 66,84% de acerto, um valor que continuou o mesmo aos testes realizados com o método de *Wrapper*.

O algoritmo do *OneR,* mostrou os piores resultados deste método, com um resultado de menor valor de apenas 56,76%. O resultado do seu maior valor também foi o menor obtido com um total de 66,57%. Isto resultou em uma variância alta de

9,81%, em consequência desta variância a média geral dos 30 *seeds* retornou um valor de 62,93%, a menor média geral dos algoritmos com a utilização deste método.

É importante perceber como este método apresentou resultados ligeiramente melhores aos testes anteriores, o algoritmo *JRip,* mostrou o resultado mais alto de todos os testes já realizados até então. Porém, o algoritmo *OneR*, não retornou resultados satisfatórios, pois sua variância de 9,81% resultou em um valor de média geral de 62,93%, valores estes ruins, se comparados com os outros algoritmos. O algoritmo *LogitBoost*, apresentou um resultado de maior valor com 68,43%, um bom resultado também, levando com consideração sua média geral de 66,18%.

6.4 RESULTADOS COM O MÉTODO *WRAPPER* E *DISCRETIZE*

Neste capítulo são apresentados os últimos testes realizados, serão mostrados os resultados dos métodos *Wrapper* e *Discretize* combinados. Assim como os anteriores, os testes são feitos com 30 *seeds* e uma média geral. Verifique os resultados retornados na tabela 25.

Tabela 25: Valores obtidos com o método de *Wrapper* e *Discretize.*

	Menor Valor	Maior Valor	Variância	Média Geral
LogitBoost	63,39%	67,9%	4,51%	65,91%
IterativeClassifierOptimizer	62,86%	67,63%	4,77%	65,65%
JRip	64,19%	67,9%	3,71%	65,77%
RandonForest	60,21%	68,43%	8,22%	64,73%
OneR	66,57%	66,57%	0%	66,57%

Fonte: Do autor.

O algoritmo *LogitBoost,* apresentou o seu maior valor de 67,9%. Levando em consideração a seu menor valor obtido com os 30 testes, que foi de 63,39%, é possível observar uma boa variância, de 4,51%, e uma média geral de 65,91%.

Com o algoritmo *IterativeClassifierOptimizer* é possível observar uma porcentagem de 62,86% no resultado do seu menor valor. Seu maior resultado obteve-se um resultado de maior valor de 67,63%, com isto a variância obtida foi de 4,77%, um valor considerado baixo para 30 testes.

Com o algoritmo *JRip*, também se obteve um valor de 67,9% no maior valor. Seu menor resultado retornou um valor de 64,19%, um resultado bom, se comparado com outros testes, que como consequência trouxe um valor de apenas 3,71% de

variância, e por sua vez uma média geral também com um valor considerável, pois a média geral no final dos 30 testes ficou em 65,77%.

Com a utilização do algoritmo *RandonForest* obteve-se a média mais baixa dos cinco algoritmos testados com este método, uma média mínima de 60,21%. Com um total de 68,43%, obteve-se o melhor valor. A variância entre os testes ficou alta, com um total de 8,22%, que resultou na menor média geral deste método, com 64,73%.

O algoritmo *OneR*, apresentou 66,57% em todos os seus resultados, ou seja, desde o *seed* 1 ao 30, a porcentagem de acerto foi a mesma. Com isto, a variância entre os valores é 0, e a média geral também ficou com o valor de 66,57%. Estes valores repetidos se dão pela funcionalidade deste algoritmo, como o *OneR* gera apenas uma regra para a elaboração das tomadas de decisões em seu funcionamento, e levando em consideração a quantidade de atributos existentes na base de dados, a regra que o algoritmo selecionou para fazer a divisão dos resultados tenha direcionado todos na mesma direção, mostrando assim a mesma porcentagem para todos os testes.

Com este método é possível perceber que seus resultados em relação aos valores mais altos foram bem consistentes, tendo o menor valor algoritmo *OneR* com 66,57%. As médias gerais também apresentaram valores bem parecidos. Com isto pode-se tirar a conclusão que a junção destes dois métodos, pode produzir resultados mais consistentes. Com exceção do *RandonForest* que teve uma variância de 8,22%, os demais algoritmos apresentaram valores de variância baixos. Conclui-se então que para testes em quantidades a solidez dos resultados podem ser reforçadas se utilizada com a junção destes dois métodos.

6.5 CONCLUSÃO DOS TESTES

Com a execução dos testes, é possível observar uma melhoria significativa nos acertos da base de conhecimento com a utilização dos métodos de *Wrapper* e *Discretize*. É importante observar que o método de *Discretize* apresentou o maior resultado obtido, um percentual de acerto de 70,02%, com a utilização do algoritmo *JRip*, valor este que como dito anteriormente, um bom resultado para a avaliação de novas instâncias com a base de conhecimento treinada. A melhoria sobre a média mínima aceitável de 50,13%, foi de 19,89% de aumento.

Este percentual de 70,02% torna o sistema de verificação de boas empresas para investimento a longo prazo na bolsa de valores eficaz, pois o sistema acerta este

percentual citado de acordo com a base de conhecimento treinada neste projeto, ou seja, com as instâncias que existem na base de dados, que por sua vez foram trabalhadas com métodos de análises fundamentalistas juntamente com os métodos de classificação em mineração de dados.

Foi possível treinar uma base de conhecimento para que acerte o percentual de 70,02%, logo, se for testada uma nova empresa que não está dentro da base de conhecimento treinada, ela será avaliada e o sistema treinado terá uma chance de 70,02% de acertá-la, retornando assim uma qualificação, se é uma boa empresa, uma empresa mediana ou uma má empresa para se investir com interesses a longo prazo.

Ao observar as médias finais das execuções dos 30 *seeds*, é notável que com a utilização do método *Wrapper*, houve uma ligeira melhora em relação a base de conhecimento aplicada somente com a seleção de atributos, esta que foi a base utilizada para comparação da melhoria com os demais testes. Na tabela 26 é possível observar as médias gerais de todos os algoritmos aplicados em todos os métodos realizados.

É importante explicar a significância dos atributos obtidos com a utilização do método de seleção de atributos, estes atributos foram escolhidos pela sua forte participação na verificação da quantidade de combinações que os mesmos participam, ajudando assim os algoritmos selecionados a encontrarem resultados melhores e mais concisos. Levando em consideração que uma parte da base de dados é composta por instâncias com atributos carregados de valores nulos, isto se faz outro motivo para estes atributos serem selecionados, pois os mesmos fazem parte de uma quantidade considerável de instâncias que compõem a base de dados.

Tabela 26: Tabela de médias gerais.

Métodos	**Algoritmos**				
	LogitBoost	**Iterative Classifier Optimizer**	**JRip**	**RandonForest**	**OneR**
Seleção de atributos	65,17	64,08	61,50	58,90	63,04
Wrapper	62,91	63,9	63,97	64,9	62,98
Discretize	66,18	65,49	67,07	64,71	62,93
Wrapper e Discretize	65,91	65,65	65,77	64,73	66,57

Fonte: Do autor.

Com a utilização do método *Wrapper* é fácil observar que houve evolução de resultados apenas nos algoritmos *JRip* e *RandonForest*. Isto porque o método

Wrapper, faz uma validação cruzada em conjuntos e subconjuntos de atributos, encontrando assim os mais relevantes para executar o algoritmo determinado em suas configurações, como explicado anteriormente, neste caso o *JRip* e *RandonForest*. O algoritmo *JRip*, funciona de maneira que compara todos os seus atributos e encontra combinações que venha a trazer um resultado o mais exato possível. Logo, com a junção deste algoritmo com o método *Wrapper*, obteve-se um bom resultado. (WEKA, 2017)

O algoritmo *RandonForest*, mostrou uma melhoria em sua média de percentual de 6%. Isto se dá ao fato de que o algoritmo funciona selecionamento um conjunto de dados mais significantes, com a combinação do método *Wrapper*, o resultado foi de melhoria na média final.

O método *Discretize* mostrou um desempenho melhor que o *Wrapper* se comparar os resultados do mesmo aos resultados das médias da base com a utilização de seleção de atributos. É possível verificar na tabela 26, que os resultados das médias aumentam em todos os algoritmos com exceção do *OneR*, que regrediu um valor de 0,05%, o que não é um valor muito considerável, e também não tem agravantes sérios de perca de resultados. O algoritmo *JRip* traz um aumento de 5,57%, o melhor resultado entre os cinco algoritmos. Um valor considerável para a utilização na base de conhecimento.

Este aumento nas médias finais na maioria dos algoritmos, se dá pela funcionalidade do método *Discretize*, de forma que se faz uma divisão de dados, como citado anteriormente, em combinação com o algoritmo *JRip*, que mescla seus dados até encontrar as melhores relações possíveis, obteve-se um bom resultado, pois como as subdivisões geradas pelo método *Discretize*, o *JRip* faz menos combinações e entre elas encontrar melhores resultados.

Por sua vez, a combinação dos métodos *Wrapper* e *Discretize* juntamente com a seleção de atributos, resultou em um aumento de médias a todos os algoritmos testados, como é possível verificar na tabela 26. Isto acontece pela combinação dos métodos, que trabalham com suas técnicas juntas, filtrando e direcionando melhor os dados dos atributos, gerando assim formas mais eficazes de treinamento da base de dados, o que resulta em melhores valores.

6.6 VALIDAÇÃO DOS RESULTADOS

Para o entendimento dos resultados finais adquiridos com os testes neste projeto, é necessário o esclarecimento da importância dos valores adquiridos com os testes no que diz respeito a sua significância e quanto ao seu funcionamento no momento em que se alcança o maior valor com os testes realizados, neste caso o 70,02% com o algoritmo *JRip.* Ou seja, o funcionamento do *software* do WEKA em relação aos valores adquiridos e qual o significado dos valores retornados. Tanto quanto também o esclarecimento dos testes com dados fictícios após a base de conhecimento pronta.

Para a elaboração de todos os testes realizados no presente trabalho, foi utilizado o método de teste *cross-validation*, com configuração de 10 fragmentos *(folds)*. Este teste consiste em dividir a base de dados em duas partes no momento de aplicação dos algoritmos. Esta divisão funciona de maneira que um número *x-1* de partes, neste caso 9, são usadas para treinamento da base de conhecimento e a outra uma parte é usada para teste. Com isto, se alcança as porcentagens de acertos que retornam dos algoritmos após o fim dos testes, derivando da quantidade de instâncias que foram usadas para teste sobre o restante das instâncias que foram utilizadas para o treinamento da base de conhecimento. (SANTOS, MIKAMI, VENDRAMIN, KAESTNER, 2009)

Uma vez entendido o funcionamento do método de *cross-validation,* subtende-se que quando o mesmo retorna uma porcentagem referente a um teste envolvendo qualquer método e algoritmo, entende-se que este mesmo já valida os resultados, ou seja, é um resultado final, pois o método *cross-validation* treina, testa e retorna um valor em forma de porcentagem que já mostra a sua eficácia em qualquer que seja a sua aplicação mais tarde.

Logo, se a base de conhecimento estiver treinada e fazer um processo de teste com um arquivo de extensão .arff externo, com uma ou mais instâncias para teste, se estará executando um método de *cross-validation* de forma manual, pois os processos que a base de conhecimento passarão serão os mesmos, treino, teste e resultado. É valido lembrar, que se o arqruivo .arff tiver um baixo número de instâncias, ele pode não retornar o valor correto atribuído a ele, pois existe uma porcentagem de acerto, e com isto existe o risco de não se acertar todas, porém subentende-se que o resultado de todos os testes realizados, devem ficar iguais ou próximos a porcentagem

retornada pelo método de *cross-validation*. Logo, o fator que valida o sistema de avaliação das instâncias submetidas ao teste é o método de teste *cross-validation*.

O Professor e Mestre em economia, Gilson Mussi dos Reis, atualmente coordenador do curso de Administração da Unisep – Campus de Francisco Beltrão, afirmou que o presente trabalho pode ser uma ferramenta excelente para o auxílio á pessoas sem muito conhecimento e que queiram investir na bolsa de valores. Também deixou claro que esta mesma ferramenta pode ser de muita utilidade para o profissional capacitado em renda variável que queira utilizá-lo como suporte em suas análises.

7.0 – CONSIDERAÇÕES FINAIS OU CONCLUSÃO

O presente trabalho de conclusão de curso abordou técnicas e métodos para a escolha de empresas para se investir na bolsa de valores a longo prazo, com a utilização de mineração de dados e análise fundamentalista. Para se chegar nos resultados foi-se necessário concluir algumas etapas pré-estabelecidas, a primeira delas, a de entender os indicadores necessários e relevantes para a construção de uma base de dados. A qual foi concluída com um estudo aprofundado com sites e diálogos com especialistas em finanças no mercado de renda variável. Após isto foi elaborada uma base de dados contendo todas as variáveis existentes dentro do mercado de ações para cada empresa cadastrada na BM&FBOVESPA. Tais variáveis que são os atributos para cada empresa, neste caso instâncias. Com a base de dados elaborada, foi criado um arquivo .arff, o qual continha toda a base de dados para ser treinada e retornar uma base de conhecimento.

Com a finalização do arquivo .arff, no WEKA, foram realizados testes em todos os algoritmos para a verificação dos melhores resultados, e filtrá-los para novos testes com métodos necessários para alavancar a porcentagem de acerto dos mesmos. Desta forma, com a definição dos melhores algoritmos e métodos analíticos, foi-se coletado os resultados dos testes. Com isto, conclui-se que a margem de acerto, neste caso, uma porcentagem de 70,02% sendo a mais alta e melhor. Pode ser muito útil para a análise de investimentos no mercado de ações, pois a porcentagem de acertos é maioria, e com uma boa vantagem sobre a margem de erros. Assim, conclui-se que este método pode ser útil em situações as quais seja necessário um filtro das empresas mais relevantes para investimento a longo prazo, e assim analisá-las e selecionar as preferidas ou mais relevantes, lembrando que o mercado de ações consiste de muitas variações nos papeis das empresas cotadas, sendo necessário a melhor análise possível antes de investir.

Levando em consideração a dificuldade de análise que se encontra no mercado de renda variável, devido à grande quantidade de fatores que podem impactar no valor das ações da empresa, como notícias, novas relações, investimentos internos, entre outras, considera-se que este método pode ser útil para pessoas sem muita experiência de campo, podendo-lhe auxiliá-la para uma possível investida que possa ser realizada em alguma empresa com ações disponíveis. Lembrando também, que isto pode ser utilizado por um especialista como uma ferramenta de apoio, para que o mesmo encontre mais confiança ao fazer indicações sobre as empresas, com isto,

este trabalho também traz benefícios ao campo do mercado de ações. Através da ferramenta desenvolvida também é possível ter uma posição sobre a legitimidade das avaliações realizadas por especialistas, verificando suas margens de acertos e comparando com a ferramenta. O que pode ser muito útil para uma possível análise sobre os especialistas ou seus métodos de análise.

Desta forma, conclui-se que o método traz resultados positivos, pois pode auxiliar uma pessoa interessada no mercado de ações em sua análise e estudo para investimento. Também é importante esclarecer a importância do projeto no campo de inteligência artificial, uma vez que traz uma nova linha científica a ser melhorada e explorada com a junção de mineração de dados e bolsa de valores, lembrando que já existem outros projetos com pesquisas relacionadas a junção destes dois ramos de estudo, mas não com as finalidades aqui concluídas.

Por fim, a elaboração deste projeto acarretou na construção de um artigo que será publicado no congresso *IEEE World Congress on Computational Intelligence*, que será realizado na cidade do Rio de Janeiro em julho de 2018. Artigo este que será usado para contar pontos entre outros fatores para a ingressão como bolsista em um mestrado de inteligência artificial na Pontifícia Universidade Católica do Paraná, (PUC-PR).

8.0 - POSSIBILIDADES DE TRABALHOS FUTUROS

Com a conclusão deste projeto, pretende-se lançar o mesmo como artigo científico para o congresso *IEEE World Congress on Computational Intelligence* que acontecerá no Rio de Janeiro em 2018.

Este projeto também será utilizado para estudos afins de um mestrado que será realizado na PUC- PR (Universidade Católica de Pontifica - Paraná) se obter-se a devida bolsa de estudos para a realização do mesmo, o que se tem boas chances de acontecer, pois a nota recebida do concurso que permite a ingressão na pós-graduação, a POSCOMP foi uma nota considerada boa, que traz boas chances de aprovação. O mesmo pode auxiliar nos estudos que irão decorrer na área estudada, já que será de alguma forma implementado novos métodos ou técnicas para alavancar resultados melhores ou extrair novos resultados no decorrer da graduação do mestrado. Parte do presente trabalho já está sendo analisado por professores da área de Inteligência Artificial na PUC-PR.

9.0 – REFERÊNCIAS BIBLIOGRÁFICAS

Goldschmidt; Bezerra. (2016). Exemplos de aplicações de data mining no mercado brasileiro. Disponível em: http://computerworld.com.br/exemplos-de-aplicacoes-de-data-mining-no-mercado-brasileiro

Veloso; Moreira; Silva; Silva. (2011). Data mining, seus benefícios, utilizações, metodologia, campo de atuação dentro de grandes e pequenas empresas. Disponível em: http://periodicos.unifacef.com.br/index.php/resiget/article/download/154/12

CORTES, Sergio da Costa; PORCARO, Rosa Maria; LIFSCHITZ, Sergio. **Mineração de Dados – Funcionalidades, Técnicas e Abordagens**. 2002. Monografia apresentada Universidade PUC-Rio para a obtenção de Doutorado em Ciência da Informação.

MARANGONI, Pedro Henrique. **Redes neurais artificiais para previsão de séries temporais no mercado acionário**. 2010. Monografia apresentada Universidade Federal de Santa Catarina para a obtenção da graduação de Ciências Econômicas.

AMO, Sandra de. **Técnicas de Mineração de Dados**. 2004. (PhD em Ciência da Computação) - XXIV Congresso da Sociedade Brasileira de Computação. Jornada de Atualização em Informática, Salvador, Brasil.

BERESTEIN, Marcelo. **Uso de Mineração de Dados na Bolsa de Valores.** 2010. Monografia apresentada Universidade do Vale do Itajai Centro de Ciências Tecnológicas da Terra e do Mar para a obtenção da graduação de Ciência da Computação.

HOSOKAWA, Eric Ossamu. **Técnica de Arvore de Decisão em Mineração de Dados**. 2011. Monografia apresentada Faculdade de Tecnologia de São Paulo para a graduação de Tecnologia em Processamento de Dados.

GRANATYR, Jones. 7 Técnicas de inteligência artificial para profissionais de TI ganharem mais dinheiro. 2017. (Doutor em Ciência da Computação).

BTG. (2017) Bolsa de valores: o que é, como funciona e como investir. Disponível em: https://www.btgpactualdigital.com/blog/investimentos/tudo-sobre-bolsa-de-valores/

Tororadar. (2016) Aprenda o que é e como funciona a Bolsa de Valores. Disponível em: https://www.tororadar.com.br/blog/bolsa-de-valores-como-funciona

ROQUE, Reginaldo do Carmo. **Estudo sobre a empregabilidade da previsão de índice BOVESPA usando redes neurais artificiais**. 2009. Monografia apresentada Universidade Federal do Rio de Janeiro para a obtenção de graduação em Departamento de Eletrônica e de Computação.

VELLEGO, Bruno Guimarães Motta. **Sistemas de negociação eletrônica automatizada**. 2011. Monografia apresentada Faculdade de Tecnologia de São Paulo para a obtenção de graduação em Processamento de Dados.

LINHARES, Bruno Guarino; PASSOS, Raphael Gonzalez. **BM&F BOVESPA: um valuation baseado no aumento da pressão competitiva no mercado de capitais brasileiro**. 2013. Monografia apresentada Universidade Federal do Rio de Janeiro – Escola Politécnica para a graduação em Engenharia de Produção.

PEASE, Brian Sigaud. **Análise fundamentalista – caso BM&FBOVESPA**. 2011. Monografia apresentada Pontifíca Universidade Católica do Rio de Janeiro para a obtenção de graduação em Economia.

ALMEIDA, Túlio Grizende e. **Otimização de carteiras de investimentos utilizando o modelo de Elton-Gruber**. 2010. Monografia apresentada Universidade Federal de Juiz de Fora para a obtenção de graduação em Engenheiro de Produção.

INFOMONEY. (2005) Como funciona o mercado de renda fixa. Disponível em: http://www.infomoney.com.br/educacao/guias/noticia/368197/como-funciona-mercado-renda-fixa

APRENDAINVESTIMENTO. (2016) Trader ou Investidor consistente? Disponível em: http://aprendainvestimentos.com/trader/

KRIEGER, Paulo Eduardo. **Uso de redes neurais para a predição da bolsa de valores**. 2012. Monografia apresentada Universidade do Vale do Itajaí Centro de Ciências Tecnológicas da Terra e do Mar para a obtenção da graduação de Ciência da Computação.

BRANCO, Gustavo Mendonça do Rio; BARROSO, Marcos André Rosendo. **Mining StockTec: Predição de preço de ações através de mineração de dados e análise de sentimentos**. 2014. Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) para a obtenção da graduação de Sistemas de Informação.

THESTREET. (2012) How Traders Are Using Text and Data Mining to Beat the Market. Disponível em: https://www.thestreet.com/story/13044694/1/how-traders-are-using-text-and-data-mining-to-beat-the-market.html

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2008.

EUQUEROINVESTIR. (2017) Quem Somos Disponível em: https://www.euqueroinvestir.com/quem-somos/

BASTTER. (2017) Grupos Disponível em: https://www.bastter.com/mercado/grupos/default.aspx

TORORADAR. (2017) ToroRadar Disponível em: https://app.tororadar.com.br/ui/#/analises/eqtl3

IBM. (2010) Mineração de dados com WEKA, Parte 1: Introdução e regressão Disponível em: https://www.ibm.com/developerworks/br/opensource/library/os-weka1/

PACHIAROTTI, Juan Francisco Beis. **Aplicação de Técnicas de Mineração de Dados no aprimoramento do Atendimento Médico em um Cenário de Plano de Saúde**. 2012. Monografia apresentada Universidade de Vila Velha – ES para a obtenção da graduação de Bacharel em Ciência da Computação.

PEREIRA, Rafael Barros. **Seleção Lazy de Atributos Para a Tarefa de Classificação.** 2009. Monografia apresentada Universidade Federal Fluminense para a obtenção de título de Mestre na área de Inteligência Artificial.

WEKA. (2017) Classe Principais Componentes. Disponível em: http://weka.sourceforge.net/doc.dev/weka/attributeSelection/PrincipalComponents.html

WEKA. (2017) Classe Discretize. Disponível em: http://weka.sourceforge.net/doc.dev/weka/filters/unsupervised/attribute/Discretize.html

DR SAED SAYAD. (2010) ZeroR Disponível em: http://chem-eng.utoronto.ca/~datamining/dmc/zeror.htm

BTGPACTUAL. (2017) Investimento a longo prazo: Vantagens e onde investir. Disponível em: https://www.btgpactualdigital.com/blog/investimentos/investimento-longo-prazo

FARIA. (2009) Quadro Cognitivo. Disponivel em: http://danuzafaria.pbworks.com/w/page/16789711/quadro%20cognitivo%2009

RESEARCHGATE. (2012) Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações Disponivel em: https://www.researchgate.net/publication/275344764_Mineracao_de_Dados_Educacionais_Conceitos_Tecnicas_Ferramentas_e_Aplicacoes

SANTOS, MIKAMI, VENDRAMIN, KAESTNER, Luciano Drosda M, Renê, Ana Cristina B. V, Celso Antonio A. **Procedimentos de validação cruzada em mineração de dados para ambiente de computação paralela**. 2009. ARTIGO REALIZADO PELA UNIVERSIDADE UTF-PR CURITIBA

DAMASCENO, Marcelo. **Introdução a Mineração de Dados utilizando o Weka**. 2017. ARTIGO REALIZADO PELO INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO RIO GRANDE DO NORTE

EXAME. (2013) Como funciona a análise fundamentalista de ações. Disponível em: https://exame.abril.com.br/seu-dinheiro/como-funciona-analise-fundamentalista-acoes-576374/

TORORADAR. (2017) O que é análise fundamentalista. Disponível em: https://www.tororadar.com.br/investimento/analise-fundamentalista/o-que-e

ANEXOS