

**Universidade Federal de Ouro Preto
Departamento de Computação e
Sistemas
Curso Engenharia de Computação**



**Uso de Técnicas de Mineração de
Dados para Encontrar Tendências
em Mercados Financeiros**

Thayse Cristina Araújo Rodrigues

**TRABALHO DE
CONCLUSÃO DE CURSO**

ORIENTAÇÃO:
Prof^a. MSc. Janniele Aparecida Soares Araujo

**Agosto, 2016
João Monlevade/MG**

Thayse Cristina Araújo Rodrigues

**Uso de Técnicas de Mineração de Dados para
Encontrar Tendências em Mercados Financeiros**

Orientadora: Prof^a. MSc. Janniele Aparecida Soares Araujo

Monografia apresentada ao Curso de Engenharia de Computação do Departamento de Computação e Sistemas, como requisito parcial para aprovação na Disciplina Trabalho de Conclusão de Curso II.

Universidade Federal de Ouro Preto
João Monlevade
Agosto de 2016

R696u

Rodrigues, Thayse Cristina Araújo.

Uso de técnicas de mineração de dados para encontrar tendências em mercados financeiros [manuscrito] / Thayse Cristina Araújo Rodrigues. - 2016.

79f.: il.: color; tabs.

Orientador: Prof. Dr. Janniele Aparecida Soares Araujo.

Monografia (Graduação). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Departamento de Computação e Sistemas de Informação.

1. Mineração de dados (Computação). 2. Mercados Financeiros. 3. Tratamento de dados. 4. Administração de dados. I. Araujo, Janniele Aparecida Soares . II. Universidade Federal de Ouro Preto. III. Título.

CDU: 004.451.5

Catálogo: ficha@sisbin.ufop.br



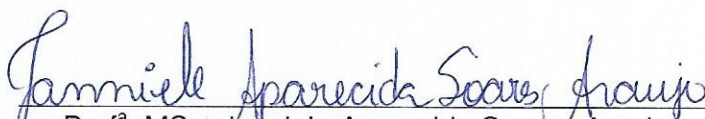
Curso Engenharia de Computação

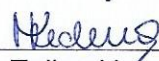
FOLHA DE APROVAÇÃO DA BANCA EXAMINADORA

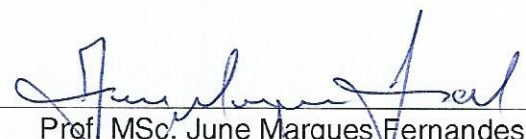
Uso de Técnicas de Mineração de Dados Para Encontrar Tendências em Mercados Financeiros

Thayse Cristina Araújo Rodrigues

Monografia apresentada ao Departamento de Computação e Sistemas da Universidade Federal de Ouro Preto como requisito parcial da disciplina CSI496 – Trabalho de Conclusão de Curso II, do curso de Bacharelado em Engenharia de Computação e aprovada pela Banca Examinadora abaixo assinada:


Prof.^a MSc. Janniele Aparecida Soares Araujo
Departamento de Computação e Sistemas - UFOP


Prof. MSc. Talles Henrique de Medeiros
Departamento de Computação e Sistemas - UFOP


Prof. MSc. June Marques Fernandes
Departamento de Engenharia de Produção - UFOP

João Monlevade, 11 de Agosto de 2016.



Curso Engenharia de Computação

TERMO DE RESPONSABILIDADE

Eu, Thayse Cristina Araújo Rodrigues, declaro que o texto do trabalho de conclusão de curso intitulado "*Uso de Técnicas de Mineração de Dados Para Encontrar Tendências em Mercados Financeiros*" é de minha inteira responsabilidade e que não há utilização de texto, material fotográfico, código fonte de programa ou qualquer outro material pertencente a terceiros sem as devidas referências ou consentimento dos respectivos autores.

João Monlevade, 11 de Agosto de 2016.


Assinatura do aluno



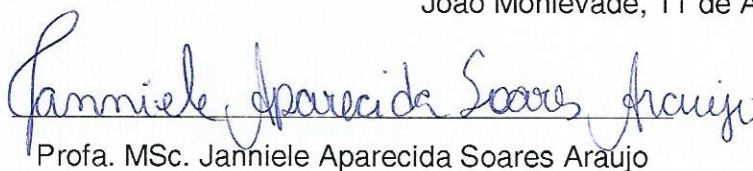
ATA DE DEFESA

Aos 11 dias do mês de Agosto de 2016, às 16 horas, na sala C304 do Instituto de Ciências Exatas e Aplicadas, foi realizada a defesa de Monografia pelo aluno **THAYSE CRISTINA ARAÚJO RODRIGUES**, sendo a Comissão Examinadora constituída pelos professores: Prof^a. MSc. Janniele Aparecida Soares Araujo, Prof. MSc. Talles Henrique de Medeiros e Prof. MSc. June Marques Fernandes.

O candidato apresentou a monografia intitulada: *"Uso de Técnicas de Mineração de Dados Para Encontrar Tendências em Mercados Financeiros"*. A comissão examinadora deliberou, por unanimidade, pela aprovação do candidato, com nota 10 (dez), concedendo-lhe o prazo de 15 dias para incorporação das alterações sugeridas ao texto final.

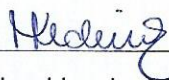
Na forma regulamentar, foi lavrada a presente ata que é assinada pelos membros da Comissão Examinadora e pelo graduando.

João Monlevade, 11 de Agosto de 2016.



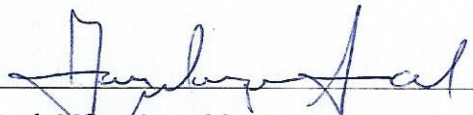
Prof. MSc. Janniele Aparecida Soares Araujo

Professor Orientador/Presidente



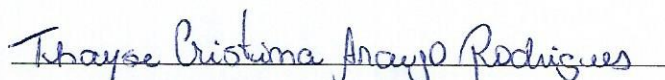
MSc. Talles Henrique de Medeiros

Professor Convidado



Prof. MSc. June Marques Fernandes

Professor Convidado



Thaysse Cristina Araújo Rodrigues

Graduando

RESUMO

O presente trabalho apresenta um estudo resultante da aplicação de mineração de dados na base de dados de resultados de empresas que atuam na bolsa de valores BM&FBovespa. Empresas de capital aberto divulgam indicadores que são usados pelos investidores para a avaliação das ações no mercado financeiro. Foram realizados experimentos com os demonstrativos de resultados, balanço patrimonial e informações de cotação dos papéis das empresas. De posse destas informações foram calculados alguns indicadores e utilizados algoritmos de classificação para analisar os resultados obtidos sobre a base de dados que foi montada. As empresas foram classificadas como investimento "Fraco", "Bom" ou "Muito Bom" comparando seus rendimentos com a taxa Selic no período pesquisado. São apresentadas ao leitor os algoritmos de classificação usados, os atributos gerados na base de dados, gráficos e tabelas para comparar o desempenho dos algoritmos. Com os indicadores fundamentalistas e os algoritmos de mineração de dados usados no trabalho foi obtida uma taxa alta de acertos para a classe "Fraco", mas poucas empresas da classe "Bom" e "Muito Bom" foram classificadas corretamente.

Palavras-chaves: Mineração de dados, algoritmo *KNN*, *Árvores de Decisão*, banco de dados, regras de classificação, mercado financeiro, bolsa de valores.

ABSTRACT

This work introduces a study resulting from application of data mining in an BM&FBovespa database with companies' results. Companies disclose their indicators into stock market and are used for investors evaluate stocks. Experiments were performed with income statements, balance sheets and price informations. Were used classification algorithms to analyze the data generated on the database. The companies were classified as "Weak", "Good" or "Very Good" comparing your incomes with the Selic rate during the studied period. This work presents the classification algorithms used, the attributes generated in the database, graphics and spreadsheets to compare the algorithms performance. With the fundamentalists indicators and data mining algorithms used in the study, was obtained a high rate of correct answers to the "Weak" class, but just a few companies of "Good" class and "Very Good" were classified correctly.

Keywords: data mining, KNN algorithm, decision trees, database, classification rules, financial market, stock exchange.

AGRADECIMENTOS

Gostaria de agradecer ao meu marido Alessandro dos Santos Alves, pelo apoio e companhia nos momentos felizes e difíceis que tive ao longo desta longa caminhada. Agradeço aos meus pais Joaquim Cassiano Rodrigues e Maria de Lourdes Cassiano Araújo pela força, paciência e carinho que sempre tiveram comigo e com meus irmãos. Agradeço a Deus pela força que diversas vezes pedi para conseguir continuar meu caminho com determinação.

Agradeço aos meus irmãos Thiago Cesar Araújo Rodrigues e Luciana Araújo Cassiano pela companhia durante toda minha vida. Agradeço a minhas amigas pelo apoio que foram inestimáveis neste período, em especial a Fabrícia Batista, Elisângela Assis, Graciela Assis, Lorena Santos, Lorraine Campos e Mabiane Sena. Apesar de termos nos distanciado um pouco elas sempre estão em doces lembranças, que me servem de força e estímulo para seguir em frente. Muito obrigada por existirem!

Agradeço a todas as pessoas que fizeram parte da minha caminhada durante o curso na UFOP, a todos os amigos, professores e funcionários que dedicam seu trabalho a manter esta importante instituição. Sou muito grata a todos os professores pelo conhecimento que dividem com os alunos. Aprendi muito aqui!

Agradecimento especial a minha orientadora professora Janniele, professor Talles e à professora Francisca, que me ajudaram a desenvolver este tema e dividir tantas informações valiosas comigo para a conclusão deste trabalho.

SUMÁRIO

1	Introdução	17
1.1	Objetivos	18
1.1.1	Objetivo Geral	18
1.1.2	Objetivos Específicos	18
2	Revisão Bibliográfica	19
2.1	Mineração de Dados	19
2.2	Tarefa de Classificação	21
2.3	Algoritmos Usados e Medidas de Desempenho	21
2.3.1	IBK	23
2.3.2	Árvores de Decisão - Algoritmo J48 e <i>REPTree</i>	25
2.3.3	Algoritmos de Comitê - <i>Random Committee</i> , <i>Random Forest</i> e <i>Bagging</i>	25
2.3.4	<i>Logit Boost</i>	26
2.3.5	Classificação via Regressão	27
2.3.6	<i>MultiClass Classify</i>	27
2.4	Teoria de Finanças	28
2.5	Bolsa de Valores BMF&Bovespa	29
2.5.1	Tipos de Ações e Códigos de Negociação	30
2.6	Trabalhos Relacionados	31
2.7	<i>WEKA</i>	34
2.8	Ambiente de Desenvolvimento	35
2.9	Indicadores Fundamentalistas e Medidas de Retorno	36
3	Desenvolvimento	42
3.1	Seleção dos Dados	42
3.2	Pré-Processamento e Transformação dos Dados	45
3.2.1	Classificação das Empresas	49
3.2.2	Banco de Dados Bovespa	51
3.2.3	Sumarização dos Dados	52
3.2.4	Dificuldades Encontradas	56
3.3	Interface Gráfica	58
3.4	Aplicação dos Algoritmos	60
3.4.1	Técnica de <i>PCA</i>	64
3.4.2	Experimentos	67

3.4.3 Estatística Kappa	73
4 Considerações Finais	76
4.1 Trabalhos Futuros	77
Referências	78

LISTA DE FIGURAS

Figura 1 – Passos do processo de <i>KDD</i> - Fonte: (Fayyad et al, 1996)	19
Figura 2 – Representação dos valores de <i>k</i>	24
Figura 3 – Resultado comparativo do uso das técnicas de Mineração de Dados - Fonte: Imandoust and Bolandraftar (2014)	34
Figura 4 – Arquivo de dados gerado em formato <i>arff</i>	35
Figura 5 – Taxa Selic, CDI e Rendimento da Poupança nos anos de 2005 a 2015. . .	50
Figura 6 – Base de dados Bovespa	51
Figura 7 – Percentual de empresas em cada classe nos períodos pesquisados. . . .	53
Figura 8 – Percentual de classificação de cada setor nos períodos analisados. . . .	53
Figura 9 – Percentual de classificação "Bom" e "Muito Bom" de cada setor no pe- ríodo mensal.	54
Figura 10 – Frequência em meses que cada empresa teve classificação "Bom" e "Muito Bom".	56
Figura 11 – Menu de interface com os usuários.	59
Figura 12 – Menu de interface com os usuários.	60
Figura 13 – Matriz de confusão do algoritmo <i>Random Tree</i>	61
Figura 14 – Matriz de confusão do algoritmo <i>Random Forest</i> aplicado aos dados do setor de Bens Industriais.	62
Figura 15 – Matriz de confusão do algoritmo <i>Random Committee</i> aplicado ao setor Financeiro e Outros.	63
Figura 16 – Distribuição de frequência dos algoritmos por colocação.	64
Figura 17 – PCA aplicado a base de dados.	65
Figura 18 – Matriz de confusão do algoritmo <i>Random Tree</i>	66
Figura 19 – <i>WEKA</i> seleção dos indicadores identificados no PCA.	67
Figura 20 – Resultados do algoritmo <i>IBK</i> aplicado aos dados semanais.	68
Figura 21 – Resultados do algoritmo <i>IBK</i> aplicado aos dados mensais.	69
Figura 22 – Resultados do algoritmo <i>IBK</i> aplicado aos dados Semestrais.	69
Figura 23 – Resultados do algoritmo <i>Random Committee</i> aplicado aos dados sema- nais.	70
Figura 24 – Resultados do algoritmo <i>Random Committee</i> aplicado aos dados mensais. .	71
Figura 25 – Resultados do algoritmo <i>Random Committee</i> aplicado aos dados Se- mestrais.	71
Figura 26 – Resultados do algoritmo <i>Random Tree</i> aplicado aos dados semanais. . .	72
Figura 27 – Resultados do algoritmo <i>Random Tree</i> aplicado aos dados mensais. . .	72
Figura 28 – Resultados do algoritmo <i>Random Tree</i> aplicado aos dados Semestrais. .	73

Figura 29 – Valores da estatística Kappa dos algoritmos aplicados aos dados semanais.	74
Figura 30 – Valores da estatística Kappa dos algoritmos aplicados aos dados mensais.	75
Figura 31 – Valores da estatística Kappa dos algoritmos aplicados aos dados semes- trais.	75

LISTA DE TABELAS

Tabela 1 – Código dos Tipos de Ação	31
Tabela 2 – Ativos	44
Tabela 3 – Empresas que tiveram rendimento acima de 80% da Selic em, pelo menos, um mês nos últimos 10 anos	55
Tabela 4 – Empresas que tiveram maior frequência de meses com rendimento acima de 120% da Selic nos últimos 10 anos	57
Tabela 5 – Empresas que tiveram maior frequência de meses com rendimento acima de 80% da Selic nos últimos 10 anos	57
Tabela 6 – Empresas que tiveram maior rendimento acumulado nos últimos 10 anos.	58
Tabela 7 – Resultado dos algoritmos de mineração aplicados aos dados totais. . . .	61
Tabela 8 – Resultado dos algoritmos de mineração aplicados aos dados do Setor de Bens Industriais.	62
Tabela 9 – Resultado dos algoritmos de mineração aplicados aos dados do Setor de Construção e Transporte.	62
Tabela 10 – Resultado dos algoritmos de mineração aplicados aos dados do Setor de Consumo Cíclico.	62
Tabela 11 – Resultado dos algoritmos de mineração aplicados aos dados do Setor Financeiro e Outros.	63
Tabela 12 – Resultado dos algoritmos de mineração aplicados aos dados do Setor de Materiais Básicos.	63
Tabela 13 – Resultado dos algoritmos de mineração aplicados aos dados do Setor de Telecomunicações.	63
Tabela 14 – Resultado dos algoritmos de mineração aplicados aos dados do Setor de Utilidade Pública.	64
Tabela 15 – Resultado dos algoritmos de mineração aplicados aos dados das empresas que tiveram pelo menos um mês com rendimento acima de 80% da Selic.	66
Tabela 16 – Resultado dos algoritmos de mineração aplicados aos dados das empresas que tiveram pelo menos um mês com rendimento acima de 80% da Selic e somente usando os atributos principais.	67
Tabela 17 – Interpretação dos valores de Kappa. Fonte: Landis and G. (1977)	74

LISTA DE ABREVIATURAS E SIGLAS

KDD - Knowledge Discovery in Databases

KNN - K-nearest neighbors ou K-vizinhos mais próximos

EMH - Hipótese de Eficiência dos Mercados

P/L - Preço/Lucro

VPA - Valor Patrimonial da Ação

P/VP - Preço / Valor Patrimonial

SGBD - Sistema de Gerenciamento de Banco de Dados

SQL - Structured Query Language

POM - *Project Object Model*

API - *Application Programming Interface*

ORM - *object-relational mapping*

JDBC - *Java Database Connectivity*

JPA - *Java Persistence API*

POJO - *Plain Old Java Objects*

HQL - *Hibernate Query Language*

JSF - *JavaServer Faces*

MVC - *Model View Controller*

BDRs - *Brazilian Deposits Receipts*

ETFs - *Exchange Traded Funds*

CCI - *Correctly Classified Instances*

MAE - *Mean Absolute Error*

RMSE - *Root Mean Squared Error*

RAE - *Relative Absolute Error*

RRSE - *Root Relative Squared Error*

P/L - Preço/Lucro

VPA - Valor Patrimonial da Ação

P/VP - Preço / Valor Patrimonial

JCP - Juros Sobre Capital Próprio

CFS - *Cash Flow Share*

P/CFS - Preço/Geração de Caixa

EV - *Enterprise Value*

EBITDA - *Earning Before Interests, Taxes, Depreciation and Amortization*

EBIT - *Earnings before interest and taxes*

ROIC - *Return Over Invested Capital*

NOPLAT - *Net Operating Profits Less Adjusted Taxes*

ROE - *Return on Equity*

Debit/Equity - Indicador que mostra o nível de endividamento da empresa

LPA - Lucro Por Ação

Selic - Sistema Especial de Liquidação e Custódia

CDI - Certificado de Depósito Interbancário

Copom - Comitê de Política Monetária do Banco Central

TP - *True Positive* (Verdadeiro Positivo)

WACC - *Weighted Average Capital Cost*(Custo Médio Ponderado de Capital)

EVA - *Enterprise Value* (Valor do Empreendimento)

1 INTRODUÇÃO

Bolsas de valores são instituições econômicas que têm como principal objetivo garantir negociações justas e equitativas entre investidores, promovendo a livre concorrência dos mesmos. Proporcionam liquidez aos ativos negociados, e transparência na fixação de preços. Possuem papel importante em sistemas econômicos, sendo ferramenta facilitadora de crescimento para empresas pela transferência de recursos, e servindo como canalizador de poupança para pequenos e grandes investidores, contribuindo, portanto, para geração de emprego e renda (Pinheiro, 2009).

A Bolsa de Valores gera diariamente grandes volumes de informação, cuja análise é complexa e não trivial para um investidor não profissional. Existem muitas variáveis que influenciam a decisão a tomar sobre a carteira de ações (vender, manter, comprar). Estas decisões têm como objetivo a maximização do lucro. Então, a proposta deste trabalho é coletar informações das empresas que negociam ações na BM&FBovespa, montar uma base de dados contendo estas informações, calcular indicadores para compor a base de dados e utilizar mineração de dados nestas informações a fim de verificar se a mineração de dados consegue prever a classificação das empresas através destes indicadores.

Uma forma reduzir os riscos de investimentos em ações é, por meio da análise das demonstrações contábeis e avaliação de desempenho, utilizando fundamentos econômicos-financeiros e de mercado para avaliar se é interessante ou não investir em uma determinada empresa. De acordo com Matias (2009), o uso de indicadores contábeis e financeiros para avaliação de resultados, com o objetivo de formar uma carteira de investimentos, ficou conhecida como análise fundamentalista.

Para a abordagem deste trabalho foram usadas regras de classificação. Classificação é a tarefa de organizar objetos em uma entre diversas categorias pré-definidas. Utilizando os índices presentes na BM&FBovespa para classificar as empresas nas classes "Muito Bom", "Bom" e "Fraco" através da comparação de seu rendimento com o rendimento da taxa Selic.

Para realizar a tarefa de classificação na Mineração de Dados existem diversas técnicas e algoritmos. Foram usados dez algoritmos de classificação neste trabalho sendo eles *Random Committee*, *Random Tree*, *Ibk*, *Random Forest*, *Bagging Meta*, *Classification Via Regression*, *Logit Boost*, *REPTree*, *J48* e *MultiClass Classifier*.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Este trabalho tem como objetivo geral classificar empresas de acordo com o cálculo de indicadores que representam liquidez, estrutura de capital, rentabilidade e índices de mercado. A partir de tais informações são utilizados algoritmos de mineração de dados e a técnica de validação cruzada para classificar as empresas.

1.1.2 Objetivos Específicos

Buscando o objetivo geral acima mencionado, é necessário atingir os objetivos específicos seguintes:

- Fazer levantamento na literatura sobre indicadores de desempenho de empresas na bolsa de valores BM&FBovespa;
- Construir base de dados com uma amostra consistente de empresas e calcular seus respectivos indicadores;
- Desenvolvimento de um software utilizado como ferramenta para construção da base de dados e exportação das informações no formato reconhecido pelo Weka;
- Pesquisa e escolha dos indicadores que serão usados no processo de classificação das empresas, definindo as melhores variáveis a serem usadas para classificar as empresas;
- Classificar as empresas usando os algoritmos *Random Committee*, *RandomTree*, *Ibk*, *Random Forest*, *Bagging Meta*, *Classification Via Regression*, *Logit Boost*, *REPTree*, *J48* e *MultiClass Classifier* e comparar seus resultados;
- Apresentação dos resultados e geração de gráficos conclusivos.

2 REVISÃO BIBLIOGRÁFICA

2.1 MINERAÇÃO DE DADOS

Avanços rápidos na tecnologia de coleta e armazenagem de dados permitem que organizações armazenem vasta quantidade de dados. Portanto, com tanta informação disponível tornou-se necessário melhorar técnicas de análise. A extração de informação útil, entretanto, tem provado ser bastante desafiadora.

O termo *Knowledge-Discovery in Databases* (Descoberta de Conhecimento em Base de Dados ou *KDD*) surgiu no primeiro *KDD work-shop em 1989* (Shapiro, 1991). Segundo Fayyad et al (1996), veio para enfatizar que conhecimento é o produto final de *Data-driven Discovery*, se popularizando na inteligência artificial e aprendizagem de máquina. Como apresentado na Figura 1, os passos básicos do *KDD* são estabelecidos como:

- Seleção (*Selection*): passo em que é selecionado o conjunto ou subconjunto de dados onde será aplicada a descoberta;
- Pré-processamento (*Preprocessing*): passo onde são eliminados ruídos;
- Transformação (*Transformation*): passo onde ocorre redução ou transformação dos dados;
- Mineração dos dados (*Data Mining*): passo onde é feita a busca por padrões interessantes, utilizando métodos específicos;
- Interpretação e avaliação dos resultados (*Interpretation/Evaluation*): passo onde são interpretados os padrões minerados, podendo ser necessário voltar a algum passo anterior.

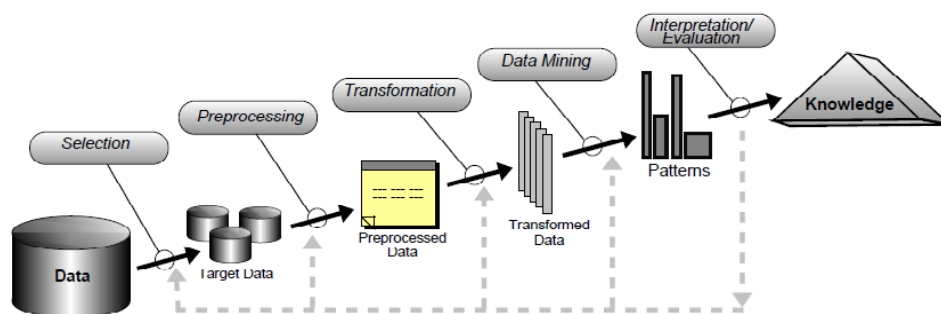


Figura 1 – Passos do processo de *KDD* - Fonte: (Fayyad et al, 1996)

Os objetivos do *KDD* são definidos pela intenção de uso do sistema. Pode-se distinguir dois tipos de objetivos:

- Verificação: onde o sistema é limitado a verificar hipóteses;
- Descoberta: onde o sistema descobre novos padrões.

O foco do presente trabalho será em descoberta. Segundo [Fayyad et al \(1996\)](#), os objetivos da descoberta são divididos em:

- Preditivos: onde sistemas encontram padrões que podem prever comportamentos futuros;
- Descritivas: onde sistemas encontram padrões com o propósito de apresentá-los de forma compreensível ao entendimento humano.

Segundo [Fayyad et al \(1996\)](#), os objetivos da predição e descrição podem ser alcançados usando métodos específicos da mineração de dados. Sendo eles:

- Classificação: função de aprendizagem que mapeia um item de dados a partir de vários itens pré-definidos;
- Regressão: função de aprendizagem que mapeia um item de dados para uma predição de valor real e descoberta de relacionamentos funcionais entre variáveis;
- Agrupamento: identificação de um conjunto finito de categorias ou grupos para descrever dados;
- Sumarização: descoberta de descrição compacta para um subconjunto de dados, derivação de resumos ou regras de associação e o uso de técnicas de visualização multivariáveis;
- Modelagem de dependência: descoberta de um modelo que descreve dependências significativas entre variáveis;
- Mudança e Detecção de desvio: descoberta da mudança mais significativa nos dados medidos anteriormente ou valores normativos.

O *KDD* refere-se a todo o processo de descobrimento, enquanto *Data Mining* ou Mineração de dados refere-se a um passo do processo que consiste na aplicação de análise sobre os dados a partir de algoritmos que respeitem as limitações da capacidade computacional produzindo assim padrões ([Camilo and Silva, 2009](#)).

Segundo Tan et al (2009) Mineração de Dados é o processo de descoberta automática de informações úteis em grandes bases de dados. As técnicas de Mineração de Dados são organizadas para agir sobre grandes bancos de dados com o intuito de descobrir padrões úteis e recentes que poderiam, de outra forma, permanecer ignorados. Elas também fornecem capacidade de previsão do resultado de uma observação futura.

O método escolhido para o trabalho é a classificação. O objetivo é classificar empresas de acordo com análise de seus indicadores calculados através de informações disponibilizadas no site da BM&FBovespa e também por consultas a outros sites de divulgação de dados financeiros.

2.2 TAREFA DE CLASSIFICAÇÃO

Classificação é a tarefa de organizar objetos em uma entre diversas categorias pré-definidas, é um problema universal que engloba muitas aplicações diferentes. Os dados de entrada são um conjunto de registros. Cada registro, também conhecido como uma instância ou exemplo, é caracterizado por uma dupla (x, y) , onde x é o conjunto de atributos e y é o atributo especial, designado como rótulo da classe também chamado de atributo alvo ou de categorização (Tan et al, 2009).

Neste trabalho os índices da bolsa de valores levantados serão conjuntos de atributos enquanto o atributo alvo é a classificação como "Muito Bom", "Bom" ou "Frac". Sendo assim, vemos que classificação é a tarefa de aprender uma função alvo $f(x)$ que mapeie cada conjunto de atributos x para um dos rótulos de classes y pré-definidas (Tan et al, 2009).

Foram usados dez algoritmos sendo eles *Random Committee*, *Random Tree*, *Ibk*, *Random Forest*, *Bagging Meta*, *Classification Via Regression*, *LogitBoost*, *REPTree*, *J48* e *MultiClassClassifier* e em seguida foi comparado o desempenho para ver qual técnica obteve melhor resultado. Estes algoritmos foram escolhidos por alguns deles, como *IBK* e Árvores de Decisão, serem os mais usados nos trabalhos correlacionados e artigos recentes sobre o assunto e por estarem disponíveis no *WEKA*. Em alguns trabalhos são citadas adaptações destes algoritmos com técnicas de análise de mercados financeiros adaptados para este problema. Estes trabalhos e seus resultados serão citados na seção 2.6.

2.3 ALGORITMOS USADOS E MEDIDAS DE DESEMPENHO

Foram usados dez algoritmos sobre a base de dados de atributos criada neste trabalho. Os algoritmos são:

- *IBK* (*Instance-Based K*) também conhecido como algoritmo *KNN* (*K-nearest neigh-*

bors: onde uma instância é classificada comparando sua distância dos K vizinhos mais próximos;

- J48: algoritmo de árvore de decisão.
- *REPTree*: árvore de decisão de aprendizado rápido.
- *Random Tree*: árvore de decisão que considera apenas alguns atributos escolhidos aleatoriamente para cada nó da árvore.
- *Random Committee*: algoritmo de comitê que constrói um grupo de classificadores básicos aleatórios.
- *Random Forest*: algoritmo de comitê composto por um conjunto de árvores de classificação, cada árvore dá um voto que indica sua decisão sobre a classe do objeto, a classe com o maior número de votos é escolhida para o objeto.
- *Logit Boost*: algoritmo para a realização de aditivo de regressão logística e executa a classificação usando um esquema de regressão, conforme a base de aprendizado, e pode lidar com problemas multiclasse.
- *Bagging Meta*: faz o *bagging* de um classificador a fim de reduzir sua variância.
- *Classification Via Regression*: algoritmo que faz a classificação usando métodos de regressão.
- *MultiClass Classifier*: um metaclassificador para a manipulação de dados multiclasse com classificadores 2-classes. Este classificador também é capaz de aplicar códigos de correção de erros de saída para aumentar a precisão.

Em cada um dos algoritmos foram utilizadas as seguintes métricas para comparar seu desempenho:

- CCI (*Correctly Classified Instances*): mostra o percentual de instâncias classificadas corretamente, é uma medida de acurácia do algoritmo que mostra o percentual de instâncias que ele classificou corretamente;
- Recall: taxa que mostra quantos dados foram classificados corretamente. Apresenta um valor por classe e é obtido dividindo a quantidade de *TP* (*True Positive*) da classe pela quantidade de instâncias desta classe (Ian H. Witten, 2011).
- Precisão (*precision*): é obtida dividindo a quantidade de *TP* (*True Positive*) da classe pela quantidade total de instâncias classificadas com esta classe (Ian H. Witten, 2011).

- *F-Measure*: seu resultado depende do domínio do problema e a fórmula de cálculo é mostrada na equação 2.1 mostrada por Ian H. Witten (2011). O autor cita que físicos usam esta medida como descrição da sensibilidade e especificidade dos testes de diagnóstico. Sensibilidade refere-se à proporção de pessoas com a doença que têm um resultado positivo, isto é, *TP (True Positive)*. Especificidade refere-se à proporção de pessoas sem a doença que têm um resultado negativo, que é $1 - FP$ (*False Positive*).

$$F - Measure = \frac{2 * recall * precision}{recall + precision} \quad (2.1)$$

Esta medida é importante para o objetivo deste trabalho, pois a intenção é verificar os algoritmos que apresentaram melhor taxa de acertos e o *F-Measure* é sensível às duas medidas adequadas para medir o desempenho dos algoritmos que são precisão (*precision*) e *recall*. Desta forma, não foram usadas estas medidas separadamente, apenas o *CCI* e o *F-Measure* já foram suficientes.

2.3.1 IBK

O algoritmo *IBK (Instance Based K)* é uma implementação do algoritmo *KNN (K-nearest neighbors* ou classificador de vizinho mais próximo) é um dos algoritmos mais utilizados para mineração de dados, por ser de fácil aplicação e trazer bons resultados. Ele representa cada exemplo como um ponto de dado em um espaço *d*-dimensional, onde *d* é o número de atributos. Dado um exemplo de teste, calculamos sua proximidade com o resto dos pontos de dados no conjunto de treinamento, usando uma medida de proximidade. Os vizinhos mais próximos *K* de um determinado exemplo *Z* se referem aos *K* pontos que estejam mais próximos de *Z* (Tan et al, 2009).

A justificativa de usar o *KNN* como método de classificação, pode ser exemplificada pelo seguinte ditado: “Se caminhar como um pato, se grasnar como um pato e se aparecer com um pato, então provavelmente é um pato” (Tan et al, 2009).

O ponto dado é classificado com base nos rótulos de classe de seus vizinhos. No caso, onde os vizinhos têm mais de um rótulo, o ponto do dado é atribuído à classe majoritária de seus vizinhos mais próximos. Assim é muito importante escolher um valor correto para *K*, pois se for pequeno demais pode ocorrer *overfitting* por causa do ruído nos dados de treinamento. Se *K* for grande demais ele pode classificar erroneamente a instância de teste. A Figura 2 mostra diferentes escolhas de *K* em um espaço bidimensional.

No exemplo mostrado na figura 2 o ponto vermelho é o ponto a ser classificado, caso seja escolhido $K = 1$ ele será classificado como cinza e para $K = 3$ ele será classificado como amarelo. O Algoritmo 1 apresenta como é realizada a classificação de cada item do conjunto de teste:

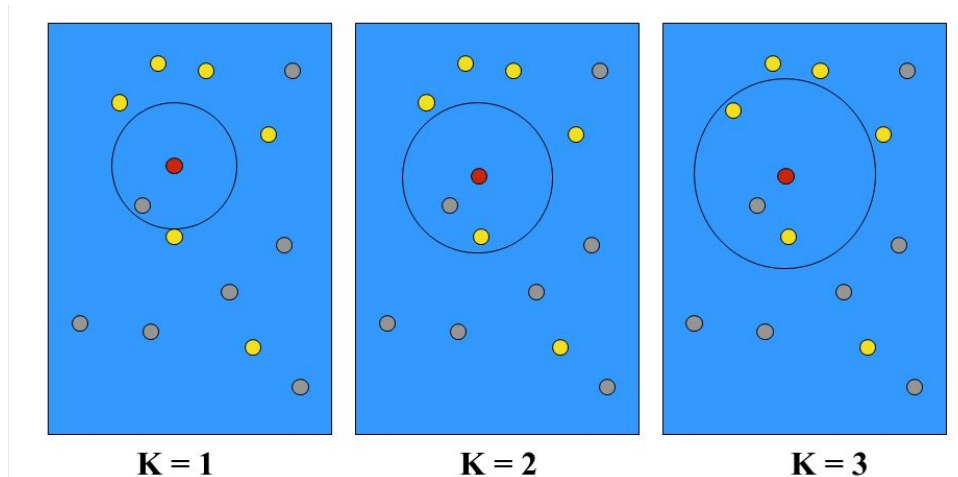


Figura 2 – Representação dos valores de k

Algoritmo 1: O algoritmo de classificação de vizinho mais próximo k.

1: Seja k o número de vizinhos mais próximos e D o conjunto de exemplos de treinamento ;

2: **para** cada exemplo de teste $z=(x',y')$ **faça**

 3: Calcule $d(x',x)$, a distância entre z e cada exemplo $(x,y) \in D$. ;

 4: Seleccion $D_z \subseteq D$ o conjunto dos k exemplos de treinamento para z ;

 5: $y' = \operatorname{argmax} \sum_{(xi,yi) \in D_z} I(v = yi)$;

fim

Onde:

- k : número de vizinhos mais próximos dado pelo usuário;
- D : conjunto de treinamento;
- z : conjunto de teste;
- $d(x',x)$: cálculo de distância entre o item de teste e todas as instâncias de treinamento;
- $y' = \operatorname{argmax} \sum_{(xi,yi) \in D_z} I(v = yi)$: votação majoritária, onde o item de teste é classificado de acordo com a classe predominante dentro dos K vizinhos mais próximos.

O Algoritmo calcula a distância (ou similaridade) entre cada exemplo de teste $z = (x', y')$ e todos os seus exemplos de treinamento $(x,y) \in D$ para determinar sua lista de vizinhos mais próximos, D_z . Assim que a lista de vizinhos mais próximos for obtida, o exemplo de teste é classificado baseado na classe majoritária dos seus vizinhos mais próximos, onde v é o rótulo de classe, y_i é o rótulo de classe para um dos vizinhos mais próximos, e $I(\cdot)$ é uma função indicadora que retorne o valor 1 se seu argumento for verdadeiro ou 0 se for falso.

Ao fim, cada item do grupo de teste é classificado conforme maior similaridade com os indivíduos do grupo de treinamento.

2.3.2 Árvores de Decisão - Algoritmo J48 e *REPTree*

A principal ideia de uma árvore de decisão é dividir os dados recursivamente em subgrupos, então cada subgrupo contém mais ou menos estados homogêneos da sua variável alvo (atributo a ser previsto). A cada divisão na árvore, todos os atributos de entrada são avaliados por seu impacto na variável alvo. Quando o processo recursivo é completado, a árvore de decisão é formada.

A divisão pode ser feita fazendo perguntas e então classificando os dados em subgrupos de acordo com a resposta à pergunta. Cada vez que recebe-se uma resposta uma questão seguinte é feita, até que se chegue a uma conclusão sobre o rótulo da classe do registro. A série de questões e suas respostas possíveis podem ser organizadas na forma de uma árvore de decisão, com sua estrutura hierárquica consistindo de nodos e arestas direcionadas (Tan et al, 2009).

Em uma árvore de decisão cada nodo folha recebe um rótulo de classe. Os nodos não terminais, que incluem os nodos raiz e outros nodos internos, contém condições de testes de atributos para separar registros que possuam características diferentes (Tan et al, 2009).

Classificar um registro de testes é direto, assim que uma árvore de decisão tenha sido construída. Começando do nodo raiz, é aplicada a condição de testes ao registro e segue-se a ramificação adequada baseada no resultado do teste. Isto levará a um outro nodo interno, para o qual uma nova condição de teste é aplicada, ou a um nodo folha. O rótulo da classe associada ao nodo folha é então atribuída ao registro (Tan et al, 2009).

O *WEKA* possui algoritmos que implementam árvores, neste trabalho são utilizados o J48 e o *REPTree*. O J48 implementa um modelo de árvore de decisão conhecido como C4.5. As árvores de decisão geradas pelo algoritmo C4.5 podem ser utilizadas para classificação e são portanto conhecidas como classificadores estatísticos (Wikipédia, 2016)

REPTree é uma árvore de decisão de aprendizado rápido. Este algoritmo constrói uma árvore de decisão/regressão com base em informações de ganho/variância e a reduz usando a poda de redução de erros (com *backfitting*). A ordenação de valores de atributos numéricos é realizada apenas uma vez. Os valores em falta são tratados dividindo as instâncias correspondentes em pedaços.

2.3.3 Algoritmos de Comitê - *Random Committee*, *Random Forest* e *Bagging*

Na vida real, quando se tem decisões importantes a tomar, muitas vezes se optam por usar um comitê. A existência de diferentes especialistas, com diferentes perspectivas

sobre o problema é, muitas vezes, uma maneira muito eficaz e robusta de tomar boas decisões. O mesmo é verdadeiro na aprendizagem de máquina, pode melhorar o desempenho preditivo ter métodos diferentes de aprendizado de máquina, todos trabalhando no mesmo problema, e depois realizar uma votação para classificar uma instância de teste desconhecida (of Computer Science at the University of Waikato, 2016).

Random Committee constrói um grupo de classificadores básicos aleatórios. Cada classificador de base é construído usando uma semente de número aleatório diferente (mas com base nos mesmos dados). A previsão final é uma média linear das previsões geradas pelos classificadores de base individuais.

Métodos de comitê visam melhorar o desempenho da classificação construindo uma combinação da saída de vários classificadores, em vez de aplicar um único modelo. O resultado da combinação é melhor do que o resultado de qualquer classificador base individual pertencente à combinação (Ramos and dos Santos, 2011).

Random Forest é um algoritmo de comitê que usa um conjunto de árvores de classificação. Cada árvore dá um voto que indica sua decisão sobre a classe do objeto e a classe com o maior número de votos é escolhida (Ramos and dos Santos, 2011).

O *Random Forest* usa algoritmos de árvore de decisão, selecionando o melhor atributo para dividir em cada interação. Faz-se a escolha de algumas das melhores opções, e aleatoriamente escolhe entre elas. Isso gera árvores diferentes a cada interação e, geralmente, apresenta um melhor desempenho.

Com *Bagging* pode-se produzir várias estruturas de decisão diferentes. Usa J48 para produzir árvores de decisão ligeiramente diferentes usando vários conjuntos do mesmo tamanho com diferentes componentes. Estes grupos são obtidos por amostragem do conjunto de treino inicial. Com o *bagging*, você amostra o conjunto "por substituição", o que significa que às vezes você pode ter a mesma instância escolhida mais de uma vez na sua amostra. São produzidos vários conjuntos de formações diferentes, e então constrói-se um modelo para cada um usando árvore de decisão com o mesmo esquema de aprendizagem de máquina, ou usando algum outro esquema de aprendizagem de máquina. Em seguida, combina-se as previsões dos diferentes modelos por votação (of Computer Science at the University of Waikato, 2016).

2.3.4 *Logit Boost*

Boosting funciona aplicando sequencialmente um algoritmo de classificação para as versões reponderadas dos dados de treinamento e, em seguida, tomar uma votação por maioria ponderada da sequência de classificadores assim produzidos. Para muitos algoritmos de classificação, esta estratégia simples resulta em melhorias significativas na performance. Em seu trabalho Friedman et al (2000) cita que esse fenômeno aparente-

mente misterioso pode ser entendido em termos de princípios estatísticos bem conhecidos, nomeados modelagem aditiva e máxima verossimilhança (*additive modeling and maximum likelihood*). Para o problema de duas classes, *Boosting* pode ser visto como uma aproximação ao aditivo de modelagem na escala logística, utilizando máxima verossimilhança de Bernoulli como critério.

Boosting é uma forma de combinar o desempenho de diversos classificadores "fracos" para produzir um comitê "poderoso". No *WEKA* essa classe executa a classificação usando um esquema de regressão, conforme a base de aprendizado, e pode lidar com problemas multiclasse.

2.3.5 Classificação via Regressão

Árvores modelo, são um tipo de árvore de decisão com funções de regressão linear para as folhas, estas formam a base de uma técnica de sucesso para a previsão de valores numéricos contínuos. Eles podem ser aplicados a problemas de classificação, utilizando um método padrão de transformação de um problema de classificação em função de um problema de aproximação (Frank et al, 1998).

Estruturalmente, uma árvore modelo assume a forma de uma árvore de decisão com funções de regressão linear em vez de valores de classe terminais nas suas folhas. Atributos numericamente valorizados desempenham um papel natural nestas funções de regressão. Atributos discretos também podem ser tratados, embora de uma forma menos natural. Isso é o contrário da situação de árvore de decisão clássica de classificação, onde os atributos discretos desempenham um papel natural. Pela simetria desta situação, Frank et al (1998) propôs um modelo de árvores usado para a classificação, sendo a base para o algoritmo *Classification Via Regression* do *WEKA*.

Classification Via Regression ou classificação via regressão, como o próprio nome já diz, faz a classificação usando métodos de regressão. Os rótulos de classe são binarizados e um modelo de regressão é construído para cada valor de classe.

2.3.6 MultiClass Classify

O processo de construção de classificadores multiclasse é dividido em dois componentes: (i) a seleção das características a serem utilizadas para treinamento e testes e (ii) a seleção do método de classificação. No *WEKA* para o desenvolvimento deste trabalho foi selecionado o *Logistic* ao executar os experimentos.

MultiClass Classifier é um metaclassificador para a manipulação de dados multiclasse com classificadores 2-classes. Este classificador também é capaz de aplicar códigos de correção de erros de saída para aumentar a precisão.

2.4 TEORIA DE FINANÇAS

O Sistema Financeiro pode ser caracterizado como o conjunto de instituições e instrumentos que possibilitam e facilitam o fluxo financeiro entre os poupadores e os tomadores de recursos na economia. Não é difícil perceber a importância desse sistema para o correto funcionamento e crescimento econômico de uma nação. Se, por exemplo, determinada empresa, que necessita de recursos para a realização de investimentos para a produção, não conseguir captá-los de forma eficiente, provavelmente ela não realizará o investimento, deixando de empregar e gerar renda (do [Investidor](#), 2016).

A Hipótese de Eficiência dos Mercados (HEM) está calcada em alguns dos seguintes pressupostos ([Fama, 1970](#)), ([Cardoso and Martins, 2004](#)):

- Os preços atuais são independentes dos preços de ontem, pois toda informação disponível já foi processada e incorporada a estes preços;
- Os investidores são considerados indivíduos racionais que estão sempre em busca da maximização de retornos, a partir do conhecimento de toda informação pública disponível;
- Os negociadores estão a par das mudanças que ocorrem no ambiente;
- Os preços são justos;
- O mercado tende a permanecer em equilíbrio, na medida em que sempre existirão vendedores e compradores; e,
- Toda nova informação impacta diretamente e de mesma maneira o mercado e seus participantes.

Sobre o comportamento de preço de ativos financeiros para a EMH, são do tipo passeio aleatório, o qual assume um padrão de saltos de crescimento e decrescimento constantes de mesma probabilidade.

A volatilidade dos preços no mercado de ações é positivamente correlacionada com a atividade do mercado e o volume de transações, um indicativo de que o mercado é heterogêneo, onde os diferentes participantes analisam eventos passados e novos com diferentes horizontes de tempo e diferentes expectativas em negociações, o que cria volatilidade.

Os esforços ao se estudarem mercados de ações é encontrar evidências, correlação, padrões de curto ou longo prazo, fractais, caos ou memória nos preços das ações. A fim de tornar o investimento em ações menos arriscado aos investidores.

Atualmente existem diversas abordagens para o estudo do comportamento do Mercado Financeiro, algumas focam apenas na tentativa de descobrir padrões de comportamento em séries temporais de preços das ações e são chamadas Análise Técnica, outras

usam toda informação disponível no mercado sobre determinada empresa e é chamada Análise Fundamentalista. De acordo com (Camilo and Silva, 2009) as escolas de análise de investimentos buscam, de forma geral, projetar resultados futuros dos investimentos disponíveis e medir os níveis de risco que essas medições podem apresentar.

O foco deste trabalho está no uso da Análise Fundamentalista e nas técnicas de Mineração de Dados. De acordo com Pinheiro (2009), pode-se conceituar Análise Fundamentalista como o estudo de toda a informação disponível no mercado sobre determinada empresa, com a finalidade de obter seu valor intrínseco, verdadeiro, do ativo financeiro, e assim formular uma recomendação de investimento.

Encontrar informações suficientes para calcular todos os indicadores econômicos e financeiros das empresas é uma tarefa difícil caso o investidor não tenha acesso às plataformas de fornecimento de informações de mercado financeiro. Estas plataformas como Economatica¹ são pagas, e pode custar ao investidor em torno de R\$2000,00 por mês para ter acesso aos dados.

2.5 BOLSA DE VALORES BMF&BOVESPA

Bolsas de valores são instituições econômicas que têm como principal objetivo garantir negociações justas e equitativas entre investidores, promovendo a livre concorrência dos mesmos. Proporcionam liquidez aos ativos negociados, e transparência na fixação de preços. Possuem papel importante em sistemas econômicos, sendo ferramenta facilitadora de crescimento para empresas pela transferência de recursos, e servindo como canalizador de poupança para pequenos e grandes investidores, contribuindo, portanto, para geração de emprego e renda (Pinheiro, 2009).

Uma ação representa a menor parte do capital social de uma empresa, e é um ativo de renda variável, ou seja, não há conhecimento prévio de rentabilidade futura. Dá direito de participação nos lucros da empresa ao investidor, pois quando este adquire uma ação, se torna dono de parte da empresa, proporcional à quantidade de ações que possui (Pinheiro, 2009).

Em 2008, a Bolsa de Valores de São Paulo e a Bolsa de Mercadorias & Futuros se integraram, dando origem a BM&FBOVESPA. Atualmente, a BM&FBOVESPA é a única bolsa que opera no Brasil, além de ser líder na América Latina e uma das maiores do mundo em valores de mercado. Seus mercados abrangem ações, contratos futuros, câmbio, operações, fundos e ETFs (fundos de índices), crédito de carbono, leilões e renda fixa pública e privada. No início de 2010, a BM&FBOVESPA ampliou sua posição no quadro acionário do CME Group, de 1,8% para 5%, o que representa um investimento de US\$ 620 milhões.

¹ <https://economica.com/>

A BM&FBOVESPA é uma companhia que administra mercados organizados de títulos, valores mobiliários e contratos derivativos, além de prestar serviços de registro, compensação e liquidação, atuando, principalmente, como contraparte central garantidora da liquidação financeira das operações realizadas em seus ambientes (BM&FBovespa, 2016).

A Bolsa oferece ampla gama de produtos e serviços, tais como: negociação de ações, títulos de renda fixa, câmbio pronto e contratos derivativos referenciados em ações, ativos financeiros, índices, taxas, mercadorias, moedas, entre outros; listagem de empresas e outros emissores de valores mobiliários; depositária de ativos; empréstimo de títulos; e licença de softwares (BM&FBovespa, 2016).

Os índices da BM&FBOVESPA são indicadores de desempenho de um conjunto de ações, ou seja, mostram a valorização de um determinado grupo de papéis ao longo do tempo. Os preços das ações podem variar por fatores relacionados à empresa ou por fatores externos, como o crescimento do país, do nível de emprego e da taxa de juros. Assim, as ações de um índice podem apresentar um comportamento diferente no mesmo período, podendo ocorrer valorização ou ao contrário, desvalorização.

Os índices são divididos em diversos grupos como: índices amplos, índices setoriais, índices de sustentabilidade, índices de governança, índices de segmento entre outros.

Segundo Cavalcante et al (2009) os índices de mercado cumprem três objetivos principais:

- São indicadores de variação de preços do mercado;
- Servem de parâmetro para avaliação de desempenho de portfólios; e,
- São instrumentos de negociação no mercado futuro.

O índice Ibovespa, que pertence ao grupo de índices amplos é o indicador do desempenho médio das cotações dos ativos de maior negociabilidade e representatividade do mercado de ações brasileiro. Como ele é um índice de retorno total, então procura refletir não apenas as variações nos preços dos ativos integrantes do índice no tempo, mas também o impacto que a distribuição de proventos, por parte das companhias emissoras desses ativos, teria no retorno do índice.

2.5.1 Tipos de Ações e Códigos de Negociação

Cada empresa que compõe a base de dados pode ter um ou mais tipos de ações sendo negociadas. Basicamente há dois tipos de ações: as ordinárias e as preferenciais. As ordinárias, simbolizadas pela sigla ON, concedem aos acionistas o direito de voto nas assembleias, bem como a participação não preferencial nos resultados da empresa. As ações

preferenciais, simbolizadas pela sigla PN, dão prioridade aos acionistas no recebimento de dividendos ou no reembolso do capital. No entanto, não concedem o direito de voto nas assembleias. Há também as ações preferenciais classes A, B, C e D, que são simbolizadas pelas siglas PNA, PNB, PNC e PND, respectivamente.

Ao negociar uma ação na bolsa de valores ela recebe um código composto de quatro letras e um número, o número corresponde ao tipo de ação e estes tipos são mostrados na tabela 1 (Vieira, 2016).

Tabela 1 – Código dos Tipos de Ação

Código	Tipo	Exemplo
3	Ordinárias	VALE3
4	Preferenciais	GGBR4
5	Preferenciais Classe A	USIM6
6	Preferenciais Classe B	ELET6
11	<i>BDRs, ETFs e Units</i>	SANB11, BOVA11

Não há uma regra específica para a ação negociada com o número 11 geralmente este número representa os recibos de ações de empresas estrangeiras negociadas na bolsa brasileira, os chamados *BDRs* (*Brazilian Deposits Receipts*). Também representa as *units*, que são ativos compostos por mais de um tipo de ação, bem como os fundos de índices, conhecido como *ETFs* (*Exchange Traded Funds*).

Então os dados de valor e quantidade de ações de cada empresa estão relacionados aos tipos de papéis que ela possui e deve-se ficar atento a estas informações ao calcular os indicadores fundamentalistas.

2.6 TRABALHOS RELACIONADOS

Nesta seção são discutidos os resultados de alguns trabalhos que tiveram objetivos relacionados à proposta de análise de empresas em mercados financeiros. Serão tratados trabalhos nacionais e internacionais de classificação de empresas para compor carteiras de ações, bem como outras formas de abordagem como os estudos de séries temporais de preços das ações.

No trabalho de Saeedmanesh et al (2010) os autores levantam a questão da grande dimensionalidade dos problemas relacionados ao Mercado Financeiro. Em um grande número de características e indicadores disponíveis quais são mais relevantes? Muitos especialistas consideram que o modelo deve ser procurado em dados históricos, e o processo de pesquisa deve ter um caráter contínuo e incremental, dependendo da eficiência do modelo usado. O problema se resume a encontrar um modelo melhor de análise, que leve a previsões mais precisas e redução da dimensionalidade (quantidade de variáveis analisadas) do problema.

Saeedmanesh et al (2010) propõe uma técnica híbrida combinando mineração de dados e técnicas de previsão. A Técnica Híbrida propõe aumentar a exatidão e a tolerância a falhas a fim de superar as principais limitações da *ANN* (*Artificial Neural Network*), uma técnica híbrida de mineração de dados tolerante a falhas para a previsão de preços na bolsa de valores chamada HDM é a proposta dos autores. Foi usado o expoente de *Hurst* para determinar o melhor período previsível. O expoente de *Hurst* foi aplicado nas séries do índice *Down Jones* de 1930 a 2004 e o maior índice obtido foi no período de 1970, então foi escolhido 1969 à 1973 como período de tratamento dos dados para pesquisa. Foi usada Teoria do Caos para determinar dimensão e atraso de tempo. A medida de desempenho adotada foi a taxa de erro e as técnicas pesquisadas foram ANN, KNN e Árvores de decisão. Nos cenários estudados pelos autores o *ANN* teve melhor desempenho, mas a Árvore de Decisão teve melhor taxa de erro para prever dados. Entretanto, uma única técnica não é suficiente para prever dados de bolsa de valores, a forma de melhorar estes resultados é combinar cálculos e modelos estatísticos de finanças com técnicas de mineração de dados. A taxa de erro da ANN foi melhorada como o uso do método que os autores denominam *5-fold-average-prediction*.

No artigo de Saeedmanesh et al (2010) foi apresentado apenas previsão de tendências dos valores de índice industrial na *Down Jones*, já no presente trabalho é usado um grupo de empresas e seus indicadores para entender melhor a influência dos indicadores no movimento dos preços.

No trabalho de Rycheski (2013) é proposto reduzir o risco na escolha de ativos para montar uma carteira de investimentos baseada em ações por meio de análise das demonstrações contábeis e avaliação de desempenho. Utilizou-se de fundamentos econômico-financeiros e de mercado para avaliar se é interessante ou não investir em determinada empresa.

Rycheski (2013) escolheu um período de 4 anos e, para a delimitação do campo de estudo, foram utilizadas apenas as empresas que compõem a carteira teórica de um índice do BM&FBovespa. Assim sendo, o índice IDIV (índice das ações que se destacaram no pagamento de dividendos nos últimos 24 meses) foi escolhido como objeto de estudo. Todos os dados contábeis, econômicos e de mercado utilizados neste levantamento foram coletados com a utilização da plataforma Economática e foram levantados dados de 2003 a 2012. Para escolher as empresas mais atrativas para compor a carteira foram utilizados indicadores que analisam a solvência, histórico de dividendos, estrutura de capital e rentabilidade. Rycheski (2013) montou uma carteira de investimento com o total de 5 ações sendo as empresas AES Tiete, Natura, Vale, Cemig e Trectebel. A carteira atingiu no período de análise, de 2008 a 2012, 138% de rentabilidade, contra um resultado de -2,96% do Ibovespa e aproximadamente 34% nas aplicações mais conservadoras. O autor concluiu que as informações oriundas das demonstrações contábeis são relevantes para

análise fundamentalista, fornecendo insumos necessários para a avaliação das opções de investimento e redução dos riscos inerentes a eles.

Existem diversos trabalhos que exemplificam formas de aplicação da Mineração de Dados no Setor Financeiro. No trabalho de [Pinho \(2008\)](#) ele usou mineração de dados para clusterizar clientes, em grupos o mais homogêneo possível, de forma a fazer estratégias mais eficazes de marketing. [Pinho \(2008\)](#) usou as Redes Som ou Mapas Auto-organizáveis como estratégia de clusterização de clientes. Os dois modelos básicos de mapas auto-organizáveis são tratados na literatura de Redes Neurais, o modelo de *Willshaw-von der Malsburg*, e o modelo de *Kohonen*. O modelo proposto e estudado na aplicação de clusterização de clientes foi o de *Kohonen*. Na conclusão de [Pinho \(2008\)](#) dividir os clientes em *clusters* é útil na diferenciação de ações de marketing. Gestores de relacionamento ao cliente podem a partir do 4º mês de relacionamento do cliente fazer ofertas personalizadas de produtos e serviços, maximizando assim o retorno de ações. Com isso, esperam-se diminuir as taxas de evasão de clientes a partir do 5º mês de relacionamento.

Na pesquisa de [Silva and Tessaro \(2013\)](#) é utilizado regras de associação da mineração de dados para identificar e analisar a correlação entre relatórios financeiros de empresas de capital aberto e a variação do preço de suas respectivas ações negociadas na Bolsa de Valores de São Paulo. Os dados utilizados neste experimento foram os relatórios financeiros (Balanço Patrimonial e Demonstração do Resultado do Exercício) dos últimos 32 trimestres a partir de agosto/2012, de 30 ações distintas. O objetivo foi correlacionar estes dados com a oscilação do preço das ações posteriores à data de divulgação de cada relatório. Definiu-se, então, que o sistema analisaria a base de dados e retornar as 50 melhores regras de associação encontradas com um nível de confiança mínimo de 30%. O sistema retorna o resultado em formato de texto com a lista das regras encontradas notou-se que o fator que mais influenciou para uma elevação no preço de uma ação foi a margem líquida. O segundo fator que mais influenciou o preço das ações para uma alta foi a combinação entre ROE e ROI.

O estudo de [Silva and Tessaro \(2013\)](#) tem uma correlação importante para o objetivo deste trabalho, pois mostra quais indicadores foram mais significativos no preço das ações.

O estudo de [Imandoust and Bolandraftar \(2014\)](#) comparou o desempenho dos três modelos para prever a direção do movimento diário dos índices no *TSE (Tehran Stock Market)*. Os métodos aplicados foram Floresta Aleatória (*Random Florest*), Árvores de Decisão e Classificadores *Naive Bayes*. Como forma de validação 80% dos dados foram usados para treinar os modelos, e os 20% remanescentes foram usados para teste e comparado seus desempenhos. A direção de mudança diária no preço do índice das ações foram categorizados como "Positivo", "Negativo" e "Sem Mudanças". Os resultados demonstram que variáveis técnica e técnica-fundamental são capazes de prever a direção do movimento

de mercados com exatidão aceitável, onde Árvores de Decisão (com precisão de 80.08% tanto em variáveis técnica quanto técnica-fundamental) superou as outras técnicas usadas. Quando variáveis fundamentais são usadas, *Random Forest*, teve melhor desempenho (61,86%) que os outros métodos. No geral, o poder preditivo do aprendizado de máquina nesta situação não é aceitável. O gráfico da Figura 3 mostra os resultados comparativos do uso das técnicas de mineração de dados é possível observar que usando indicadores técnicos e fundamentais as Árvores de Decisão (*DT*) têm melhor desempenho. Entretanto, usando apenas indicadores fundamentais *Random Forest* possui melhor desempenho.

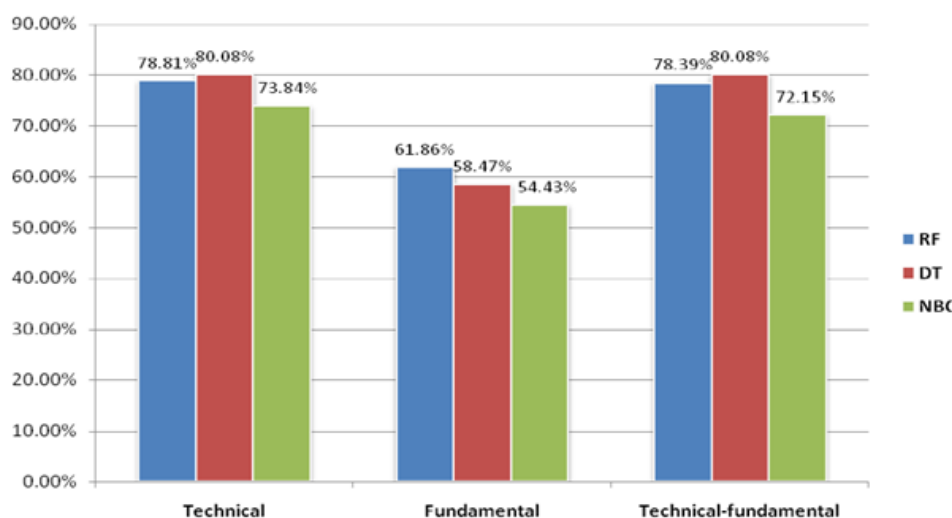


Figura 3 – Resultado comparativo do uso da técnicas de Mineração de Dados - Fonte: [Imandoust and Bolandraftar \(2014\)](#)

Pelos trabalhos apresentados, é possível observar que, diversas técnicas de mineração de dados e metodologias de cálculo de análise financeira podem ser usados para trazer informações para os investidores e empresas. A aplicação dos algoritmos de mineração de dados são amplas, entretanto a escolha dos métodos, cálculos e variáveis influencia fortemente os resultados.

2.7 WEKA

*WEKA*² é uma ferramenta de *software* livre, desenvolvida na linguagem de programação *Java*, que possui um conjunto de algoritmos de aprendizado de máquina para mineração de dados. Os algoritmos *RandomCommitee*, *RandomTree*, *Ibk*, *RandomForest*, *Bagging Meta*, *ClassificationRegression*, *LogitBoost*, *REPTree*, *J48* e *MultiClassClassifier* implementados no *WEKA* foram utilizados a fim de avaliar a eficiência destes algoritmos sobre a base de dados que foi desenvolvida neste trabalho. A ferramenta *WEKA* trabalha com dados no formato *arff* composto basicamente por um cabeçalho com os atributos e

² <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>

todos os valores que ele pode assumir e pelas instâncias, que representam a base de dados de maneira adaptada.

Todos os atributos deverão ser inseridos no cabeçalho seguindo a estrutura:

@attribute ‘nome do item’ {valores que ele pode assumir}

Esses valores que o atributo assume podem ser categóricos ou numéricos. No caso específico desse trabalho, foram usados valores numéricos para os atributos referentes aos índices. A Figura 4 mostra um dos arquivos de dados gerados para este trabalho, veja que o cabeçalho contém informações de como o arquivo é composto, seguido de informações dos atributos e só depois são dispostas as instâncias de dados a serem usadas nos algoritmos.

Figura 4 – Arquivo de dados gerado em formato arff.

Tendo criado um arquivo *arff* de forma correta, basta utilizar a ferramenta, fazer as devidas manipulações e gerar as classificações do grupo de teste.

2.8 AMBIENTE DE DESENVOLVIMENTO

No ambiente de desenvolvimento desse trabalho foram utilizados: *Eclipse Java EE IDE for Web Developers*, versão *Mars.1 Release (4.5.1)* para desenvolvimento do programa usando linguagem Java, o *MySQL* como SGBD (Sistema de Gerenciamento de Banco de Dados), o *Hibernate* para mapeamento objeto relacional e *Apache Maven* para gerenciamento de dependências.

*MySQL*³ é um Sistema Gerenciador de Banco de Dados que utiliza linguagem SQL (*Structured Query Language*). Ele oferece facilidades no gerenciamento da base de dados e interface de fácil utilização. É um software livre com excelente desempenho e estabilidade.

³ <http://www.mysql.com/downloads/>

Apache Maven é uma ferramenta da *Apache Software Foundation* para gerenciamento de dependências e automação de *build*, principalmente em projetos Java. Um projeto que usa *Maven* possui um arquivo XML (*pom.xml*) que descreve o projeto, suas dependências, detalhes do *build*, diretórios, *plugins* requeridos, etc. O uso do *Apache Maven* deixa o desenvolvimento mais ágil, pois ele faz o gerenciamento de dependências do projeto.

Mapeamento objeto relacional (*object-relational mapping*, *ORM*, *O/RM* ou *O/R mapping*) é uma técnica de programação para conversão de dados entre banco de dados relacionais e linguagens de programação orientada a objetos. Em um ambiente ORM, as aplicações interagem com APIs e o modelo de classes de domínio e os códigos SQL/JDBC são abstraídos. Os comandos SQL são automaticamente gerados a partir dos metadados que relacionam objetos a banco de dados.

A *Java Persistence API (JPA)* é um framework para persistência em Java, que oferece uma API de mapeamento objeto-relacional e soluções para integrar persistência com sistemas corporativos escaláveis. Com *JPA*, os objetos são *POJO (Plain Old Java Objects)*, ou seja, não é necessário nada de especial para tornar os objetos persistentes. Basta adicionar algumas anotações nas classes que representam as entidades do sistema e começar a persistir ou consultar objetos.

O projeto do *Hibernate ORM* possui alguns módulos, sendo que o *Hibernate Entity Manager* é a implementação da *JPA* que encapsula o *Hibernate Core* que é a base para o funcionamento da persistência, com APIs nativas e metadados de mapeamentos em arquivos *XML*. Possui uma linguagem de consultas chamada *HQL (Hibernate Query Language)* parecido com *SQL* e um conjunto de interfaces para consultas usando critérios (*Criteria API*).

2.9 INDICADORES FUNDAMENTALISTAS E MEDIDAS DE RETORNO

Os indicadores conseguem dar indícios de uma boa compra, porém não garantem a certeza de sucesso. Os indicadores utilizados neste trabalho serão explicados no decorrer deste capítulo e divididos de acordo com sua classificação.

Aqui estão as principais medidas de retorno, sendo elas valorização em bolsa e dividendos:

- “*Pay-Out*” é a taxa de distribuição do lucro da empresa para os acionistas na forma de dividendos ou juros sobre o capital próprio. A legislação exige que seja distribuído, no mínimo, 25% do lucro líquido após as deduções legais. Exemplificando, se uma empresa tem lucro de R\$ 1,00/ação e distribui R\$ 0,35/ação como dividendos, seu

pay-out seria de 35%. Então para este indicador quanto maior melhor (Corretora, 2015).

- *Dividend Yield*: revela o percentual do preço da ação que voltou aos acionistas sob a forma de dividendos no último balanço Halfield (2007). O indicador é uma divisão de quanto a empresa paga de dividendo por ação pelo preço da ação:

$$DividendYield = DividendoPorAção/PreçodaAção$$

Esse pagamento pode ser anual, semestral, mensal, semanal, etc. Depende de empresa para empresa. Como este indicador é uma fração que varia conforme o preço da ação, uma porcentagem muito alta pode ser consequência de uma ação muito barata, por tanto não compare o *Dividend Yield* de duas empresas sem antes ver o preço da ação. Ele pode ser visto como o rendimento da ação independente da valorização em bolsa, então quanto maior melhor.

- Juros Sobre Capital Próprio (JCP): As empresas podem pagar para os acionistas como *Dividend Yield* ou JCP, com o JCP ela tem um limite a ser distribuído mas tem benefícios fiscais. Em contrapartida o acionista paga 15% de Imposto de Renda sobre o valor recebido, o que não acontece ao receber com *Dividend Yield*. As empresas sempre otimizam as duas possibilidades buscando economia fiscal e vantagens para os acionistas (Corretora, 2015).

Agora vamos explicar os indicadores fundamentalistas mais utilizado para análise de ações. São eles:

- P/L (Preço/Lucro): segundo Cavalcante et al (2009) é um dos indicadores mais utilizados para analisar uma ação. Ele é o número de anos necessário para o retorno do capital investido. O P/L é um indicador bastante simples e é um bom critério de comparação entre empresas de um mesmo setor. A fórmula é simples:

$$P/L = (PreçoPorAção/LucronoúltimoanoPorAção)$$

Onde Lucro Por Ação é o lucro da empresa no último ano ou a expectativa de lucro por ano da empresa, dividido pelo número de ações disponíveis no mercado. Quanto menor este valor mais rápido será o retorno do seu investimento, porém quando os valores são muito altos ou negativos não é aconselhável utilizá-lo. Mais informações sobre este índice também pode ser vista no site InvestidorJovem (Jovem, 2015).

- VPA (Valor Patrimonial da Ação): o cálculo do VPA é bem simples, basta dividir o Patrimônio líquido pelo número de ações que a empresa disponibiliza na bolsa. É como se somássemos os valores dos prédios, instalações, equipamentos, terrenos, estoque, caixa, etc e dividisse pelo número de ações. Imagine uma situação hipotética em que uma empresa precisasse fechar as portas. Se ela vendesse todo o patrimônio

e pagasse todas as suas dívidas, o dinheiro que sobrasse seria dividido pelos acionistas. A quantidade de dinheiro que cada um receberia por ação seria o valor do VPA. Por este motivo o ideal é ter o VPA próximo ou até maior que o preço da ação. Representa a divisão do Patrimônio Líquido da empresa pelo seu número total de ações (Martins, 2009):

$$VPA = \text{Patrimônio Líquido} / \text{Número Total de Ações}$$

- **Liquidez Corrente:** é uma relação entre o que a empresa tem em dívida (empréstimos, impostos, fornecedores, etc) com o que ele tem em “propriedades e dinheiro” (caixa, estoques, clientes, etc) no curto prazo. Indica quantas vezes os ativos mais líquidos superam os passivos. A prática costuma recomendar que esse índice seja superior a um, isto é, que o total do Ativo Circulante supere o Passivo Circulante. É um indicador de solvência da empresa, ou seja, a capacidade de uma empresa cumprir seus compromissos de curto prazo (Halfield, 2007).
- **P/VP (Preço / Valor Patrimonial):** comparação entre o valor de mercado da empresa (Preço) e o Valor Patrimonial da Empresa (Patrimônio Líquido) (Halfield, 2007). Ações negociadas com preço muito acima de seu VPA expõe o investidor a um maior risco, uma vez que a “porção expectativa” do preço da ação é muito grande e a “porção patrimônio líquido” é muito pequena. Nesses casos, uma mudança de expectativa causada por uma crise pode causar um grande choque no preço do papel, traduzindo-se em uma grande perda ao investidor. Não existe uma relação correta, mas vamos usar valores de P/VP menores que 2, para diminuir os riscos inerentes a perdas por crises (Martins, 2009).

De acordo com Cavalcante et al (2009) este indicador mede o quanto a ação está cotada em relação ao seu valor patrimonial, então quanto menor este indicador mais subavaliada está a ação.

- **Preço/Geração de Caixa (P/CFS):** este indicador oferece uma informação complementar à do P/L, sendo visto o tempo teórico de retorno esperado à luz apenas do efeito caixa do resultado do exercício (Corretora, 2015):

$$P/CFS = \text{Cotação da Ação} / \text{Geração de Caixa por Ação}$$

De acordo com Cavalcante et al (2009) a vantagem deste indicador em relação ao P/L é que considera o fluxo apenas da operação da empresa, excluídos componentes não operacionais que podem distorcer a análise. Quanto menor o valor mais subavaliada está a ação.

- **EV/EBITDA** é dado pela fórmula (*Valor do Empreendimento*) / *Geração de Caixa na Atividade* (Corretora, 2015).

EBITDA (*Earning Before Interests, Taxes, Depreciation and Amortization*) representa a geração operacional de caixa da companhia, ou seja, o quanto uma empresa gera de recursos apenas em suas atividades operacionais, sem levar em consideração os efeitos financeiros e de impostos (InfoMoney, 2015). Aqui há um confronto entre a capacidade de geração interna de caixa da empresa, a partir de suas atividades próprias, e o valor presente dos fluxos de caixa projetados para a empresa. Podem ser usados dados históricos ou projetados, a depender de cada caso. Assim como os outros múltiplos, quanto menor, em princípio é mais atrativo.

De acordo com Cavalcante et al (2009) a relação EV/EBITDA indica qual o prazo para o retorno do capital total (próprio e de terceiros).

- ROIC (*Return Over Invested Capital* - Retorno Sobre Capital Investido) é um dos principais indicadores fundamentalistas, sintetiza diversas características da empresa incluindo a força de sua vantagem competitiva. Ele refere-se apenas ao retorno sobre o capital próprio. O que o ROIC tenta medir é qual o retorno que cada real investido na empresa traz, independente da forma com que esse investimento é financiado, então quanto maior o ROIC melhor. O ROIC é calculado como (Zeidler, 2015):

$$ROIC = NOPLAT / \text{Capital Investido}$$

Onde: *NOPLAT* (*Net Operating Profits Less Adjusted Taxes*) é o Lucro Líquido Operacional menos impostos ajustados. É calculado com $EBIT * (1 - T)$, onde T é a alíquota do imposto de renda; Capital Investido = capital próprio mais capital de terceiros.

O ROIC pode ser usado para comparar empresas de setores diferentes e com diferentes planos de negócios. Para exemplificar, vamos supor que se deseja montar um novo negócio, e para isso é preciso investir R\$ 8,8 milhões. No final do ano o lucro operacional antes dos impostos (*EBIT*) é de R\$ 2 milhões. Se a alíquota de imposto de renda for de 34%, então o *NOPLAT* é de $2 * (1 - 34\%) = R\$1,32 \text{ milhões}$. O *ROIC* neste caso é de $1,32 / 8,8 = 15\%$, ou seja, o investimento retornou este ano 15% dos R\$ 8,8 milhões investidos inicialmente. Esse retorno independe do número de empréstimos tomado para financiar o investimento.

- *ROE* (*Return on Equity*): este indicador mostra a porcentagem de retorno que a empresa dá sobre o dinheiro de seus acionistas. É calculado como (Wawrzyniac, 2015):

$$ROE = \text{Lucro Líquido} / \text{Patrimônio Líquido}$$

Este indicador é muito útil para medir a lucratividade de diferentes companhia no mesmo setor. Em outras palavras ele mostra quanto que a companhia gerou para cada Real investido por seus acionistas. É comum encontrar um ROE alto para

companhias em crescimento, enquanto que companhias mais estáveis possuam um ROE menor. No entanto, o ROE é uma indicação clara de como a empresa está remunerando seus acionistas (TORORADAR, 2016).

- Indicador também conhecido como indicador *Debt-Equit*, este número mostra o nível de endividamento da empresa (Corretora, 2015). Calculado como:

$$D/CS = \text{Capital de terceiros} / \text{capital próprio}$$

Embora o nível de endividamento varie bastante entre diferentes setores, vale tomar cuidado com empresas que possuam altos níveis de endividamento, especialmente se o setor está passando por momentos ruins. Este é o primeiro sinal de que a empresa pode ter se endividado mais do que deveria. Não é válido para comparar empresas de diferentes setores. Enquanto que alguns setores são muito dependentes de financiamentos (ex: construtoras), outros praticamente não precisam (ex: elétricas).

- Liquidez em Bolsa: ações com baixa frequência de negócios nos pregões trazem um risco a mais para o investimento, pois não há garantias de que você vai conseguir negociá-las sem dificuldades. Para que um mercado seja eficiente, a boa liquidez dos títulos negociados é uma premissa básica.
- Valor de Mercado: é representado pela cotação das suas ações em bolsa, multiplicado pelo número total de ações que compõe seu capital. O importante neste índice é analisar se está tendo aumento do valor de mercado de uma ação ou se está tendo queda. O importante é manter-se estável ou com ganhos (Corretora, 2015).
- Geração de Caixa na Atividade: também chamado de *EBITDA (Earnings Before Interest, Taxes, Depreciation and Amortization)* traz os efetivos impactos de caixa. Observando a partir da Receita Líquida, o *EBITDA* vem a ser o resultado da seguinte seqüência Corretora (2015): (+) Receita Líquida (-) Custo dos Produtos Vendidos (-) Despesas da Atividade (c/ vendas, administrativas e outras diretamente ligadas às operações) (=) Lucro da Atividade (ou *EBIT*) (+) Depreciação e Amortização (valor correspondente ao período sob análise, que pode ser encontrado na Demonstração de Fluxo de Caixa) (=) *EBITDA*
- Fluxo de Caixa Operacional: propicia uma visão adequada do que efetivamente entrou ou saiu do caixa no que se refere às vendas, aos custos e às despesas. É igual ao Ebitda (-) Imposto de Renda (+ ou -) Variação da Necessidade de Capital de Giro Corretora (2015).
- Fluxo de Caixa Livre (ou, em inglês, *Free Cash Flow*): é obtido simplesmente descontando do Fluxo de Caixa Operacional os investimentos da empresa no ativo imobilizado, habitualmente denominados como Capex (*Capital Expenditures*). O Fluxo de Caixa Livre quanto maior melhor, valores negativos são ruins Corretora (2015).

- LPA (Lucro por Ação): representa a divisão do lucro líquido pelo número total de ações da empresa [Corretora \(2015\)](#). Calculado como:

$$LPA = \text{Lucro Líquido} / \text{Número Total de Ações}$$

- A Geração de Caixa por Ação (ou *CFS Cash Flow/Share*): pode utilizar vários conceitos de geração de caixa, sendo que aqui optou-se por indicar o de geração de caixa na atividade, ou EBITDA ([Corretora, 2015](#)):

$$CFS = \text{Geração de Caixa na Atividade} / \text{Número Total de Ações}$$

- Ibovespa e o β (Beta) (Índice do BM&FBovespa): empresas de maior risco são aquelas cujas ações sobem ou caem em Bolsa de forma mais agressiva que o índice do mercado (mais de uma vez a variação do Ibovespa). Para calcular o risco foi criado o coeficiente ([Corretora, 2015](#)):

$$\beta = \text{Covariância Retorno do Mercado e Retorno da Ação} / \text{Variância Retorno do Mercado}$$

O beta da carteira de ações padrão, IBOVESPA, é sempre igual a 1, uma vez que ela é a base para o cálculo comparativo. O beta desta carteira é o beta médio de todos os títulos disponíveis. Desta forma, conclui-se que:

$\beta = 1$ Ativo médio. Sua variação tende a acompanhar perfeitamente o mercado. Quando o IBOVESPA valoriza 5%, o ativo valoriza na mesma proporção.

$\beta < 1$ Ativo defensivo. Possui oscilações inferiores ao mercado e no mesmo sentido. Quando o IBOVESPA valoriza 5%, o ativo tende a valorizar menos do que 5

$\beta > 1$ Ativo agressivo. Possui oscilações maiores do que o mercado e no mesmo sentido.

- Avaliação do Setor: analisar e avaliar se o setor está com boas expectativas para o período de aplicação.

Estes são os indicadores citados na bibliografia relacionada a mercados financeiros e utilizados na base de dados do presente trabalho. Os valores destes indicadores foram usados como atributos nos algoritmos e a classe a qual a empresa pertence vai depender do rendimento da empresa em relação à Selic.

3 DESENVOLVIMENTO

O primeiro problema enfrentado por qualquer investidor é como compor uma carteira de ações. Primeiramente ele precisa escolher as empresas nas quais pretende investir e para isto é importante fazer análise dos indicadores de saúde financeira desta empresa. Neste momento surge a dúvida em todos os investidores inexperientes, como fazer esta escolha e em qual site obter informações?

Sites de investimento como o *InfoMoney* alertam para montar carteira de ações com empresas que possuem bons fundamentos, endividamento coerente com os ganhos e estrutura operacional eficiente. É importante observar também se a empresa está bem posicionada em seu mercado e quais as perspectivas para o setor e para a empresa ([InfoMoney, 2005](#)).

O *InfoMoney* recomenda que ao decidir o período de aplicação e os ativos (empresas) seria bom aplicar no máximo 20% do valor que será investido em cada ativo, pois assim diversifica a aplicação e reduz os riscos de perda, já que perdas em uma ação X podem ser minimizada por ganhos em uma ação Y .

Além do ativo a ser escolhido o prazo também é uma variável-chave. Pode-se investir a curto, médio ou longo prazo. Investimentos em curto prazo exigem mais atenção ao mercado, enquanto investimentos de médio e longo prazo têm maior tempo para tomadas de decisão e o risco de crises políticas internas ou externas é menor.

Existem dezenas de indicadores fundamentalistas para dizer se uma empresa é boa ou não. Estes indicadores são explicados na seção 2.9 e entender todos eles é uma tarefa difícil e que leva tempo. Assim, neste trabalho, será usado a mineração de dados para verificar se é possível obter um bom desempenho dos algoritmos analisando estes indicadores. Caso a mineração de dados apresente bom desempenho ela pode auxiliar investidores, já que o mercado financeiro apresenta muitas variáveis para analisar antes de decidir em qual empresa investir.

3.1 SELEÇÃO DOS DADOS

Os ativos (ações) escolhidos para compor este trabalho foram selecionados usando informações do site da BM&FBovespa na guia empresas listadas¹, ao clicar em cada empresa é possível ver quais tem ativos para negociar no mercado a vista. Esta pesquisa foi feita entre 15 e 24 de Setembro de 2015, e foram selecionadas quais empresas tiveram

¹ <http://www.bmfbovespa.com.br/cias-listadas/empresas-listadas/BuscaEmpresaListada.aspx?Idioma=pt-br>

ativos negociados entre Agosto de 2014 e Setembro de 2015. Então entraram no estudo todas as empresas que fazem parte do índice ibovespa, sendo que estes são os de maior negociabilidade e representatividade do mercado de ações brasileiro, e também empresas que tem menor volume de negociações mas que tiveram ativos sendo negociados no último ano. Atendendo a estas premissas foram selecionadas 335 empresas.

Os ativos são divididos em setores, subsetores e segmento, pode ser visto no site da BM&FBovespa². Os setores são:

- Bens Industriais;
- Construção e Transporte;
- Consumo Cíclico;
- Consumo Não Cíclico;
- Financeiro e Outros;
- Materiais Básicos;
- Petróleo, Gás e Biocombustíveis;
- Tecnologia da Informação;
- Telecomunicações;
- Utilidade Pública.

No total foram selecionadas 335 empresas, de diversos setores, que são negociadas na BM&FBovespa. Os dados de balanço patrimonial e demonstrativos de resultados foram extraídos do site Fundamentus³. Das 335 empresas selecionadas inicialmente, 311 tinham informações na base de dados do site *Fundamentus*, então os dados foram buscados no site no período compreendido entre 24 de Setembro de 2015 e 05 de Outubro de 2015.

As 311 empresas mostradas na tabela 2 participaram deste estudo, mas muitas destas empresas tem mais de um papel para compra no mercado de ações, então todos os papéis são levados em consideração através da análise dos seus proventos.

Através do site Fundamentus capturou-se os dados de Balanço Patrimonial e Demonstrativo de resultados dos últimos 10 anos para as empresas da tabela 2, assim os dados do presente trabalho são de 2005 a 2015. Algumas amostras dos dados foram comparadas com as informações contidas no site da BM&FBovespa e estavam compatíveis.

² <http://www.bmfbovespa.com.br/cias-listadas/empresas-listadas/BuscaEmpresaListada.aspx?Idioma=pt-br>

³ <http://www.fundamentus.com.br/index.php>

Tabela 2 – Ativos

Índ.	Nome de Pregão	Índ.	Nome de Pregão	Índ.	Nome de Pregão	Índ.	Nome de Pregão
1	AES ELPA	79	SABESP	157	IDEIASNET	234	PORTOBELLO
2	AES TIETE	80	COPASA	158	IGB S/A	235	POSITIVO INF
3	AFLUENTE	81	SANEPAR	159	IGUATEMI	236	PROFARMA
4	ALIANSC	82	SID NACIONAL	160	J B DUARTE	237	PRUMO
5	ALPARGATAS	83	COTEMINAS	161	INDS ROMI	238	QGEF PART
6	ALUPAR	84	SANTANENSE	162	INEPAR	239	QUALICORP
7	AMBEV S/A	85	CIELO	163	INEPAR TEL	240	RAIADROGASIL
8	AMPLA ENER	86	COBRASMA	164	IMC S/A	241	RANDON PART
9	ARTERIS	87	ALFA CONSORC	165	INVEST BEMGE	242	RECRUSUL
10	AZEVEDO	88	LIX DA CUNHA	166	IOCHP-MAXION	243	REDE ENERGIA
11	B2W DIGITAL	89	SULTEPA	167	ITAUUNIBANCO	244	REDENTOR
12	BAHEMA	90	CONTAX	168	ITAUSA	245	PET MANGUINH
13	BANESTES	91	COSAN LTD	169	ITAUTECH	246	RENAR
14	BAUMER	92	COSAN LOG	170	JBS	247	RENOVA
15	BBSEGURIDADE	93	COSAN	171	JEREISSATI	248	LE LIS BLANC
16	ABC BRASIL	94	CPFL ENERGIA	172	LA FONTE TEL	249	RJCP
17	ALFA INVEST	95	CPFL RENOVAV	173	JHSF PART	250	RODOBENSIMOB
18	AMAZONIA	96	CR2	174	JOAO FORTES	251	ROSSI RESID
19	BRADESCO	97	CREMER	175	JOSAPAR	252	RUMO LOG
20	BRASIL	98	CSU CARDSYST	176	JSL	253	SANSUY
21	DAYCOVAL	99	TRAN PAULIST	177	KARSTEN	254	SANTOS BRP
22	BANRISUL	100	CVC BRASIL	178	KEPLER WEBER	255	SAO CARLOS
23	BICBANCO	101	CYRELA REALT	179	KLABIN S/A	256	SAO MARTINHO
24	INDUSVAL	102	CYRE COM-CCP	180	KROTON	257	SPTURIS
25	MERC INVEST	103	D H B	181	LATAM AIRLN	258	SARAIVA LIVR
26	MERC BRASIL	104	DASA	182	LIGHT S/A	259	SCHULZ
27	BANCO PAN	105	DIMED	183	LINX	260	SENIOR SOL
28	PINE	106	DIRECIONAL	184	LOCALIZA	261	SER EDUCA
29	SANTANDER BR	107	DOHLER	185	LOG-IN	262	ALIPERTI
30	SOFISA	108	DTCOM-DIRECT	186	LOJAS AMERIC	263	SLC AGRICOLA
31	BEMATECH	109	DUFREY AG	187	LOJAS RENNEN	264	SMILES
32	BIC MONARK	110	GER PARANAP	188	LOPES BRASIL	265	SIERRABRASIL
33	BIOMM	111	DURATEX	189	LUPATECH	266	SONDOTECNICA
34	BIOSEV	112	ECORODOVIAS	190	M.DIASBRANCO	267	SOUZA CRUZ
35	BMFBOVESPA	113	ENERGIAS BR	191	MAGAZ LUIZA	268	SPRINGER
36	BOMBRIIL	114	ACO ALTONA	192	METAL LEVE	269	SPRINGS
37	BR MALLS PAR	115	ELEKEIROZ	193	MANGELS INDL	270	SUL AMERICA
38	BR PROPERT	116	ELEKTRO	194	ESTRELA	271	SUZANO PAPEL
39	BRADESPAR	117	ELETROPAR	195	MARCOPOLO	272	SWEETCOSMET
40	BR BROKERS	118	ELETROPOLAU	196	MARFRIG	273	TIME FOR FUN
41	BR INSURANCE	119	EMAE	197	LOJAS MARISA	274	TARPON INV
42	BRASILAGRO	120	EMBRAER	198	MENDES JR	275	TECTOY
43	BRASKEM	121	ENCORPAR	199	MERC FINANC	276	TEC BLUMENAU
44	BRASMOTOR	122	ENERGISA MT	200	METALFRIO	277	TECHNOS
45	BRB BANCO	123	ENERGISA	201	METAL IGUACU	278	TECNISA
46	BRF SA	124	ENEVA	202	MET DUQUE	279	TECNOSOLO
47	BTG PACTUAL	125	EQUATORIAL	203	GERDAU MET	280	TEGMA
48	CAMBUCCI	126	ESTACIO PART	204	RIOSULENSE	281	TEKA
49	CCR SA	127	ETERNIT	205	METISA	282	TEKNO
50	CCX CARVAO	128	EUCATEX	206	MILLS	283	TELEBRAS
51	CELUL IRANI	129	EVEN	207	MINERVA	284	TELEF BRASIL
52	ELETROBRAS	130	EVORA	208	MINUPAR	285	TEMPO PART
53	CELESC	131	EXCELSIOR	209	MMX MINER	286	TEREOS
54	CELPA	132	EZTEC	210	MONT ARANHA	287	TEX RENAUX
55	CHIARELLI	133	FER HERINGER	211	MRV	288	TIM PART S/A
56	CESP	134	TECEL S JOSE	212	MULTIPLAN	289	TOTVS
57	CETIP	135	FIBAM	213	MULTIPLUS	290	TRIUNFO PART
58	P.ACUCAR-CBD	136	FIBRIA	214	MUNDIAL	291	TRACTEBEL
59	CEG	137	ALFA FINANC	215	NADIR FIGUEI	292	TAESA
60	DOC IMBITUBA	138	FLEURY	216	NATURA	293	TREVISIA
61	COELBA	139	FORJA TAURUS	217	NORDON MET	294	TRISUL
62	CEB	140	PARCORRETORA	218	NUTRIPLANT	295	TUPY
63	CEMIG	141	FRAS-LE	219	ODONTOPREV	296	ULTRAPAR
64	CELPE	142	ANIMA	220	NOVA OLEO	297	UNICASA
65	COELCE	143	GAFISA	221	OI	298	UNIPAR
66	COSERN	144	GENERALSHOPP	222	OGX PETROLEO	299	USIMINAS
67	CEEE-D	145	GERDAU	223	OSX BRASIL	300	VALE
68	CEEE-GT	146	GOL	224	OUROFINO S/A	301	VALID
69	FERBASA	147	GP INVEST	225	PANATLANTICA	302	V-AGRO
70	CEDRO	148	GPC PART	226	PARANA	303	VIAVAREJO
71	COMGAS	149	GRAZZIOTIN	227	PARANAPANEMA	304	VIGOR FOOD
72	HABITASUL	150	GRENDENE	228	PDG REALT	305	VIVER
73	CIA HERING	151	GUARARAPES	229	PETROBIO	306	VULCABRAS
74	LOCAMERICA	152	HAGA S/A	230	PETROBRAS	307	WEG
75	MELHOR SP	153	HELBOR	231	PETTENATI	308	WETZEL S/A
76	COPEL	154	HERCULES	232	PLASCAR PART	309	WHIRLPOOL
77	PAR AL BAHIA	155	HOTEIS OTHON	233	PORTO SEGURO	310	WILSON SONS
78	PROVIDENCIA	156	HYPERMARCAS			311	WLM IND COM

Os papéis (ativos) de cada empresa foi obtido no site da BM&FBovespa⁴. Foram inseridos na base de dados apenas os papeis das empresas listadas na tabela 2.

As cotações destes papéis foram obtidas na base histórica de cotações da BM&FBovespa

⁴ http://www.bmfbovespa.com.br/pt_br/servicos/market-data/consultas/mercado-a-vista/titulos-negociaveis/

⁵. Ao buscar os dados obteve-se um arquivo-texto onde cada linha corresponde aos dados de um papel em um dia do ano escolhido, ao todo onze arquivos contemplando os anos de 2005 à 2015.

Todos os dados obtidos estavam disponíveis em planilhas Excel ou arquivos de texto, então foi necessário entendê-los e criar as tabelas no MySQL para receber estes dados. Os dados foram lidos destes arquivos e assim enviados ao banco de dados.

3.2 PRÉ-PROCESSAMENTO E TRANSFORMAÇÃO DOS DADOS

A base de dados foi criada com o nome bovespa usando o *Hibernate* para gerenciamento e criação das entidades e relacionamentos. Na base de dados temos as seguintes informações:

- Informações de Setor (10 instâncias), Subsetor (43 instâncias) e Segmento (93 instâncias) aos quais as empresas pertencem.
- Todos os dados de Balanço Patrimonial das empresas de 2005 à 2015 que gerou 11209 instâncias no banco de dados.
- Todos os dados de Demonstrativos de Resultados das empresas de 2005 à 2015 que gerou 11209 instâncias no banco de dados.
- Os Papéis destas empresas que são negociados na Bovespa que gerou 468 instâncias no banco de dados.
- Para cada Papel existe a cotação diária de 2005 à 2015 com preço de abertura, preço máximo, preço mínimo, preço médio, preço final, preço da melhor oferta de compra, preço da melhor oferta de venda, número de negócios efetuados, quantidade total de títulos negociados e volume total de títulos. Para a tabela de Cotações há um total de 49624 instâncias.

Como era de se esperar os dados de cotações são muitos e precisa de tratamento para reduzir a quantidade de registros a ser utilizada. Na verdade, todos estes dados foram usados para criar os índices discutidos na seção 2.9.

Uma empresa pode ter vários papéis sendo negociados, para cada papel existem várias informações de preço como preço de abertura, preço de fechamento, preço médio no dia, preço mínimo e preço máximo. O preço de fechamento é o foco principal deste

⁵ http://www.bmfbovespa.com.br/pt_br/servicos/market-data/historico/mercado-a-vista/series-historicas/

trabalho, pois ele é o valor mais importante do dia para os analistas de mercado. Ele informa o valor final de negociação de um papel durante o dia e também é um fator importante para definir o preço de abertura no dia seguinte. Como a empresa pode ter vários papéis pegamos a média, variância e desvio padrão durante a semana do papel de maior valor da empresa. O valor mínimo é o mínimo preço de fechamento durante a semana de todos os papéis da empresa e o valor máximo é o máximo preço de fechamento de todos os papéis da empresa.

Foi criada uma tabela de atributos no banco de dados para cada período analisado. Então ficaram três tabelas nomeadas como Atributos_semana, Atributos_mensais e Atributos_semestrais. Nestas tabelas estão consolidados os ganhos da ação no período, preço mínimo, médio, máximo e desvio padrão e também os indicadores calculados para cada empresa.

A coluna ganho possui valor positivo caso a ação tenha tido valorização no período ou negativo caso ela tenha desvalorizado, preço fechamento será o preço de fechamento da ação no último dia do período. Para este cálculo de ganho usamos o preço de fechamento do papel em cada dia da semana.

Os cálculos utilizados para cada indicador são:

- *Pay-Out* = $(\text{tabela BalancoPatrimonial campo pasDividendosJCPPagar} / \text{tabela DemonstrativosResultados campo lucroPrejuizoPeriodo}) * 100$.

Valor percentual referente a 3 meses e quanto maior melhor (Pereira, 2015).

- *Dividend Yield* = $((\text{tabela BalancoPatrimonial campo pasDividendosJCPPagar} / \text{por tabela Empresa campo nroAcoes}) / \text{tabela Cotacao campo precoFechamento na data pesquisada}) * 100$.

Valor dado em percentual, quanto maior melhor.

Estes são os indicadores fundamentalistas mais utilizados:

- *P/L (Preço/Lucro)* = $\text{tabela Cotação campo precoFechamento no dia consultado} / (\text{tabela DemonstrativosResultados campo lucroPrejuizoPeriodo} / \text{tabela Empresa campo nroAcoes})$.
- *VPA (Valor Patrimonial da Ação)* = $\text{tabela BalancoPatrimonial campo patrimonio-Liquido} / \text{tabela Empresa campo nroAcoes}$.
- *Liquidez Corrente* = $\text{tabela BalancoPatrimonial campo ativoCirculante} / \text{tabela BalancoPatrimonial campo passivoCirculante}$.

- $P/VP = \text{tabela Cotação campo precoFechamento no dia consultado} / \text{VPA}$ mostrado no item anterior.
 - $EBIT = (+)$ Receita Líquida: *tabela DemonstrativosResultados campo receitaLiquidaVendasServicos*.
 - $(-)$ Custo dos Produtos Vendido: *tabela DemonstrativosResultados campo custoBensServicosVendidos*.
 - $(-)$ Despesas da Atividade (com vendas, administrativas e outras diretamente ligadas às operações): *DemonstrativosResultados campo despesasComVendas + campo despesasGeraisAdm*.
- Cálculo retirado da cartilha disponibilizada por [Corretora \(2015\)](#).
- Geração de Caixa na Atividade ou $EBITDA = EBIT (+) \text{ Depreciação e Amortização}$. Cálculo retirado da cartilha disponibilizada por [Corretora \(2015\)](#).
 - $EV/EBIT: ((\text{tabela Cotação campo precoFechamento no dia consultado} * \text{tabela Empresa campo nroAcoes}) + \text{tabela BalancoPatrimonial campo pasEmprestimosFinanciamentos} - \text{tabela BalancoPatrimonial campo ativoCirculante}) / EBIT$). Valor do empreendimento foi obtido em [Tibúrcio \(2012\)](#).
 - $NOPLAT = EBIT - \text{tabela DemonstrativoResultado campo iRDiferido}$.
 - $ROIC = NOPLAT / \text{tabela BalancoPatrimonial campo atuInvestimentos}$.
 - $ROE = \text{tabela BalancoPatrimonial campo reservasLucros somar o valor dos quatro últimos balanços} / \text{tabela BalancoPatrimonial campo patrimonioLiquido}$.
 - Indicador $Debt/Equity = \text{Capital de terceiros} / \text{capital próprio}$
- Onde,
- capital de terceiros = *tabela BalancoPatrimonial campo passivoCirculante + campo passivoNaoCirculante*;
- Capital próprio = *tabela BalancoPatrimonial campo patrimonioLiquido*.
- Liquidez em Bolsa é um indicador muito importante para investimentos de curto prazo, como o foco deste trabalho é em investimentos de longo prazo optou-se por não levar em conta este indicador.
 - Valor de Mercado = *tabela Cotacao campo precoFechamento * número total de ações que compõe seu capital* .
 - Fluxo de Caixa Operacional = $Ebitda (-) \text{ tabela DemonstrativosResultados campo iRDiferido (+ ou -) Necessidade de Capital de Giro}$.

Onde,

Necessidade do Capital de Giro = *Ativo Operacional* – *Passivo Operacional*;

Ativo Operacional = *tabela BalancoPatrimonial campo atvContasReceber + campo atvEstoques*.

Passivo Operacional = *tabela BalancoPatrimonial campo pasObrigacoesSociaisTrabalhistas + campo pasFornecedores + campo pasTributosDiferidos*.

- Fluxo de Caixa Livre (ou, *Free Cash Flow*) = *Fluxo de Caixa Operacional* - *tabela BalancoPatrimonial campo atvImobilizado*. Fórmula retirada do trabalho de [Carneiro \(2011\)](#).

Onde:

EBIT (-) *tabela DemonstrativoResultados campo iRDiferido*

(+) *Despesas com depreciação, amortização e exaustão*

(=) *Fluxo de Caixa Operacional (FCO)*

- LPA (Lucro por Ação) = *tabela DemonstrativosResultados campo lucroPrejuizoPeriodo / Número Total de Ações*.
- A Geração de Caixa por Ação (ou *CFS: Cash Flow/Share*) = *EBITDA / Número Total de Ações*.
- β : embora seja um indicador importante para investidores não existe, na base de dados desenvolvida neste trabalho, dados suficientes para calcular com detalhes este indicador, logo não foi utilizado neste trabalho.
- Avaliação do Setor: este dado é um conhecimento que o investidor deve adquirir ao pesquisar o mercado das ações que deseja investir, logo não será utilizada neste trabalho.

Conforme cita [Cavalcante et al \(2009\)](#) o raciocínio básico no uso destes índices que são múltiplos como P/VP, P/EBITDA, P/L e Ev/EBIT é extremamente simples:

- Calculam-se os múltiplos (indicadores) de várias empresas.
- As ações que apresentam os menores múltiplos estão subavaliadas em relação àquela de maiores múltiplos, por proporcionarem retorno em menor tempo elas são opções de investimento.
- As ações com maiores múltiplos estão superavaliadas e, portanto, são opções de venda.

A ideia do investidor é sempre obter o maior retorno com o menor risco. O ganho de capital provido pelas ações têm sempre duas componentes:

- Resultados distribuídos pela companhia: dividendos, bonificações, juros sobre capital próprio.
- Resultados obtidos com operações de mercado: ganho líquido obtido com compra e venda de ações.

Após calcular todos os indicadores que servirão de atributos a base de dados foi normalizada para preparar as informações para a execução dos algoritmos. Os dados foram normalizados com os valores indo de 0 a 1 seguindo a equação 3.1.

$$Valor_{norm} = \frac{ValorAtual_a - MenorValor_a}{MaiorValor_a - MenorValor_a} \quad (3.1)$$

Onde a indica a série de dados do atributo a ser normalizado.

3.2.1 Classificação das Empresas

Para saber se um investimento está tendo bom retorno é indicado comparar este investimento com o retorno dos investimentos mais conservadores. Olhar primeiro a taxa básica de juros da economia que é a Selic (Sistema Especial de Liquidação e Custódia), outra taxa a ser olhada é a CDI (Certificado de Depósito Interbancário) que normalmente fica um pouco abaixo da Selic. Taxas acima da Selic são indicadores de bom investimento, mas deve-se ficar atento ao risco.

Após algumas pesquisas em sites especializados em investimentos, a classificação das empresas foi vinculada à taxa Selic, pois ela é usada como referência na maioria dos investimentos de renda fixa.

Esta comparação é citada por [Cavalcante et al \(2009\)](#) como prêmio pelo risco, onde ele é a diferença entre a taxa de juros de uma aplicação com risco e de uma aplicação sem risco. Dessa forma, se os investimentos em renda fixa rendem, por exemplo, 12% ao ano, um investidor poderá preferir investir em renda variável se seu rendimento projetado estiver na casa dos 20% (com um prêmio de risco potencial de 8%).

Os valores históricos das taxas foram obtidos no site do Banco Central do Brasil⁶. Após buscar o histórico destas taxas foi possível gerar o gráfico da Figura 5 que mostra os valores da taxa Selic, CDI e rendimento da Caderneta de Poupança no período pesquisado que é de 2005 a 2015. É possível observar que a Selic está sempre igual ou acima da CDI e sempre maior que a taxa de rendimento da caderneta de poupança.

⁶ <https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries>

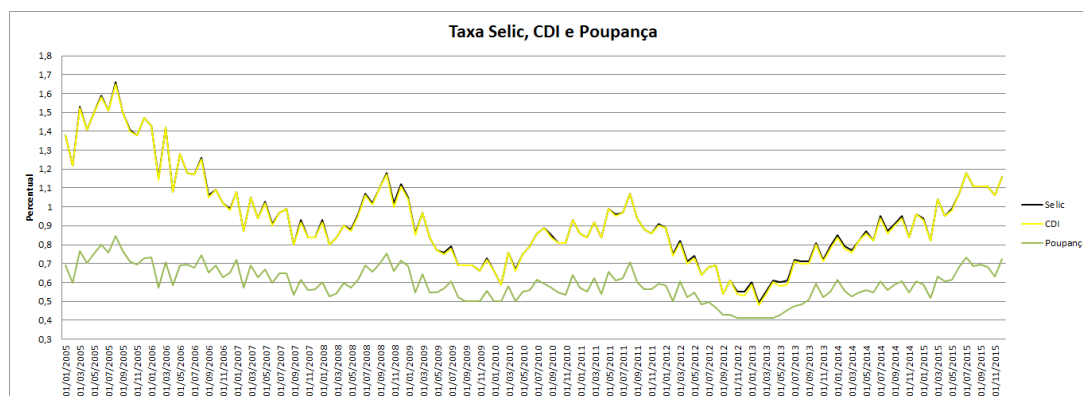


Figura 5 – Taxa Selic, CDI e Rendimento da Poupança nos anos de 2005 a 2015.

Conforme cita o site [InfoMoney](#) (2007) a taxa Selic é a taxa de financiamento no mercado interbancário para operações de um dia, ou *overnight*, que possuem lastro em títulos públicos federais, títulos estes que são listados e negociados no Sistema Especial de Liquidação e Custódia (Selic). Esta taxa serve de referência para todas as demais taxas de juros da economia brasileira. Esta taxa não é fixa e varia praticamente todos os dias, mas dentro de um intervalo muito pequeno, já que, na grande maioria das vezes, ela tende a se aproximar da meta Selic, que é determinada mensalmente pelo Copom (Comitê de Política Monetária do Banco Central).

O CDI (Certificado de Depósito Interbancário) foi criado para lastrear as operações de curtíssimo prazo entre bancos. A taxa média diária da CDI é calculada com base nas operações de emissão de certificados de um dia e é sempre muito próxima da taxa de juro básica da economia que é a Selic. Por ser muito importante no mercado interbancário o CDI acaba servindo de referência para outras taxas praticadas pelos bancos e também é utilizada como referencial (*benchmark*) para a rentabilidade das aplicações financeiras, principalmente de renda fixa.

Diante destas informações para o presente trabalho decidiu-se usar como base para classificação das ações a taxa Selic, pois assim é possível comparar o rendimento das empresas do mercado de ações ao rendimento das aplicações de renda fixa que possuem risco menor. Portanto, será feita a classificação trimestral levando em consideração o rendimento das ações das empresas no trimestre:

- Fraco: empresas cujas ações tiveram rendimento abaixo de 80% da taxa Selic no trimestre.
- Bom: empresas cujas ações tiveram rendimento entre 80% e 120% da taxa Selic.
- Muito bom: rendimento maior que 120% da taxa Selic.

3.2.2 Banco de Dados Bovespa

Após levantar todos os indicadores, dados importantes para cálculo dos indicadores e formas de classificação foi construída a base de dados com o nome Bovespa.

A base de dados de foi construída de forma que ficasse mais fácil e rápido para exportar os dados. Portato, não foi feita a otimização do banco de dados.

A Figura 6 mostra o modelo da base de dados gerada com todos os dados coletados das ações em negociação em 2015.

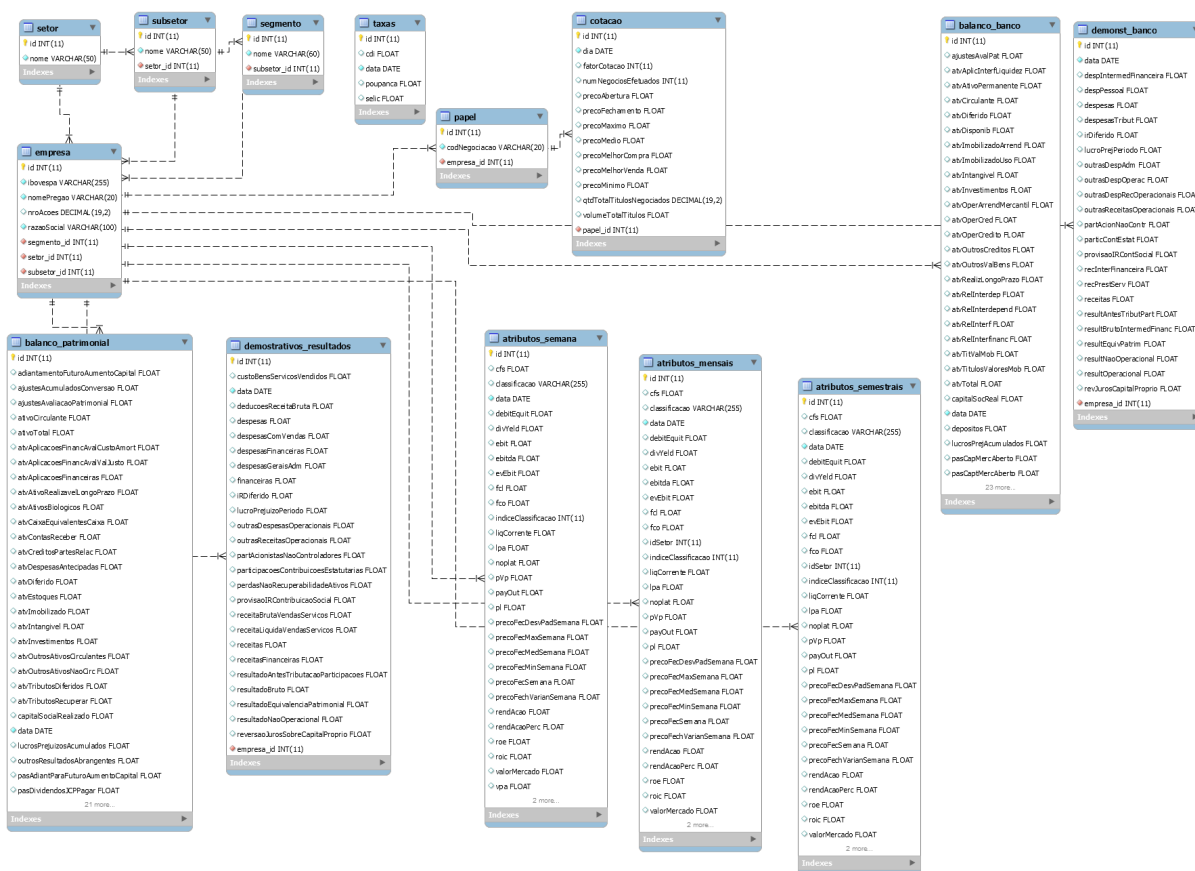


Figura 6 – Base de dados Bovespa

A base de dados possui:

- O setor, subsetor e segmento ao qual uma empresa pertence.
- As empresas com seus respectivos papéis que são negociados na bolsa de valores BM&FBovespa.
- Os papéis com suas respectivas cotações que são dados diários.
- O balanço patrimonial de cada empresa que são dados trimestrais.

- O demonstrativo de resultado de cada empresa que são dados trimestrais.
- O balanço patrimonial e demonstrativo de resultados dos bancos, que são dados trimestrais e tem uma estrutura diferente das outras empresas.
- O histórico das taxas Selic, CDI e Rendimento da Caderneta de Poupança dos quais pode-se comparar os rendimentos das ações.
- De posse de todos estes foram gerados os atributos semanais, mensais e semestrais das empresas para analisar com os algoritmos de mineração de dados.

Após finalizar o banco de dados várias análises foram feitas a fim de entender melhor nossos dados, é disto que trata a próxima seção.

3.2.3 Sumarização dos Dados

Antes de rodar os algoritmos é necessário entender a base de dados e extrair informações estatísticas do número de dados. Esta seção é destinada a fazer análises estatísticas em cima dos indicadores calculados e com as classificações geradas.

O programa desenvolvido nos dá a opção de exportar dados em três períodos diferentes:

- análise por semana: os atributos serão gerados por semana e estarão disponíveis junto com a classificação para cada empresa na semana na tabela `AtributosSemana`;
- análise por mês: os atributos serão gerados por mês e estarão disponíveis junto com a classificação para cada empresa na semana na tabela `AtributosMensais`;
- análise por semestre: os atributos serão gerados por semestre e estarão disponíveis junto com a classificação para cada empresa no semestre na tabela `AtributosSemestre`;

Após gerar todos os atributos e classificação das empresas foi possível verificar que:

- 139233 instâncias de atributos que foram gerados a partir da análise das ações de cada empresa por semana. Destas 98,48% são investimentos classificados como fraco, 0,74% são investimentos classificados como bons e 0,78% são investimentos classificados como muito bom.
- 32423 instâncias de atributos que foram gerados a partir da análise das ações de cada empresa por mês. Destas 97,75% são investimentos classificados como fraco, 1,03% são investimentos classificados como bons e 1,22% são investimentos classificados como muito bom.

- 4223 instâncias de atributos que foram gerados a partir da análise das ações de cada empresa por semestre. Destas 98,25% são investimentos classificados como fraco, 0,66% são investimentos classificados como bons e 1,09% são investimentos classificados como muito bom.

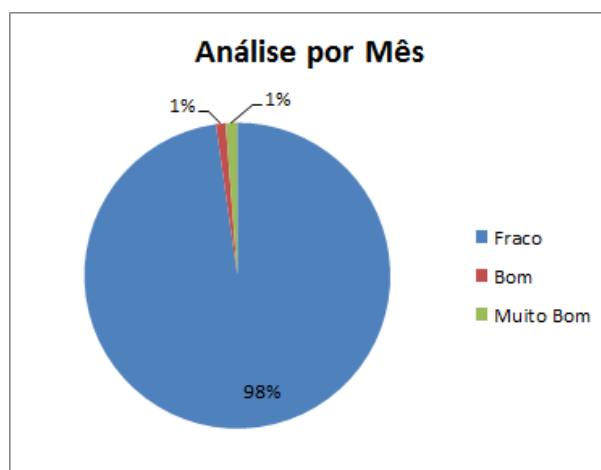


Figura 7 – Percentual de empresas em cada classe nos períodos pesquisados.

A Figura 7 mostra uma distribuição dos percentuais em 10 anos de análise. Veja que os dados são desanimadores, pois apenas 2% das empresas tiveram classificação Bom e Muito Bom. 98% dos ativos analisados no período tiveram desempenho fraco comparado com a Selic conforme citado no ítem 3.2.1 deste trabalho.

Sobre estes períodos de rendimentos classificados como "Bom" e "Muito Bom", ou seja, rendimento superior a 80% da taxa Selic, realizou-se uma análise por setor para verificar como está a distribuição dentro dos setores. A Figura 8 mostra o percentual de empresas em cada classe nos períodos analisados separado por setor.

Setor Id	Nome do Setor	Semana			Mês			Semestre		
		Fraco	Bom	Muito Bom	Fraco	Bom	Muito Bom	Fraco	Bom	Muito Bom
1	Bens Industriais	98,13	0,87	1,00	97,65	1,08	1,26	97,69	0,58	1,73
2	Construção e Transporte	98,76	0,54	0,70	97,74	0,83	1,44	98,09	0,44	1,47
3	Consumo Cíclico	98,16	0,80	1,04	97,27	1,24	1,50	97,59	1,02	1,40
4	Consumo Não Cíclico	99,53	0,30	0,17	99,34	0,49	0,16	99,78	0,00	0,22
5	Financeiro e Outros	98,21	0,94	0,85	98,02	0,46	1,52	98,66	0,22	1,11
6	Materiais Básicos	97,50	1,26	1,24	96,62	1,46	1,92	97,05	1,38	1,57
7	Petróleo Gás e Combustíveis	98,10	1,03	0,88	98,17	1,61	0,23	98,78	1,22	0,00
8	Tecnologia da Informação	99,88	0,08	0,04	99,80	0,00	0,20	97,75	1,12	1,12
9	Telecomunicações	99,16	0,37	0,48	96,42	2,61	0,98	98,39	0,00	1,61
10	Utilidade Pública	98,98	0,53	0,49	97,81	1,35	0,84	99,32	0,68	0,00

Figura 8 – Percentual de classificação de cada setor nos períodos analisados.

A Figura 9 mostra um gráfico com o percentual por setor de rendimentos acima de 80% da Selic, ou seja, classificados como "Bom" e "Muito Bom". Pode-se observar que os setores de Telecomunicações, Materiais Básicos e Consumo Cíclico tiveram mais empresas classificadas como "Bom" e "Muito Bom" nos últimos 10 anos.

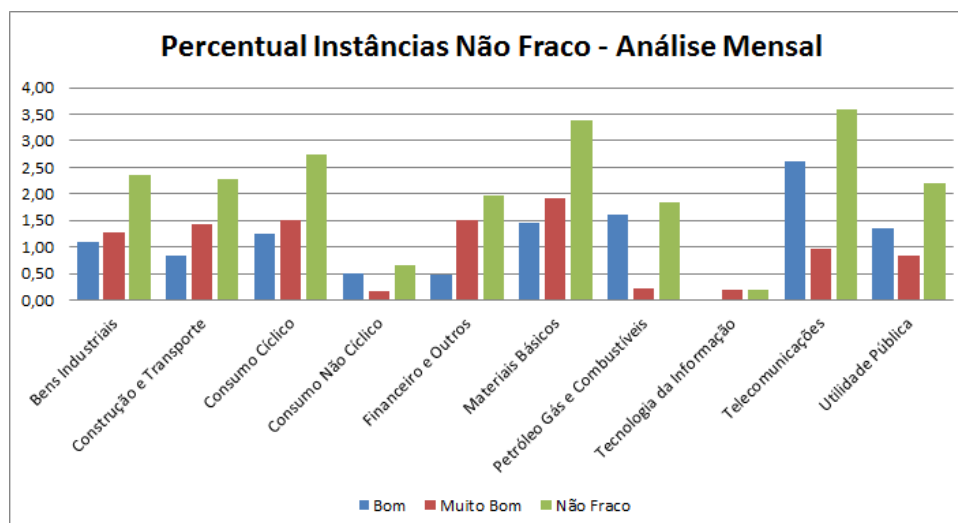


Figura 9 – Percentual de classificação "Bom" e "Muito Bom" de cada setor no período mensal.

Com estes gráficos já é possível observar que poucas empresas brasileiras apresentam bons rendimentos em suas ações. Portanto, para não perder dinheiro o investidor precisa pesquisar muito antes de investir. Olhando para estes dados já surge a dúvida: Quais são estas empresas que renderam acima de 80% da Selic? Em qual período isto aconteceu?

Consultando a base de dados construída neste trabalho é possível responder a estas perguntas. Foi feita a busca por todas as empresas e suas respectivas datas com classificação "Bom" e "Muito Bom". Estes dados foram exportados para o Excel e filtrados a fim de identificar quais empresas estão dentro destes 2% que renderam mais de 80% da Selic. 110 empresas fazem parte deste grupo, sendo elas mostradas na Tabela 3.

Foram feitas mais pesquisas para saber com que frequência as empresas tiveram desempenho acima de 80% da Selic, a fim de entender quais são mais recorrentes como "Bom" e "Muito Bom". Os gráficos mostrados na figura 10 apresentam a frequência em meses que cada empresa teve rendimentos acima de 80% da Selic.

As empresas com maior incidência de classificação "Muito Bom" (rendimento acima de 120% da taxa Selic) estão mostradas na Tabela 4. A tabela 5 mostra as empresas que tiveram pelo menos por 6 meses nessa situação nos últimos 10 anos.

Analisando apenas o critério de quantidade de meses com rendimentos muito altos, acima de 120% da taxa Selic, a Tabela 4 mostra ações da Telebras, JB Duarte, Metal Iguaçu, Tectoy, Unipar e Doc Imbituba como melhores opções. Foram feitas mais análises sobre a base de dados para ver se este mesmo grupo aparece analisando sobre outras perspectivas.

Fazendo a soma de frequências "Boas" e "Muito Boas" pode-se identificar as empresas que tiveram por pelo menos 6 meses, nestes últimos 10 anos, rendimentos acima de

Tabela 3 – Empresas que tiveram rendimento acima de 80% da Selic em, pelo menos, um mês nos últimos 10 anos

Índ.	Nome de Pregão	Índ.	Nome de Pregão	Índ.	Nome de Pregão
1	ALFA CONSORC	38	ENERGISA	75	NORDON MET
2	ALPARGATAS	39	ESTRELA	76	NOVA OLEO
3	AMPLA ENERG	40	EUCATEX	77	OI
4	AZEVEDO	41	EXCELSIOR	78	OSX BRASIL
5	B2W DIGITAL	42	FERBASA	79	PANATLANTICA
6	BIOMM	43	FORJA TAURUS	80	PET MANGUINH
7	BR BROKERS	44	FRAS-LE	81	PETROBRAS
8	BRADSPAR	45	GAFISA	82	PLASCAR PART
9	BRASKEM	46	GER PARANAP	83	RANDON PART
10	BRASMOTOR	47	GERDAU MET	84	RECRUSUL
11	CAMBUCI	48	GPC PART	85	REDE ENERGIA
12	CCX CARVAO	49	GRAZZIOTIN	86	RIOSULENSE
13	CEB	50	GUARARAPES	87	RJCP
14	CEDRO	51	HAGA S/A	88	ROSSI RESID
15	CEEE-D	52	HERCULES	89	SANSUY
16	CEEE-GT	53	HOTEIS OTHON	90	SARAIVA LIVR
17	CELESC	54	IDEIASNET	91	SONDOTECNICA
18	CELPA	55	IGB S/A	92	SPRINGER
19	CELPE	56	INEPAR	93	SULTEPA
20	CELUL IRANI	57	ITAUSA	94	TARPON INV
21	CESP	58	J B DUARTE	95	TEC BLUMENAU
22	CHIARELLI	59	JEREISSATI	96	TECEL S JOSE
23	CIELO	60	JOAO FORTES	97	TECNOSOLO
24	COBRASMA	61	KARSTEN	98	TECTOY
25	COELCE	62	KEPLER WEBER	99	TEKA
26	CONTAX	63	KLABIN S/A	100	TELEBRAS
27	COSERN	64	KROTON	101	TEX RENAUX
28	COTEMINAS	65	LA FONTE TEL	102	TREVISA
29	CYRE COM-CCP	66	LIX DA CUNHA	103	UNIPAR
30	D H B	67	MANGELS INDL	104	USIMINAS
31	DIMED	68	MARCOPOLO	105	V-AGRO
32	DOC IMBITUBA	69	MELHOR SP	106	VALE
33	ELEKEIROZ	70	MENDES JR	107	VULCABRAS
34	ELEKTRO	71	METAL IGUACU	108	WETZEL S/A
35	ELETROBRAS	72	MINUPAR	109	WHIRLPOOL
36	ELETROPAULO	73	MMX MINER	110	WLM IND COM
37	ENCORPAR	74	MUNDIAL		

80%. Isto é mostrado na Tabela 5.

As empresas que tiveram maior frequência de meses com rendimento acima de 80% da Selic, foram Telebrás, Metal Iguaçu, Unipar, Tecnosolo, Tectoy, Recrusul, CEB, JB Duarte, Doc Imbituba e Chiarelli.

Os grupos são compatíveis como opções de investimento. Foram realizadas consultas para obter a valorização das ações, nos últimos 10 anos, de todas as empresas, e foi feita a Tabela 6 de sumarização ordenada por ordem decrescente com relação aos ganhos.

Com método de comparação foi feita uma consulta na base de dados para saber quanto a taxa Selic rendeu de 2006 em diante. A Selic rendeu de 2006 à 2015 110,10%. A média mensal deste período deu 0,89%. Na Tabela 6 é possível observar que a Tectoy, a JB Duarte, Elekeiroz e Mundial tiveram excelentes rendimentos nos últimos 10 anos.

Então, apenas por análises estatísticas aplicadas a base de dados e levando em consideração somente o rendimento, seria indicado a um investido acompanhar os preços das ações das empresas Tectoy, a JB Duarte, Elekeiroz e Mundial em busca de boas

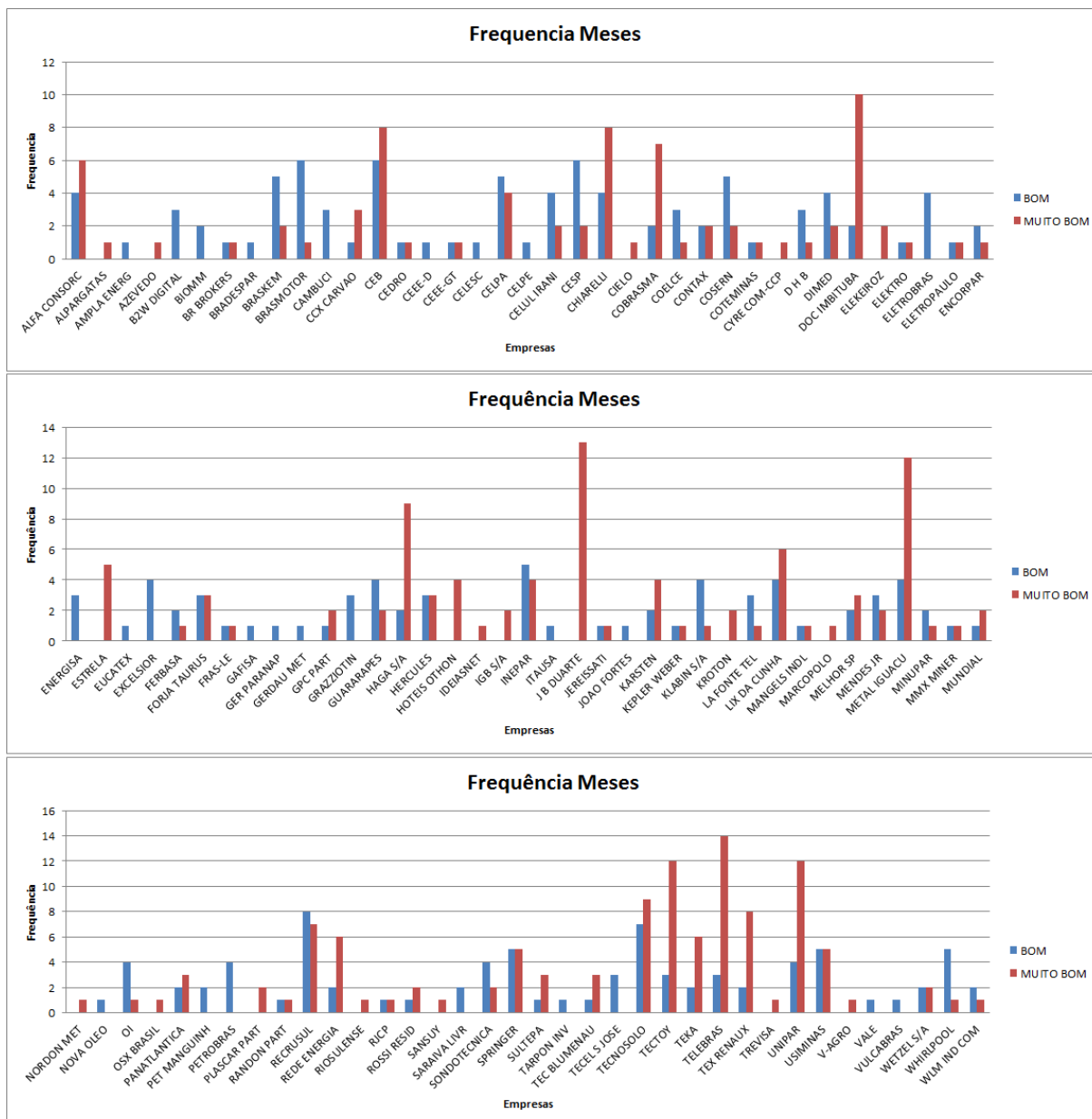


Figura 10 – Frequência em meses que cada empresa teve classificação "Bom" e "Muito Bom".

oportunidades de compra.

Com esta base de dados é possível fazer diversas análises, mas este não é o objetivo deste trabalho, com estes levantamentos já foi possível entender a base de dados de forma eficiente para partir para o treinamento dos algoritmos de mineração de dados.

3.2.4 Dificuldades Encontradas

Ao fazer a organização dos dados foi identificado que a estrutura do balanço patrimonial e demonstrativos de resultados dos bancos é diferente das demais empresas, portanto, por questão de tempo para o desenvolvimento do trabalho, os bancos foram re-

Tabela 4 – Empresas que tiveram maior frequência de meses com rendimento acima de 120% da Selic nos últimos 10 anos

Índ.	Nome de Pregão	Bom	Muito Bom
1	TELEBRAS	3	14
2	J B DUARTE	0	13
3	METAL IGUACU	4	12
4	TECTOY	3	12
5	UNIPAR	4	12
6	DOC IMBITUBA	2	10
7	HAGA S/A	2	9
8	TECNOSOLO	7	9
9	CEB	6	8
10	CHIARELLI	4	8
11	TEX RENAUX	2	8
12	COBRASMA	2	7
13	RECRUSUL	8	7
14	ALFA CONSORC	4	6
15	LIX DA CUNHA	4	6
16	REDE ENERGIA	2	6
17	TEKA	2	6

Tabela 5 – Empresas que tiveram maior frequência de meses com rendimento acima de 80% da Selic nos últimos 10 anos

Índ.	Nome de Pregão	Bom	Muito Bom	Soma
1	TELEBRAS	3	14	17
2	METAL IGUACU	4	12	16
3	UNIPAR	4	12	16
4	TECNOSOLO	7	9	16
5	TECTOY	3	12	15
6	RECRUSUL	8	7	15
7	CEB	6	8	14
8	J B DUARTE	0	13	13
9	DOC IMBITUBA	2	10	12
10	CHIARELLI	4	8	12
11	HAGA S/A	2	9	11
12	TEX RENAUX	2	8	10
13	ALFA CONSORC	4	6	10
14	LIX DA CUNHA	4	6	10
15	SPRINGER	5	5	10
16	USIMINAS	5	5	10
17	COBRASMA	2	7	9
18	CELPA	5	4	9
19	INEPAR	5	4	9
20	REDE ENERGIA	2	6	8
21	TEKA	2	6	8
22	CESP	6	2	8
23	BRASKEM	5	2	7
24	COSERN	5	2	7
25	BRASMOTOR	6	1	7
26	KARSTEN	2	4	6
27	FORJA TAURUS	3	3	6
28	HERCULES	3	3	6
29	CELUL IRANI	4	2	6
30	DIMED	4	2	6
31	GUARARAPES	4	2	6
32	SONDOTECNICA	4	2	6
33	WHIRLPOOL	5	1	6

tirados da análise, pois os cálculos de indicadores para eles seriam alterados. Entretanto, foi criada uma tabela de Demonstrativo de Resultados e Balanço Patrimonial para os bancos para que futuramente eles possam ser usados para análises.

Os dados de mercado financeiro são muito volumosos e a variação de valores entre as empresas mudam muito, isto torna os dados muito diversificados e dificulta a organi-

Tabela 6 – Empresas que tiveram maior rendimento acumulado nos últimos 10 anos.

Nome de Pregão	idSetor	Percentual de Ganhos	Meses	Rendimento Médio Mensal
TECTOY	3	653,18	117	5,59
J B DUARTE	5	261,04	117	2,24
ELEKEIROZ	6	226,75	114	1,99
MUNDIAL	3	144,05	115	1,26
TELEBRAS	5	58,82	117	0,51
CONTAX	1	53,93	117	0,47
UNIPAR	6	53,32	117	0,46
KEPLER WEBER	1	48,03	116	0,42
PLASCAR PART	1	36,84	102	0,37
HAGA S/A	2	30,98	111	0,28
BRASMOTOR	3	30,15	108	0,28
METAL IGUACU	6	28,62	115	0,25

zação e tratamento destes valores.

Para uma pessoa comum que decide investir na bolsa de valores e não tem acesso a sites de divulgação de informações como o Economática é difícil conseguir buscar todas as informações das empresas de forma a fazer uma análise fundamentalista completa. A maior dificuldade encontrada foi obter o histórico de quantidade de ações e dados de depreciação e amortização de cada empresa. A falta de praticidade para conseguir estes valores no site da BM&FBovespa não permitiu que estes dados fossem incluídos no trabalho.

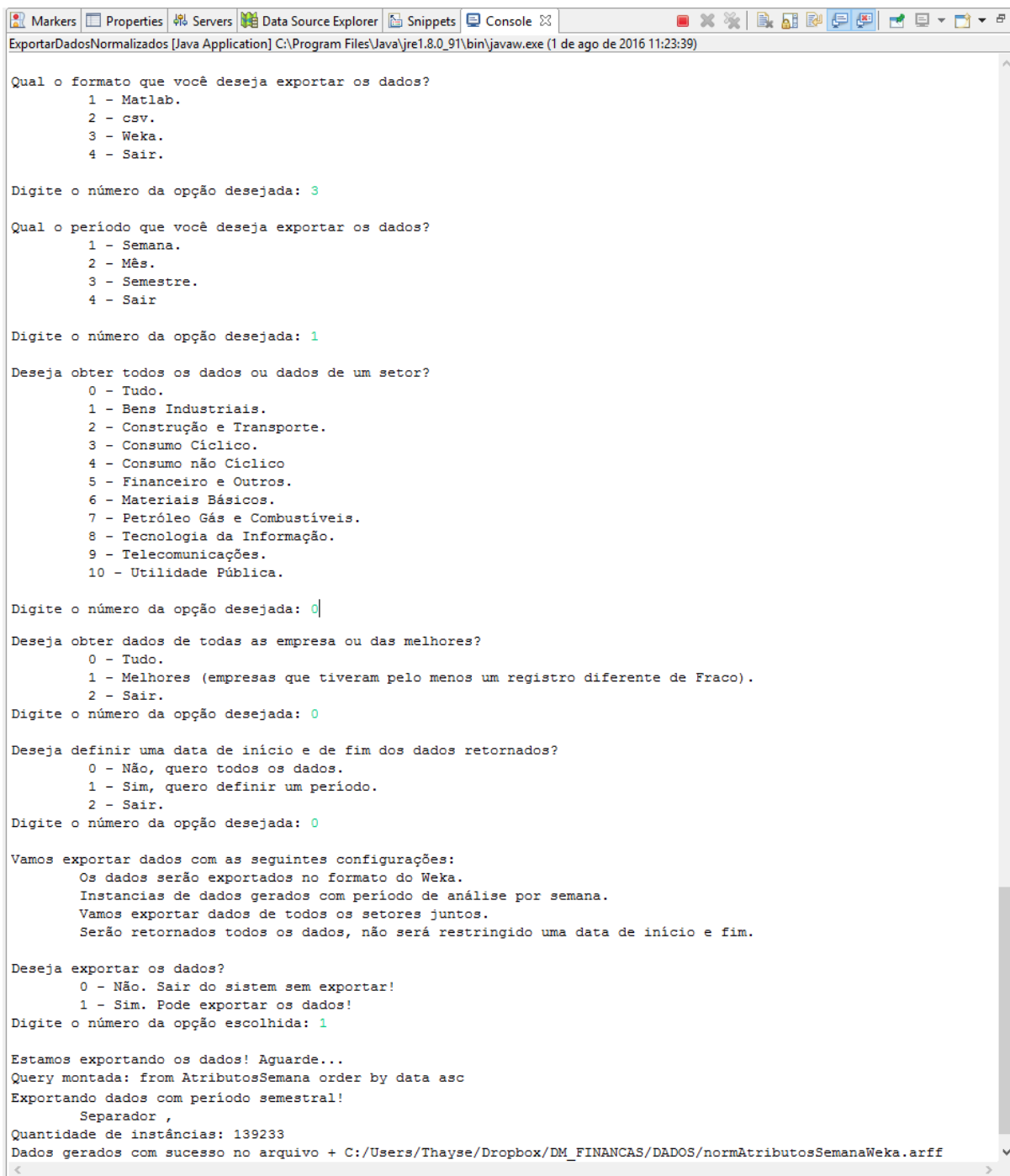
Existem muitos sites que divulgam informações de mercado financeiro e livros com abordagens diversas de análise de mercados financeiros. Normalmente a bibliografia cita índices como liquidez em bolsa e beta como importantes para análise de ativos, entretanto, alguns especialistas como Warren Buffet indicam não dar tanto peso a estes indicadores pois, investindo a longo prazo, eles não vão fazer diferença e são muito complicados para o investidor obter seus resultados. Então é difícil obter informações suficientes para realizar boas análises.

A etapa de pré-processamento e transformação dos dados foi a etapa que demandou mais tempo no desenvolvimento do trabalho.

3.3 INTERFACE GRÁFICA

Foi desenvolvido um menu de interface para que o usuário escolha os dados e o formato em que ele deseja exportar as informações. Prevendo que o usuário pode querer escolher outros programas como o matlab para aplicar algoritmos sobre os dados foram desenvolvidas as seguintes opções:

- Em qual formato o usuário deseja exportar os dados: Matlab, csv ou *WEKA*;
- Qual período o usuário deseja: semanal, mensal ou semestral;
- Se o usuário deseja exportar todos os dados ou os dados de algum setor;



```
ExportarDadosNormalizados [Java Application] C:\Program Files\Java\jre1.8.0_91\bin\javaw.exe (1 de ago de 2016 11:23:39)

Qual o formato que você deseja exportar os dados?
  1 - Matlab.
  2 - csv.
  3 - Weka.
  4 - Sair.

Digite o número da opção desejada: 3

Qual o período que você deseja exportar os dados?
  1 - Semana.
  2 - Mês.
  3 - Semestre.
  4 - Sair

Digite o número da opção desejada: 1

Deseja obter todos os dados ou dados de um setor?
  0 - Tudo.
  1 - Bens Industriais.
  2 - Construção e Transporte.
  3 - Consumo Cíclico.
  4 - Consumo não Cíclico
  5 - Financeiro e Outros.
  6 - Materiais Básicos.
  7 - Petróleo Gás e Combustíveis.
  8 - Tecnologia da Informação.
  9 - Telecomunicações.
  10 - Utilidade Pública.

Digite o número da opção desejada: 0

Deseja obter dados de todas as empresa ou das melhores?
  0 - Tudo.
  1 - Melhores (empresas que tiveram pelo menos um registro diferente de Fraco).
  2 - Sair.

Digite o número da opção desejada: 0

Deseja definir uma data de início e de fim dos dados retornados?
  0 - Não, quero todos os dados.
  1 - Sim, quero definir um período.
  2 - Sair.

Digite o número da opção desejada: 0

Vamos exportar dados com as seguintes configurações:
  Os dados serão exportados no formato do Weka.
  Instancias de dados gerados com período de análise por semana.
  Vamos exportar dados de todos os setores juntos.
  Serão retornados todos os dados, não será restringido uma data de início e fim.

Deseja exportar os dados?
  0 - Não. Sair do sistem sem exportar!
  1 - Sim. Pode exportar os dados!

Digite o número da opção escolhida: 1

Estamos exportando os dados! Aguarde...
Query montada: from AtributosSemana order by data asc
Exportando dados com período semestral!
  Separador ,
Quantidade de instâncias: 139233
Dados gerados com sucesso no arquivo + C:/Users/Thayse/Dropbox/DM_FINANCAS/DADOS/normAtributosSemanaWeka.arff
```

Figura 11 – Menu de interface com os usuários.

- Deseja obter dados de todas as empresas ou daquelas de melhor desempenho;
- Deseja limitar uma data de início e de fim para exportar os dados;
- Deseja exportar os dados: sim ou não.

Os dados exportados são mostrados na figura 12. Conforme exigido pelo *WEKA* ele gera o cabeçalho e o arquivo já está com os valores normalizados.

```

1 % Description Bovespa dataset.
2 % 1. Title: Bovespa Data
3 % 2. Source Information
4 %
5 % Thaysse Cristina Araújo Rodrigues
6 % Graduando em Engenharia de Computação 2016
7 % Universidade Federal de Ouro Preto
8 % ICEA-João Monlevade-MG
9 % 3. Number of instances: 139233
10 @relation Bovespa
11 @attribute "Id" numeric
12 @attribute "PayOut" real
13 @attribute "DivYield" real
14 @attribute "Pl" real
15 @attribute "Vpa" real
16 @attribute "LiqCorrente" real
17 @attribute "Pvp" real
18 @attribute "Ebit" real
19 @attribute "Ebitda" real
20 @attribute "Evebit" real
21 @attribute "Noplat" real
22 @attribute "Roi" real
23 @attribute "Roe" real
24 @attribute "DebitEquit" real
25 @attribute "ValorMercado" real
26 @attribute "Fco" real
27 @attribute "Fcl" real
28 @attribute "Lpa" real
29 @attribute "Cfs" real
30 @attribute "IdSetor" real
31 @attribute "Type" { FRACO, BOM, "MUITO BOM"}
32 @data
33 33,0,0.8734268855,0.0000045686,0.0175521094,0.0943525978,0.0000836944,0.6465848000,0.2480180109,0.2480180109,0.5344270693,0.2485874105,0.1877077477,0.4662016373,0.0009306645,0E-10,0.1840770810,0.4820148772,0.2047838492,0.4398580991,0.33333334,FRACO
34 44,0,0.8721907285,0.0000045686,0.0175521094,0.0956653294,0.0003126121,0.6465848000,0.2946486725,0.2946486725,0.5344270692,0.2952108753,0.1879329392,0.4668995420,0.0009395420,0E-10,0.1258215304,0.4933091967,0.2047689239,0.4814097373,0.22222222,FRACO
35 55,0,0.8734268855,0.0000045686,0.0175521094,0.0942863430,0.0001895353,0.6465848000,0.2481609053,0.2481609053,0.5344270690,0.2487359148,0.1877579373,0.4662016373,0.0009415826,0E-10,0.1041145340,0.4820327474,0.2046710387,0.4403363898,0.11111111,FRACO
36 48,0,0.8734268855,0.0000045686,0.0175521094,0.0933084846,0.0001752842,0.6465848000,0.2482703184,0.2482703184,0.5344270694,0.2488402675,0.1878483466,0.4662016373,0.0009332150,0E-10,0.1842151439,0.4820763937,0.2041725270,0.4415309446,0.22222222,FRACO
37 103,0,0.8734268855,0.0000045686,0.0175521094,0.0444242374,0.0000412954,0.6465848000,0.2497951788,0.2497951788,0.5344270697,0.2503643906,0.1898386601,0.4662099862,0.0009343189,0E-10,0.1048277727,0.4824659900,0.1988133290,0.5479208131,0.0,FRACO
38 112,0,0.8736226602,0.3544602546,0.0175521094,0.0946793304,0.0002668435,0.6465848000,0.2488831806,0.2488831806,0.5344270690,0.2495663676,0.1878447896,0.4662327135,0.000378014,0E-10,0.1047211417,0.4811450383,0.2049560437,0.4407640626,0.11111111,FRACO

```

Figura 12 – Menu de interface com os usuários.

Com a escolha do formato e dos dados que podem ser exportados é possível usar algoritmos no Matlab sobre estes dados ou usar como entrada para qualquer software que recebe formatos csv. Abre-se muitas possibilidades de análise com diversas ferramentas diferentes que não foram usadas neste trabalho mas pode ser usado para trabalhos futuros.

3.4 APLICAÇÃO DOS ALGORITMOS

Como visto na seção 3.2.3 a análise por mês foi a que teve maior percentual de dados "Bons" e "Muito Bons" para analisar, sendo 2,25% dos dados distribuídos nestas duas classes e o restante sendo classificadas como "Fraco". Desta forma, optou-se utilizar os dados mensais para aplicar os algoritmos.

Ao aplicar os algoritmos foram usadas algumas medidas para comparar o desempenho, estas medidas foram mostradas no item 2.3. A ferramenta *WEKA* nos mostra estes indicadores, sendo elas: *CCI* (*Correctly Classified Instances*), *F-Measure* de cada algoritmo em cada classe. Para facilitar a visualização em um mesmo gráfico, os dados de *CCI* foram normalizadas.

O método de validação usado na execução dos algoritmos foi Validação Cruzada de k partes. Conforme cita Tan et al (2009) esta abordagem segmenta os dados em k partições iguais, onde durante cada execução uma destas partições é escolhida para teste, enquanto que as outras são usadas para treinamento. Este procedimento é repetido k vezes, de modo que cada partição seja usada para teste exatamente uma vez. O erro total

é encontrado pela soma do erro de todas as k execuções. Para este trabalho foi escolhido o valor de k igual a dez.

Ian H. Witten (2011) cita que testes extensivos, sobre vários conjuntos de dados diferentes, com diferentes técnicas de aprendizagem, têm mostrado que dez é o número certo de divisões para obter a melhor estimativa de erro, e há também algumas evidências teóricas que corroboram com esta afirmação. Embora estes argumentos não são conclusivos, e o debate continua em aprendizado de máquina e círculos de mineração de dados sobre o melhor esquema para avaliação, Validação Cruzada com k igual a dez tornou-se o método padrão em termos práticos.

Em todos os experimentos foram geradas tabelas contendo os dados, estas tabelas mostram os dados ordenados pelo maior valor da coluna "Soma Bons". Em todos os campos desta tabela quanto mais próximo de 1 o valor estiver significa que é a melhor opção. Todos os experimentos mostraram facilidade dos algoritmos para classificar dados da classe "Fraco", então a coluna "Soma Bons" é a mais importante, pois ela mostra a soma do desempenho do algoritmo para reconhecer instâncias das classes "Bom" e "Muito Bom".

No primeiro experimento executamos os algoritmos sobre todos os dados, com todos os indicadores calculados. Os resultados obtidos estão na Tabela 7:

Tabela 7 – Resultado dos algoritmos de mineração aplicados aos dados totais.

Algoritmo	CCI	Fraco	Bom	Muito Bom	Soma Bons
<i>RandomTree</i>	0,9626	0,981	0,042	0,123	0,165
<i>RandomCommitee</i>	0,9672	0,984	0,043	0,12	0,163
<i>Ibk</i>	0,9633	0,982	0,043	0,111	0,154
<i>RandomForest</i>	0,97	0,985	0,025	0,112	0,137
<i>ClassificationRegression</i>	0,9773	0,989	0,008	0,007	0,015
<i>REPTree</i>	0,9775	0,989	0	0,014	0,014
<i>Bagging Meta</i>	0,9775	0,989	0,009	0	0,009
<i>J48</i>	0,9775	0,989	0	0	0
<i>LogitBoost</i>	0,9775	0,989	0	0	0
<i>MultiClassClassifier</i>	0,9774	0,989	0	0	0

```

      a      b      c  <-- classified as
21728  177   201 |   a = FRACO
   213    9    11 |   b = BOM
   233   10   32 |   c = MUITO BOM

```

Figura 13 – Matriz de confusão do algoritmo Random Tree.

Na tabela 7 resultados foram ordenados pelo maior valor da coluna "Soma Bons" e pode-se observar que, o algoritmo *Random Tree* apresentou o melhor desempenho para este experimento. Os resultados mostram aproximadamente 96% a 97,7% de instâncias classificadas corretamente, mas como somente 2% destes dados não são dados "Fracos" é apresentada na figura 13 a matriz de confusão do algoritmo *Random Tree*. É possível observar que poucos dados da classe "Bom" e "Muito Bom" foram classificados corretamente, portanto os valores *F-measure* mostrados nas colunas "Bom" e "Muito Bom" confirmam esta conclusão.

As empresas pertencem a setores diferentes, são dez setores na BM&FBovespa, sendo eles: Bens Industriais, Construção e Transporte, Consumo Cíclico, Consumo não Cíclico, Financeiro e Outros, Materiais Básicos, Petróleo Gás e Combustíveis, Tecnologia da Informação, Telecomunicações e Utilidade Pública.

Os algoritmos foram executados dividindo as empresas por seus setores. Os algoritmos que não conseguiram reconhecer nenhuma instância de "Bom" e "Muito Bom" no teste anterior, ou seja, algoritmos que apresentaram valor zero na coluna "Soma Bons" da Tabela 7, foram retirados das próximas análises, sendo eles *LogitBoost*, J48 e *Multiclass Classifier*.

Tabela 8 – Resultado dos algoritmos de mineração aplicados aos dados do Setor de Bens Industriais.

Algoritmo	CCI	Fraco	Bom	Muito Bom	Soma Bons
<i>RandomForest</i>	96,86	0,984	0,125	0,118	0,243
<i>RandomCommitee</i>	96,5	0,982	0,107	0,113	0,22
<i>RandomTree</i>	96,28	0,981	0,075	0,1	0,175
<i>ClassificationRegression</i>	97,65	0,988	0,061	0,054	0,115
<i>Bagging Meta</i>	97,65	0,988	0	0	0
<i>REPTree</i>	97,61	0,988	0	0	0
<i>Ibk</i>	97,7	0,988	0	0	0

```

a    b    c  <-- classified as
2676 15   13 | a = FRACO
27   3    0 | b = BOM
32   0    3 | c = MUITO BOM

```

Figura 14 – Matriz de confusão do algoritmo *Random Forest* aplicado aos dados do setor de Bens Industriais.

Tabela 9 – Resultado dos algoritmos de mineração aplicados aos dados do Setor de Construção e Transporte.

Algoritmo	CCI	Fraco	Bom	Muito Bom	Soma Bons
<i>RandomTree</i>	96,43	0,982	0,036	0,122	0,158
<i>RandomCommitee</i>	96,56	0,983	0	0,112	0,112
<i>Ibk</i>	96,5	0,983	0	0,089	0,089
<i>RandomForest</i>	96,83	0,984	0	0,076	0,076
<i>REPTree</i>	97,76	0,989	0	0,036	0,036
<i>Bagging Meta</i>	97,71	0,988	0	0	0
<i>ClassificationRegression</i>	97,63	0,988	0	0	0

Tabela 10 – Resultado dos algoritmos de mineração aplicados aos dados do Setor de Consumo Cíclico.

Algoritmo	CCI	Fraco	Bom	Muito Bom	Soma Bons
<i>RandomTree</i>	95,7	0,978	0,043	0,108	0,151
<i>RandomCommitee</i>	95,84	0,979	0,042	0,099	0,141
<i>Ibk</i>	95,68	0,978	0,043	0,075	0,118
<i>RandomForest</i>	95,35	0,982	0,025	0,069	0,094
<i>REPTree</i>	97,24	0,986	0	0,058	0,058
<i>Bagging Meta</i>	97,27	0,986	0	0	0
<i>ClassificationRegression</i>	97,22	0,986	0	0	0

As tabelas com informações de desempenho dos algoritmos por setor mostra que o melhor desempenho foi obtido aplicando o *Random Forest* ao setor de "Bens Industriais".

Tabela 11 – Resultado dos algoritmos de mineração aplicados aos dados do Setor Financeiro e Outros.

Algoritmo	CCI	Fraco	Bom	Muito Bom	Soma Bons
<i>RandomCommittee</i>	97,1	0,985	0	0,229	0,229
<i>Ibk</i>	96,76	0,984	0	0,225	0,225
<i>RandomForest</i>	97,22	0,986	0	0,182	0,182
<i>RandomTree</i>	96,67	0,983	0	0,164	0,164
<i>REPTree</i>	98,02	0,99	0	0,1	0,1
<i>Bagging Meta</i>	98,02	0,99	0	0	0
<i>ClassificationRegression</i>	97,89	0,989	0	0	0

```

      a   b   c  <-- classified as
2298  5  25 |   a = FRACO
    10  0   1 |   b = BOM
    28  0   8 |   c = MUITO BOM
    
```

Figura 15 – Matriz de confusão do algoritmo *Random Committee* aplicado ao setor Financeiro e Outros.

Tabela 12 – Resultado dos algoritmos de mineração aplicados aos dados do Setor de Materiais Básicos.

Algoritmo	CCI	Fraco	Bom	Muito Bom	Soma Bons
<i>RandomTree</i>	94,18	0,971	0,054	0,081	0,135
<i>RandomForest</i>	95,57	0,978	0,069	0,057	0,126
<i>RandomCommittee</i>	94,82	0,974	0,057	0,049	0,106
<i>Ibk</i>	94,14	0,971	0,052	0,043	0,095
<i>REPTree</i>	96,66	0,983	0	0,038	0,038
<i>ClassificationRegression</i>	96,66	0,983	0	0,038	0,038
<i>Bagging Meta</i>	96,6	0,983	0	0	0

O segundo melhor desempenho foi obtido aplicando *Random Committee* aos dados do setor "Financeiro e Outros". O terceiro melhor desempenho foi obtido aplicando o algoritmo *IBK* aos dados do setor "Financeiro e Outros". Por este motivo a matriz de confusão destes testes são mostradas nas figuras 14 e 15.

Não foram realizados testes para o setor de Tecnologia da Informação, pois das 493 instâncias, apenas um é "Muito Bom" as outras 492 instâncias são empresas classificadas como investimento "Fraco". Os testes realizados com os setores "Consumo Não Cíclico" e "Petróleo Gás e Combustíveis" nenhum dos algoritmos conseguiram reconhecer instâncias diferentes de "Fraco". Por este motivo não foram mostradas as tabelas de dados destes setores.

Com os resultados, por setor, organizados por maior valor da coluna "Soma Bons" foi calculado a distribuição de frequências dos algoritmos para saber qual deles se manteve

Tabela 13 – Resultado dos algoritmos de mineração aplicados aos dados do Setor de Telecomunicações.

Algoritmo	CCI	Fraco	Bom	Muito Bom	Soma Bons
<i>RandomTree</i>	95,77	0,978	0,182	0	0,182
<i>RandomForest</i>	95,76	0,978	0,182	0	0,182
<i>RandomCommittee</i>	95,76	0,978	0,167	0	0,167
<i>Bagging Meta</i>	96,42	0,982	0	0	0
<i>REPTree</i>	96,42	0,982	0	0	0
<i>ClassificationRegression</i>	96,42	0,982	0	0	0
<i>Ibk</i>	94,79	0,973	0	0	0

Tabela 14 – Resultado dos algoritmos de mineração aplicados aos dados do Setor de Utilidade Pública.

Algoritmo	CCI	Fraco	Bom	Muito Bom	Soma Bons
<i>RandomTree</i>	96,56	0,983	0,081	0,087	0,168
<i>Ibk</i>	96,62	0,983	0,061	0,082	0,143
<i>RandomForest</i>	96,94	0,985	0,034	0,047	0,081
<i>RandomCommittee</i>	96,78	0,984	0,032	0,047	0,079
<i>Bagging Meta</i>	97,81	0,989	0	0	0
<i>REPTree</i>	97,81	0,989	0	0	0
<i>ClassificationRegression</i>	97,75	0,989	0	0	0

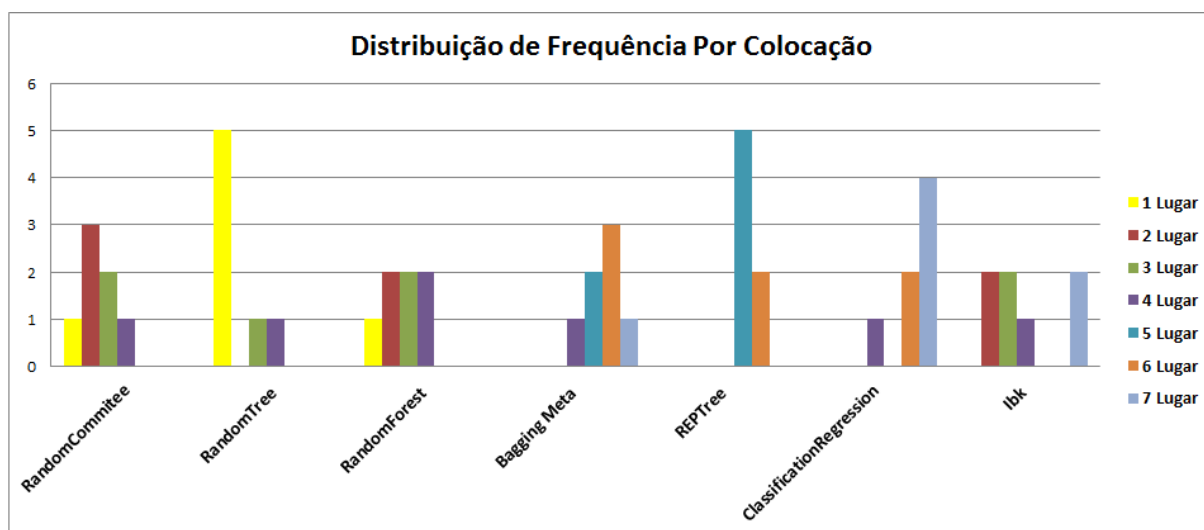


Figura 16 – Distribuição de frequência dos algoritmos por colocação.

mais vezes como melhor opção entre os setores. A figura 16 mostra estes resultados. Pode-se observar que o melhor desempenho (maior valor de "Soma Bons") foi obtido com o algoritmo *Random Tree*, seguido pelo *Random Committee*, *Random Forest* e *IBK* respectivamente, então estes são os quatro algoritmos que mostraram melhor desempenho com os dados por setor.

3.4.1 Técnica de PCA

Pode-se fazer mais algumas divisões dos dados para análise, uma delas se baseia no fato de que os indicadores usados são divididos em funções, sendo: liquidez, rentabilidade e estrutura de capital. Conforme cita o trabalho de (Ramires, 2011) os índices de cada grupo são:

- Liquidez: Liquidez corrente.
- Rentabilidade: PayOut, Divided Yeld, ROE, ROIC, Valor de Mercado.
- Estrutura de Capital: Debt/Equity, EBIT, EBITDA, Noplat, FCO (Fluxo de Caixa Operacional) e FCL (Fluxo de Caixa Livre).

- Múltiplos de mercado: P/L (Preço por Lucro), VPA (Valor Patrimonial da Ação), P/VP (Preço por Valor Patrimonial da Ação), Ev/Ebit, LPA (Lucro por ação).

Para entender melhor os dados foi aplicada a técnica de *PCA* (*Principal Component Analysis*) na base de dados. O *PCA* é uma das abordagens mais comuns para redução da dimensionalidade. Conforme cita (Tan et al, 2009) usa técnica de álgebra linear para projetar os dados de um espaço de alta dimensionalidade para um de dimensionalidade menor. Os dois primeiros componentes capturam tanta variação nos dados quanto possível. Nos permite identificar quais são os atributos mais importantes para explicar a classe a qual cada instância pertence, sendo assim, ele possibilita reduzir a dimensionalidade do problema.

A técnica de *PCA* foi feita no *WEKA* e a figura 17 mostra os resultados da análise. Pode-se observar que o grupo Ebitda, Ebit, Noplat, FCO e FCL explicam 80,89% da variação da classificação. Note que estes indicadores pertencem ao grupo de variáveis que compõem a Estrutura de Capital das empresas. O grupo Ev/Ebit, PL, PVP, valor de mercado e Liquidez Corrente juntos explicam 70% da classificação. E o terceiro grupo mais representativo é VPA, LPA, CFS, Liquidez corrente e Valor de Mercado juntos explicam 63% das variações das classes.

```
Ranked attributes:
0.8089  1  0.522Ebit+0.522Ebitda+0.522Noplat+0.302Fco-0.275Fcl...
0.702   2  0.633EvEbit-0.561Pl-0.532pVp+0.027ValorMercado-0.015LiqCorrente...
0.6285  3 -0.699Vpa-0.523Lpa-0.42Cfs-0.189LiqCorrente+0.087ValorMercado...
0.5597  4 -0.717Fcl-0.675Fco-0.116Cfs-0.098DebitEquit-0.045Roic...
0.4984  5 -0.677ValorMercado-0.637IdSetor+0.198Roic+0.196Cfs-0.14LiqCorrente...
0.4433  6 -0.692DebitEquit-0.431Roe-0.369Cfs+0.318LiqCorrente+0.206Lpa...
0.3895  7 -0.566Roe-0.563LiqCorrente+0.365Cfs+0.234PayOut+0.227Roic...
0.336   8 -0.867DivYeld+0.297Lpa+0.244Roic+0.237Roe-0.128Cfs...
0.2835  9  0.953PayOut+0.164DivYeld+0.147Roe-0.107Cfs+0.101LiqCorrente...
0.2312 10 -0.894Roic-0.321DivYeld-0.198Roe-0.169ValorMercado+0.149PayOut...
0.1802 11 -0.672LiqCorrente+0.425Roe+0.371Lpa-0.325DebitEquit-0.225Cfs...
0.1309 12 -0.592DebitEquit-0.444Lpa+0.433Roe+0.429Cfs+0.161IdSetor...
0.0883 13 -0.698IdSetor+0.642ValorMercado+0.214Cfs-0.113Roic+0.106LiqCorrente...
0.0535 14  0.753pVp-0.651Pl+0.058EvEbit+0.033Vpa-0.031ValorMercado...
0.0214 15 -0.705Vpa+0.494Lpa+0.425Cfs+0.182LiqCorrente+0.173DivYeld...
```

Figura 17 – PCA aplicado a base de dados.

Isto significa que usar o conjunto de indicadores Ebit, Ebitda, Noplat, FCO, FCL, Ev/Ebit, PL, PVP, Valor de Mercado e Liquidez Corrente já traz informações necessárias para classificar as empresas.

No item 3.2.3 foi feito um estudo estatístico em cima da base de dados a fim de entendê-la melhor e também de conseguir preparar os dados de forma satisfatória para aplicar a mineração. As principais empresas são aquelas citadas na análise estatística da Tabela 3.

Então foram usados os dados destas empresas que possuem melhor desempenho, descartando aquelas empresas que estão muito abaixo do desempenho esperado pelos investidores em ações. Fazendo estes filtros foram usados 10289 instâncias de dados com a seguinte distribuição:

- 9781 instâncias classificadas como "FRACO", ou seja, 95,06%;
- 233 instâncias classificadas como "BOM", ou seja, 2,38%;
- 275 instâncias classificadas como "MUITO BOM", ou seja, 2,67%;

Os resultados obtidos com a execução dos algoritmos estão na tabela 15. Comparando esta tabela com a tabela 7 é possível observar que teve melhora no índice de reconhecimento de empresas da classe "Bom" e "Muito Bom", mas teve queda no indicador de reconhecimento de empresas da classe "Fraco" e redução do *CCI*. Isto acontece devido à redução das instâncias da classe "Fraco". A Figura 18 mostra a matriz de confusão do algoritmo *Random Tree*. Nos resultados pode-se observar pouca classificação correta para as empresas da classe "Bom" e "Muito Bom".

Tabela 15 – Resultado dos algoritmos de mineração aplicados aos dados das empresas que tiveram pelo menos um mês com rendimento acima de 80% da Selic.

Algoritmo	CCI	Fraco	Bom	Muito Bom	Soma Bons
<i>RandomTree</i>	91,93	0,959	0,055	0,135	0,19
<i>RandomCommitee</i>	92,526	0,962	0,056	0,121	0,177
<i>Ibk</i>	91,77	0,958	0,051	0,121	0,172
<i>RandomForest</i>	93,25	0,965	0,046	0,102	0,148
<i>REPTree</i>	95,02	0,974	0,009	0	0,009
<i>ClassificationRegression</i>	95,01	0,974	0	0,007	0,007
<i>Bagging Meta</i>	95,06	0,975	0	0	0

```

      a    b    c  <-- classified as
9413 182 186 |   a = FRACO
 212  12   9 |   b = BOM
 230  11  34 |   c = MUITO BOM

```

Figura 18 – Matriz de confusão do algoritmo *Random Tree*.

Foram feitos experimentos usando os atributos de maior representatividade mostrado pelo *PCA*. As colunas selecionadas são mostradas na figura 19. Os resultados obtidos com a aplicação dos algoritmos é visto na tabela 16.

Comparando os resultados obtidos na tabela 16 com os resultados da tabela 15 é possível observar que os resultados apresentaram pouca diferença, não sendo considerado como perda de qualidade dos resultados. Analisando os valores da primeira linha de cada tabela pode-se perceber que a variação é mínima pequena, demonstrando que realmente não é preciso usar todos os atributos para aplicar os algoritmos de classificação.

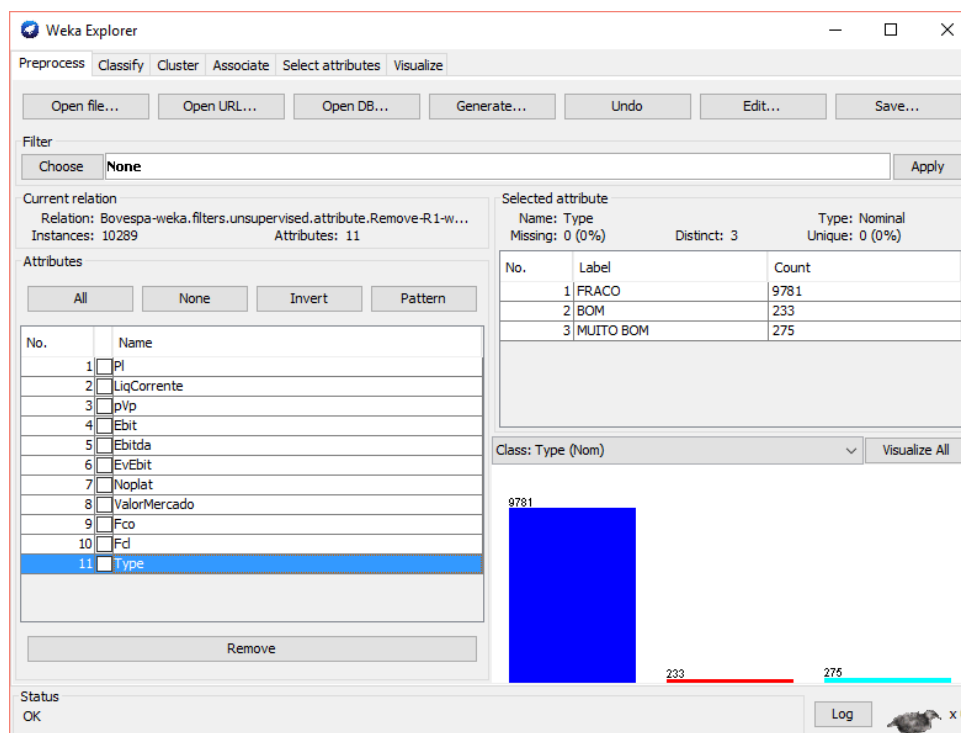


Figura 19 – WEKA seleção dos indicadores identificados no PCA.

Tabela 16 – Resultado dos algoritmos de mineração aplicados aos dados das empresas que tiveram pelo menos um mês com rendimento acima de 80% da Selic e somente usando os atributos principais.

Algoritmo	CCI	Fraco	Bom	Muito Bom	Soma Bons
<i>RandomTree</i>	91,87	0,958	0,065	0,117	0,182
<i>RandomCommittee</i>	92,72	0,963	0,058	0,114	0,172
<i>Ibk</i>	91,93	0,959	0,051	0,12	0,171
<i>RandomForest</i>	93,81	0,968	0,02	0,073	0,093
<i>Bagging Meta</i>	95,05	0,975	0	0	0
<i>REPTree</i>	95,04	0,975	0	0	0
<i>ClassificationRegression</i>	95,01	0,974	0	0	0

3.4.2 Experimentos

Com os resultados anteriores identificou-se os algoritmos que obtiveram melhor desempenho aplicado aos dados, conforme mostra a figura 16, são eles *Random Tree*, *Random Committee*, *Random Forest* e *IBK*. Foram realizados experimentos mais detalhados com os algoritmos *Random Tree*, *Random Committee* e *IBK* nos três períodos existentes na base de dados, sendo eles semanal, mensal e semestral. A escolha do *IBK* ao invés do *Random Forest* é devido ao fato dele ter um método de aprendizado diferente das árvores de decisão, que é o *KNN* explicado na seção 2.3.1.

Foram usados nestes experimentos apenas os atributos mais representativos indicados na seção 3.4.1 que são: Ebit, Ebitda, Noplat, FCO, FCL, Ev/Ebit, PL, PVP, Valor de Mercado e Liquidez Corrente.

A máquina usada para realização dos testes foi Computador Dell Inspiron 14 3000 Series, Processador Intel Core i5-4210U CPU de 1.70GHz, RAM de 8GB, Sistema Ope-

racional Windows 10 de 64 bits e processador com base em x64.

Os gráficos apresentados nesta seção seguem a mesma lógica presente nas tabelas dos experimentos das seções anteriores, quanto mais próximo de 1 as tendências estiverem melhor é o desempenho do algoritmo.

Para o algoritmo *IBK*

- Variáveis de resposta: serão os itens de desempenho medidos no sistema que são *CCI* e *F-Measure*.
- Fatores: período usado para gerar os indicadores e o valor de K .
- Níveis: período sendo semanal, mensal ou semestral; valores de K .

O número de experimentos obtido foi:

$$n = 16(\text{valores de } K) * 3(\text{variação do período}) = 48 \text{ experimentos}$$

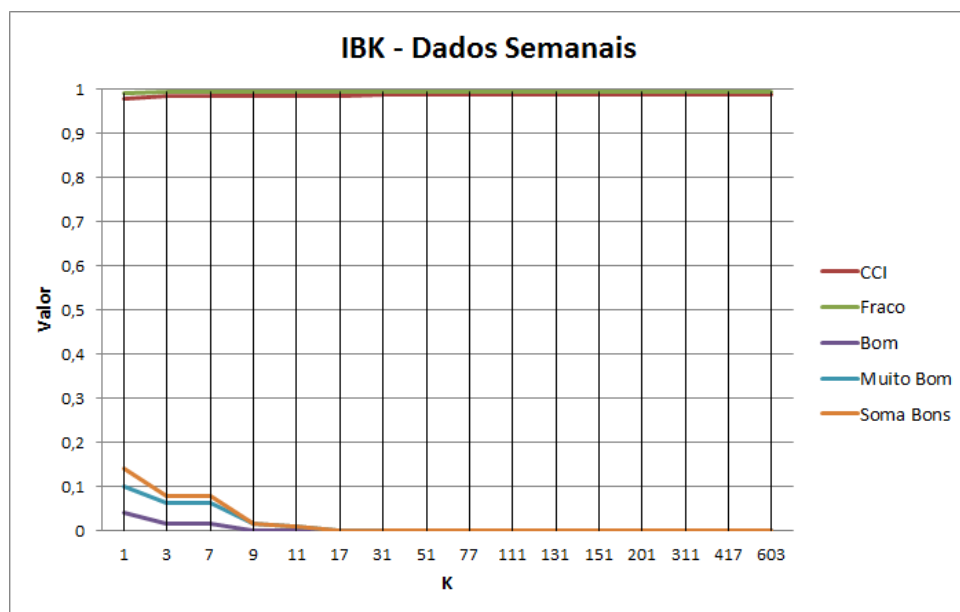


Figura 20 – Resultados do algoritmo *IBK* aplicado aos dados semanais.

No gráfico da figura 20 pode-se observar que variando o valor de K ainda aparecem poucas instâncias da classe "Bom" e "Muito Bom" classificadas corretamente. O valor de K que apresentou melhor desempenho foi $K = 1$. Isto é um indicador de que há muita variabilidade nos dados dentro da mesma classe.

No gráfico da figura 21 pode-se observar que variando o valor de K ainda aparecem poucas instâncias da classe "Bom" e "Muito Bom" classificadas corretamente. O valor de K que apresentou melhor desempenho também foi $K = 1$.

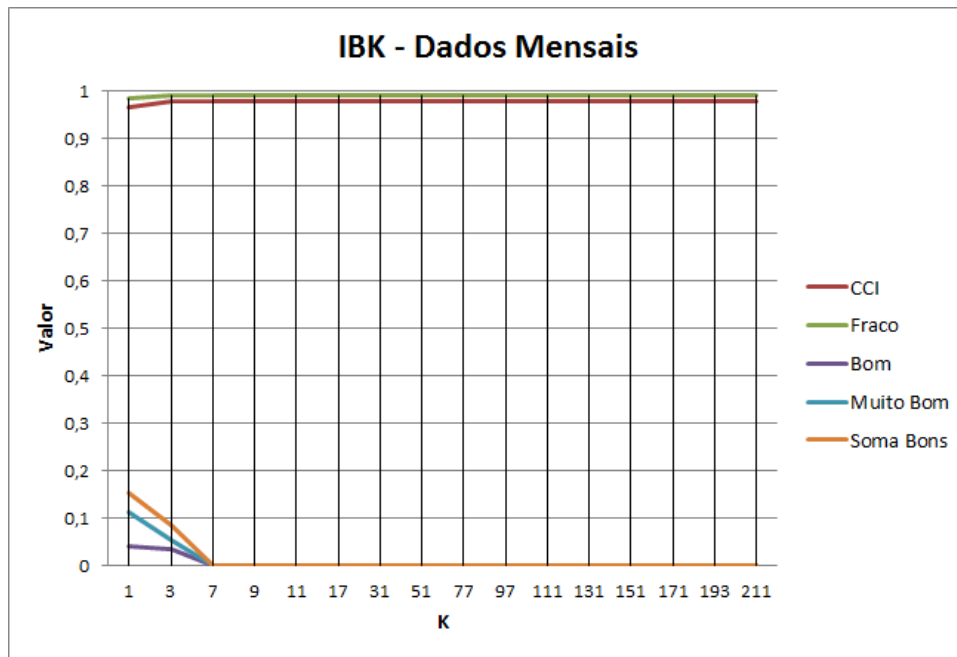


Figura 21 – Resultados do algoritmo *IBK* aplicado aos dados mensais.

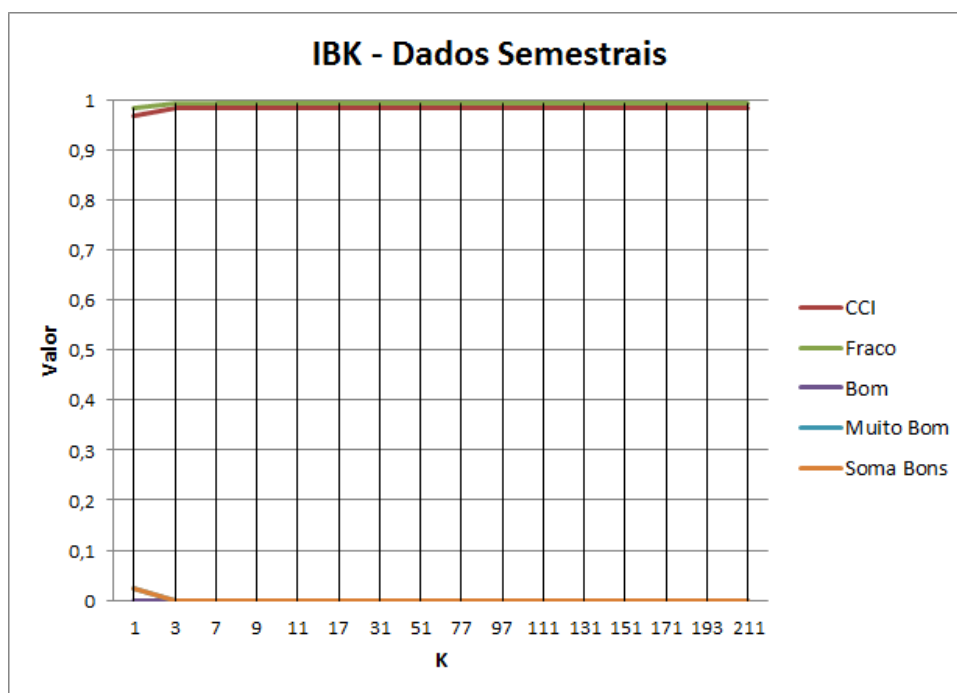


Figura 22 – Resultados do algoritmo *IBK* aplicado aos dados Semestrais.

No gráfico da figura 22 pode-se observar que variando o valor de K ainda aparecem poucas instâncias da classe "Bom" e "Muito Bom" classificadas corretamente. O valor de K que apresentou melhor desempenho também foi $K = 1$.

Para o algoritmo *Random Committee*:

- Variáveis de resposta: serão os itens de desempenho medidos no sistema que são

CCI e *F-Measure*.

- Fatores: período usado para gerar os indicadores e a quantidade de *seed*.
- Níveis: período sendo semanal, mensal ou semestral e quantidade de *seed*.

O número de experimentos obtido foi:

$$n = 16(\text{valores de seed}) * 3(\text{variação do período}) = 48 \text{ experimentos}$$

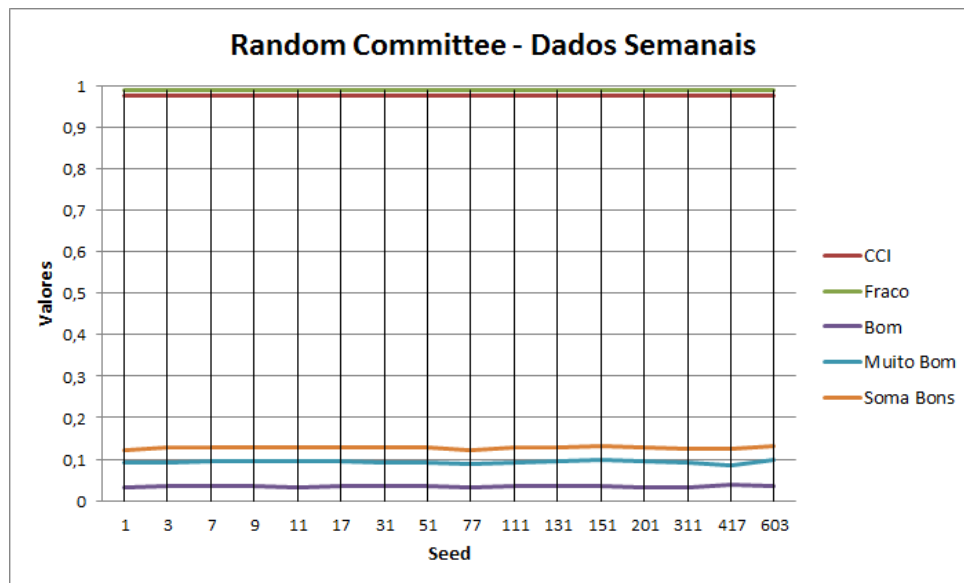


Figura 23 – Resultados do algoritmo *Random Committee* aplicado aos dados semanais.

No gráfico da figura 23 pode-se observar que variando a quantidade de *seed* observa-se pouca variação nos resultados. Poucas instâncias da classe "Bom" e "Muito Bom" são classificadas corretamente. A quantidade de *seed* que apresentou melhor desempenho 151.

No gráfico da figura 24 pode-se observar que variando a quantidade de *seed* observa-se mais variação nos resultados do que os obtidos com dados semanais. Poucas instâncias da classe "Bom" e "Muito Bom" são classificadas corretamente. A quantidade de *seed* que apresentou melhor desempenho 9.

No gráfico da figura 24 pode-se observar que variando a quantidade de *seed* observa-se mais variação nos resultados do que os obtidos com dados semanais. Poucas instâncias da classe "Bom" e "Muito Bom" são classificadas corretamente. A quantidade de *seed* que apresentou melhor desempenho 151.

Para o algoritmo *Random Tree*:

- Variáveis de resposta: serão os itens de desempenho medidos no sistema que são *CCI* e *F-Measure*.
- Fatores: período usado para gerar os indicadores e a quantidade de *seed*.

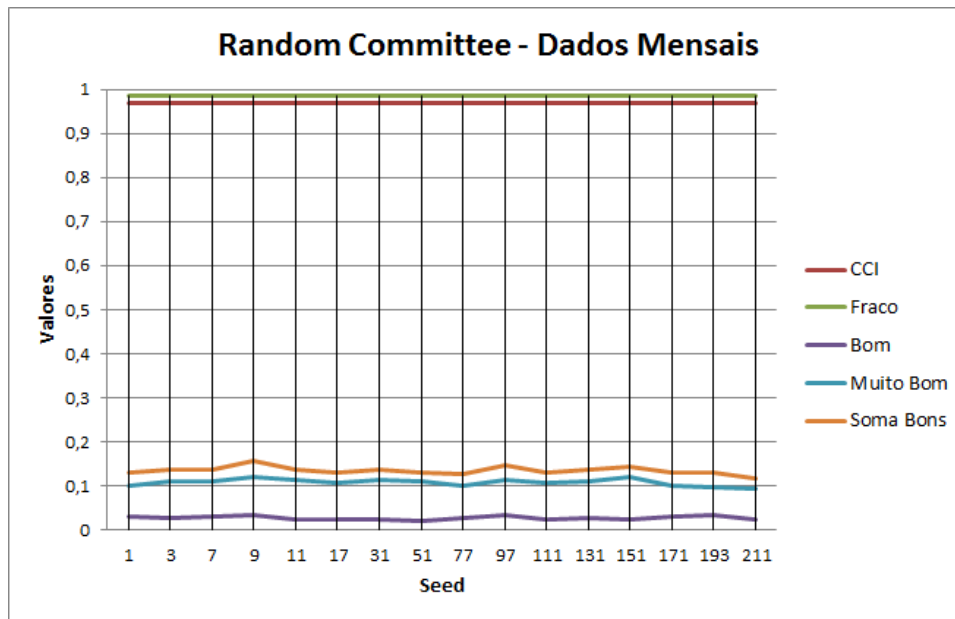


Figura 24 – Resultados do algoritmo *Random Committee* aplicado aos dados mensais.

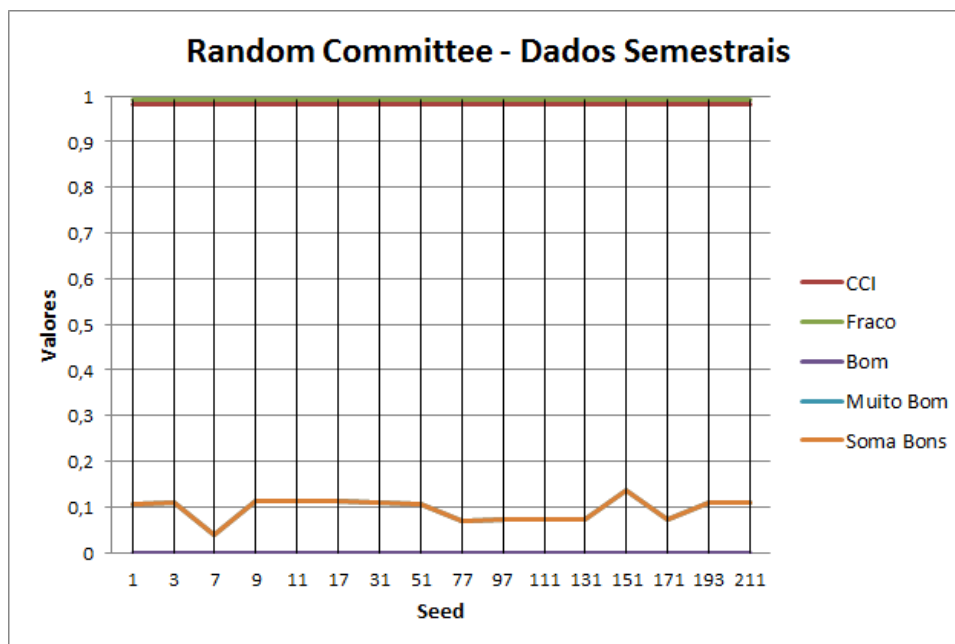


Figura 25 – Resultados do algoritmo *Random Committee* aplicado aos dados Semestrais.

- Níveis: período sendo semanal, mensal ou semestral e quantidade de *seed*.

O número de experimentos obtido foi:

$$n = 16(\text{valores de seed}) \times 3(\text{variação do período}) = 48 \text{ experimentos}$$

No gráfico da figura 26 pode-se observar que variando a quantidade de *seed* observa-se pouca variação nos resultados. Poucas instâncias da classe "Bom" e "Muito Bom" são classificadas corretamente. A quantidade de *seed* que apresentou melhor desempenho 9.

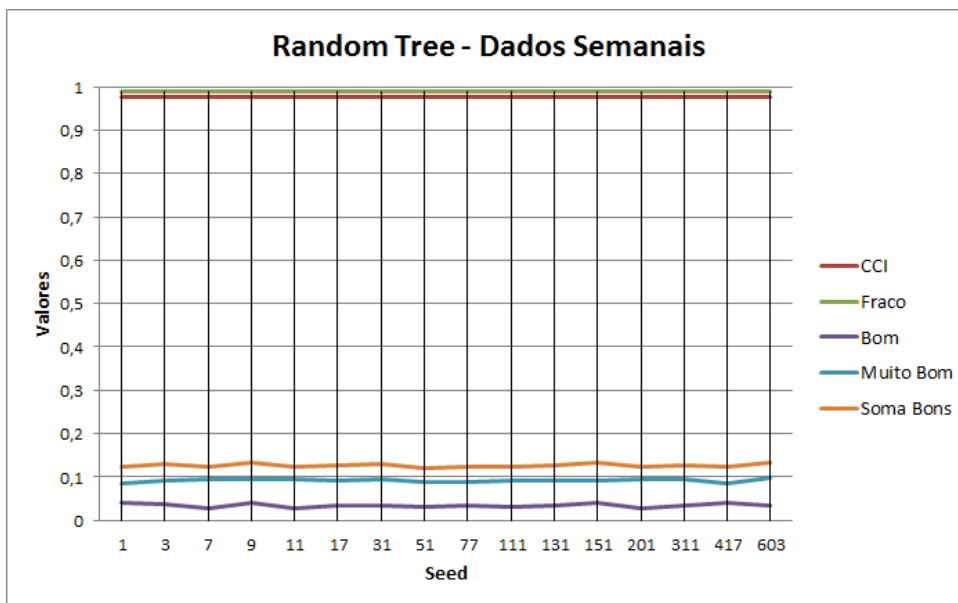


Figura 26 – Resultados do algoritmo *Random Tree* aplicado aos dados semanais.

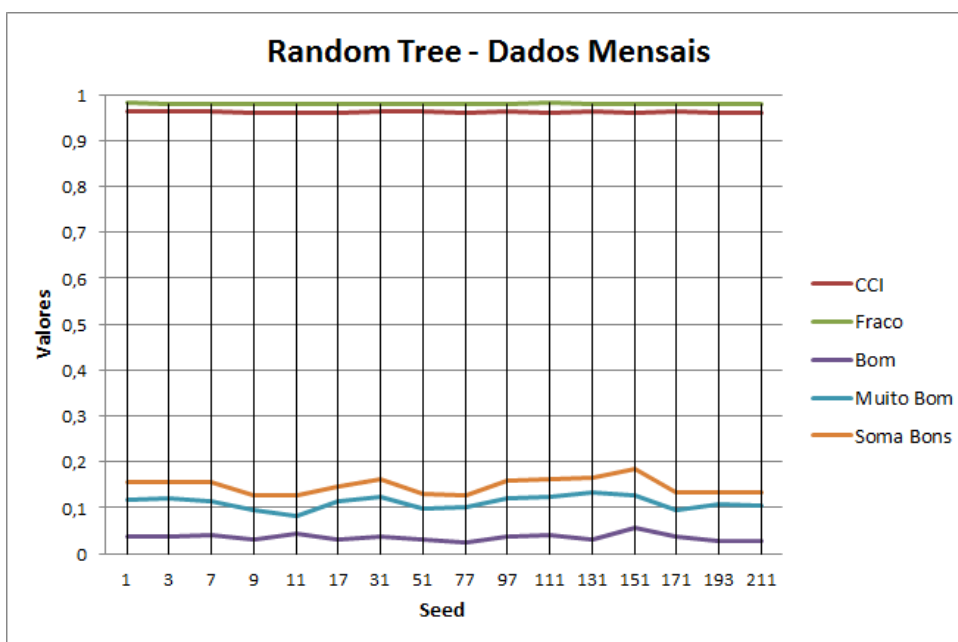


Figura 27 – Resultados do algoritmo *Random Tree* aplicado aos dados mensais.

No gráfico da figura 27 pode-se observar que variando a quantidade de *seed* observa-se pouca variação nos resultados. Poucas instâncias da classe "Bom" e "Muito Bom" são classificadas corretamente. A quantidade de *seed* que apresentou melhor desempenho 151.

No gráfico da figura 28 pode-se observar que variando a quantidade de *seed* observa-se mais variação nos resultados. Poucas instâncias da classe "Bom" e "Muito Bom" são classificadas corretamente. A quantidade de *seed* que apresentou melhor desempenho 131 e 211.

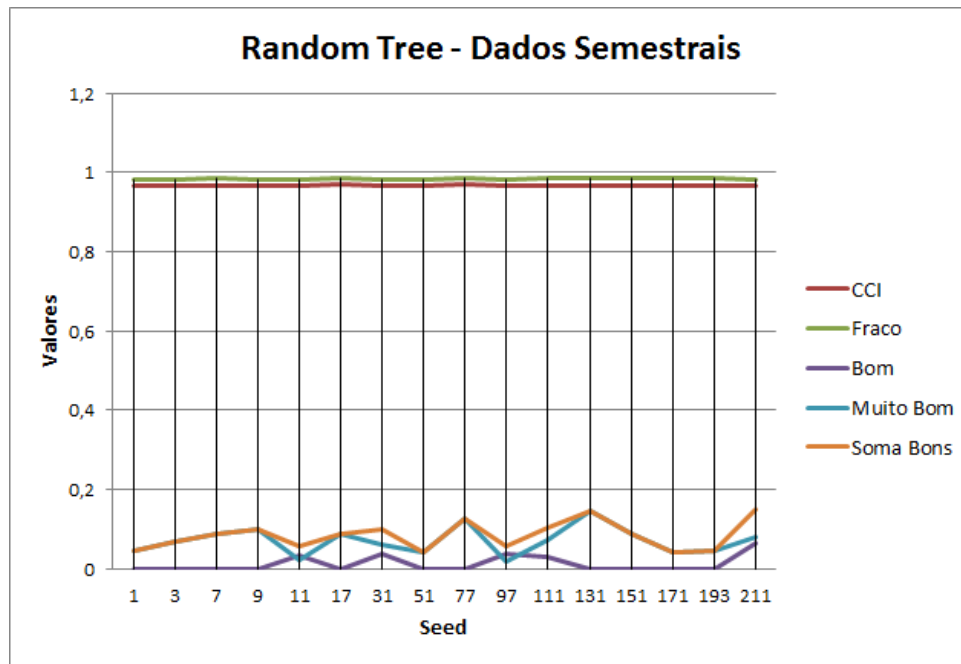


Figura 28 – Resultados do algoritmo *Random Tree* aplicado aos dados Semestrais.

3.4.3 Estatística Kappa

A medida chamada estatística Kappa leva o valor esperado em conta, deduzindo-o com os sucessos da previsão e expressa o resultado como uma proporção do total para um preditor perfeito. O valor máximo de Kappa é de 100%, e o valor esperado para um preditor aleatório nas mesmas condições é 0. Em resumo, a estatística Kappa é utilizada para medir o acordo entre classificações previstas e observadas de um conjunto de dados e pode ser usado para medir a concordância entre classificadores. (Ian H. Witten, 2011).

No trabalho de Landis and G. (1977) ele mostra a fórmula de K que pode ser vista na 3.2. O autor cita que K é diretamente análogo com o Coeficiente de Correlação Intraclasse obtido com modelos ANOVA para medidas quantitativas e pode ser usado como uma medida de confiabilidade de múltiplas determinações sobre mesmas questões.

$$k = \frac{\pi_0 - \pi_e}{1 - \pi_e} \quad (3.2)$$

Onde:

- π_0 é uma probabilidade observada de concordância.
- π_e é uma probabilidade esperada hipotética de concordância.

Landis and G. (1977) propôs uma análise interessante, a fim de manter a nomenclatura consistente ao descrever a força relativa de acordo associada à estatística Kappa, as seguintes descrições foram atribuídas aos intervalos correspondentes de Kappa:

Tabela 17 – Interpretação dos valores de Kappa. Fonte: Landis and G. (1977)

Values of Kappa	Strength of Agreement
<0	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1	Almost Perfect

Os gráficos das Figuras 29, 30 e 22 mostram os valores da estatística Kappa para os algoritmos *Random Committee*, *Random Tree* e *IBK* aplicados aos dados semanais, mensais e semestrais mostrados no item 3.4.2.

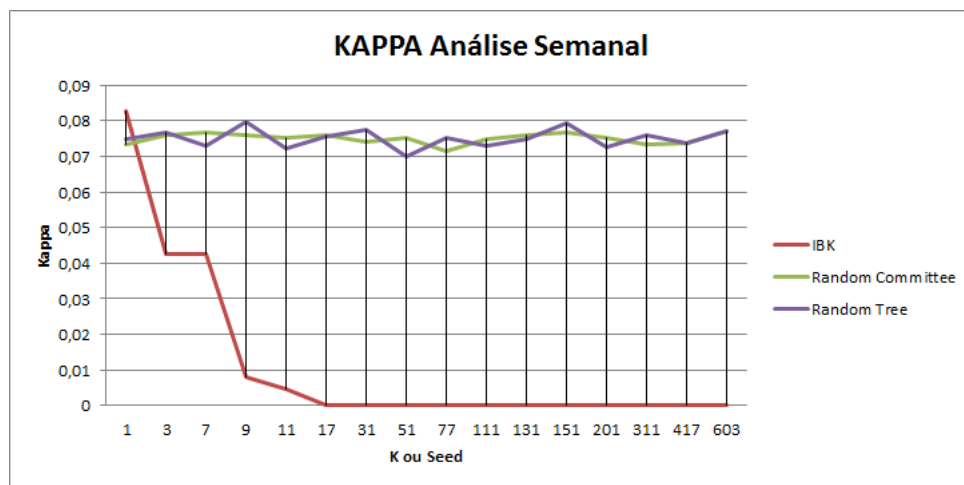


Figura 29 – Valores da estatística Kappa dos algoritmos aplicados aos dados semanais.

Nos dados semanais os três algoritmos tiveram desempenho semelhante, todos na faixa considerada pela tabela 17 como leve concordância (*Slight*). As considerações feitas no item 3.4.2 foram confirmadas sendo que o *IBK* apresentou melhor desempenho com $k = 1$, *Random Committee* apresentou melhor desempenho com $seed = 151$, *Random Tree* apresentou melhor desempenho com $seed = 9$ e $seed = 151$.

Nos dados mensais o *Random Tree* com $seed = 151$ apresentou melhor desempenho, mas mesmo assim ainda continua na faixa considerada pela tabela 17 como leve concordância (*Slight*).

Nos dados semestrais o *Random Committee* com $seed = 151$ apresentou melhor desempenho, mas mesmo assim ainda continua na faixa considerada pela tabela 17 como leve concordância (*Slight*). É possível observar que a estatística Kappa também confirma os resultados obtidos na seção 3.4.2.

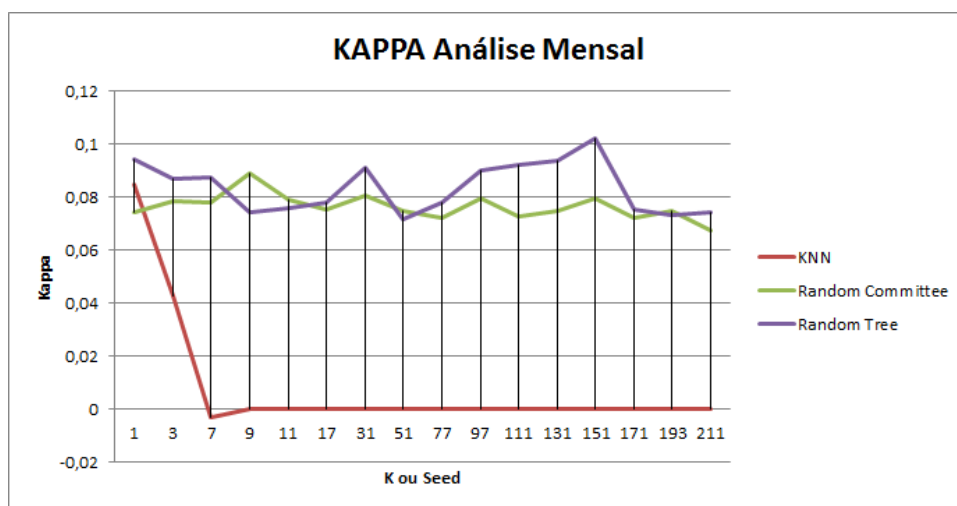


Figura 30 – Valores da estatística Kappa dos algoritmos aplicados aos dados mensais.

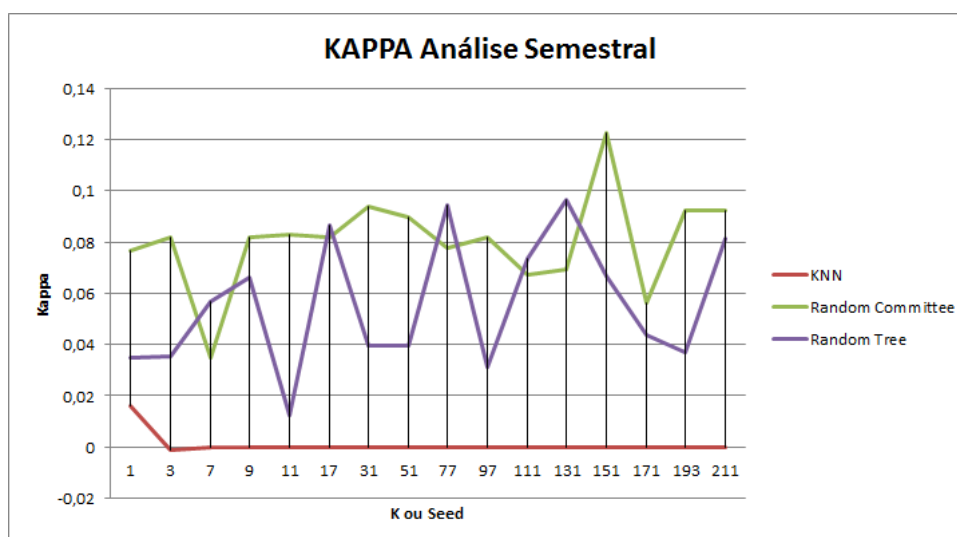


Figura 31 – Valores da estatística Kappa dos algoritmos aplicados aos dados semestrais.

4 CONSIDERAÇÕES FINAIS

O trabalho apresentou aplicação de algoritmos de mineração de dados em informações da bolsa de valores BM&FBovespa. Iniciou com uma pesquisa à bibliografia da área de mercados financeiros a fim de identificar os indicadores usados para avaliação de empresas, seguindo a análise fundamentalista. Em seguida todas as informações para cálculo dos indicadores foram buscadas no site da BM&FBovespa e outros sites destinados à disponibilização de informações das empresas de capital aberto.

Durante o levantamento dos dados foi observado a dificuldade enfrentada por investidores para obter o histórico de informações das empresas. Os dados foram obtidos em vários formatos e de fontes diferentes, assim a próxima etapa foi a organização e limpeza destas informações, consolidando-as em uma base de dados única para que pudessem ser exportadas já normalizadas e no formato desejado pelo usuário. As classes foram definidas comparando o rendimento das empresas com a taxa Selic, onde as empresas foram classificadas como "Fraco", "Bom" ou "Muito Bom".

A análise estatística dos dados mostrou que as empresas que tiveram maior rendimento nos últimos dez anos foram Tectoy, JB Duarte, Elekeiroz e Mundial e estão na Tabela 6. Empresas que apresentaram maior frequência de meses com rendimento acima de 120% da Selic foram Telebras, JB Duarte, Metal Iguaçu, Tectoy e Unipar e são mostradas na Tabela 4.

Foram gerados os indicadores com três períodos distintos, sendo semanal, mensal e semestral. Os dados foram exportados no formato *arff* que é o padrão de entrada de dados da ferramenta *WEKA* onde existem diversos algoritmos de mineração de dados. Foram analisados o desempenho de dez algoritmos de mineração de dados, destinados à tarefa de classificação, aplicados nestes indicadores.

No primeiro experimento realizado foram usadas todas as instâncias de dados gerados com o período mensal e aplicou-se os algoritmos *Random Committee*, *Random Tree*, *Ibk*, *Random Forest*, *Bagging Meta*, *Classification Via Regression*, *Logit Boost*, *REPTree*, *J48* e *MultiClass Classifier*. O algoritmo que apresentou melhor desempenho foi o *Random Tree* entretanto, o percentual de reconhecimento das classes "Bom" e "Muito Bom" foi baixo.

No segundo experimento realizado os dados mensais foram divididos por setor, foram usados os algoritmos *Random Committee*, *Random Tree*, *Ibk*, *Random Forest*, *Bagging Meta*, *Classification Via Regression* e *REPTree* sobre cada setor. Os setores "Tecnologia da Informação", "Consumo não Cíclico" e "Petróleo Gás e Combustíveis" não apresentaram reconhecimento de instâncias diferente de "Fraco". O experimento que apresentou me-

lhor desempenho foi o algoritmo *Random Forest* aplicado ao setor de "Bens Industriais", mesmo assim a taxa de reconhecimento das classes diferentes de "Fraco" foi baixo.

Foi aplicada a técnica de *PCA* sobre os dados a fim de obter os componentes principais. Com estes componentes foram realizados testes com os três algoritmos que tiveram melhor desempenho nos testes anteriores sendo eles *Random Committee*, *Random Tree* e *IBK*. Os experimentos foram realizados com estes algoritmos sobre os dados semanais, mensais e semestrais. Conforme mostra os gráficos do item 3.4.2 e os gráficos do item 3.4.3. Usando todos os dados a taxa de reconhecimento (*F-Measure*) apresentou valor alto para reconhecimento de empresas da classe "Fraco", mas baixa taxa de reconhecimento de empresas da classe "Bom" e "Muito Bom". Estes resultados foram confirmados pela estatística Kappa que também apresentou taxa de acerto inferior a 0,2 na maioria dos experimentos demonstrando baixa previsão das classes na abordagem proposta por este trabalho.

4.1 TRABALHOS FUTUROS

Algumas informações não foram possíveis de ser obtidas devido ao tempo para conclusão do trabalho pode-se adicionar a esta base de dados informações de WACC (Custo Médio Ponderado de Capital) e EVA (*Enterprise Value*). Continuando este trabalho, pode-se explorar diversas possibilidades como aplicar Redes Neurais sobre os dados de cotação das ações, fazendo assim um estudo enfatizando a análise técnica. Pode-se também incluir novos indicadores a fim de melhorar a análise e também usar análise técnica e análise fundamentalista com algoritmos de Redes Neurais e Mineração de dados a fim de obter melhor desempenho.

REFERÊNCIAS

- BM&FBovespa (2016) Sobre a bm&fbovespa URL http://www.bmfbovespa.com.br/pt_br/a-bm-fbovespa/institucional/quem-somos/, Último acesso em 2 de Agosto, 2016 30
- Camilo CO, Silva JC (2009) Mineração de dados: Conceitos, tarefas, métodos e ferramentas. Universidade Federal de Goiás 20, 29
- Cardoso RL, Martins VA (2004) Teoria avançada da contabilidade. Org. IUDÍCIBUS, S.; LOPES, A.B. Atlas 28
- Carneiro RB (2011) O fluxo de caixa como instrumento de gerenciamento financeiro na empresas URL <http://www.unicampsciencia.com.br/pdf/50bff4f521455.pdf>, Último acesso em 2 de Agosto, 2016 48
- Cavalcante F, Misumi J, Rudge L (2009) Mercado de Capitais. Campus 30, 37, 38, 39, 48, 49
- Corretora B (2015) Apostila guia de análise fundamentalista URL https://www.bradesco.corretora.com.br/static_files/Corretora/PDF/Apostila_Guia%20de%20An%C3%A1lise%20Fundamentalista.pdf, Último acesso em 2 de Setembro, 2015 37, 38, 40, 41, 47
- Fama E (1970) Efficient capital markets: A review of theory and empirical work. Journal of Finance, v48, mar 1970, p 383-417 28
- Fayyad U, Shapiro GP, Padhraic S (1996) Knowledge discovery and data mining: Towards a unifying framework. Second International Conference on Knowledge Discovery and Data Mining KDD-96 12, 19, 20
- Frank E, Wang Y, Inglis S, Holmes G, Witten IH (1998) Using model trees for classification. Machine Learning 32(1):63–76, DOI 10.1023/A:1007421302149, URL <http://dx.doi.org/10.1023/A:1007421302149> 27
- Friedman J, Hastie T, Tibshirani R, et al (2000) Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics 28(2):337–407 26
- Halfield M (2007) Investimentos: como administrar melhor o seu dinheiro. Fundamento Educacional ltda 37, 38

- Ian H Witten MAH Eibe Frank (2011) Data Mining: Practical Machine Learning Tools and Techniques. Elsevier 22, 23, 61, 73
- Imandoust S, Bolandraftar M (2014) 5 - forecasting the direction of stock market index movement using three data mining techniques: the case of tehran stock exchange. S Bafandeh Imandoust Int Journal of Engineering Research and Applications 12, 33, 34
- InfoMoney E (2005) Investimento em ações: saiba como montar e gerir sua carteira de longo prazo URL <http://www.infomoney.com.br/educacao/guias/noticia/387929/investimento-acoes-saiba-como-montar-gerir-sua-carteira-longo-prazo>, Último acesso em 2 de Agosto, 2016 42
- InfoMoney E (2007) Entenda o que é e como a selic afeta a economia brasileira e seu bolso URL <http://www.infomoney.com.br/educacao/guias/noticia/125180/entenda-que-como-selic-afeta-economia-brasileira-seu-bolso>, Último acesso em 2 de Agosto, 2016 50
- InfoMoney E (2015) Ebitda: entenda o conceito e o cálculo desse importante indicador de desempenho URL <http://www.infomoney.com.br/educacao/guias/noticia/318552/ebitda-entenda-conceito-calculo-desse-importante-indicador-desempenho>, Último acesso em 2 de Setembro, 2015 39
- do Investidor P (2016) Introdução - o mercado de valores mobiliários URL http://www.portaldoinvestidor.gov.br/menu/Menu_Investidor/introducao_geral/introducao_mercado.html, Último acesso em 2 de Agosto, 2016 28
- Jovem I (2015) O indicador pl URL <http://www.investidorjovem.com.br/o-indicador-pl>, Último acesso em 2 de Setembro, 2015 37
- Landis JR, G KG (1977) The measurement of observer agreement for categorical data. Biometrics 33, no 1: 159-74 14, 73, 74
- Martins A (2009) Série bolsa de valores com análise fundamentalista 2: Vpa e p/vp URL <http://iniciantenabolsa.com/serie-analise-fundamentalista-2-vpa-e-pvp/#ixzz3g7Q4gjb4>, Último acesso em 2 de Setembro, 2015 38
- Matias AB (2009) Análise Financeira Fundamentalista de Empresas. Atlas 17

- Pereira D (2015) Dividendos: o guia absolutamente completo URL <http://viverdeinvestimento.com/value-investing/dividendos>, Último acesso em 2 de Agosto, 2016 46
- Pinheiro JL (2009) Mercado de capitais: fundamentos e técnicas. Atlas 17, 29
- Pinho AG (2008) Mineração de dados com mapa de kohonen: Uma abordagem no setor financeiro. Revista Pensamento Contemporâneo em Administração 33
- Ramires KK (2011) A análise de investimentos de b. graham w. buffett e p. fisher aplicada ao mercado de capitais brasileiro 64
- Ramos AL, dos Santos CN (2011) Combinando algoritmos de classificação para detecção de intrusão em redes de computadores 26
- Rycheski LM (2013) A importância das demonstrações contábeis na elaboração de uma carteira de investimentos: é possível aplicar com segurança? Universidade Federal do Rio Grande do Sul 32
- Saeedmanesh M, Izadi T, Ahvar E (2010) Hdm: A hybrid data mining technique for stock exchange prediction. Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol I 31, 32
- Shapiro GP (1991) Knowledge discovery in real databases: A report on the ijcai-89 workshop. AI Magazine Winter 19
- Silva GM, Tessaro N (2013) Análise de correlação entre indicadores financeiros e variações de preços de ações utilizando mineração de dados. Revista Organização Sistêmica 33
- Tan P, Steinbach M, Kumar V (2009) Introdução ao Data Mining - Mineração de Dados. Editora Ciência Moderna Ltda 21, 23, 25, 60, 65
- Tibúrcio C (2012) Valor do empreendimento URL <http://www.contabilidade-financeira.com/2012/03/valor-do-empreendimento-enterprise.html>, Último acesso em 2 de Agosto, 2016 47
- TORORADAR (2016) Roe - retorno sobre o patrimônio líquido URL <http://www.tororadar.com.br/investimento/analise-fundamentalista/roe-retorno-sobre-o-patrimonio-liquido>, Último acesso em 2 de Agosto, 2016 40
- Vieira D (2016) Tipos de ações e seus códigos de negociação URL <http://daltonvieira.com/tipos-de-acoes-e-seus-codigos-de-negociacao>, Último acesso em 2 de Agosto, 2016 31

- of Computer Science at the University of Waikato D (2016) Lesson 4.6 on ensemble learning URL <http://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/transcripts/Transcript4-6.txt>, Último acesso em 2 de Agosto, 2016 26
- Wawrzenciac D (2015) 5 indicadores fundamentalistas mais importantes em uma análise URL <http://blog.bussoladoinvestidor.com.br/indicadores-fundamentalistas-mais-importantes/>, Último acesso em 10 de Setembro, 2015 39
- Wikipédia ael (2016) Algoritmo c4.5 URL https://pt.wikipedia.org/wiki/Algoritmo_C4.5, Último acesso em 2 de Agosto, 2016 25
- Zeidler R (2015) Analise o roic antes de investir URL <http://www.infomoney.com.br/blogs/blog-numeros-falam/post/3316788/analise-roic-antes-investir>, Último acesso em 2 de Setembro, 2015 39