UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

DIONE APARECIDO DE OLIVEIRA SANGA

MINERAÇÃO DE TEXTOS PARA O TRATAMENTO AUTOMÁTICO EM SISTEMAS DE ATENDIMENTO AO USUÁRIO

DISSERTAÇÃO

Curitiba
2017

DIONE APARECIDO DE OLIVEIRA SANGA

MINERAÇÃO DE TEXTOS PARA O TRATAMENTO AUTOMÁTICO EM SISTEMAS DE ATENDIMENTO AO USUÁRIO

Dissertação submetida ao Programa de Pós-Graduação em Computação Aplicada da Universidade Tecnológica Federal do Paraná como requisito parcial para a obtenção do título de Mestre em Computação Aplicada.

Área de concentração: *Sistemas Inteligentes e Lógica.*

Orientador: Prof. Dr. Celso Antônio Alves Kaestner

Curitiba
2017

Dados Internacionais de Catalogação na Publicação

s225m
2017

Sanga, Dione Aparecido de Oliveira
Mineração de textos para o tratamento automático em sistemas de atendimento ao usuário / Dione Aparecido de Oliveira Sanga --2017.
93 f.: il.; 30 cm.

Disponível também via Word Wide Web.
Texto em português, com resumo em inglês.
Dissertação (Mestrado) – Universidade Tecnológica Federal do Paraná. Programa de Pós-graduação em Computação Aplicada. Área de concentração: Sistemas Inteligentes e Lógica, Curitiba, 2017.
Bibliografia: f. 91-93.

1. Mineração de Dados (Computação). 2. Sistemas de recuperação da informação. 3. Telecomunicações – Serviços ao cliente. 4. Algoritmos. 5. Banco de dados da Web. 6. Processamento eletrônico de dados. 7. Comunicação e tecnologia. 8. Inteligência artificial. 9. Computação – Dissertações. I. Kaestner, Celso Antônio Alves, orient. II. Universidade Tecnológica Federal do Paraná. Programa de Pós-graduação em Computação Aplicada. III. Título.

CDD: Ed. 22 – 621.39

Biblioteca Central da UTFPR, Campus Curitiba

ATA DE DEFESA DE DISSERTAÇÃO DE MESTRADO Nº 54

Aos 22 dias do mês de agosto de 2017, realizou-se na sala B106 a sessão pública de Defesa da Dissertação de Mestrado intitulada **"Mineração de textos para o tratamento automático em sistemas de atendimento ao usuário"**, apresentado pelo aluno **Dione Aparecido de Oliveira Sanga** como requisito parcial para a obtenção do título de Mestre em Computação Aplicada, na área de concentração "Sistemas Inteligentes e Lógica", linha de pesquisa "Aprendizagem de máquina e Mineração de dados".

Constituição da Banca Examinadora:

Celso Antônio Alves Kaestner– UTFPR _____

Julio Cesar Nievola – PUCPR _____

Laudelino Cordeiro Bastos - UTFPR _____

Robinson Vida Noronha – UTFPR _____

Em conformidade com os regulamentos do Programa de Pós-Graduação em Computação aplicada e da Universidade Tecnológica Federal do Paraná, o trabalho apresentado foi considerado _____(aprovado/reprovado) pela banca examinadora. No caso de aprovação, a mesma está condicionada ao cumprimento integral das exigências da banca examinadora, registradas no verso desta ata, da entrega da versão final da dissertação em conformidade com as normas da UTFPR e da entrega da documentação necessária à elaboração do diploma, em até___dias desta data.

Ciente (assinatura do aluno): _____

(para uso da coordenação)

A Coordenação do PPGCA/UTFPR declara que foram cumpridos todos os requisitos exigidos pelo programa para a obtenção do título de Mestre.

Curitiba PR,_____/_____/_____ _____

"A Ata de Defesa original está arquivada na Secretaria do PPGCA".

AGRADECIMENTOS

A Deus que permitiu e me deu forças para realizar mais este sonho e etapa da minha vida, pois sem a vontade dele nada disso seria possível.

A minha família que nunca me deixou de acreditar naquilo que sempre sonhei e pelo incentivo proporcionado.

A minha namorada Angélica pela paciência, incentivo e carinho neste período e pelo companheirismo nos momentos mais difíceis, que me impulsionou a seguir em frente e nunca desistir daquilo que sempre acreditei.

Ao meu orientador Prof. Celso Antônio Alves Kaestner, que mesmo após cumprir seu dever junto a instituição honrou com seu compromisso e conduziu-me na realização deste trabalho com sua disponibilidade, colaboração, dedicação e sabedoria.

A instituição pelo ambiente e organização incontestável para o processo de aprendizagem.

A empresa a qual desenvolvi a pesquisa e meus colegas de trabalho que direta ou indiretamente fizeram parte deste trabalho.

Por fim deixo uma pequena mensagem a qual tenho admiração:

“A vida é para nós o que concebemos dela. Para o rústico cujo campo lhe é tudo, esse campo é um império. Para o César cujo império lhe ainda é pouco, esse império é um campo. O pobre possui um império; o grande possui um campo. Na verdade, não possuímos mais que as nossas próprias sensações; nelas, pois, que não no que elas veem, temos que fundamentar a realidade da nossa vida.”

RESUMO

A explosão de novas formas de comunicação entre empresas e clientes proporciona novas oportunidades e meios para que empresas possam tirar proveito desta interação. A forma como os clientes interagem com as empresas tem evoluído nos últimos anos, devido ao aumento dos dispositivos móveis e o acesso à internet: clientes que tradicionalmente solicitavam atendimento via telefone migraram para meios de atendimento eletrônicos, sejam eles via app´s dos smartphones ou via portais de atendimento a clientes. Como resultado desta transformação tecnológica do meio de comunicação, a Mineração de Textos tornou-se uma atrativa forma das empresas extraírem conhecimento novo a partir do registro das interações realizadas pelos clientes. Dentro deste contexto, o ambiente de telecomunicações proporciona os insumos para a realização de experimentos devido ao grande volume de dados gerados diariamente em sistemas de atendimento a clientes. Esse trabalho tem por objetivo analisar se o uso de Mineração de Textos aumenta a acurácia dos modelos de Mineração de Dados em aplicações que envolvem textos livres. Para isso é desenvolvido uma aplicação que visa a identificação de clientes propensos a saírem de ambientes internos de atendimento (CRM) e migrarem para órgãos regulamentadores do setor de telecomunicações. Também são abordados os principais problemas encontrados em aplicações de Mineração de Textos. Por fim, são apresentados os resultados da aplicação de algoritmos de classificação sobre diferentes conjuntos de dados, para a avaliação da melhoria obtida com a inclusão da Mineração de Textos para este tipo de aplicação. Os resultados obtidos mostram um ganho consolidado na melhoria da acuraria na ordem de 32%, fazendo da Mineração de Textos uma ferramenta útil para este tipo de problema.

Palavras-chave: Mineração de dados, Mineração de Textos, Classificação, Telecomunicações, Atendimento a Clientes.

ABSTRACT

The explosion of new forms of communication between companies and new opportunities and means for companies to take advantage of this interaction. The way customers interact with companies has evolved in the recent years due to the increase in mobile devices and Internet access: clients who traditionally requested phone service migrated to electronic means of service, whether via smartphone app's or via customer service portals. As a result of this technological transformation of the communication medium, text mining has become an attractive form for companies to extract new knowledge from the register of interactions carried out by customers. Within this context, the telecommunications environment provides the inputs for conducting experiments due to the large volume of data generated daily in customer service systems. This job aims to analyze if the use of text mining increases the accuracy of data mining models in applications involving free texts. For this purpose, an application is developed that aims to identify clients likely to leave internal service environments (CRM) and migrate to regulatory agencies in the telecommunications sector [Baeza, Ricardo e Berthier ,1999]. Also addressed are the main problems encountered in text mining applications. Finally, the results of the application of classification algorithms on different data sets are presented for the evaluation of the improvement obtained with the inclusion of text mining for this type of application. The results obtained show a consolidated gain in the improvement of the acuraria in the order of 32%, making the mining of texts a useful tool for this type of problem.

Key-Words: Data Mining, Text Mining, Classification, Telecommunications, Customer Service.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

ANN	Artificial Neural Network
AR	Association Rules
CRM	Customer Relationship Management
DMEL	Data Mining by Evolutionary Learning
DM	Data Mining
DW	Data Warehouse
ETL	Extract Transform Load
IDA	Indice de Desempenho no Atendimento
K-NN	k Nearest Neighbors
KDD	Knowledge Discovery in Database
KDT	Knowledge Discovered in Texts
LABIC	Laboratory of Computational Intelligence
ML	Machine Learning
MLP	Multilayer Perceptron
NLP	Natural Language Processing
ODC	Órgão de Defesa do Consumidor
PCA	Principal Component Analysis
PPGCA	Programa de Pós-graduação em Computação Aplicada
RMT	Random Matrix Theory
ROC	Receiver Operating Characteristic Curve
SLA	Service Level Agreement
SOM	Self-Organizing Maps
SVM	Support Vector Machines
TF-IDF	Term Frequency Inverse Document Frequency
URA	Unidade de Resposta Audível
USP	Universidade de São Paulo
UTFPR	Universidade Tecnológica Federal do Paraná
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

1. INTRODUÇÃO

A popularização do acesso à internet e o aumento exponencial de dispositivos móveis como *smartphones* e *tablets*, alterou a forma como empresas e seus clientes se comunicam. Nesse novo cenário surgiram diversas aplicações que utilizam-se de mecanismos como realidade virtual e Inteligência Artificial para facilitar essa interação.

Com a evolução dos meios de comunicação entre empresas e clientes surge a oportunidade de tirar melhor proveito das informações geradas. Dentro deste contexto a Mineração de Dados é um dos meios mais apropriados para a extração de conhecimento novo neste cenário. Em particular, no caso de sistemas que permitem aos usuários a manifestação por meio de textos livres, a Mineração de Textos – que utiliza ferramentas advindas das áreas de Processamento de Linguagem Natural e de Recuperação de Informações – surge como alternativa adequada ao tratamento das informações armazenadas.

Um mercado que sempre acompanhou esta revolução tecnológica de perto e pode tirar proveito destas informações é o de telecomunicações: este setor sempre esteve alinhado com as principais tendências tecnológicas e possui os insumos para a aplicação da mineração devido ao grande volume de dados gerados diariamente em centrais de relacionamento com o cliente.

Um dos desafios deste setor é identificar clientes que não recebem tratamento apropriado após suas reclamações em centrais de atendimento e migram para órgãos de defesa do consumidor. Essa possível identificação “a priori” permitiria a tomada de decisões que evitassem que um cliente insatisfeito saísse do ambiente interno e migrasse para ambientes externos de atendimento, tais como os que são regulados por órgãos de defesa do consumidor. O mecanismo já amplamente utilizado pela área de telecomunicações que pode apresentar bons resultados na identificação de clientes insatisfeitos é a Mineração de Dados (*Data Mining* - DM). Nesse contexto, a Mineração de Dados possui diversos algoritmos que podem explorar dados, a fim de classificar se clientes podem ou não migrar do ambiente interno para o ambiente externo de atendimento.

A Mineração de Dados fornece grande potencial para ajudar empresas a encontrarem tendências importantes em suas enormes bases de dados.

Ferramentas de Mineração de Dados podem responder perguntas de negócios que tradicionalmente poderiam levar muito tempo para serem respondidas. Lejeune (2001) [Lejeune et al. 2001] abordou técnicas de Mineração de Dados que permitiram a transformação de dados brutos em conhecimento para o negócio através da aplicação de análise de dados e técnicas algorítmicas [Hung et al. 2006].

A indústria de telecomunicações gera e armazena uma enorme quantidade de dados [Weiss et al. 2005], estes que são insumos básicos para a Mineração de Dados. Geralmente empresas de telecomunicações, registram todas as atividades sobre o ciclo de vida dos seus clientes, como chamadas realizadas e recebidas, contatos com as centrais de relacionamento com o cliente (CRM – *Customer Relationship Management*), entre outros. Os dados gerados a partir de Centrais de Relacionamento com o Cliente, representam um recurso valioso para as empresas de telecomunicações, tais dados podem ser utilizados para diversos fins por meio da extração de conhecimento novo. Existem diversos canais de contato com o cliente que podem gerar informações, dentre eles podemos citar: telefonemas, mensagens instantâneas, e-mails, formulários web, etc [Pallotta et al. 2013].

O setor de Telecomunicações foi um dos primeiros a adotar a tecnologia de Mineração de Dados em larga escala, portanto, são diversas as aplicações desenvolvidas para esta área de negócio. Estas aplicações podem ser divididas em três principais áreas: *marketing* e retenção de clientes, isolamento de falhas de rede e detecção de fraudes [Weiss et al. 2005]. Para essas aplicações a Mineração de Dados possui diferentes tarefas com algoritmos específicos que permitem a extração de conhecimento novo para os mais variados contextos de negócio.

Um ponto importante identificado na etapa de levantamento de trabalhos correlatos é que não foram encontrados trabalhos similares ao contexto em que é aplicado a Mineração de Dados nesta pesquisa. Para auxiliar na análise de identificação de potenciais clientes que migram do ambiente interno para o externo uma alternativa promissora a ser utilizada é a Mineração de Textos, que tem a função de enriquecer os conjuntos de dados extraindo informações de textos livres para o emprego de algoritmos de Mineração de Dados. Este é o diferencial desta pesquisa.

A Mineração de Textos permite a transformação de dados textuais não estruturados em atributos estruturados, que propiciam após a aplicação de

algoritmos de Mineração de Dados o conhecimento útil, muitas vezes inovador para algumas organizações. O seu uso permite a extração de conhecimento novo a partir de dados brutos não estruturados [Rezende, Marcacini e Moura, 2011]. Nessa abordagem, a Mineração de Textos torna-se o objeto de estudo principal para o enriquecimento de conjuntos de dados desta pesquisa; esse enriquecimento busca fornecer melhores condições para os algoritmos de classificação da Mineração de Dados que sejam capazes de identificar potenciais clientes que saiam do ambiente interno e migrem para o ambiente externo de atendimento.

1.1. CONTEXTUALIZAÇÃO

A transformação que está acontecendo na forma de comunicação entre empresas e clientes é dada em grande parte pela evolução tecnológica, que permite cada vez mais a interação entre os envolvidos. Está evolução leva a transformações sociais onde clientes, que estavam limitados a um único canal de comunicação, passam a ter a possibilidade de acessar outras formas de comunicação devido ao acesso a novos mecanismos que permitem esta interação.

Do ponto de vista técnico, a grande maioria das empresas não armazenam em seus bancos de dados contatos telefônicos entre clientes e empresa, e quando o fazem a extração de conhecimento para estes ambientes é cara e complexa. Contudo, com estes novos formatos de comunicação entre empresa e os clientes o armazenamento de dados textuais são mais simples, permitindo a aplicação de técnicas de Mineração de Textos para a adequação dos dados à tarefa de mineração.

A quantidade de dados que são armazenados é tão grande que a análise manual dos dados se torna impossível. A necessidade de lidar com tais volumes de dados levaram ao desenvolvimento de sistemas robustos e inteligentes. Estes sistemas automatizados desempenham funções importantes, tais como a identificação de padrões escondidos, a classificação de dados e o agrupamento de perfis de clientes.

Nas aplicações desenvolvidas para Mineração de Dados em telecomunicações a grande quantidade de dados gerada apresenta vários problemas interessantes. Um dos principais problemas diz respeito a escala: as

bases de dados em telecomunicações podem conter bilhões de registros e estão entre os maiores bancos de dados do mundo. Uma segunda questão é que os dados brutos não estão adequados para a aplicação de Mineração de Dados na maioria das vezes, sendo necessário a aplicação de diversas técnicas de pré-processamento para a adequação de seu uso na Mineração de Dados [Weiss et. al., 2005].

Outro ponto interessante tratando de Mineração de Dados para telecomunicações é que muitas aplicações são voltadas para prever eventos muito raros, como falhas de componentes de rede ou uma instância de fraude telefônica, portanto, raridade é outra questão que deve ser tratada. Por fim, o desempenho em tempo real é outro ponto de atenção, modelos de detecção de fraude por exemplo devem executar de maneira *online* para realizar adequadamente sua função [Weiss et. al., 2005].

Nesse contexto, a Mineração de Dados é apresentada como uma das etapas do KDD que é definida como um processo que utiliza a Matemática, a Estatística, a Inteligência Artificial e técnicas de aprendizado de máquina (*Machine Learning* - ML) para extrair e identificar informações úteis, implícitas e previamente desconhecidas, a partir de grandes bases de dados e posteriormente, utilizar de forma adequada os conhecimentos adquiridos [Femina et al. 2015].

Portanto, são diversos os desafios a serem enfrentados para o desenvolvimento de aplicações úteis, que gerem conhecimento novo a partir dos dados disponíveis. Dentro do escopo deste trabalho, são muitos os pontos que precisam ser superados, porém o KDD fornece meios adequados para a realização deste trabalho, cujo objetivo é verificar se a inclusão de Mineração de Textos dentro do contexto analisado gera uma acurácia maior nos modelos de classificação aplicados aos conjuntos de dados.

1.2. MOTIVAÇÃO

Observando esta constante mudança na forma de interação entre clientes e empresas, fica claro a necessidade do desenvolvimento ou adaptação das ferramentas existentes para o melhor uso das informações geradas a partir deste novo formato de comunicação. Diante deste exposto, é necessário que a tecnologia

empregada, seja capaz de superar diversos desafios, tais como: a escalabilidade, o pré-processamento de textos e a descoberta de conhecimento em textos não estruturados.

O formalismo que disponibiliza todos os recursos necessários citados acima é a Mineração de Textos, pois ela fornece um conjunto de métodos usados para navegar, organizar, achar e descobrir informações em bases textuais. Esse formalismo pode ser visto como uma extensão da Mineração de Dados pois é focada na análise de textos.

A Mineração de Textos surgiu a partir da necessidade de se descobrir, de forma automática, informações (padrões e anomalias) em textos. O uso deste formalismo permite recuperar informações, extrair dados, resumir documentos, descobrir padrões, associações e regras e realizar análises qualitativas ou quantitativas em documentos de texto [Gupta et al. 2009]. Sendo assim, estas funcionalidades encaixam-se perfeitamente com as necessidades deste novo formato de comunicação entre clientes e empresas, e passa a ser o formalismo adotado para a extração de conhecimento neste trabalho.

Outro ponto motivador é que as empresas buscam incansavelmente formas de melhorarem seus resultados e ganhos de capital. Este estudo propõe a utilização da Mineração de Dados para encontrar clientes críticos de empresas de telecomunicações que possam sair do ambiente interno de atendimento e que possam migrar para órgãos de defesa do consumidor. Muitos estudos comprovam o sucesso de companhias que obtiveram retorno com o estudo de seu banco de dados, comprovando o investimento em técnicas, hardware e ferramentas analíticas a fim de trabalhar como mineradores nos seus grandes bancos de dados em vez de mantê-los apenas como repositório de dados [Weiss et al. 2005].

Telecomunicações foi uma das primeiras áreas a utilizar a Mineração de Dados para o descobrimento de conhecimento sobre os dados [Weiss et al. 2005]. Posteriormente outras áreas iniciaram a aplicação de modelos de mineração para a extração de conhecimento sobre diversas origens de dados.

Um dos fatores que motivam o uso de técnicas para realização de descobrimento de conhecimento novo é a escalabilidade dos bancos de dados atuais, ou seja, atualmente apenas o esforço de equipes de pessoas não é o suficiente para realização do trabalho de preparar, analisar e classificar informações

para descobrir fatos novos que possam contribuir com o objeto em questão. Dessa forma, a utilização de algoritmos e técnicas automatizam o processo de preparação do ambiente, deixando apenas o trabalho de análise das informações por parte dos especialistas para descobrir novos fatos que possam ser utilizados em tomadas de decisões.

1.3. OBJETIVOS

1.3.1. OBJETIVO GERAL

O objetivo é avaliar o quanto a Mineração de Textos é útil em tarefas de Mineração de Dados que envolvem textos livres escritos por diversos usuários. Isto é especialmente importante devido a evolução dos meios de comunicação, pois existe uma tendência de aumento de aplicações que façam o uso desta forma de comunicação. Em particular se tratará do problema da classificação de clientes que migraram do ambiente interno para o ambiente externo de atendimento em empresas de telecomunicações.

1.3.2. OBJETIVOS ESPECÍFICOS

Para o desenvolvimento desta pesquisa, seguem os objetivos específicos alinhados com o objetivo geral:

- Atuar na fase de preparação de dados desenvolvendo um modelo capaz de utilizar dados estruturados e não-estruturados (textuais);
- Propor um modelo de classificação baseado na relevância das entradas mistas utilizando dados textuais e não-textuais;
- Identificar e analisar por meio de experimentos qual algoritmo melhor se adapta para resolver o problema de pesquisa;
- Provar por meio de experimentos que modelos ajustados as informações fornecidas pelos clientes em forma de texto livre são superiores a modelos tradicionais que não utilizam-se de dados não-estruturados.

1.4. METODOLOGIA

Esse trabalho pode ser classificado como uma pesquisa experimental, pois implica na intervenção sistêmica no ambiente pesquisado de forma a observar se as alterações provocadas produzem os resultados esperados acerca das modificações executadas [Wazlawick et al. 2014]. Trata-se de uma pesquisa quantitativa, uma vez que a abordagem adotada para análise do método proposto ocorrerá por meio dos resultados mensuráveis obtidos com os experimentos executados.

O método científico adotado para a pesquisa é o método dedutivo, pois com base no conhecimento técnico e científico já formalmente conhecido é possível o desenvolvendo e avaliação de uma solução computacional que ofereça suporte consistente com base nas premissas estabelecidas [Gerhardt e Silveira, 2009].

Os procedimentos e técnicas empregadas nos experimentos foram selecionadas com base no levantamento bibliográfico realizado para o estado da arte. Durante essa fase foram identificadas as principais técnicas utilizadas em projetos de Mineração de Dados voltados para a área de telecomunicações. Com base nas aplicações desenvolvidas é possível identificar as diversas soluções que apoiam o processo de descoberta de conhecimento novo em diversos segmentos de telecomunicações (*marketing*, fraude, falha de rede e atendimento aos clientes). Os resultados são analisados e avaliados comparando-os com os métodos tradicionais para a comprovação do conceito e análise da proposta.

1.5. ESTRUTURA DO TRABALHO

Este documento está organizado da seguinte forma. No capítulo 2 é apresentado o referencial teórico, são conhecidos os principais conceitos relacionados à Mineração de Dados e à Mineração de Textos que são utilizados nessa pesquisa.

O capítulo 3 apresenta aplicações desenvolvidas utilizando a Mineração de Dados e Textos voltadas para o setor de telecomunicações. Os dados utilizados nestas aplicações são sempre voltados para o ciclo de vida dos clientes dentro de aplicações nestas empresas, onde diversas tarefas são abordadas com o uso destas tecnologias.

O capítulo 4 apresenta os fluxos de atendimento em ambientes de centrais de relacionamento e a organização dos órgãos de defesa do consumidor. Isto é importante para melhor compreender os dados utilizados nesta pesquisa, que objetivam o desenvolvimento de um modelo de classificação conforme o objetivo geral desta pesquisa.

O capítulo 5 apresenta os experimentos realizados, indicando os algoritmos de classificação selecionados e os diferentes métodos utilizados para a ponderação dos dados não-estruturados utilizados na pesquisa. Na sequência são apresentados os resultados obtidos com a aplicação destas variações sobre duas bases distintas, a base que utiliza dados não-estruturados e a que não os utiliza.

O capítulo 6 finaliza o documento com as considerações finais, apresentando as conclusões e as contribuições desse trabalho. Nesse capítulo ainda são apresentadas as oportunidades identificadas, que possam futuramente ser exploradas dando continuidade a esta pesquisa.

2. FUNDAMENTAÇÃO TEÓRICA

Esta seção tem por objetivo apresentar técnicas e métodos utilizados em Mineração de Dados e em Mineração de Textos que estão presentes nos trabalhos correlacionados e são utilizadas neste trabalho.

2.1. DESCOBERTA DE CONHECIMENTO

Na década de 1980, devido ao avanço em tecnologias de hardware dos computadores e em seus meios de armazenamento, surgiu a possibilidade de se utilizar novas técnicas e ferramentas para a análise de dados. Os métodos utilizados até então estavam limitados à geração de relatórios informativos que não extraiam conhecimento novo para o apoio à tomada de decisão. Essa ampla área passa a ser denominada Descobrimento de Conhecimento em Banco de Dados ou *Knowledge Discovery in Databases* (KDD) [Fayyad et al. 1996].

O KDD é o conjunto de técnicas e métodos de extração de conhecimento que abrange desde a seleção dos dados até a análise dos resultados que foram obtidos na etapa de Mineração de Dados (ver Figura 1). De acordo com Witten [Witten et al. 2000], a etapa de preparação dos dados para o uso na mineração dos dados é a responsável por consumir a maior parte dos esforços investidos em todo o processo. Cabena [Cabena et al. 1998] estima que a etapa de pré-processamento dos dados pode consumir até 60% dos recursos utilizados em projetos de Mineração de Dados.

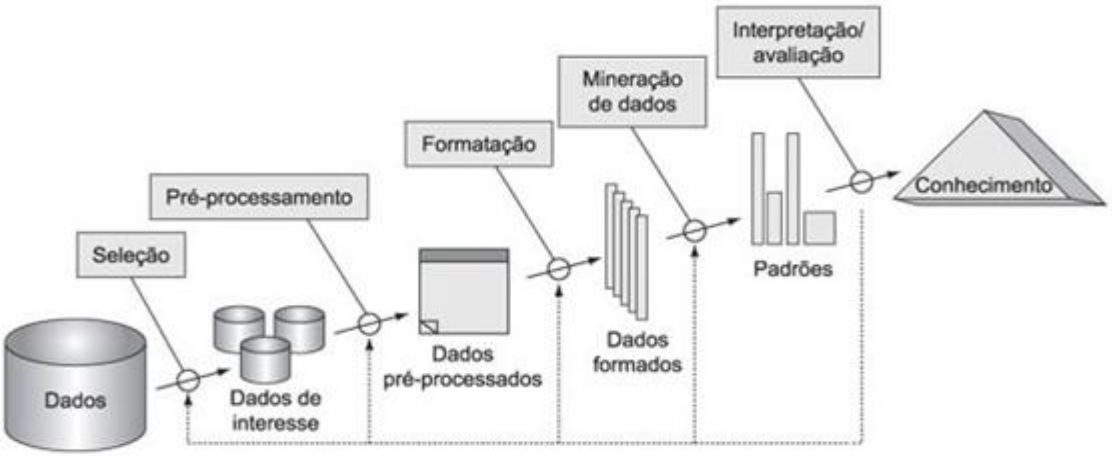


Figura 1 – Etapas do KDD [Fayyad et al. 1996]

As etapas de KDD citadas por Fayyad [Fayyad et al. 1996] são apresentadas na Figura 1 e descritas a seguir.

- Entendimento do domínio: focaliza no entendimento do que se deseja extrair através do processo;
- Pré-processamento: escolhe os atributos relevantes e realiza ajustes como discretização, conversão, normalização, tratamento de ruídos e valores ausentes e a normalização dos dados para a construção de um conjunto de dados apto para a etapa de Mineração de Dados;
- Seleção da tarefa de Mineração de Dados: seleciona a tarefa de Mineração de Dados que melhor se enquadra nos objetos do projeto (classificação, regressão, agrupamento, etc);
- Seleção de algoritmos: escolhe o algoritmo ou processo computacional mais adequado para desempenhar a tarefa objetivada;
- Mineração de Dados: executa o algoritmo ou algoritmos conforme a técnica e os métodos selecionados;
- Interpretação dos resultados: analisa de forma heurística dos resultados obtidos.
- Consolidação: valida o conhecimento adquirido com novos indicadores ou os compara com resultados obtidos por outros meios.

Embora bancos de dados estejam em constante evolução com o objetivo de facilitar o manuseio dos dados, ainda há diversos problemas que podem ser encontrados quanto ao ambiente. Esses problemas normalmente estão relacionados à forma como os bancos de dados são utilizados e não com a estrutura com a qual a informação está modelada. Matheus [Matheus et al. 1993] apresenta desafios constantemente encontrados no processo de KDD, e menciona soluções práticas para alguns deles.

- Dinâmica dos dados: as informações em bancos de dados estão em frequente mudança, a validade de amostragens interfere na validade do conhecimento, sendo necessário identificar os períodos em que a análise é praticada;
- Ruído: dados discrepantes prejudicam a geração de conhecimento novo e somente com amostras maiores pode-se facilmente identificar os *outliers* (valores discrepantes);

- Dados faltantes ou incompletos: dados nulos ou a falta de informações devido a falhas no projeto em banco de dados impedem a construção de modelos e análises contundentes;
- Padronização de medidas: dados com dois ou mais tipos de medidas (como por exemplo metros e centímetros) para a mesma informação causam dependências herdadas, prejudicando as análises devido a falsa correlação;
- Volume de dados: a grande quantidade de registros obriga a sua seleção randômica para a geração de amostras;
- Sumarização dos dados: é necessário que os conjuntos de dados utilizados representem os diferentes contextos em que a informação está inserida.

Alguns dos pontos mencionados são facilmente contornados com técnicas de KDD, porém há problemas que só podem ser identificados e tratados sabendo-se o propósito e objetivos do projeto de Mineração de Dados.

2.1.1. SELEÇÃO DE DADOS

Uma das primeiras atividades práticas no processo de KDD é a seleção dos dados. Geralmente são diversas as fontes de dados dentro de um mesmo ambiente computacional de uma organização ou ainda pode existir situações onde diversas fontes de dados de diferentes origens devem ser mapeadas e extraídas para um banco de dados único permitindo a integração dos dados. Sem a compreensão das diferentes fontes de dados, dificilmente aplicações úteis possam ser desenvolvidas, deixando uma grande lacuna entre o que se espera e o produto entregue [Weiss et al.2005]. Portanto, é necessário o entendimento dos objetivos ao qual a pesquisa é desenvolvida, o contexto em que a informação está inserida, a complexidade das diferentes fontes de dados e suas tecnologias para a eficiente coleta de dados para o desenvolvimento de aplicações úteis.

2.1.2. PRÉ-PROCESSAMENTO

O propósito do pré-processamento dos dados é transformar os dados de entrada brutos em um formato apropriado para análises subsequentes e estabelecer bases para a Mineração de Dados. Ou seja, antes da descoberta de conhecimento novo o conjunto de dados deve ser previamente preparado; a Figura 2 apresenta as etapas desta atividade. Em casos onde esta atividade é ignorada ou não efetivamente executada os resultados finais normalmente são insatisfatórios. Dessa forma, os resultados obtidos com a execução dos algoritmos estão atrelados à efetiva preparação dos dados e à extração correta de suas características [Zhang et al. 2007].

Os principais objetivos da etapa de pré-processamento são identificar dados corrompidos ou ruidosos, atributos irrelevantes e valores desconhecidos. Outras atividades comumente realizadas na etapa de pré-processamento são o uso de técnicas de discretização, binarização, construção de algoritmos de transformação e criação de variáveis, e o pré-processamento de dados não estruturados que está presente na Mineração de Textos.

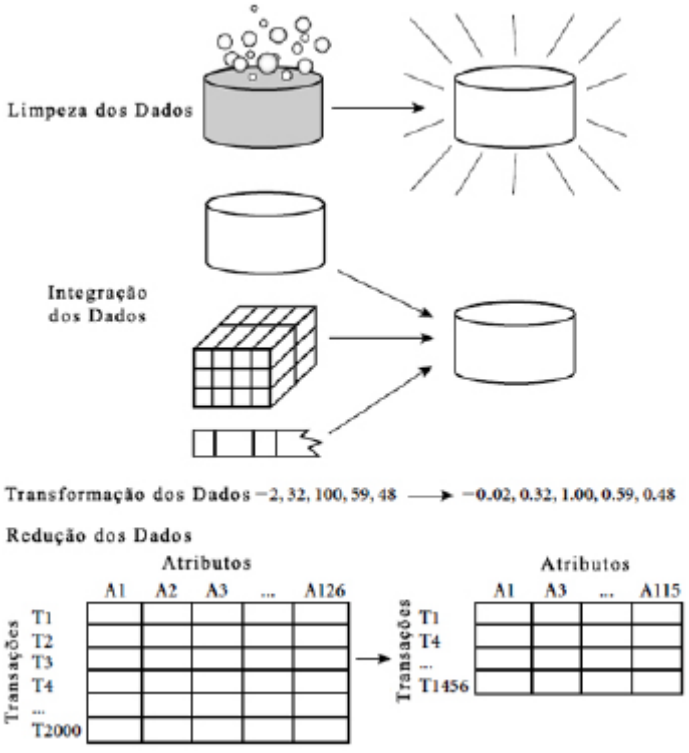


Figura 2 – Atividade do pré-processamento

A redução de dimensionalidade – número de atributos no conjunto de dados – é uma importante técnica utilizada na etapa de pré-processamento, pois essa delimita a extensão dos dados utilizados, o que gera diversos benefícios em projetos de Mineração de Dados. Um benefício chave é que os algoritmos de Mineração de Dados funcionam melhor se a dimensionalidade for menor. Isto ocorre em parte porque a redução de dimensionalidade pode eliminar características irrelevantes e reduzir o ruído. Outros benefícios são os de permitir uma melhor visualização dos dados e gerar modelos mais compreensíveis.

A discretização é uma técnica importante para alguns algoritmos de aprendizado de máquina, em especial para algoritmos de classificação que requerem a transformação de atributos contínuos em atributos categóricos. A aplicação dessa técnica na etapa de pré-processamento permite que algoritmos de classificação apresentem melhores resultados [Antunes e Oliveira, 2001].

A transformação de variáveis refere-se a transformação aplicada aos valores dos atributos. Um exemplo disto são os métodos de normalização de dados – ajustar a escala dos valores de um atributo entre 0 e 1 – ou a criação de novos atributos à partir de atributos já existentes. Esse tipo de operação é justificada pois, além de expressar relacionamentos conhecidos entre atributos existentes, pode reduzir o conjunto de dados simplificando o processamento de algoritmos [Fayyad et al. 1996].

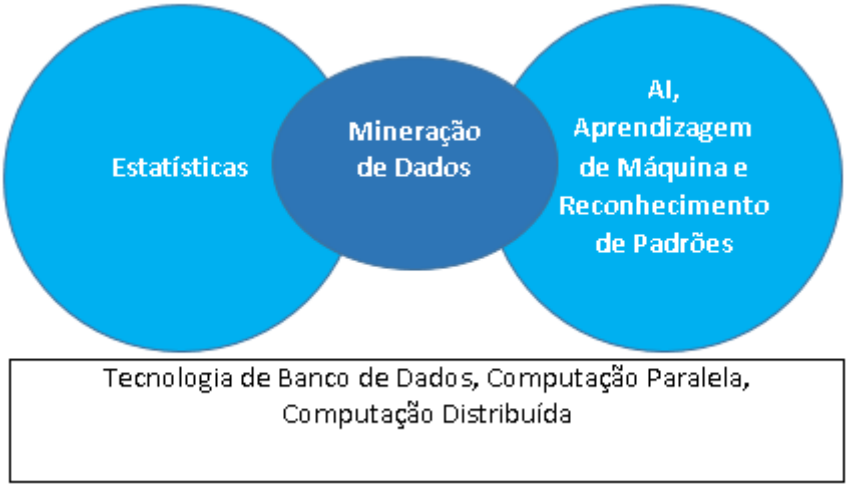
Segundo Zhang [Zhang et al. 2007], o maior tempo gasto em um projeto de Mineração de Dados é consumido com a preparação dos dados: estima-se que 80% do tempo do projeto é gasto na preparação dos dados. Portanto, na grande maioria dos projetos de Mineração de Dados os dados brutos devem ser processados de tal maneira que possam fornecer melhores condições ao conjunto de dados visando facilitar à compreensão dos modelos e a execução dos algoritmos [Tan et al. 2006].

2.1.3. MINERAÇÃO DE DADOS PROPRIAMENTE DITA

As etapas descritas anteriormente garantem a limpeza e a preparação dos dados utilizados na Mineração de Dados, e sua não aplicação pode levar à descoberta de padrões sem sentido e inválidos [Fayyad et al. 1996]. Mineração de

Dados é a etapa do processo de KDD que consiste na aplicação de algoritmos específicos, que extraem padrões a partir dos dados [Fayyad et al. 1996].

Esta etapa utiliza conceitos como os de amostragem, estimativa e teste de hipóteses, algoritmos de buscas, técnicas de modelagem e teorias de Inteligência Artificial, reconhecimento de padrões e aprendizagem de máquina para obter conhecimento útil [Tan et al. 2006]. Além disto a Mineração de Dados rapidamente adotou ideias de outras áreas, como otimização, computação evolutiva, teoria da informação, processamento de sinais, visualização e recuperação de informações



[Tan et al. 2006].

Figura 3 – Mineração de Dados como uma confluência de muitas disciplinas

A escolha da técnica utilizada na etapa de Mineração de Dados está intimamente ligada ao tipo de tarefa adotada no projeto. Isto torna necessário distinguir o que é uma tarefa e o que é uma técnica de mineração. A tarefa está relacionado com o que se busca nos dados, se é encontrar similaridades entre dois objetos, classificar itens ou prever a variação de valores. As técnicas de Mineração de Dados consistem na especificação de métodos que garantam descobrir os padrões estabelecidos e está fortemente ligado com a tarefa de Mineração de Dados estabelecida [Goldshmidt et al. 2005].

As tarefas de Mineração de Dados geralmente são divididas em duas categorias: tarefas de previsão e tarefas descritivas. O objetivo das tarefas de previsão é prever um valor futuro de um atributo alvo baseado em valores de outros atributos. Nas tarefas descritivas o objetivo é identificar padrões que demonstrem os relacionamentos dos dados. Na Figura 4 são apresentados resumidamente as principais tarefas de Mineração de Dados.

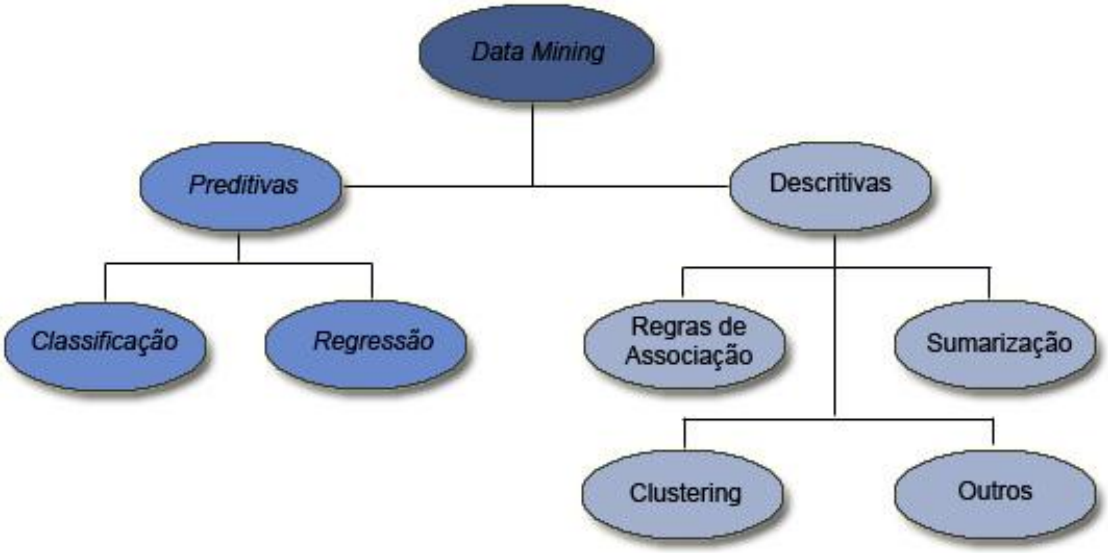


Figura 4 - Tarefas de Mineração de Dados

Associação: É a tarefa utilizada para identificar quais atributos estão relacionados, ou seja, descobre padrões que descrevem características comuns entre atributos de dados. Os padrões descobertos são normalmente representados na forma de regras de implicação ou subconjuntos de características extraindo os padrões interessantes de uma forma eficiente. A análise de associação pode envolver, por exemplo, a identificação de páginas Web que sejam acessadas simultaneamente [Tan et al. 2006].

Classificação: A classificação pode ser definida como a tarefa de aprendizado de uma função *f* que mapeie cada conjunto de atributos *x* para rótulos de classes *y* pré-determinadas [Tan et al. 2006]. Após encontrada essa função a mesma pode ser aplicada a novos registros para prever qual a classe correspondente às novas entradas. A classificação é executada em duas etapas: aprendizagem e classificação. Na etapa de aprendizagem os algoritmos são treinados com dados de teste que permitem obter a correta classificação das entradas, e na etapa de classificação entradas desconhecidas são apresentadas às estruturas de decisão geradas pelos algoritmos e classificadas conforme o aprendizado [Deulkar et al. 2016].

Regressão: Tarefa similar a classificação, porém na regressão a variável alvo é contínua e tem como principal objetivo apresentar uma previsão a partir de dados históricos contidos em uma base de dados, ou seja, compreende a busca por uma função que mapeie os registros de um banco de dados em valores reais. Estatística e Redes Neurais, dentre outras áreas, oferecem ferramentas para a implementação da tarefa de regressão [Michie e Spiegelhalter, 1994].

Agrupamento: O objetivo da tarefa de agrupamento é identificar e aproximar registros similares. Um agrupamento ou *cluster* é o conjunto de registros similares entre si que forma determinado cluster e é distante de outros grupamentos que possuem características diferentes. A análise de agrupamentos procura verificar a existência de diferentes grupos dentro de um determinado conjunto de dados. Portanto, o objetivo dessa tarefa não é classificar, estimar ou predizer o valor de uma variável, mas sim identificar os grupos de dados similares.

Detecção de Desvios ou *Outliers* é a identificação dos registros considerados anormais, ou seja, que não atendem ao padrão considerado normal. O objetivo dos algoritmos de detecção de desvios é identificar valores verdadeiramente fora do padrão e evitar rotular erroneamente objetos normais como anômalos. Na prática algoritmos dessa tarefa devem ter uma alta taxa de detecção e uma baixa taxa de alarme falso [Tan et al. 2006].

2.1.4. MÉTRICAS DE AVALIAÇÃO DOS RESULTADOS

A última etapa do processo de KDD tem por objetivo realizar a interpretação e avaliação dos resultados obtidos a fim de identificar se os objetivos iniciais foram alcançados. Com a interpretação podem surgir padrões, relacionamentos e descoberta de novos fatos antes desconhecidos, de forma que esta fase também busca identificar e eliminar resultados não legítimos da Mineração de Dados. Caso os resultados obtidos não satisfaçam os objetivos iniciais é possível retornar as etapas anteriores para a realização de ajustes e correções, caso contrário os resultados podem ser incorporados a outros sistemas, documentados ou utilizados em processos de tomada de decisão [Fayyad et al. 1996].

Considerada uma fase importante no processo de KDD, nesta etapa é necessária a participação de especialistas nos dados, conhecedores do negócio e tomadores de decisão. Diversas ferramentas e técnicas podem ser utilizadas para a análise dos resultados ou modelos obtidos. Visando obter confiabilidade nos modelos testes e validações devem ser aplicadas e calculados indicadores para medir a qualidade dos resultados. São exemplos de técnicas de validação: *cross validation, suplied test set, use training set, percentage split*, e de indicadores de

avaliação: matriz de confusão, índice de correção e incorreção de instâncias mineradas, estatística *kappa*, erro médio absoluto, erro relativo médio, precisão, *F-measure*, dentre outros [Witten et al. 2005].

2.2. ALGORITMOS DE CLASSIFICAÇÃO

Detalha-se a seguir a tarefa de Mineração de Dados que será empregada neste trabalho. A classificação é a tarefa de Mineração de Dados mais comumente aplicada, que emprega um conjunto de exemplos pré-classificados, para desenvolver um modelo que possa classificar registros futuros. O processo de classificação de dados envolve a aprendizagem e classificação, como visto anteriormente. A precisão das regras de classificação é apurada na fase de testes, onde os dados são apresentados para o modelo com sua classificação pré-definida [Deulkar et al. 2016]. Os principais algoritmos de classificação são descritos a seguir.

2.2.1. ÁRVORES DE DECISÃO

As Árvores de Decisão constituem uma técnica capaz de extrair um conjunto de decisões organizadas em uma estrutura hierárquica. Consiste em nós que formam uma árvore com um ponto raiz, o que significa que o nó raiz é o ponto de partida. Os nós que possuem arestas de saídas são chamados de nós internos ou de teste. Os nós que estão localizados nas folhas da árvore são chamados de nós terminais ou nós de decisão.

As Árvores de Decisão podem incorporar em seus testes valores tanto nominais como numéricos, e são de fácil interpretação. Cada caminho da raiz da árvore de decisão com uma das suas folhas pode ser interpretado como regra. Normalmente, a complexidade da árvore é medida por um dos seguintes métodos: o número total de nós, número total de nós folhas, profundidade e número de atributos utilizados.

2.2.2. NAÏVE BAYES

O classificador Naïve Bayes pertence a uma família de classificadores probabilísticos simplificados, apoiados na aplicação da Regra de Bayes sobre a probabilidade de ocorrência de cada classe de atributos na base, assumindo a independência entre os atributos. O classificador Naïve Bayes é denominado ingênuo (Naïve) por assumir que os atributos são condicionalmente independentes. Para a aplicação do algoritmo as probabilidades necessárias são estimadas com base nas frequências correspondentes obtidas a partir da base de treinamento.

2.2.3. MÁQUINAS DE VETORES DE SUPORTE

As Máquinas de Vetores de Suporte (*Support Vector Machines - SVM*) formam um conjunto de métodos para aprendizado supervisionado, aplicáveis a problemas de classificação e regressão [Maimon et al. 2010]. SVM apresenta bons resultados em diversas aplicações práticas, inclusive em bases de dados com muitas dimensões e, portanto, possui uma certa imunidade à “maldição da dimensionalidade”. O método baseia-se na construção de hiperplanos separadores para as classes em um espaço de atributos de dimensão muito superior ao do problema original, obtido por meio de transformações matemáticas (*kernels*) adequados.

O tempo de treinamento geralmente é rápido, e são altamente precisos, devido à sua capacidade de modelar complexos limites de decisão linear. SVM´s possui propensão menor a *overfitting* do que outros métodos.

2.2.4. K-NN

Classificadores K-Vizinhos mais Próximos são baseados em métodos de aprendizagem por analogia, ou seja, são obtidos comparando-se a tupla a classificar com tuplas de treinamento que são semelhantes. As tuplas de treinamento são descritas por *n* atributos, e portanto cada tupla representa um ponto em um espaço *n*-dimensional. Quando uma nova tupla desconhecida é apresentada, o classificador K-Vizinhos mais Próximo procura as K tuplas de treinamento que estão mais

próximas da tupla desconhecida. Estas tuplas de treinamento K são os K “vizinhos mais próximos” da tupla desconhecida. De forma geral a classe da tupla desconhecida é obtida como sendo a da maioria das classes das K tuplas mais próximas.

2.2.5. REDES NEURAIS

O estudo de redes neurais artificiais (ANN) foi inspirado em tentativas de simular sistemas neurais biológicos, onde uma ANN é composta de um conjunto interconectado de nós e de ligações direcionados. As Redes Neurais são geralmente construídas sobre o modelo básico de neurônio denominado perceptron. Cada perceptron é constituído por várias entradas e uma saída. As entradas formam a estimulação do perceptron, e o valor de saída é obtido pela comparação entre a soma ponderada das entradas e um limiar predefinido.

As Redes Neurais Artificiais multicamadas possuem uma estrutura mais complexa, sendo formada por vários perceptrons interconectados. A rede multicamadas pode conter diversas camadas intermediárias entre as camadas de entrada e de saída, tais camadas intermediárias também são conhecidas como camadas ocultas. Desenvolver um modelo de redes neurais artificiais não é uma tarefa trivial, pois sua construção envolve uma série de fatores, como o entendimento dos dados, e o balanceamento da divisão dos dados para teste e treinamento, entre outros.

2.3. DESCOBERTA DE CONHECIMENTO COM MINERAÇÃO DE TEXTOS

A Mineração de Textos ou descoberta de conhecimento a partir de textos (KDT) utiliza técnicas de recuperação de informações e de processamento de linguagem natural (*Natural Language Processing* - NLP) em conjunto com algoritmos e métodos de KDD, Mineração de Dados, Aprendizado de Máquina e Estátistica [Hotho, Nürnberger e Paaß, 2005]. Mineração de Textos refere-se genericamente ao processo de extração de padrões ou conhecimentos interessantes e não-triviais de documentos de textos não estruturados. Acredita-se que a extração de conhecimento de textos tenha potencial comercial mais elevado do que a

extração de conhecimento sobre dados. No entanto, a Mineração de Textos é uma tarefa bem mais complexa do que a Mineração de Dados, pois, envolve lidar com dados de texto que são inerentemente não-estruturados e distorcidos [Tan et al. 1999]. A Figura 5 apresenta KDT como um campo multidisciplinar que envolve a recuperação de informação, análise de texto, extração de informações, agrupamento, classificação, visualização, banco de dados, aprendizado de máquina e Mineração de Dados.

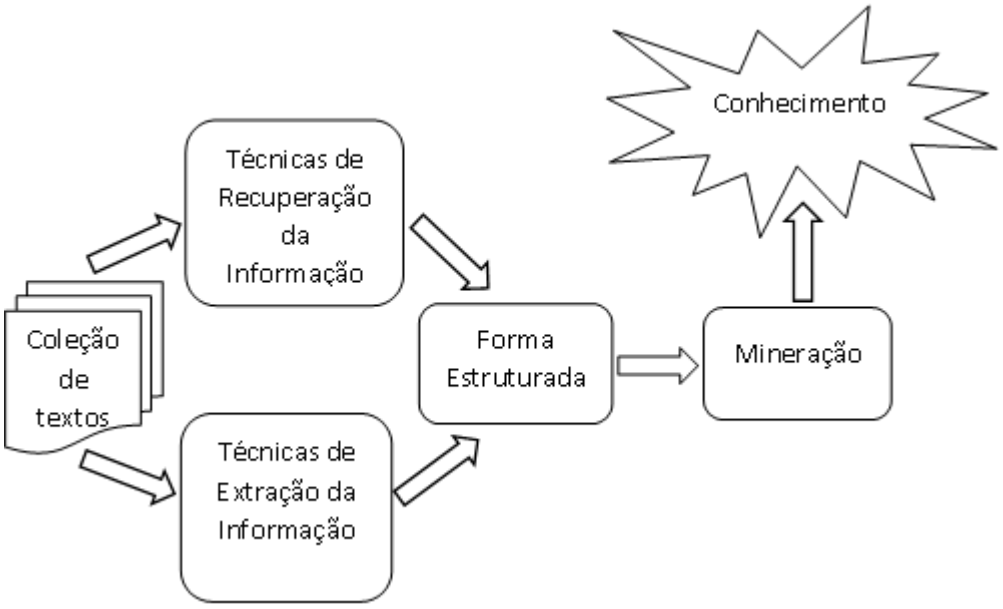


Figura 5 – O processo de Mineração de Textos

A Mineração de Textos é uma técnica emergente no campo de Mineração de Dados [Kaur e Aggarwal, 2013]. Os seres humanos têm a capacidade de distinguir e aplicar padrões ao texto e podem facilmente superar obstáculos que os computadores não podem facilmente resolver, tais como, gírias, variações de grafia e significado contextual. No entanto, embora as capacidades linguísticas humanas permitam compreender dados não estruturados, não temos a capacidade do computador para processar grandes volumes de textos ou em altas velocidades.

2.4. ETAPAS DA MINERAÇÃO DE TEXTOS

O uso de dados não estruturados em projetos de Mineração de Dados envolve a aplicação de diversas técnicas de pré-processamento tais como a radicalização, a remoção de *Stop Words*, a conversão de termos e outros para

tornar esses dados à forma estruturada, tornando-os aptos ao uso nos algoritmos de Mineração de Dados [Kaur e Aggarwal, 2013].

O descobrimento de conhecimento textual (KDT) extrai conceitos explícitos, implícitos e relações semânticas utilizando técnicas de processamento de linguagem natural (NLP). Para isso, conforme Figura 6, é necessário a execução de uma série de etapas para que os dados se tornem apropriados para o seu uso em projetos de Mineração de Dados. Estas etapas são: remoção de pontuação, remoção de números, conversão do texto para um caso único (maiúsculas ou minúsculas), remoção (*Stop words*) e radicalização das palavras (*Stemming*).

O pré-processamento tem o objetivo de converter documentos desestruturados em uma forma estruturada, resultando geralmente em uma tabela atributo-valor. A aplicação de tais técnicas promove a eliminação de ruídos sobre os dados e aumenta a precisão, viabilizando a diminuição da dimensionalidade e gerando uma tabela atributo-valor mais coerente.

Figura 6 – Etapas aplicadas no pré-processamento da Mineração de Textos

A remoção de *Stop-Words* é a etapa executada no pré-processamento de dados não-estruturados que consiste na identificação de termos frequentes em textos e que não geram informação relevante para a base de dados, ou seja, sem conteúdo semântico representativo no texto. Sua remoção tem como finalidade a redução dos termos analisados no documento e a diminuição do número de palavras

armazenadas na base de dados, sendo portanto um meio de diminuição da dimensionalidade da tabela atributo-valor. Os termos irrelevantes são aqueles que aparecem com muita frequência e se tornam desnecessários para as tarefas de busca e categorização de textos.

Algoritmos de radicalização ou *stemmer* permitem a remoção das variações de uma palavra, permanecendo apenas o radical correspondente do termo, ou seja, as variações de uma palavra são simplificadas a uma forma comum. Tais variações incluem plurais, gerúndios, sufixos de terceira pessoa, sufixos de tempo passado, etc, por exemplo, o verbo “Trabalhar” que pode ter muitas variações, tais como: trabalhou, trabalhando, trabalhei, trabalhaste estas variações são reduzidas por *stemming* ao radical “trabalh”. É válido ressaltar que o radical resultante da radicalização não é necessariamente igual a sua raiz linguística. Ao final do processo o armazenamento é melhorado e ocorre a redução da dimensionalidade, pois menos termos são armazenados [Baeza, Ricardo e Berthier ,1999].

2.5. REPRESENTAÇÃO VETORIAL

Para a representação estruturada de textos o modelo espaço-vetorial é o mais amplamente utilizado [Baeza, Ricardo e Berthier ,1999]. Neste modelo, cada termo corresponde a um radical obtido no pré-processamento textual e também a um atributo na base estruturada. Estes atributos são associados um vetor e cada um dos termos possui um valor associado que indica o seu grau de importância. No vetor estão todos os termos considerados da coleção e não aqueles presentes no documento. Os termos que não aparecem no elemento textual recebem grau de importância zero [Tan et al. 1999]. Na Tabela 1 é apresentado um exemplo de matriz (termo x documento) que é utilizada nos experimentos deste trabalho, e que considera para valor de cada termo a frequência com que o mesmo aparece em cada documento.

Tabela 1 – Exemplo de matriz de termo documento

	Atendimento	Boleto	Cobrança	Dúvida	Erro	Informação	Plano	Reclamação
Documento1	2	0	0	0	1	1	1	3
Documento2	0	1	1	0	0	1	4	0
Documento3	1	0	0	0	2	1	1	3
Documento4	1	0	1	1	0	1	7	1
Documento5	0	0	2	0	1	1	2	4
Documento6	3	1	0	0	0	1	1	1
Documento7	0	0	1	0	1	1	1	2
Documento8	3	0	0	0	2	1	0	2
Documento9	1	0	1	0	0	1	2	1
Documento10	0	3	4	0	5	0	1	6

O peso de um termo pode ser calculado de diversas formas. Uma forma comum é o método booleano que pondera os termos com dois valores possíveis: zero ou um, o valor zero é assumido quando não existe a menção do termo no documento e o valor um é assumido quando o documento possui o termo em questão no documento, independentemente da quantidade de vezes que ele é citado. A frequência do termo (*Term Frequency*) é outra medida utilizada que consiste na frequência (número de vezes) que o termo é encontrado no documento. Por fim, o TF-IDF (*Term Frequency – Inverse Document Frequency*) é uma medida que leva em consideração a frequência do termo no documento e o número de documentos da coleção em que o termo aparece.

Essencialmente o TF-IDF funciona determinando a frequência relativa das palavras em um documento específico em comparação com a proporção inversa dessa palavra em todo o corpus do documento. As palavras que são comuns em um único ou pequeno grupo de documentos tendem a ter números TF-IDF mais altos do que palavras comuns, como artigos e preposições.

O procedimento para a implementação do TF-IDF tem algumas variações em diferentes aplicações, mas a abordagem geral funciona da seguinte forma: dada uma coleção de documentos D, um termo w e um documento individual $d \in D$, o peso é calculado por

$$f_{w,d} * \log(|D|/f_{w,D}) \quad (2),$$

Onde $f_{w,d}$ é igual ao número de vezes que w aparece em d, $|D|$ este é o tamanho da coleção de documentos, e $f_{w,D}$ é o número de documentos em que w aparece em D (Salton & Buckley, 1988, Berger, et al, 2000). Existem algumas situações diferentes que podem ocorrer para cada termo, dependendo dos valores de $f_{w,d}$, $|D|$, e $f_{w,D}$.

Suponha que $|D| \sim f_{w,}$ é só $f_{w,}$ D, ou seja, o tamanho do corpus é aproximadamente igual à frequência de w sobre D. Se $1 < \log(|D| / f_{w\,D}) < c$ para alguma constante muito pequena c, então w_d será menor que $f_{w,d}$, mas ainda positivo. Isso implica que w é relativamente comum em todo o corpus, mas ainda tem alguma importância em D. Por exemplo, este poderia ser o caso se TF-IDF examinaria a palavra "Jesus" sobre o Novo Testamento. Este é também o caso de palavras extremamente comuns, que por si só não possuem significado relevante em uma consulta. Tais palavras comuns recebem assim um escore TF-IDF muito baixo, tornando-os essencialmente insignificantes para a ponderação.

Finalmente, suponha que $f_{w,d}$ seja grande e $f_{w,D}$ seja pequeno. Então, $\log(|D| / f_{w,D})$ será bastante grande, e, portanto também será grande. Este é o caso de maior interesse, uma vez que termos com alta ponderação são importantes em d mas não são comuns em D, tendo portanto um grande poder discriminatório.

2.6. CONSIDERAÇÕES FINAIS

Neste capítulo foi apresentado a fundamentação teórica necessária para fundamentar esta pesquisa. Foram apresentadas as diferentes etapas do KDD e suas funções, como também as diferentes tarefas encontradas na Mineração de Dados e as técnicas comumente utilizadas na etapa de pré-processamento. Também foi apresentado o conceito da Mineração de Textos que busca a obtenção de conhecimento novo a partir de informações não-estruturadas e as principais técnicas disponíveis para a realização dessa tarefa.

Os pontos abordados nesse capítulo estão relacionados com as técnicas utilizadas nos trabalhos correlatos que serão apresentados no próximo capítulo.

3. TRABALHOS CORRELATOS

Neste capítulo são apresentados alguns trabalhos encontrados na literatura sobre Mineração de Dados e Mineração de Textos em ambientes de Telecom e de centrais de atendimento ao cliente para esse setor. Os trabalhos abordados tratam de assuntos como marketing, fraude, e *churning* e utilizam diversas técnicas e métodos encontrados na Mineração de Dados e de Textos, e serviram para embasar este trabalho.

Chang [Chang, Ling Wang, 2009] aplicou tecnologias de *data warehouse* (DW) e Mineração de Dados sobre dados de CRM para analisar o comportamento e identificar o perfil clientes e o modelo de crescimento em ambientes de Internet e *e-commerce.* Parte dos dados é formada por reclamações e demandas de clientes através de e-mails, ou seja, dados não-estruturados. Nas primeiras etapas são aplicadas regras de seleção pré-estabelecidas na integração dos dados primitivos para decidir se os dados são mantidos ou descartados e decidir a qual subconjunto cada dado pertence. Após esta limpeza e organização dos dados, os mesmos são organizados em grupos de temas relacionados através de métodos de transformação de dados.

No estudo as classes alvos foram definidas conforme o enquadramento dos clientes em categorias pré-definidas, determinadas pelo seu interesse: investimentos, informações da indústria, dados empresariais, informações de produtos, recrutamento, etc.

Para o experimento foi utilizado o algoritmo C4.5, uma extensão do algoritmo ID3 de Ross Quinlan, para gerar a árvore de decisão. O critério do ID3 para selecionar atributos é o “ganho de informação”, e seus métodos são baseados na teoria da informação. Ele mede a quantidade de informação em cada classe e calcula a quantidade média de informação, ou entropia, no conjunto de treinamento, a fim de expressar o seu nível de complexidade. A acurácia obtida nos experimentos realizados foi de 88%. O experimento contribuiu com o estabelecimento de categorias através da classificação de textos, além de possibilitar pela análise de conteúdo transformar dados de textos para dados estruturados que auxiliam no processo de descoberta de conhecimento.

Hadden [Hadden et al. 2006] avaliou o desempenho de três técnicas a fim de identificar o modelo mais adequado para a predição de *churn* utilizando dados de reclamações dos clientes. Foram utilizados dados não estruturados em conjunto com os classificadores Redes Neurais, Árvores de Decisão e Regressão. O conjunto de dados utilizado no experimento é formado por três grupos de variáveis com duas classes possíveis para a classificação dos dados, que são *churn* e *no-churn*. O primeiro grupo de variáveis representa as estimativas feitas pela empresa para a resolução da reclamação. No segundo grupo estão representadas as informações das reclamações do cliente. O terceiro grupo reúne as informações sobre uma falha ou reparo. A partir deste conjunto de dados foram aplicadas as técnicas descritas acima com diversas configurações nos seus modelos. Os resultados são interessantes para Redes Neurais, onde duas variações foram utilizadas na arquitetura da rede: a bayesiana e a *feed-forward*, com diferentes funções de ativação. A arquitetura bayesiana apresentou melhores resultados que a *feed-forward*. Porém dentro de todas as técnicas utilizadas nos experimentos foi árvore de decisão que obteve os melhores resultados, seguido de Regressão e Redes Neurais. Na acurácia obtida nas classes *churn* e *no-churn* existem algumas particularidades, como as Redes Neurais serem melhores na predição de *churn* do que as outras técnicas e a Regressão atingir índices superiores a 90% na predição de *no-churn*. Mas, como a precisão geral da acurácia é dada pela predição de *churn* e *no-churn,* a técnica com melhor resultado foi árvore de decisão com 82%. Das vinte e quatro variáveis selecionadas para o experimento apenas quatorze foram utilizadas nas técnicas avaliadas. Árvore de decisão utiliza sete variáveis do conjunto de dados, sendo que duas são utilizadas em Redes Neurais e outras duas na Regressão Linear. A partir dessa informação é definido que o tipo de reclamação, número de reclamações, número de compromissos perdidos e se um pedido foi feito são variáveis importantes para todas as técnicas. A pesquisa forneceu *insights* interessantes sobre a previsão de rotatividade de clientes em empresas de telecomunicações e as diferentes tecnologias disponíveis para a tarefa de previsão.

Ahn [Ahn et al. 2011] desenvolveu um modelo heterogêneo para facilitar o aumento de *cross-selling* no mercado de telecomunicações móveis. Ou seja, seu modelo utiliza dados demográficos dos clientes e padrões dos mesmos como idade, média de tempo nas chamadas de voz, tipos de chamada de voz e tipo do plano do

cliente entre outros. Esses dados mais as informações de produtos ou serviços anteriormente utilizados são utilizados para encontrar novos produtos e serviços com alto potencial de vendas. As classes alvos do experimento indicam se um cliente vai adquirir um novo produto ou não. Para isso foram determinadas três classes alvos: perspectivas pouco prováveis, chances médias e perspectivas altamente prováveis. Várias técnicas de Mineração de Dados foram aplicadas sobre o conjunto de dados para a realização do experimento, que foi dividido em duas etapas.

Na primeira etapa técnicas de classificação são aplicadas, tais como Regressão Logística, Redes Neurais Artificiais e Árvores de Decisão. As técnicas são aplicadas de forma independente, onde cada modelo produz as probabilidades de sua predição. Na segunda etapa o modelo considera todas essas probabilidades usando algoritmo genético e toma a decisão final para um cliente-alvo se ele ou ela vai adquirir um novo produto. Na configuração do algoritmo genético é utilizada uma população de 100 indivíduos e com uma taxa de *crossover* definida em 0,5 e taxa de mutação em 0,06. Como resultado do desenvolvimento do experimento é possível verificar que é possível aplicar o modelo em outras áreas que utilizam técnicas de Mineração de Dados para *cross-selling*.

O experimento produziu resultados satisfatórios na identificação de clientes com alto potencial para a aquisição de novos produtos ou serviços e possibilitou a economia no envio de propagandas de *marketing* para clientes não alvo da empresa. A melhor acurácia foi obtida com a combinação de três classificadores heterogêneos (Regressão Logística, Árvore de Decisão e Rede Neurais) chegando a 66%.

Adwan [Adwan et al. 2014] propõe em seu trabalho o uso de redes neurais perceptron multi-camadas MLP com aprendizagem *back-propagation* para a previsão de *churn* em uma empresa de telecomunicações da Jordânia. Diferentes topologias MLP com diferentes configurações foram utilizadas para construir os modelos de classificação de *churn*. Foram investigadas duas abordagens diferentes para a identificação de variáveis importantes. A primeira baseia-se na métrica de calcular o conjunto de variáveis removendo-as uma a uma e a segunda na contribuição das variáveis aos pesos na rede.

O conjunto de dados utilizado nos experimentos possui onze atributos, estes atributos indicam se um cliente possui o serviço 3G ou não, a taxa de consumo mensal, registros de SMS´s locais e internacionais como também a quantidade de minutos gastos em ligações locais e internacionais. Por fim, o último atributo do conjunto classifica se o registro pertence da classe dos clientes que abandonaram os serviços da empresa ou não.

O experimento também estudou o efeito da alteração do número de épocas e o número de neurônios na camada oculta do modelo. Foi identificado que a melhor configuração para o problema da pesquisa é a rede com 4 neurônios na camada oculta, a qual alcançou 62% de acurácia com um número de 5000 épocas. Dos dois modelos de abordagens utilizados no experimento, a abordagem que calcula a contribuição das variáveis aos pesos da rede apresentou melhor resultado.

Lin [Lin et al. 2014] com base em um conjunto de dados de clientes de telecomunicações aplica técnicas de redução de dimensionalidade e redução de dados para compreender o melhor procedimento para estas duas importantes etapas da fase de pré-processamento dos dados. O conjunto de dados inicial possui 173 atributos divididos em duas classes, dos quais 34761 registros são clientes que se desligaram da empresa e 16545 são registros de clientes que não se desligaram.

Para o experimento foram construídos oito modelos de predição combinando técnicas de estatística multivariada (*Principal Component Analysis*), regras de associação (*Association Rules* - AR) e mapas auto-organizáveis (*Self-Organizing Maps* - SOM). A fim de testar os subconjuntos de dados foram utilizadas Redes Neurais M*ultilayer Perceptron* com o algoritmo de aprendizagem B*ack-Propagation*. Na configuração das Redes Neurais foram consideradas quatro configurações diferentes nas camadas ocultas com 8, 12, 16 e 24 neurônios e quatro variações incluindo 50, 100, 200 e 300 épocas. Ainda é utilizado o método de validação cruzada, que divide o conjunto de dados em dez partes iguais onde qualquer nove dos dez subconjuntos são selecionados para o treinamento e a parte restante é utilizada para testar o modelo. Em seguida, a distribuição da acurácia média e dos erros pode ser obtida.

A avaliação do desempenho dos modelos de predição é dada através da matriz de confusão que incida a quantidade de *no-churning* que é classificado como *churning* e de *churning* que é classificado como *no-churning*. Nos resultados dos

experimentos ficou claro que não existe um modelo que seja melhor que os outros em todos os métodos de avaliação. Na acurácia de predição o método que considera primeiro a redução de dados e a redução de dimensionalidade tanto utilizando SOM + PCA ou SOM + AR produziu os melhores resultados com a Rede Neural utilizando oito neurônios na camada oculta e cinquenta épocas, obtendo 98,99% e 99,01% de taxa de acurácia. Na avaliação de qual modelo apresenta melhores resultados na taxa de erros de previsão os resultados são similares, porém SOM + PCA apresenta melhores resultados do que SOM + AR.

No quesito que avalia as melhores taxas de redução de dados e performance de predição AR produz melhores resultados que PCA e torna o modelo MLP ligeiramente melhor em termos de precisão e acerto de *churning* classificados como *no-churning*. Portanto, o objetivo da redução de dados e a redução da dimensionalidade é disponibilizar conjuntos de dados mais “limpos” e/ou mais representativos, filtrando as características irrelevantes e eliminando amostras de dados com ruídos. Dessa forma, o experimento que usou a redução de dados seguida pela redução de dimensionalidade produziu um “melhor” conjunto de dados para a construção de um modelo de predição ideal, onde o custo de treinamento foi amplamente reduzido se comparado à utilização do conjunto de dados originais.

Tan [Tan et al. 2000] apresenta em seu trabalho uma abordagem que combina o uso de Mineração de Dados e Mineração de Textos para a melhoria do custo de chamadas de serviços. O caso de uso dos experimentos aconteceu em uma empresa que oferece suporte telefônico para produtos de controle industrial, como sistema de controle distribuído, válvulas automáticas e sensores. O objetivo era obter informações sobre a natureza dos problemas tratados e o custo esperado de diferentes tipos de solicitação de serviço. No trabalho foram combinadas técnicas de Recuperação de Informação e Aprendizado de Máquina em um novo método de categorização de campos híbridos de formato fixo e texto livre.

A base utilizada foi coletada durante um ano e possui cerca de 20 mil casos. Os atributos selecionados contêm informações do número de requisições de serviços, tipo do problema encontrado, número de funcionários envolvidos na solução do problema, o produto reclamado e o tempo que o cliente levou para resolver o problema em questão. Para os campos de textos livres as informações foram escritas pelos funcionários da empresa que descreveram o problema em

questão. Os algoritmos utilizados foram Árvores de Decisão C4.5 e Naïve Bayes, pois estes algoritmos são comumente utilizados na categorização de textos e são de fácil entendimento. Nos experimentos realizados foi utilizado a validação cruzada para dar maior confiabilidade nos resultados. Observou-se que nos conjuntos de dados onde foram incorporados as informações de textos livres houveram pequenas melhorias nas taxas de acurácias dos modelos; a taxa de acurácia foi de 53% para C4.5 e 79% para o Naïve Bayes. A conclusão do trabalho é que a incorporação de dados de textos livres pode contribuir com o aumento de acurácia dos modelos de classificação de dados.

Ye [Ye et al. 2012] realiza a segmentação de clientes de empresas de telecomunicações através do algoritmo de agrupamento *K-means*. A utilização de *K-means* no experimento foi devida aos seguintes fatores: (1) o algoritmo fornece uma boa solução para o problema de agrupamento com a utilização de atributos numéricos; (2) é relativamente escalável e eficiente no processamento de grandes conjuntos de dados; (3) não é sensível a entrada de novos dados embora seja sensível a ruídos os dados estão completos; (4) o algoritmo é rápido na sua modelagem e seus resultados são de fácil entendimento.

O objetivo do estudo é segmentar centenas de milhares de clientes segundo as dimensões de valores e comportamentos, para entender as características de consumo de diferentes grupos de clientes, fornecendo uma base analítica para estratégias de marketing e para o desenvolvimento de novos negócios. Os dados utilizados nos experimentos contêm: produto do cliente, tempo de acesso à rede, quantidade de reclamações, informações sobre benefícios do cliente e informações de atendimento do cliente, como a consulta de tarifas, a consulta de serviço e o aviso de tarifas, dados de duração de ligações e valores das tarifas cobradas.

A conclusão do trabalho é que a segmentação dos clientes foi realizada com sucesso auxiliando no processo de tomada de decisão da empresa. Foram identificadas características para diferentes grupos, estes grupos são: características por custo total de faturas, por chamadas de longa distância, chamadas locais e características de negócio. Cada um dos grupos identificados possui características que os determinam e fornecem informações competitivas para a empresa em questão, tornando o uso da ferramenta desenvolvida indispensável para a agregação de valor ao negócio.

3.1. OUTRAS APLICAÇÕES DE MINERAÇÃO DE DADOS EM TELECOMUNICAÇÕES

Wu [Wu et al. 2014] desenvolve uma solução para prever o comportamento fraudulento em empresas de telecomunicações utilizando o algoritmo rede neural de *Kohonen*. Foram comparados três tipos de algoritmos no experimento: redes neurais de *Kohonen,* agrupamento em duas etapas*,* e *K-means.* As Redes Neurais de *Kohonen* apresentaram os melhores resultados sendo o procedimento mais eficiente em encontrar valores discrepantes (*outliers)*. O processo utilizado no experimento foi o CRISP-DM que auxiliou nas etapas de construção, avaliação e aplicação do modelo. Por fim o modelo se mostrou útil na tarefa de identificação de clientes que possuem comportamento fraudulento dentro de empresas de telecomunicações.

Zaman [Zaman et al. 2015] apresenta a solução E-Stream, um software de predição em redes de telecomunicações. Sua arquitetura permite o processamento de grandes volumes de dados para identificação de melhores ações corretivas sobre falhas de rede, configuração, contabilidade, desempenho e segurança. O software é baseado em componentes baseados na redução de dados, correlação, filtros, predição e recomendação. Um dos princípios utilizados nos filtros aplicados na arquitetura do sistema é baseado na teoria de matriz aleatória (*random matrix theory-RMT*). De acordo com a RMT, é possível separar o sinal verdadeiro do ruído aleatório de uma matriz de correlação. O filtro espectral analisa o espaço da matriz de correlação dos eventos observados e decompõe a matriz em duas partes, uma exibindo a forte estrutura correlativa entre os eventos e a outra com fraca condição espectral que pode ser tratada como ruído. Algoritmos de Regras de Associação são utilizados para explorar a relação sequencial entre os eventos de dados de rede. Este componente do E-Stream, através de técnicas de mineração de regras de associação sobre eventos de rede, identifica como as sequências de eventos estão associados a incidentes reais e forma um padrão do evento. Com o desenvolvimento dessa solução é possível automatizar o processo de gerenciamento de redes heterogêneas, padrão que vem tornando-se cada vez mais comum no setor de telecomunicações. A aplicação desenvolvida atua como uma ferramenta de apoio no processo de tomada de decisão, podendo ser automatizada para pequenas tomadas de decisão menos expressivas e deixando apenas as decisões de maior severidade para os especialistas.

4. FLUXO E ANÁLISE DE RECLAMAÇÕES EM TELECOMUNICAÇÕES

Este capítulo trata da aplicação desenvolvida no contexto deste trabalho e apresenta o fluxo de reclamações em ambientes de telecomunicações. Para isto é realizada uma análise preliminar sobre as reclamações, com o objetivo de apresentar quais são os principais motivos e fatores que levam clientes a reclamarem em ambientes de telecomunicações e posteriormente migrarem para órgãos de defesa do consumidor. Por fim, é apresentada a proposta para o uso destes dados e a seleção das principais entradas, baseado nos dados disponíveis no CRM da empresa, para a classificação de reclamações e os métodos utilizados para a formatação adequada do conjunto de dados utilizado nos experimentos.

4.1. LIMITAÇÃO DA ÁREA DE PESQUISA

Telecomunicações é conhecida pela grande quantidade de dados gerados e a complexidade na manutenção de grandes redes. Essas características tornam-se desafios para a entrega de serviços de qualidade para milhares de clientes que estão conectados diariamente. Para que se atinja o objeto de estudo dessa pesquisa são utilizados dados de reclamações de uma empresa de telecomunicações. Nos experimentos são eliminados todos os registros gerados por meio de atendimento eletrônico, tais como formulários *online* e registros de Unidade de Resposta Audíveis (URA) entre outros. A eliminação desse tipo de registro busca a obtenção de uma base de dados onde exista apenas a interação entre clientes e atendentes ou entre clientes e sistemas com registros textuais dessas interações. Esses registros são a matéria prima dos experimentos que compreendem essa pesquisa.

A base selecionada para os experimentos corresponde a uma pequena fração de um banco de dados de produção de uma empresa de grande porte. No conjunto de dados utilizado busca-se representar fielmente a base de dados real da empresa estudada, para que os modelos desenvolvidos e resultados obtidos com os experimentos possam ser úteis para a aplicação de tais técnicas para fins comerciais no ambiente de produção da empresa.

A Figura 7 apresenta a quantidade de registros de clientes que acionaram o CRM da empresa e migraram para a Anatel, comparado à quantidade de registros da base completa do órgão no período selecionado. Analisando a base de produção da empresa verifica-se que uma pequena fração dos clientes que procuraram atendimento via CRM migraram para a Anatel. Essa pesquisa limita-se a identificar clientes que solicitaram atendimento via ambiente interno da empresa e migraram para a Anatel.

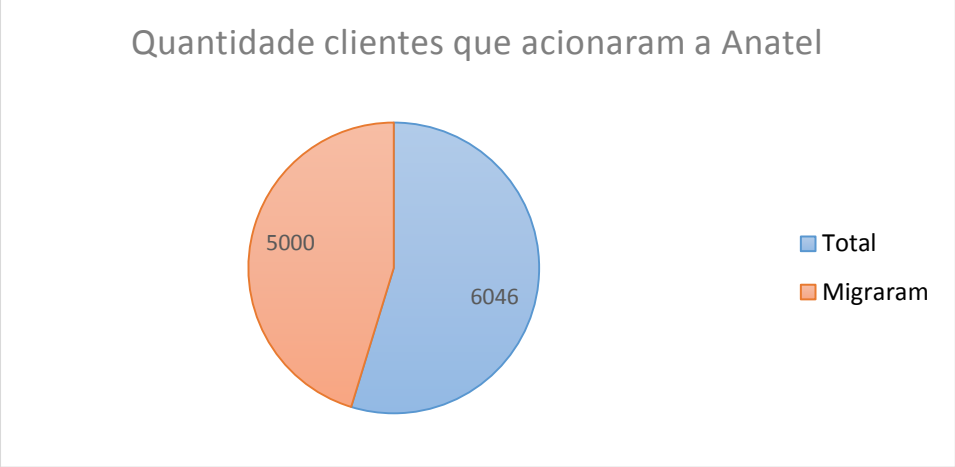


Figura 7 – Clientes que solicitaram atendimento na Anatel

Um ponto que não foi abordado nesta pesquisa é a identificação de clientes que acionam órgãos de defesa do consumidor sem ao menos ter solicitado atendimento por algum canal de atendimento da empresa, já que o mesmo está condicionado à classificação de clientes que não receberam atendimento adequado em ambiente interno e migraram para a Anatel.

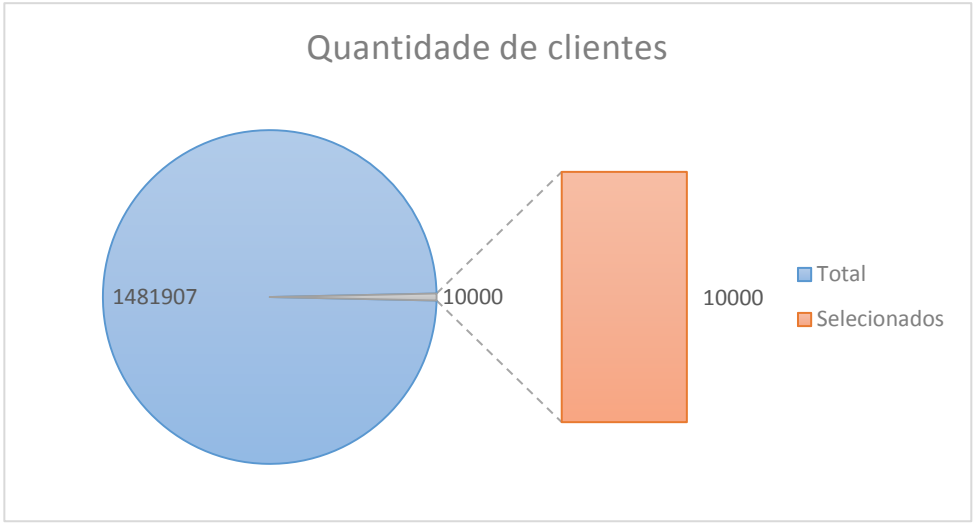


Figura 8 – Total de clientes que solicitaram atendimento via CRM x clientes selecionados

4.2. FLUXO DAS RECLAMAÇÕES EM TELECOMUNICAÇÕES

Os dados utilizados pela aplicação desenvolvida neste trabalho são provenientes das reclamações de clientes em uma empresa de telecomunicações, com isso nada mais importante do que conhecer como se dá o início de uma reclamação até a solução apresentada para a reclamação do cliente.

O cliente, quando precisa de algum atendimento sobre os serviços contratados, pode acionar a empresa de duas formas: canais eletrônicos como formulários web e o portal do cliente ou contato via telefone na central de atendimento que corresponde ao atendimento de primeiro nível da empresa. Para os contatos realizados via central de atendimento o cliente primeiramente é direcionado a URA que realiza o atendimento eletrônico, disponibilizando diversas opções para o cliente via teclado para resolver suas solicitações. Caso não seja possível finalizar o atendimento via URA o cliente tem sua reclamação encaminhada a um atendente da equipe de CRM. A partir desse ponto, independente da forma de contato do cliente, os fluxos de atendimento são similares. A única diferença é que os clientes que solicitaram atendimento via telefone estão em contato direto com o atendente e as solicitações via formulário são deslocadas sistematicamente para um atendimento de segundo nível.

Em seguida os atendentes devem analisar as solicitações dos clientes e verificar se conseguem finalizar o atendimento sanando todas as dúvidas, caso não seja possível é realizado o escalonamento para as áreas responsáveis. Estas áreas cumprem rigorosamente o prazo estabelecido pelo Acordo de Nível de Serviço (*Service Level Agreement* - SLA) em suas atividades. Caso o escalonamento seja de ordem técnica é enviado um técnico de campo que fica responsável por atender o cliente em sua residência, caso contrário a demanda é escalonada para a áreas administrativas que resolvem problemas com faturas, descontos não concedidos, etc.

A Figura 9 ilustra de forma geral o processo de atendimento em centrais de relacionamento com os clientes, ou seja, apresenta as opções de atendimento disponíveis e como as atividades são encaminhas dentro dos workflows possíveis de trabalho. Tais atividades são executadas sistematicamente visando sempre a automatização dos processos de negócio. Casos onde não é possível as soluções de forma automática são realizados os escalonamentos com as demais áreas envolvendo os responsáveis até que se consiga uma solução.

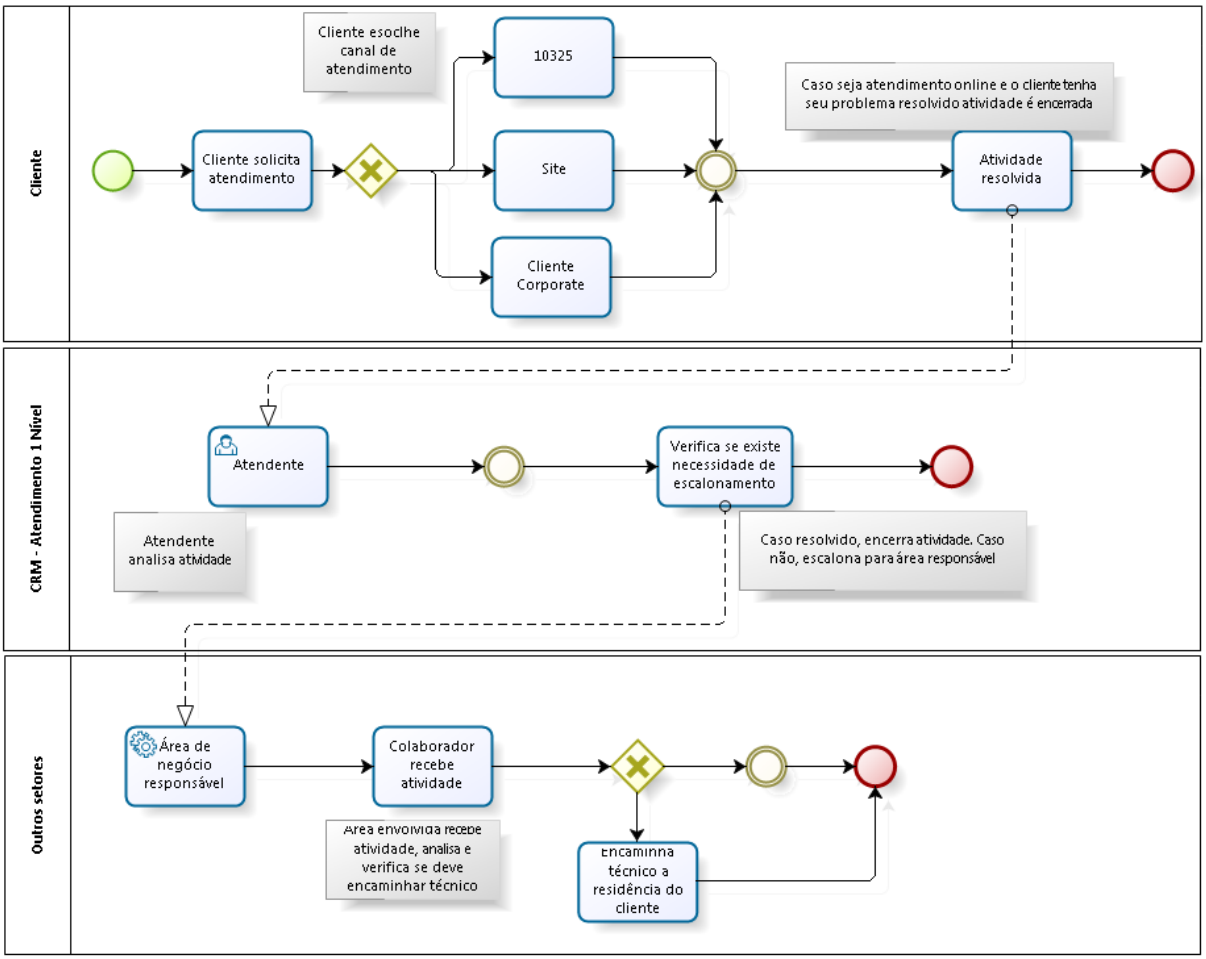


Figura 9 – Fluxo de atendimento em CRM

As demais áreas que recebem escalonamentos de centrais de relacionamento com o cliente possuem autonomia própria para organizar suas atividades conforme a necessidade. Essas áreas recebem o nome de *BackOffice*, pois não estão em contato direto com o cliente: apenas recebem as atividades encaminhadas do CRM e devolvem a solução ao CRM, que possui pessoal treinado e instruído em como relacionar-se com os clientes. Um ponto importante é que na atual estrutura da empresa é que para o tratamento de uma pendência existe a

integração de diferentes ferramentas para a sua solução devido à complexidade dos sistemas envolvidos no cenário atual.

Os processos de cada área de negócio (*BackOffice*) são muito bem definidos, porém são inúmeras estas áreas e estão separadas geograficamente distantes uma das outras o que torna praticamente impossível o mapeamento de uma maneira geral. Por esse motivo algumas áreas necessitam de muitos funcionários para realizar a gestão e o planejamento, de forma a viabilizar o acompanhamento e funcionamento conforme as metas da organização. Fica claro que grandes empresas que fornecem serviços de atendimento ao cliente necessitam de muitas pessoas com diferentes conhecimentos para fornecer um bom nível de atendimento ao cliente final, além de grandes ambientes físicos e de infraestrutura computacional com tecnologia de ponta para comportar sistemas de informação adequados para realizar a automação dessas atividades.

A severidade das reclamações de um cliente – definida pela gravidade, prejuízo causado ou rigor definido para uma reclamação –geralmente inicia com um índice baixo, ou seja, considerado normal. Porém, em algumas situações quando o cliente sofre prejuízos devido à falta de seus serviços ou problemas causados por esse motivo e não tem seus problemas resolvidos nas primeiras solicitações de atendimento, a sua insatisfação começa a aumentar, sendo assim, a severidade de suas reclamações tende a crescer.

Quando o cliente fica muito insatisfeito normalmente procura outras instâncias de atendimento. Nesse caso são os órgãos de defesa do consumidor, como a Agência Nacional de Telecomunicações (Anatel), o Programa de Proteção e Defesa do Consumidor (PROCON), o site “www.consumidor.gov.br” que é um canal eletrônico público para a solução alternativa de conflitos de consumo por meio de internet, etc. Esses órgãos têm poder para aplicar multas e penalizações sobre as empresas com muitas reclamações, além de divulgarem mensalmente *rankings* de qualidade de serviço empresas.

Cada órgão de defesa do consumidor possui suas regras e forma de comunicação com as empresas reclamadas, ou seja, a forma como as reclamações são tratadas são distintas de órgão para órgão. Quando o cliente aciona um desses órgãos, normalmente é estipulado um tempo limite para que essa reclamação seja respondida. Essa resposta deve ser dada pela empresa reclamada em tempo hábil e

caso o cliente não concorde com a mesma pode reabrir a reclamação, que novamente é direcionada à empresa reclamação. Esse fluxo pode se repetir por várias vezes até que o problema seja resolvido.

As empresas de telecomunicações que tratam reclamações recebidas de órgãos de defesa do consumidor, normalmente possuem um grande contingente de pessoas e sistemas para tratar os diferentes segmentos de clientes envolvidos. Quanto maior a segmentação de produtos e serviços que uma empresa fornece, mais complexo será o tratamento dessas reclamações, o que facilmente exigirá um contingente maior de pessoas, tornando o custo para o tratamento dessas demandas muito alto.

A Figura 10 apresenta o fluxo empregado pelos departamentos responsáveis de tratar as reclamações oriundas dos órgãos de defesa do consumidor (ODC). Os órgãos normalmente disponibilizam meios automáticos para todas as prestadoras tratarem as suas reclamações que estão no órgão, ou seja, receber as demandas, tratar e responder as mesmas. No caso da Anatel, que é o órgão escolhido para o desenvolvimento dos experimentos, a troca de informações acontece diariamente por meio de arquivos XML em que são recebidas as reclamações e enviadas as respostas ao órgão.

A Anatel disponibiliza para os clientes, no momento de registrar a reclamação, uma árvore de motivos que classifica as reclamações. Ou seja, quando o cliente inicia o processo de abertura de uma reclamação é solicitado que ele informe valores em campos pré-definidos que classificam uma reclamação como por exemplo “cobrança” ou “cobrança após cancelamento”. A classificação dessas reclamações é utilizada tanto pela Anatel quanto pela empresa para identificar o que o cliente realmente reclama, pois usando somente com o campo observação, onde o cliente descreve sua solicitação fica difícil a classificação dessas demandas. Outra característica importante nas reclamações registradas na Anatel é o preenchimento do campo denominado “serviço” que corresponde ao produto/serviço que o cliente reclama, e.g., banda larga, telefone fixo, TV, etc. Essas informações são muito úteis para as prestadoras, pois permitem ações que automatizem o tratamento das reclamações.

As prestadoras geralmente carregam as informações recebidas dos órgãos em seus sistemas próprios, o que permite realizar procedimentos automatizados

conforme as estratégias da empresa e integração com outras áreas. Após o carregamento das reclamações nos sistemas de tratamento, as reclamações são distribuídas entre os assistentes, onde finalmente é iniciado o seu processo de tratamento.

Cada órgão utiliza um prazo de resposta e estabelece uma meta diferente que deve ser cumprida pelas operadoras. No caso da Anatel é utilizado o prazo de cinco dias úteis para a resposta das reclamações. Caso a reclamação não seja respondida no prazo de cinco dias a reclamação é contabilizada como fora do prazo no Índice de Desempenho no Atendimento (IDA), indicador utilizado pelo órgão que tem como meta 85% de resolutividade.

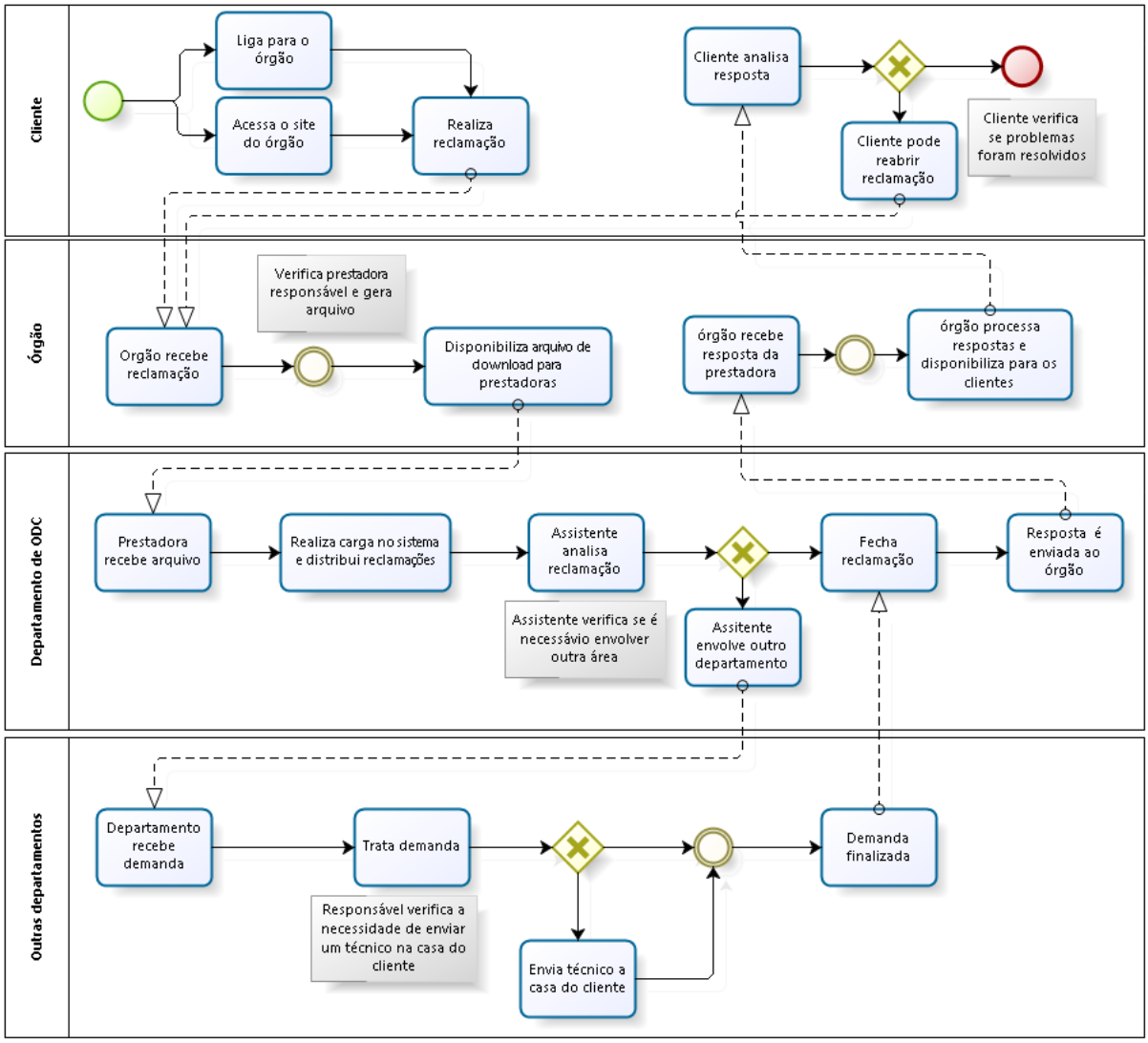


Figura 10 – Fluxo de atendimento em ODC

Diariamente é realizado o upload das respostas das reclamações, e durante um período de quinze dias o cliente pode reabrir a reclamação realizada solicitando

a revisão de algum item não atendido inicialmente, o que faz a prestadora ser penalizada pela solução incompleta do problema. Clientes que recebem atendimento a reclamações originadas nos órgãos de defesa do consumidor recebem uma marcação e são classificados como críticos por possuírem risco de *churn* (o cliente que pede cancelamento de plano e migra para outra operadora). Esse tipo de marcação serve para desde campanhas de *marketing* que visam a retenção ou a priorização de demandas na empresa.

Por fim, existe situações onde o prazo estabelecido pelo órgão para o atendimento das reclamações de um determinado serviço não é atingido, ou existe um número alto de reclamações. Nestes casos a prestadora pode ser notificada e sofrer penalidades como ficar impedida de vender em determinada região, fornecer serviços gratuitamente, ou até mesmo multas financeiras.

A principal diferença entre as reclamações atendidas via CRM e as recebidas pelos setores responsáveis por atender reclamações oriundas dos órgãos de defesa do consumidor são as prioridades dadas: devido ao cumprimento obrigatório do prazo estipulado para que se atinjam adequadamente os indicadores, as reclamações tratadas em ODC recebem prioridade em relação as reclamações de outros setores da empresa.

Outro ponto é a autonomia que o setor que atende ODC possui para tratar uma reclamação em relação a central de atendimento a clientes. O departamento que trata as reclamações de ODC normalmente possui maiores acessos por exemplo às ferramentas de desconto e prioridades diferentes comparadas as demandas tratadas por CRM, sendo mais eficaz nos atendimentos. Contudo, o modelo adotado pelo departamento que trata ODC exige um grande nível de controle e relatórios para monitoramento da qualidade e prazo das tratativas. Outro ponto que merece atenção são as fraudes que podem acontecer devido ao grande portfólio de ferramentas e acessos: é comum ocorrerem casos onde são identificados descontos e ajustes indevidos, caracterizando uma fraude dentro da própria empresa.

4.3. ANÁLISE DA BASE DE DADOS

Conforme enfatizado por Fayyad, primeiramente é necessário a compreensão do domínio da aplicação visando identificar os procedimentos adequados das etapas do KDD, conforme os objetivos relacionados ao que se quer obter [Fayyad et al. 1996]. Sendo assim, com base nos dados extraídos para os experimentos foram elaborados gráficos que apresentam características dos dados, a fim de se ter uma maior compreensão dos conjuntos de dados utilizados.

Os principais motivos das reclamações de clientes em ambientes de CRM são apresentados na Figura 11. São identificados que 4 motivos principais – Retenção, Informação, Defeito Adsl e Defeito – que correspondem a 63,14% das reclamações recebidas no período.

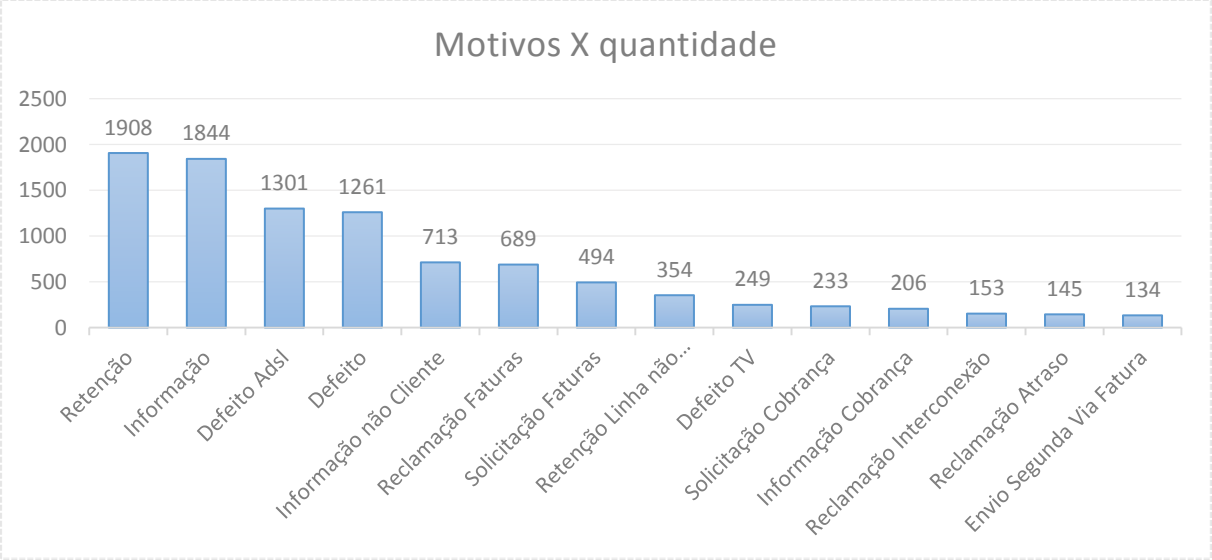


Figura 11 – Principais motivos de reclamações em CRM

Outra característica observada nos dados é que as maiores concentrações de reclamações em ambientes de CRM envolvem clientes com contratos superiores a dois anos. Porém, clientes que recém firmam contratos de prestação de serviços com a operadora também têm uma quantidade expressiva de solicitações de atendimento nesse ambiente.

Ao observar uma relação entre os clientes que recém contrataram serviços da empresa com clientes que possuem contrato superior a dois anos, constatou-se que os motivos de reclamação dos clientes é basicamente o mesmo. Essa comparação pode levar ao entendimento de que clientes que recém contrataram

serviços passam por dificuldades no início do contrato e acionam diversas vezes os ambientes de CRM. Da mesma forma clientes com mais de dois anos de contrato tendem a acionar a empresa por diversas vezes e em um número médio superior ao dos outros clientes. É necessário avaliar esse cenário a fim de identificar o que leva clientes com mais tempo na empresa a acionar tanto ambientes de CRM; uma hipótese para essa característica são os problemas contratuais que podem ocorrer ao longo do contrato.

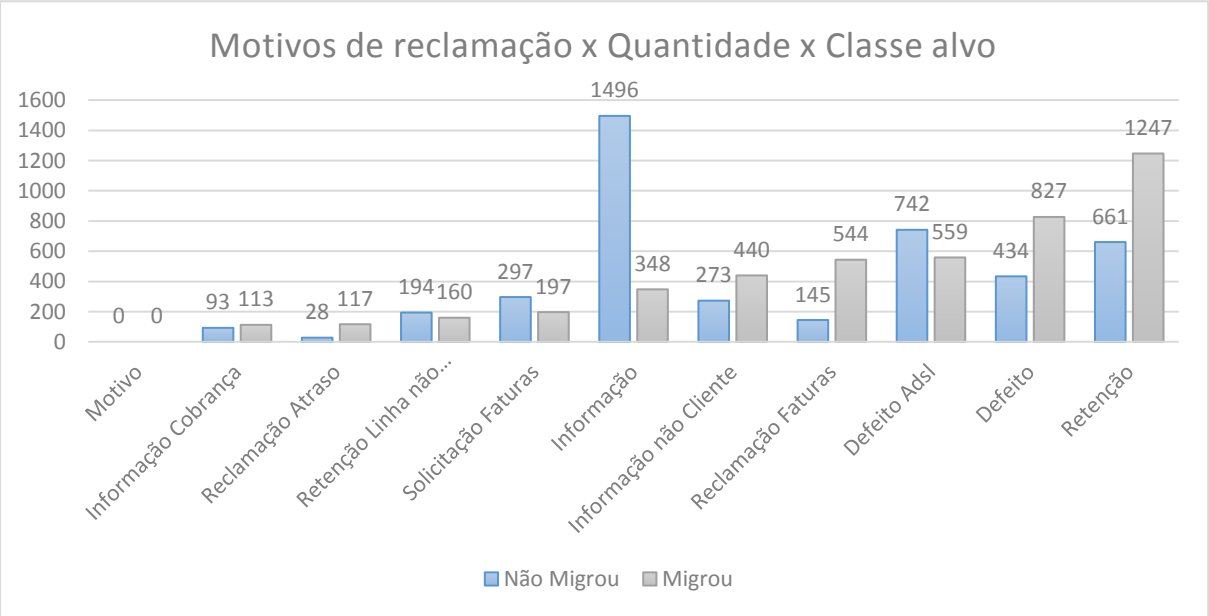


Figura 12 – Quantidade de reclamações por tempo de instalação

A comparação os motivos das reclamações dos clientes com a informação do cliente no âmbito de se ele migrou ou não para a Anatel, conforme indicado à Figura 12, permite identificar comportamentos diferentes que podem auxiliar na construção de modelos de classificação.

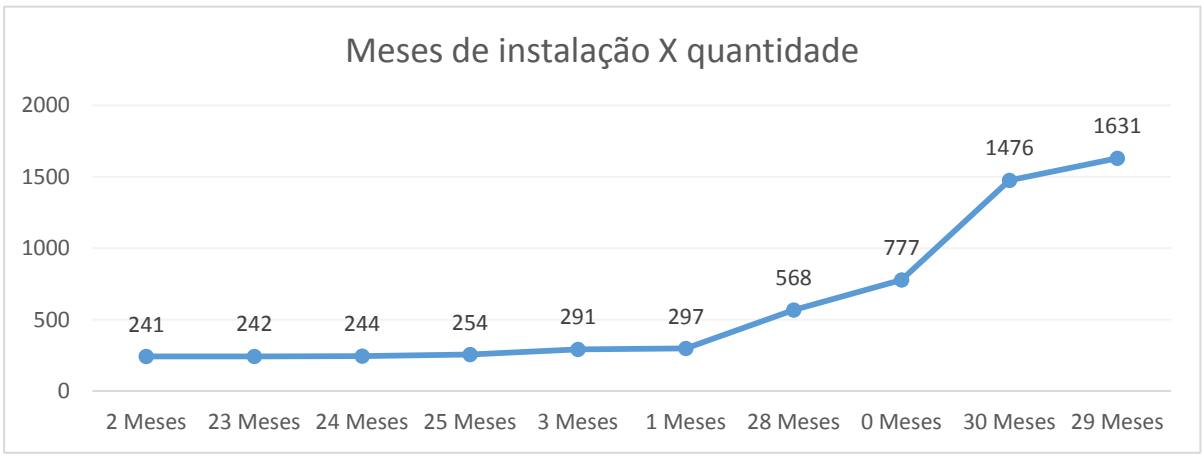


Figura 13 – Comparação entre motivo da reclamação X quantidade X Classe alvo

A quantidade de clientes que não migraram para a Anatel e cuja solicitação de atendimento foi classificada com o motivo “Informação” é significativamente superior aos clientes que tiveram a mesma solicitação de atendimento e migraram para a Anatel. Esse tipo de análise é importante para gerar os modelos de classificação, pois é possível determinar pesos maiores para motivos de reclamação cuja recorrência seja maior para clientes que migram para a Anatel.

Solicitações de atendimento em CRM que são classificados com o motivo “Retenção” têm uma probabilidade muito maior de ocasionar uma migração para um órgão de defesa do consumidor. Sendo assim, quando for identificado que o cliente reclama várias vezes com o motivo “Retenção“ o mesmo pode ser considerado como propenso a migrar para um órgão de defesa do consumidor.

Analisando o perfil dos clientes (Figura 14) que solicitaram atendimento via CRM, pode-se perceber que a faixa etária que mais reclama e aciona órgãos de defesa do consumidor está é de 30 e 40 anos, ou seja, o cliente dentro dessa faixa etária pode ser classificado como um cliente mais exigente. Quando identifica que suas solicitações não são atendidas o mesmo efetua a migração para um órgão de defesa do consumidor.

Clientes que solicitam atendimento via CRM, independentemente do serviço prestado pela empresa, querem ser prontamente atendidos em suas solicitações. Clientes que não são atendidos nas primeiras demandas se sentem lesados e procuram outras formas de atendimento.

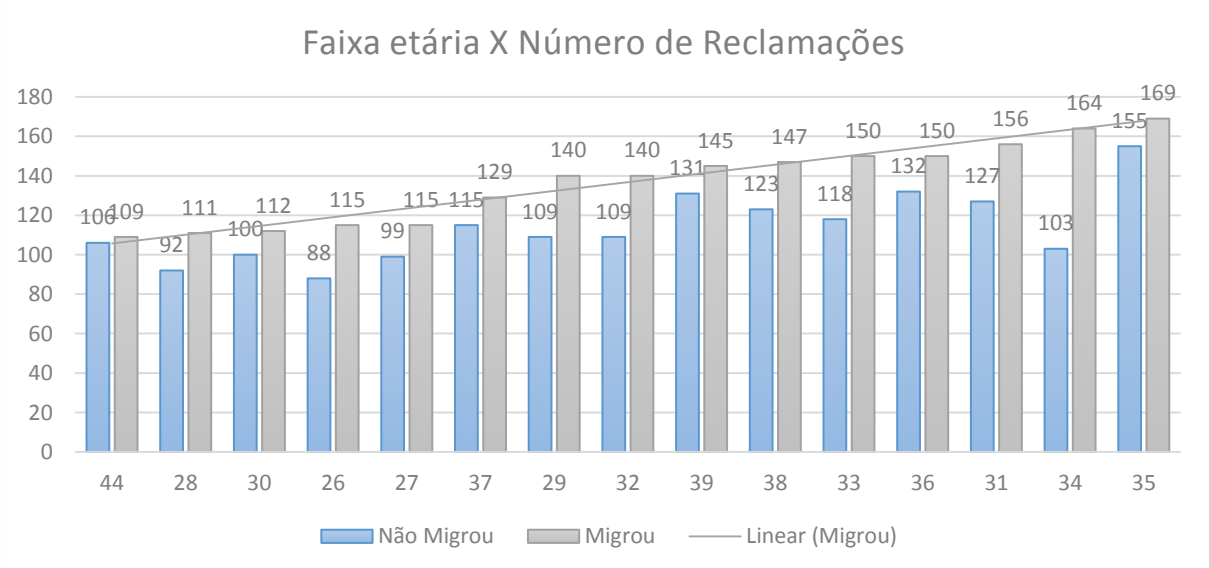


Figura 14 – Faixa etária de clientes que solicitam atendimento

Na Figura 15 é possível visualizar que clientes com mais de nove reclamações têm no mínimo 75% de chances de migrar para a agência reguladora. Esse tipo de informação pode auxiliar no atendimento, pois clientes com poucas solicitações de atendimento, que são em grande número, têm menor chance de migrar para a Anatel. Desta forma os esforços para encontrar clientes propensos a migrar para órgãos de defesa do consumidor podem ser concentradas em clientes que possuem grande número de reclamações em CRM.

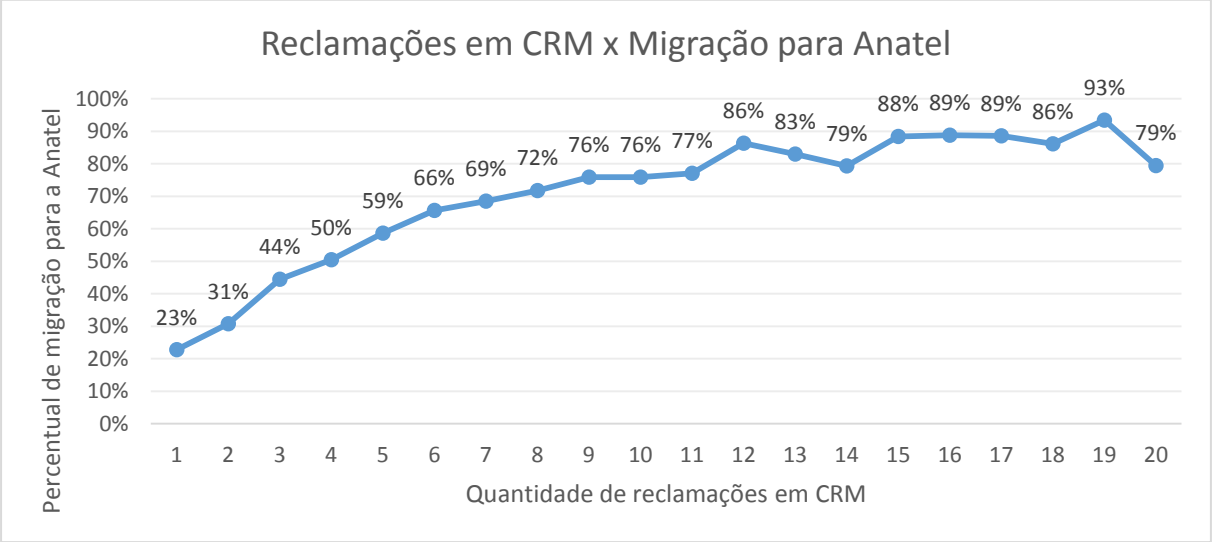


Figura 15 – Quantidade de reclamações em CRM X percentual de clientes que migraram para a Anatel

4.4. FORMAÇÃO DAS BASES UTILIZADAS

A técnica utilizada para a seleção dos registros que formam a base de dados deste trabalho foi a forma randômica, sempre respeitando as características da base completa do ambiente disponibilizado para que não houvessem resultados distintos na aplicação em diferentes conjuntos de dados. Os dados utilizados são de uso autorizado pela empresa, e os experimentos foram executados de forma a garantir o anonimato dos clientes e colaboradores, mantendo a confidencialidade e a privacidade dos mesmos.

Os dados utilizados na pesquisa são provenientes de duas fontes de dados distintas (Figura 16). Esses dados foram enriquecidos com informações adicionais de sistemas relacionados ao atendimento. A primeira base, denominada “base original sem atributos textuais”, contêm os registros das informações do ambiente de CRM da empresa. Esses registros correspondem a todos os contatos realizados

pelos clientes que podem ter sido feitos nos diferentes canais de atendimento da empresa. O período escolhido para a extração dos dados que formam a base para os experimentos compreende de 1 de julho de 2015 a 31 de julho de 2015, e são extraídos de forma aleatória, ou seja, sem distinção de região, cliente ou tipo de problema, procurando assim representar fielmente o banco de dados da empresa.

O único filtro aplicado na extração da base original sem atributos textuais é a eliminação de registros gerados pela interação de URA e atendimentos eletrônicos em que não existiu a interação entre cliente e atendente, pois esse tipo de registro não auxilia no processo de geração de conhecimento novo.

Foram selecionados nessa base informações de CRM de 10 mil clientes da empresa, o que resultou em um banco de dados com 56.970 registros. Dentre os atributos extraídos incluem-se o tipo da reclamação – informação, defeito, solicitação – a data de ocorrência do evento, o produto reclamado, a origem e o atendente, entre outros. Uma forma de enriquecer esses dados é inserir dados de cadastro dos clientes. Sendo assim foram incluídas informações tais como idade, sexo, ocupação, estado civil e entre outros.

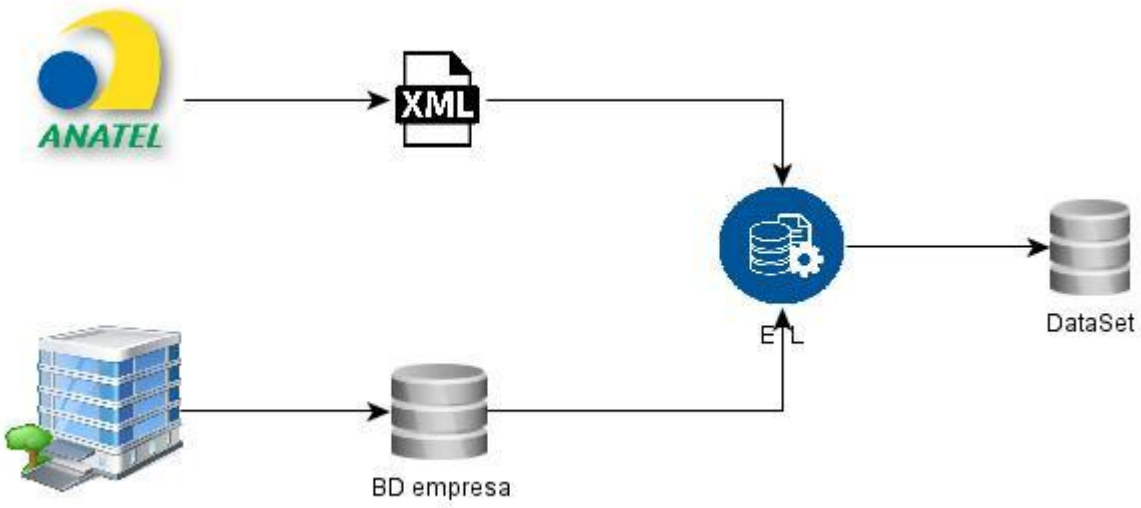


Figura 16 – Processo de criação da base de dados

A segunda base utilizada para a criação do conjunto de dados dos experimentos é denominada “base com atributos textuais”, pois ao conjunto de dados são adicionados novos atributos gerados pela Mineração de Textos. Em ambas as bases é necessário indicar em cada registro a classe correspondente. O órgão regulador disponibiliza diariamente para cada operadora de telecomunicação

do país, por meio do sistema Focus[1], uma base atualizada com as informações das reclamações recebidas pelo órgão e qual o status destas reclamações. Essa base contém informações como a identificação do cliente, produto reclamado, motivo da reclamação e prazo para atendimento, além do histórico da reclamação. Com isto é possível adicionar a classe (“migrou” e “não migrou”) a cada registro base, tornando-a adequada aos experimentos de classificação.

4.4.1. CRIAÇÃO DE ATRIBUTOS DERIVADOS

Os atributos que formam os conjuntos de dados para o uso em projetos de Mineração de Dados também podem ser criados à partir da etapa de pré-processamento, com a execução de algoritmos escritos com base no conhecimento adquirido. Esses atributos enriquecem o conjunto de dados original de tal forma que permite obter ganhos para as etapas seguintes do processo de mineração.

No capítulo 5 a seguir são apresentados os atributos derivados criados que são utilizados nos experimentos deste trabalho.

[1]https://sistemas.anatel.gov.br/sis/cadastrosimplificado/pages/acesso/login.xhtml

4.4.2. KDD APLICADO AO PROBLEMA

Para chegar à etapa de execução dos algoritmos de classificação é necessária a aplicação de diversas técnicas de pré-processamento de forma a permitir a construção de um conjunto de dados apto à execução dos algoritmos de Mineração de Dados. A aplicação desses métodos está alinhada conforme a proposta de Fayyad [Fayyad et al. 1996], e é apresentada à Figura 17 no contexto deste trabalho.

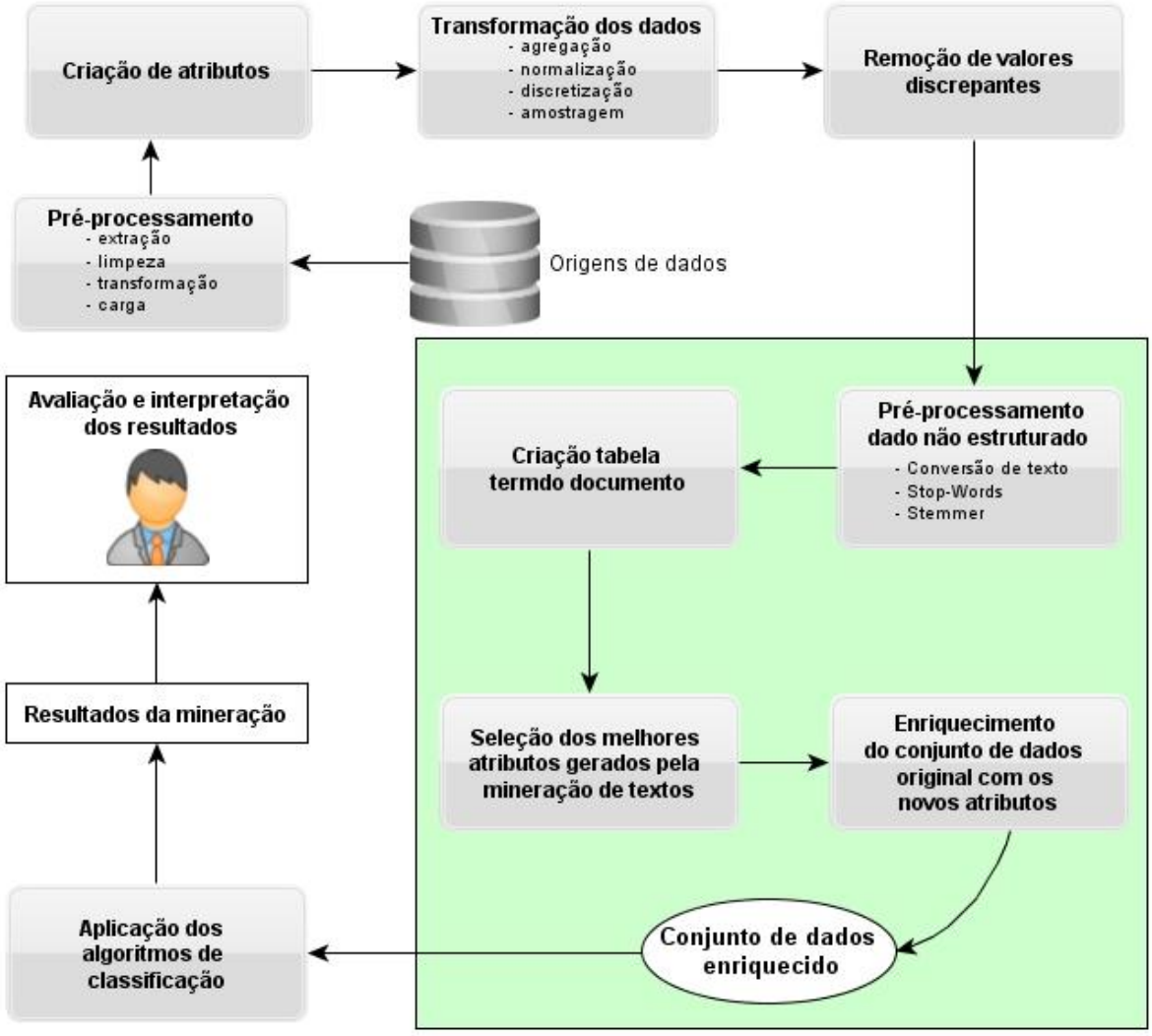


Figura 17 – Tarefas executadas para o desenvolvimento dos experimentos.

(1) Pré-processamento

Nessa etapa são realizadas atividades de extração, carga, limpeza e atualização dos dados, conforme os procedimentos tradicionais empregados em Mineração de Dados [Fayyad et al. 1996].

(2) Transformação dos dados

Na transformação dos dados são realizadas tarefas de agregação, normalização, discretização e amostragem dos dados, também seguindo os procedimentos tradicionais empregados em Mineração de Dados, como descrito em Han [Han et al. 2011].

(3) Remoção de Valores Discrepantes

Nessa etapa é verificada a necessidade de remoção de valores discrepantes (*outliers*). Para isso foi utilizado o cálculo da amplitude interquartil. Os limites superiores e inferiores são calculados e os valores fora destes limites são considerados discrepantes.

(4) Seleção do Conjunto de Dados

Com os dados previamente pré-processados é o momento de realizar a seleção do conjunto de dados para as próximas etapas do processo de mineração, de forma que o conjunto de dados selecionado possa representar o universo dos dados da aplicação.

(5) Criação de Atributos

O objetivo da criação de novos atributos é criar índicadores quantitativos que sejam simples e fáceis de interpretar, e que possam capturar informações importantes em um conjunto de forma mais eficiente do que os atributos originais. No capítulo 5 é apresentada a forma como são criados os novos atributos para o conjunto de dados em questão.

(6) Aplicação dos Algoritmos

Após os dados pré-processados e preparados deve-se aplicar os algoritmos de classificação sobre a base de dados. Nos experimentos realizados utilizou-se a técnica de validação cruzada, em que o conjunto de dados é divido em dez partes iguais, nove delas utilizadas para o treinamento e uma para formar o conjunto de testes.

Os experimentos foram realizados utilizando-se o ambiente WEKA (*Waikato Enviroment for Knowledge Analysis*), que possui diversos recursos que compreendem as etapas do KDD, ou seja, desde a seleção, normalização, discretização, execução dos algoritmos até a análise dos dados. Os algoritmos adotados para os experimentos neste trabalho são Árvores de Decisão, *Naïve*

Bayes, K-NN, *support vector machines* e redes neurais. Após a aplicação dos algoritmos os resultados são utilizados para avaliação e interpretação.

(7) Resultados da Mineração de Dados

Para avaliar os resultados dos algoritmos de classificação são utilizados recursos que mensuram a acurácia dos modelos obtidos, ou seja avaliam quanto o modelo inferido pelo classificador é adequado para aplicação em novas instâncias de dados sem que ocorra sobreajuste (*overfitting*) na validação dos resultados. Outro elemento que é empregado é o uso das matrizes de confusão, que permite a identificação detalhada dos resultados das classes alvos sobre a execução dos algoritmos de Mineração de Dados.

4.4.3. PRÉ-PROCESSAMENTO DE DADOS TEXTUAIS PARA O PROBLEMA EM QUESTÃO

Neste estudo as técnicas de Mineração de Textos são aplicadas aos dados não estruturados provenientes das reclamações efetuadas pelos clientes em ambientes de CRM. A aplicação dessas técnicas permite a formatação das informações de maneira estruturada, para que possam ser utilizadas no processo de descoberta de conhecimento.

Nos experimentos realizados efetuou-se o pré-processamento textual com auxílio da linguagem R[2]. Essa linguagem de programação possui um ambiente de desenvolvimento que permite a aplicação de comandos como um meio conveniente para a análise exploratória de dados. A linguagem R é utilizada neste trabalho para a aplicação das seguintes etapas típicas da Mineração de Textos:

- Transformação dos caracteres em *LowerCase*;
- Remoção de caracteres espúrios, pontuação e números;
- Remoção de *StopWords*;
- Obtenção de radicais (*Stemming*);
- Construção da matriz (termos x documentos).

No que se refere às palavras comuns (*stopwords*), utilizou-se no projeto uma lista com os principais termos que não geram conhecimento novo, obtido por outros trabalhos desenvolvidos com a Mineração de Textos. Para o item obtenção de

[2]https://www.r-project.org/

radicais foi utilizado o *Stemmer* desenvolvido pelo *Laboratory of Computational Intelligence*[3] (LABIC) da Universidade São Paulo (USP), e que permite a identificação da raiz das palavras removendo seus prefixos, sufixos e terminações para textos escritos em português. A matriz (termo x documentos) foi obtida de acordo com os procedimentos descritos no Capítulo 2, empregando diversos esquemas de ponderação.

Figura 18 - Nuvem de termos obtidos após o pré-processamento textual

Os textos elaborados pelos usuários e presentes na base de dados foram, portanto, submetidos à etapas de pré-processamento textual, como descrito anteriormente. O resultado gerado ao final desta etapa é a matriz de (termos x documentos) dos dados processados. A matriz gerada possui 2152 termos, obtidas das reclamações com textos livres envolvidas nesta pesquisa. Foram selecionados os termos com maior representação nesta matriz, que foi reduzida a 137 termos. Estes são os termos que passaram a compor o conjunto de dados enriquecido pela Mineração de Textos nos experimentos. A Figura 18 apresenta na forma de nuvem de palavras os termos obtidos ao final do processo de tratamento dos textos.

[3]http://www.nilc.icmc.usp.br/nilc/tools/stemmer.html

A execução dos algoritmos é efetuada sobre dois conjuntos de dados previamente preparados. Essa execução tem por objetivo mensurar se o conjunto de dados proposto, com a inclusão dos atributos textuais, obtêm resultados superiores ao conjunto de dados que utiliza apenas as entradas tradicionais. Essa comparação visa comprovar a eficácia da inclusão de técnicas de Mineração de Textos sobre dados não estruturados como as reclamações dos clientes.

Para isso são construídos dois conjuntos de dados. A base de dados original sem atributos textuais, que possui apenas os atributos referentes ao ciclo de vida do cliente dentro da empresa é constituída de 17 atributos. Neste conjunto de dados não foi aplicado qualquer técnica que faça o uso de informações referente ao texto gerado no atendimento deste cliente, ou seja, sem utilizar os dados obtidos por atendimentos que gerem informações textuais.

A segunda base é formada pelos mesmos atributos da base anterior, enriquecida com as informações referentes aos atributos textuais obtidos dos atendimentos recebidos no ambiente de CRM da empresa. A aplicação destas técnicas de processamento textual resultou no acréscimo de 137 atributos à base, de forma que o número final de atributos considerados passou a ser de 154.

O desempenho dos diferentes algoritmos de classificação sobre os conjuntos de dados também foi avaliado. No caso que está sendo considerado o conjunto de dados proposto possui grande quantidade de atributos provenientes da inclusão da Mineração de Textos. Sendo assim, é importante avaliar se os algoritmos sofrem algum tipo de perda de performance sobre conjuntos de dados com dimensionalidade maior, problema identificado no estado da arte como crucial para alguns projetos de Mineração de Dados.

5. EXPERIMENTOS REALIZADOS E ANÁLISE DOS RESULTADOS

Esse capítulo apresenta os experimentos que são efetuados sobre os dois conjuntos de dados apresentados anteriormente, visando a comprovação de que modelos de dados enriquecidos pelas informações textuais dos clientes possuem performance superior ao modelo que não utiliza tais informações. A acurácia dos algoritmos de classificação sobre os diferentes conjuntos de dados também é avaliada, a fim de identificar quais os algoritmos que melhor se adequam à tarefa em questão.

5.1. EXPERIMENTOS REALIZADOS

Durante o desenvolvimento desta pesquisa foi desenvolvido o artigo intitulado como “Classificação Automática das Reclamações de Clientes de uma Empresa de Telecomunicações” [de Oliveira Sanga et al. 2017]. Esse artigo foi apresentado na 8º edição do *Computer on the Beach* e apresentou a aplicação de diversos algoritmos de classificação sobre dois conjuntos de dados que são analisados e comparados sob diferentes formas de ponderação dos termos gerados pela Mineração de Textos.

5.1.1. ATRIBUTOS ORIGINAIS

Do conjunto de dados inicial que utiliza apenas as informações do ciclo de vida dos clientes e dados cadastrais foram extraídos 17 atributos, estes são provenientes da base pura da empresa e, do conhecimento adquirido sobre o negócio em questão, além das análises efetuadas, onde foram encontradas as principais características e fatos relevantes dos dados.

Para a criação do conjunto de dados utilizado nos experimentos é necessário identificar os clientes que migraram ou não para a Anatel. Isso é feito por meio do cruzamento com os dados da base original, ou seja, é comparado o documento (CPF ou CNPJ) de uma base com a fornecida pela agência reguladora para encontrar clientes que migraram ou não para o órgão.

Tabela 2 - Conjunto de atributos utilizados nos experimentos

Nome do Atributo	**Tipo**	**Origem**
Reclamações no CRM	Discreto	Base Original
Reclamações em 12 meses	Discreto	Base Original
Número de reclamações na Anatel	Discreto	Base Original
Tempo de instalação em meses	Discreto	Base Original
Linha de outra operadora	Discreto	Base Original
Idade	Discreto	Base Original
Sexo	Categórico	Base Original
Ocupação	Categórico	Base Original
Dependentes	Discreto	Base Original
Estado civil	Categórico	Base Original
Classe social	Categórico	Base Original
Tipo de residência	Categórico	Base Original
Escolaridade	Categórico	Base Original
Contexto das reclamações	Discreto	Base Original
Peso da reclamação mais critica	Discreto	Base Original
Soma do peso das reclamações	Discreto	Base Original
Cidade	Nominal	Base Original
Classe Alvo	Discreto	Base Anatel
Total		**17**

5.1.2. ATRIBUTOS DERIVADOS

Para esse estudo três novos atributos foram gerados a partir da execução de algoritmos escritos na linguagem Java com base no conhecimento adquirido na etapa de análise dos dados. Segue a descrição desses novos atributos:

- Número de reclamações na Anatel: esse é o mais simples dos atributos derivados, sua criação depende única e exclusivamente da soma da quantidade de reclamações que um cliente fez no período analisado. Durante a etapa de análise dos dados foi identificado forte relação entre clientes que migraram para a Anatel e a quantidade de reclamações dos mesmos, justificando a criação dessa nova entrada.
- Peso da reclamação mais crítica: esse atributo é criado a partir da identificação da reclamação mais crítica do cliente. Para isso é necessário realizar o mapeamento de todas as categorizações da reclamação – que pode ser definida pelo conjunto de quatro colunas em banco de dados que tipificam uma reclamação – disponíveis nos sistemas de CRM. A variável utilizada no algoritmo que faz a

identificação do peso da reclamação mais grave do cliente é iniciada com o valor 0. Em seguida executa-se uma busca nas reclamações dos clientes no período estudado de forma iterativa, onde cada reclamação é quantificada de acordo com uma tabela previamente mapeada que indica o peso de cada categorização. Caso este valor seja maior do que o valor armazenado na variável de apoio o valor é substituído. Ao finalizar o laço a variável armazena o valor da reclamação mais grave do cliente, e é inserida no conjunto de dados dos experimentos como um novo atributo.

- Soma do peso das reclamações: a última entrada gerada consiste na soma dos pesos das reclamações feitas por um usuário. O cálculo é efetuado em uma iteração similar à efetuada para o cálculo do atributo anterior. A criação desse novo atributo permite a identificação da severidade do conjunto de reclamações de um cliente. Clientes com a mesma quantidade de reclamações em ambientes de CRM podem ter valores diferentes para essa entrada, pois o teor da severidade das reclamações de um cliente pode ser menor do que outro. Esse tipo de abordagem permite apresentar ao classificador um atributo que apresenta pesos diferentes para clientes com a mesma quantidade de reclamações, mas com severidades diferentes. Por exemplo, clientes que tem reclamações relacionadas a informação tem um peso menor do que clientes que reclamam sobre defeitos.

```
Função SOMA-TIPO-RECL(listaReclamações) retorna Hash<Tipo, Integer>
Hash< Tipo, Integer > hash  =  new HashMap< Tipo, Integer>();
Se NÃO-VAZIA(listaReclamações)
    Para cada reclamação na listaReclamações
        Tipo tipo = new Tipo (reclamação.tipo1, reclamação.tipo2,
reclamação.tipo3, reclamação.tipo4);
        Inteiro total = 0;
Se (hash.existe(tipo)
total = hash.get(tipo);
end se;
        total++;
```

```
            hash.put(tipo, total);
    end cada;
      end se;
      retorna hash;
    end funcão;
```

Figura 19 – Cálculo iterativo do atributo derivado “Soma de reclamações”

5.1.3. ALGORITMOS DE CLASSIFICAÇÃO UTILIZADOS

Essa seção apresenta alguns dos parâmetros utilizados nos algoritmos de classificação empregados nos experimentos a fim de esclarecer quais as configurações utilizadas durante os experimentos.

- Para o algoritmo redes neurais o modelo adotado é o MLP *back-propagation,* que se baseia na retropropagação dos erros para realizar os ajustes de pesos das camadas intermediárias. Nos experimentos a rede neural utilizada usou taxa de aprendizado 0,3 com o número de épocas configurado em 500 e momentum 0,2, com 5 neurônios em 1 camada oculta.
- Para os experimentos com o algoritmo K-NN o valor de *k* foi configurado com 15 vizinhos mais próximos.
- Para as Máquinas de Vetores de Suporte o kernel utilizado foi o *PolyKernel* com expoente 1, pois apresentou os melhores resultados nos testes iniciais.

Conforme citado anteriormente, nessa pesquisa foi utilizado para a execução dos algoritmos o ambiente WEKA com os parâmetros originais (*default*) da aplicação. A ferramenta fornece meios que permitem a fácil configuração e aplicação dos algoritmos de Mineração de Dados. A execução de cada algoritmo em cada conjunto de dados acontece 10 vezes devido ao uso da técnica de validação cruzada com fator 10 apresentada na mesma seção 4.4.2, ou seja, acontecem 100 execuções dos algoritmos sobre cada um dos conjuntos de dados.

5.1.4. EXPERIMENTOS COM A BASE INICIAL

Os experimentos realizados sobre a base inicial, isto é, sem os atributos textuais, tem por objetivo estabelecer um valor inicial de acurácia para os diferentes algoritmos de classificação, além de indicar o procedimento mais adequado para a tarefa desejada. A Figura 20 apresenta a acurácia obtida para todos os algoritmos com esse conjunto de dados.

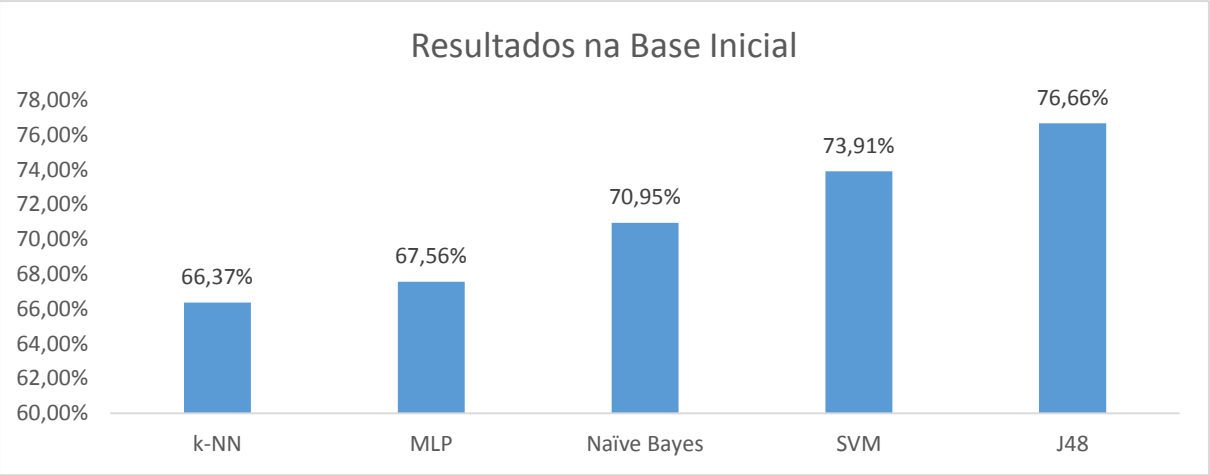


Figura 20 – Acurácia obtida nos experimentos com a base inicial

A seguir são apresentadas as matrizes de confusão dos algoritmos e suas medidas de precisão. A precisão é definida pela fórmula TP/FP (TP+FP), onde TP (verdadeiros positivos) é o número de casos classificados como verdadeiros e que realmente o são e FP (falsos positivos) é o número de casos indicados como verdadeiros, mas que na verdade são falsos.

A Tabela 3 apresenta a matriz de confusão obtida no experimento com a base inicial e o algoritmo árvore de decisão J48.

Tabela 3 – Matriz de confusão obtida pelo algoritmo árvore de decisão sob a base inicial

J48		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	3946	1054
	Migrou	1280	3720

Na Tabela 4 são destacadas três medidas de desempenho da árvore de decisão aplicada sobre o conjunto de dados inicial, permitindo a visualização de maiores detalhes dos resultados obtidos.

Tabela 4 - Desempenho árvore de decisão J48 sobre a base inicial

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
Não Migrou	0,789	0,256	0,755
Migrou	0,744	0,211	0,779
Média	0,767	0,233	0,767

SVM foi o algoritmo com a segunda melhor taxa de acurácia no modelo de dados inicial, obtendo o valor de 73,91%. Na Tabela 5 é apresentada a matriz de confusão e em seguida as principais medidas de desempenho do algoritmo.

Tabela 5 - Matriz de confusão obtida pelo algoritmo SVM sobre a base inicial

SVM		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	3774	1226
	Migrou	1383	3617

Na Tabela 6 são destacadas três medidas de desempenho de SVM aplicada sobre o conjunto de dados inicial, permitindo a visualização dos resultados obtidos.

Tabela 6 - Medidas de desempenho do algoritmo SVM sobre a base inicial

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
Não Migrou	0,755	0,277	0,732
Migrou	0,723	0,245	0,747
Média	0,739	0,261	0,739

A Tabela 7 apresenta a matriz de confusão do algoritmo Naïve Bayes, que obteve a terceira melhor taxa de acurácia com 70,95% de classificação correta.

Tabela 7 - Matriz de confusão obtida pelo algoritmo Naïve Bayes sobre a base inicial

Naïve Bayes		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	4503	497
	Migrou	2408	2592

A Tabela 8 apresenta três medidas de desempenho do algoritmo Naïve Bayes, permitindo a visualização dos resultados obtidos.

Tabela 8 - Medidas de desempenho do algoritmo Naïve Bayes sob a base inicial

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
Não Migrou	0,901	0,482	0,652
Migrou	0,518	0,099	0,839
Média	0,710	0,291	0,745

A Tabela 9 apresenta a matriz de confusão do algoritmo rede neural MLP, que obteve a quarta melhor taxa de acurácia com 67,56% de classificação correta.

Tabela 9 - Matriz de confusão obtida Rede Neural MLP sobre a base inicial

Rede neural MLP		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	2858	2142
	Migrou	1102	3898

A Tabela 10 apresenta três medidas de desempenho da rede neural MLP, permitindo a visualização dos resultados obtidos.

Tabela 10 - Medidas de desempenho da rede neural MLP sobre a base inicial

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
Não Migrou	0,572	0,220	0,722
Migrou	0,780	0,428	0,645
Média	0,676	0,324	0,684

A Tabela 11 apresenta a matriz de confusão do algoritmo K-NN, que obteve a quinta melhor taxa de acurácia com 66,37% de classificação correta.

Tabela 11 - Matriz de confusão obtida pelo algoritmo K-NN sobre a base inicial

K-NN		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	3371	1629
	Migrou	1734	3266

A Tabela 12 apresenta três medidas de desempenho do algoritmo K-NN, permitindo a visualização dos resultados obtidos.

Tabela 12 - Medidas de desempenho do algoritmo K-NN sobre a base inicial

Classe	Taxa de verdadeiro positivo	Taxa de falso positivo	Precisão
Não Migrou	0,674	0,347	0,660
Migrou	0,653	0,326	0,667
Média	0,664	0,336	0,664

5.1.5. EXPERIMENTOS COM A BASE QUE UTILIZA A MINERAÇÃO DE TEXTOS E FOI PONDERADA PELA FREQUÊNCIA DOS TERMOS

Nesta etapa dos experimentos os algoritmos são executados com o conjunto de dados enriquecido pela Mineração de Textos e onde se utiliza a ponderação dos termos dada pela frequência. Na Figura 21 são apresentados os resultados obtidos para os diferentes algoritmos de classificação nessa situação.

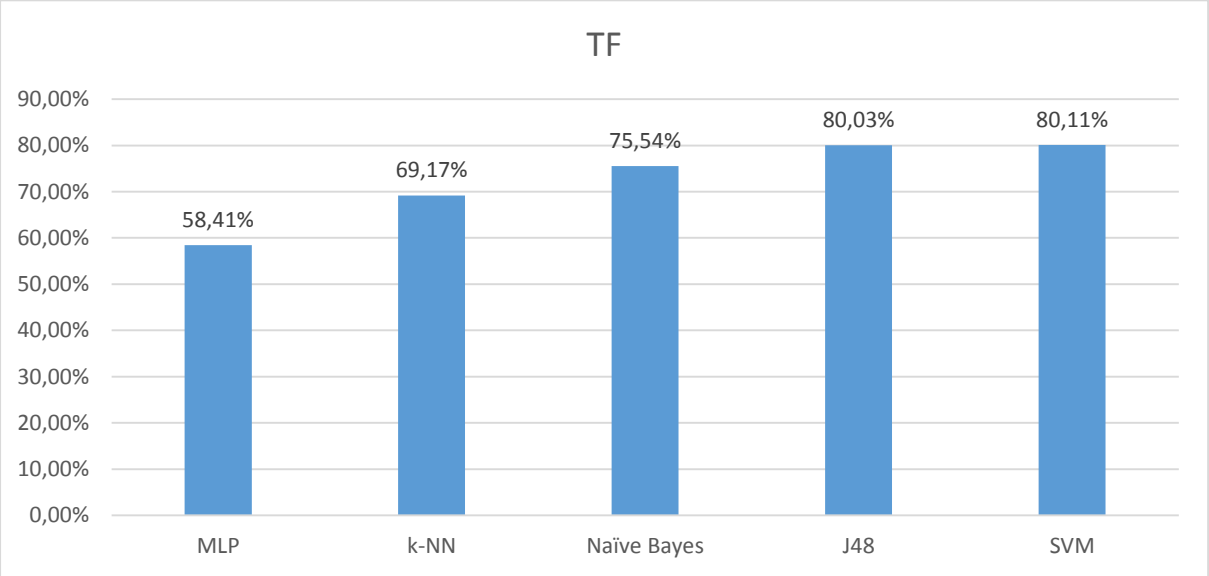


Figura 21 – Acurácia obtida na base com Mineração de Textos e ponderada pela Frequência dos Termos

A seguir são apresentadas as matrizes de confusão dos algoritmos e suas medidas de precisão.

A Tabela 13 apresenta a matriz de confusão do experimento com a base enriquecida e ponderada pela frequência dos termos obtida pelo algoritmo SVM.

Tabela 13 – Matriz de confusão obtida pelo algoritmo SVM na base com Mineração de Textos e ponderada pela Frequência dos Termos

SVM		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	4544	456
	Migrou	1533	3467

A Tabela 14 apresenta três medidas de desempenho do algoritmo SVM, permitindo a visualização dos resultados obtidos.

Tabela 14 - Medidas de desempenho do algoritmo SVM na base com Mineração de Textos e ponderada pela Frequência dos Termos

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
Não Migrou	0,909	0,307	0,748
Migrou	0,693	0,091	0,884
Média	0,801	0,199	0,816

A Tabela 15 apresenta a matriz de confusão do experimento com a base enriquecida e ponderada pela frequência dos termos obtida pelo algoritmo J48.

Tabela 15 - Matriz de confusão obtida pelo algoritmo Árvore de Decisão J48 na base com Mineração de Textos e ponderada pela Frequência dos Termos

J48		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	4311	689
	Migrou	1308	3692

A Tabela 16 apresenta três medidas de desempenho do algoritmo J48, permitindo a visualização dos resultados obtidos.

Tabela 16 - Medidas de desempenho do algoritmo j48 na base com Mineração de Textos e ponderada pela Frequência dos Termos

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
Não Migrou	0,862	0,262	0,767
Migrou	0,738	0,138	0,843
Média	0,800	0,200	0,805

A Tabela 17 apresenta a matriz de confusão do experimento com a base enriquecida e ponderada pela frequência dos termos obtida pelo Naïve Bayes.

Tabela 17 - Matriz de confusão obtida pelo algoritmo Naïve Bayes na base com Mineração de Textos e ponderada pela Frequência dos Termos

Naïve Bayes		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	4281	719
	Migrou	1727	3273

A Tabela 18 apresenta três medidas de desempenho do algoritmo Naïve Bayes, permitindo a visualização dos resultados obtidos.

Tabela 18 - Medidas de desempenho obtidas pelo algoritmo Naïve Bayes na base com Mineração de Textos e ponderada pela frequência dos termos

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
Não Migrou	0,856	0,345	0,713
Migrou	0,655	0,144	0,820
Média	0,755	0,245	0,766

A Tabela 19 apresenta a matriz de confusão do experimento com a base enriquecida e ponderada pela frequência dos termos obtida pelo algoritmo K-NN.

Tabela 19 - Matriz de confusão obtida pelo K-NN na base com Mineração de Textos e ponderada pela frequência dos termos

K-NN		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	4503	497
	Migrou	2586	2414

A Tabela 20 apresenta três medidas de desempenho do algoritmo K-NN, permitindo a visualização dos resultados obtidos.

Tabela 20 - Medidas de desempenho obtidas pelo algoritmo K-NN na base com Mineração de Textos e ponderada pela Frequência dos Termos

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
Não Migrou	0,901	0,517	0,635
Migrou	0,483	0,099	0,829
Média	0,692	0,308	0,732

A Tabela 21 apresenta a matriz de confusão do experimento com a base enriquecida e ponderada pela frequência dos termos obtida pelo algoritmo Redes Neurais MLP.

Tabela 21 - Matriz de confusão obtida pelo algoritmo Rede Neural MLP na base com Mineração de Textos e ponderada pela frequência dos termos

MLP		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	3342	1658
	Migrou	2501	2499

A Tabela 22 apresenta três medidas de desempenho da Rede Neural MLP, permitindo a visualização dos resultados obtidos.

Tabela 22 - Medidas de desempenho da Rede Neural MLP na base com Mineração de Textos e ponderada pela Frequência dos Termos

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
Não Migrou	0,668	0,500	0,572
Migrou	0,500	0,332	0,601
Média	0,584	0,416	0,587

5.1.6. EXPERIMENTOS COM A BASE QUE UTILIZA A MINERAÇÃO DE TEXTOS E FOI PONDERADA POR TF-IDF

Outro meio de avaliar a base que possui dados textuais é ponderando-a com outras métricas. Nessa seção são apresentados os resultados da execução dos algoritmos com a base enriquecida e ponderada pelo método *TF-IDF*.

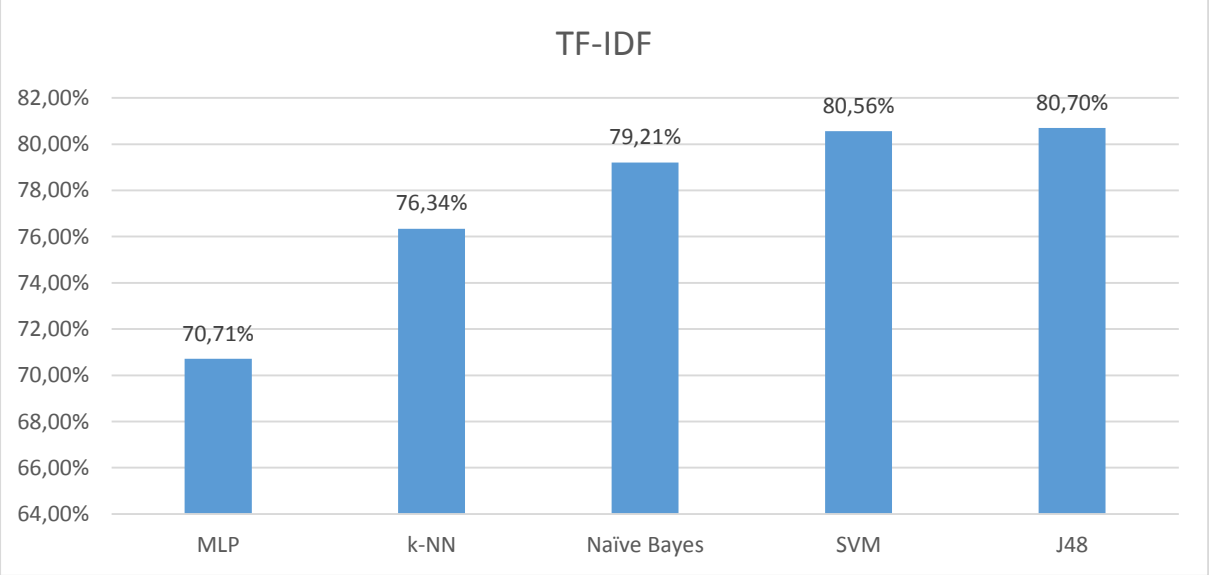


Figura 22 – Acurácia obtida na base enriquecida e ponderada por *TF-IDF*

A Tabela 23 apresenta a matriz de confusão do experimento com a base enriquecida e ponderada por *TF-IDF* obtida pelo algoritmo J48.

Tabela 23 - Matriz de confusão obtida pelo algoritmo Árvore de Decisão J48 na base com Mineração de Textos e ponderada por *TF-IDF*

J48		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	4356	644
	Migrou	1286	3714

A Tabela 24 apresenta três medidas de desempenho do algoritmo J48, permitindo a visualização dos resultados obtidos.

Tabela 24 - Medidas de desempenho obtidas pelo algoritmo J48 na base com Mineração de Textos e ponderada por *TF-IDF*

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
Não Migrou	0,871	0,257	0,772
Migrou	0,743	0,129	0,852
Média	0,807	0,193	0,812

A Tabela 25 apresenta a matriz de confusão do experimento com a base enriquecida e ponderada por *TF-IDF* obtida pelo algoritmo SVM.

Tabela 25 – Matriz de confusão obtida pelo algoritmo SVM na base com Mineração de Textos e ponderada por *TF-IDF*

SVM		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	4470	530
	Migrou	1414	3586

A Tabela 26 apresenta três medidas de desempenho do algoritmo SVM, permitindo a visualização dos resultados obtidos.

Tabela 26 - Medidas de desempenho obtidas pelo algoritmo SVM na base com Mineração de Textos e ponderada por *TF-IDF*

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
Não Migrou	0,894	0,283	0,760
Migrou	0,717	0,106	0,871
Média	0,806	0,194	0,815

A Tabela 27 apresenta a matriz de confusão do experimento com a base enriquecida e ponderada por *TF-IDF* obtida pelo algoritmo Naïve Bayes.

Tabela 27 - Matriz de confusão obtida pelo algoritmo Naïve Bayes na base com Mineração de Textos e ponderada por *TF-IDF*

Naïve Bayes		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	4250	750
	Migrou	1329	3671

A Tabela 28 apresenta três medidas de desempenho do algoritmo Naïve Bayes, permitindo a visualização dos resultados obtidos.

Tabela 28 - Medidas de desempenho obtidas pelo algoritmo Naïve Bayes na base com Mineração de Textos e ponderada por *TF-IDF*

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
Não Migrou	0,850	0,266	0,762
Migrou	0,734	0,150	0,830
Média	0,792	0,208	0,796

A Tabela 29 apresenta a matriz de confusão do experimento com a base enriquecida e ponderada por *TF-IDF* obtidas pelo algoritmo K-NN.

Tabela 29 - Matriz de confusão obtida pelo algoritmo K-NN na base com Mineração de Textos e ponderada por *TF-IDF*

K-NN		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	4669	331
	Migrou	2035	2965

A Tabela 30 apresenta três medidas de desempenho do algoritmo K-NN, permitindo a visualização dos resultados obtidos.

Tabela 30 - Medidas de desempenho obtidas pelo algoritmo K-NN na base com Mineração de Textos e ponderada por *TF-IDF*

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
Não Migrou	0,934	0,407	0,696
Migrou	0,593	0,066	0,900
Média	0,763	0,237	0,798

A Tabela 31 apresenta a matriz de confusão do experimento com a base enriquecida e ponderada por *TF-IDF* obtida pelo algoritmo redes neurais MLP.

Tabela 31 - Matriz de confusão obtida pela Rede Neural MLP na base com Mineração de Textos e ponderada por *TF-IDF*

MLP		**Classes previstas**	
		Não Migrou	**Migrou**
Classes corretas	**Não Migrou**	3957	1043
	Migrou	1886	3114

A Tabela 32 apresenta três medidas de desempenho da rede neural MLP, permitindo a visualização dos resultados obtidos.

Tabela 32 - Medidas de desempenho obtidas pela Rede Neural MLP na base enriquecida e ponderada por *TF-IDF*

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
Não Migrou	0,791	0,377	0,677
Migrou	0,623	0,209	0,749
Média	0,707	0,293	0,713

Para finalizar a Tabela 33 apresenta um resumo das precisões obtidas nos experimentos pelos diversos algoritmos nas diferentes bases: (1) inicial, (2) enriquecida pela Mineração de Textos e ponderada por TF e (3) enriquecida pela Mineração de Textos e ponderada por TF-IDF.

Tabela 33 - Tabela comparativa das precisões médias dos resultados

	Rede Neural	K-NN	Naïve Bayes	SVM	J48
Base sem Mineração de Textos	0,684	0,664	0,745	0,739	0,767
Base com Mineração de Textos e Ponderada pela Frequência dos Termos	0,587	0,732	0,766	0,816	0,805
Base com Mineração de Textos e Ponderada por TF-IDF	0,713	0,798	0,796	0,815	0,812

5.2. ANÁLISE DOS RESULTADOS

Inicialmente a análise é feita pela verificação dos melhores resultados para cada conjunto de dados, à fim de observar qual algoritmo obteve o melhor desempenho em cada conjunto de dados. Paralelamente se busca identificar se a

inclusão da Mineração de Textos para o problema em questão gera conjuntos de dados melhores para os algoritmos de classificação. Os algoritmos com os melhores resultados nos experimentos foram a Árvore de Decisão J48 e às Máquinas de Vetores de Suporte, ambos alcançaram acurácias entre 75 e 80% em todos os conjuntos de dados.

Analisando a Figura 23 observa-se que o conjunto de dados que apresentou os melhores resultados para os experimentos utilizou TF-IDF para a ponderação dos novos atributos gerados pela Mineração de Textos. Em seguida, o conjunto de dados que utiliza a Mineração de Textos e ponderado pela Frequência dos Termos, e finalmente a base de dados inicial, ou seja, a base que não foi enriquecida com a inclusão de atributos gerados pela Mineração de Textos obteve os piores resultados. Portanto, a inclusão de técnicas de Mineração de Textos e a incorporação dos termos na forma de novos atributos sobre o problema em questão gerou melhores condições de mineração, e permitiu aos algoritmos a obtenção de melhores taxas de acurácia.

Na execução dos algoritmos sobre o conjunto de dados inicial, sem inclusão de novos atributos gerados pela Mineração de Textos, o algoritmo que apresentou maior taxa de acurácia foi a Árvore de Decisão J48 com 76,66%, classificando corretamente 7666 dos 10 mil registros. Os resultados para os demais algoritmos por ordem da taxa de acurácia são: SVM com 73,91%, *Naïve Bayes* com 70,95%, redes neurais MLP com 67,56% e K-NN com 66,37%.

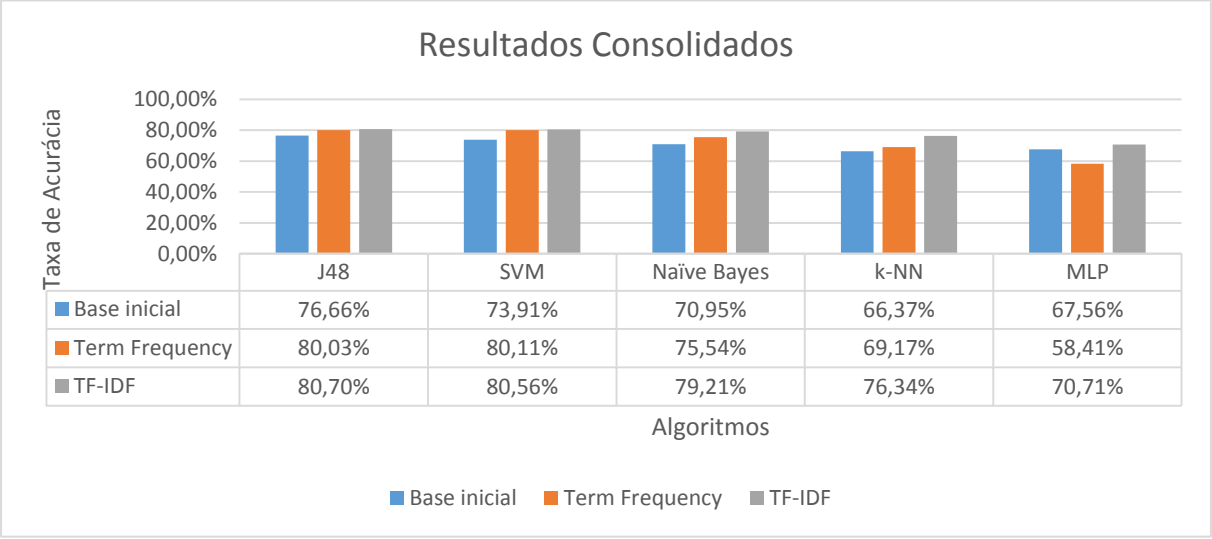


Figura 23 - Resultados consolidados dos experimentos

Na comparação geral dos algoritmos sobre esse conjunto de dados, redes neurais MLP foi o algoritmo que gerou mais verdadeiros positivos do que falsos positivos, porém a quantidade de falsos negativos é muito alta em relação a quantidade de verdadeiros negativos. Na média geral os melhores resultados foram obtidos pela Árvore de Decisão J48, que obteve 78,9% na taxa de verdadeiro negativo e 74,4% na taxa de verdadeiro positivo na base inicial.

Na base ponderada pela frequência dos termos, o algoritmo com melhor desempenho é o SVM com 80,11% de taxa de acurácia, isto é, o algoritmo classificou corretamente 8011 das 10 mil instâncias do conjunto de dados. Em seguida a Árvore de Decisão J48 ficou com 80,03%, a terceira melhor taxa ficou com o algoritmo Naïve Bayes que obteve 75,54% e finalmente K-NN com 69,17% e redes neurais MLP com 58,41% de acurácia.

Na comparação geral dos algoritmos sobre esse conjunto de dados, a Árvore de Decisão J48 foi o algoritmo que gerou mais verdadeiros positivos do que falsos positivos. No entanto em geral SVM é o melhor algoritmo sobre esse conjunto de dados, pois existe um equilíbrio maior entre o número verdadeiros positivos e verdadeiros negativos, e a quantidade de acertos se equilibra, evitando o *overfitting*.

Na base ponderada por *TF-IDF* o algoritmo com o melhor desempenho foi a Árvore de Decisão J48 com 80,70% de taxa de acurácia, classificando corretamente 8070 das 10 mil instâncias. Os demais algoritmos obtiveram os seguintes resultados: o segundo melhor desempenho foi o do algoritmo SVM com 80,56%, o Naïve Bayes ficou com 79,21% obtendo o terceiro melhor desempenho, e na sequência K-NN com 76,34% e MLP com 70,71% de taxa de acurácia.

Na comparação geral dos resultados essa configuração foi onde os algoritmos obtiveram os melhores resultados, pois todos conseguiram uma taxa de acurácia acima dos 70%. Esse conjunto de dados pode ser considerado a melhor configuração para esse tipo de problema devido aos resultados obtidos. Árvore de decisão J48 e SVM ficaram com resultados muito próximos, porém árvore de decisão J48 obteve 74,28% de verdadeiros positivos contra 71,72% de verdadeiros positivos de SVM, confirmando dessa forma sua superioridade sobre os demais algoritmos nesse conjunto de dados.

Os resultados obtidos comprovam a hipótese deste trabalho: a inclusão dos atributos textuais obtidos a partir do texto livre produzido pelo usuário em sua

reclamação com a aplicação de técnicas de Mineração de Textos permite aumentar a acurácia obtida pelos classificadores, proporcionando ganhos nos resultados.

6. CONCLUSÕES E TRABALHOS FUTUROS

6.1. CONCLUSÕES

Nesse trabalho destacou-se que o modo como clientes interagem com as empresas tem mudado devido a propagação de novos meio de comunicação. Estes meios permitem o acesso as empresas por meio de chats, aplicativos, redes sociais, SMS´s ou mesmo até e-mails. Essa nova característica permite que as empresas façam o uso destas informações para extraírem conhecimento novo a partir dos dados gerados. Considerando este novo cenário de relacionamento entre as empresas e seus clientes, este trabalho buscou comprovar que o uso de Mineração de Textos permite que se obtenham melhores resultados nas tarefas de descoberta de informação.

Para comprovar tal hipótese, a área escolhida para o experimento foi a de telecomunicações, área que já utiliza os benefícios da Mineração de Dados para diferentes contextos de aplicações. A aplicação utilizada neste trabalho busca classificar a severidade das reclamações recebidas por uma empresa de telecomunicações, identificando os clientes que saem do ambiente interno de atendimento e migram para órgãos de defesa do consumidor.

A abordagem empregada no estudo visou o enriquecimento do conjunto de dados original por meio da Mineração de Textos onde foram criados novos atributos com base nas reclamações efetuadas na forma de texto livre por clientes em ambientes de CRM. Após a aplicação das técnicas de Mineração de Textos sobre as reclamações dos clientes os termos mais citados são identificados e passam a compor o conjunto de dados original onde são aplicados os algoritmos de classificação. Dessa forma, o conjunto de dados original passou de 16 para 154 atributos, proporcionando uma base que inclui informações referentes aos textos livres informados no atendimento do cliente.

Na comparação dos resultados obtidos após os experimentos executados com diferentes algoritmos e conjuntos de dados, observou-se que o enriquecimento do conjunto de dados original por meio de técnicas de Mineração de Textos com as informações extraídas das reclamações dos clientes em ambientes de CRM é adequado. Os resultados obtidos se mostraram superiores aos obtidos com o

conjunto de dados original, gerando modelos de classificação com taxas de acurácia superiores e comprovando assim a eficácia da proposta sugerida nesse trabalho.

Nos resultados obtidos com a execução dos experimentos ficou claro que os algoritmos Máquina de Vetores de Suporte e Árvores de Decisão J48 obtiveram os melhores resultados (80,56% e 80,70%) respectivamente sendo a melhor solução para a classificação dos dados para o problema em questão. Além disto, quando os novos atributos gerados pela Mineração de Textos foram ponderados pela técnica *TF-IDF* todos os algoritmos de classificação obtiveram taxas de acurácias superiores aos 70% o que faz desse conjunto de dados o de melhor desempenho nos experimentos. Também foi com esse conjunto de dados que foram obtidas as melhores taxas de acurácia, sendo que a Árvore de Decisão J48 alcançou 80,70% de acurácia e o *SVM* obteve 80,56%.

Com a abordagem proposta nesse trabalho, a introdução da Mineração de Textos possibilitou o desenvolvimento de um modelo de classificação que permite a extração de conhecimento novo a partir dos dados fornecidos via aplicações desenvolvidas para as novas formas de interação entre clientes e empresas. Portanto, esta pesquisa constata que é possível o desenvolvimento de aplicações com maior potencial de assertividade em modelos de classificação. Esse ganho garante aos tomadores de decisão a disponibilidade dos conhecimentos adequados para que sejam tomadas decisões confiáveis em seus respectivos negócios.

6.2. TRABALHOS FUTUROS

Este trabalho apresentou um abordagem computacional que está alinhada com os novos meios de interação entre clientes e empresas. Contudo, para a agregação de conhecimento novo e para o fornecimento de informações mais confiáveis e que possam gerar melhores informações, é possível explorar outros meios que podem compor os dados nesta pesquisa.

Outro ponto pode ser destacado é que tanto o comportamento social dos clientes de grandes empresas quanto os meios tecnológicos estão em constante evolução. Essas características devem provocar contínuas mudanças na forma como as informações são disponibilizadas para a utilização em ferramentas de Mineração de Dados. É importante ficar atento para que estas mudanças sejam seguidas adequadamente.

Outra forma de ampliação do escopo desta pesquisa é o desenvolvimento de conjuntos de dados que façam o uso de informações de redes sociais (Facebook, Twitter e outros). O uso dessas informações deve permitir o enriquecimento dos conjuntos de dados de tal forma que seja possível a inclusão de novos atributos que possam melhorar os resultados já alcançados. Além disso, as atuais ferramentas das empresas podem ser adaptadas de forma que os dados disponibilizados para a Mineração de Dados estejam em maior conformidade com estas aplicações, reduzindo desta forma os esforços da etapa de pré-processamento dos dados.

Do ponto de vista computacional e das técnicas aplicadas nos experimentos, existe a possibilidade do desenvolvimento e aplicação de novas formas de ponderação dos termos gerados pela aplicação de técnicas de Mineração de Textos. Nos experimentos apresentados nesse trabalho foi possível identificar que diferentes formas de ponderação resultam em modelos de classificação com resultados distintos, justificando e motivando dessa forma o desenvolvimento de novas formas de ponderação dos termos, visando a obtenção de melhores resultados para o problema em questão.

Por fim, outras técnicas que podem ser utilizas na fase de pré-processamento dos dados são as abordagens de seleção de atributos por meio de procedimentos como o *filter* e *wrapper* [Fayyad et al. 1996][Kohavi et al. 1997]. Estas técnicas permitem a identificação dos melhores atributos para um conjunto de dados. Para esse tipo pesquisa essa é uma abordagem interessante, pois contribui para a redução de dimensionalidade, problema conhecido em Mineração de Dados e que devido ao grande número de atributos gerados pela Mineração de Textos está presente nestas aplicações.

Referências Bibliográficas

Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., Ghatasheh, N. (2014). Predicting customer churn in telecom industry using multilayer perceptron neural networks: Modeling and analysis. *Life Science Journal*, 11(3), 75-81.

Ahn, H., Ahn, J. J., Oh, K. J., Kim, D. H. (2011). Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques. *Expert System with Applications*, 38(5), 5005-5012

Almana A. M., Aksoy M. S., Alzahrani R. (2014). A Survey On Data Mining Techniques In Customer Churn Analysis For Telecom Industry. *Journal of Engineering Research and Applications,* (4):165-171.

Antunes, C. M., & Oliveira, A. L. (2001, August). Temporal data mining: An overview. In *KDD workshop on temporal data mining* (Vol. 1, p. 13).

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: from concept to implementation.* Prentice-Hall, Inc.

Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. *Modern information retrieval*. Vol. 463. New York: ACM press, 1999.

Chang, C. W., Lin, C. T., Wang, L. Q. (2009). Mining the text information to optimizing the customer relationship management. *Expert Systems with applications*, 36(2), 1433-1443.

de Oliveira Sanga, D. A., & Kaestner, C. A. A. (2017). Classificação Automática das Reclamações de Clientes de uma Empresa de Telecomunicações. *Anais do Computer on the Beach*, 230-238.

Deulkar, Miss Deepa S., and R. R. Deshmukh. Data Mining Classification. *Imperial Journal of Interdisciplinary Research* 2.4 (2016).

Dogan, N., & Tanrikulu, Z. (2013). A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness. *Information Technology and Management*, 14(2), 105-124.

Fayyad, U., Piatetsky-Shapiro, G., e Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine,* 17(3):37.

Femina, B. T., Sudheep, E. M. (2015). An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour. *Procedia Computer Science*, (46):725-731.

Gerhardt, T. E., & Silveira, D. T. (2009). *Métodos de pesquisa*. Plageder.

GOLDSCHMIDT, R., & PASSOS, E. (2005). Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações. *Rio de Janeiro: Campus*.

Gupta, V., Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, 1(1), 60-76.

Hadden, J., Tiwari, A. Roy, R., Ruta, D. (2006). Churn Prediction using complaints data. In *Proceedings Of World Academy Of Science*, Engineering and Technology.

Hotho, A., Nürnberger, A., & Paaß, G. (2005, May). A Brief Survey of Text Mining. In Ldv Forum (Vol. 20, No. 1, pp. 19-62).

Huang, Y., Huang, B. Q., & Kechadi, M. T. (2010). A new filter feature selection approach for customer churn prediction in telecommunications. *Industrial Engineering and Engineering Management (IEEM), 2010 IEEE International Conference on (pp. 338-342).* IEEE.

Hung, S. Y., Yen D. C., Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, (31):515-524.

Kaur, R., Aggarwal, S. (2013). Techniques for mining text documents. *International Journal of Computer Applications,* 66(18).

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, *97*(1-2), 273-324.

Lejeune, M. A. (2001). Measuring the impact of data mining on churn management. *Internet Research*, *11*(5), 375-387.

Lin, W. C., Tsai, C. F., Ke, S. W. (2014). Dimensionality and data reduction in telecom churn prediction. *Kybernetes*, (43)5, 737-749.

Maimon, O., & Rokach, L. (2010). *Data mining and Knowledge discovery handbook*. New York: Springer.

Piatetsky-Shapiro, G., Matheus, C., Smyth, P., & Uthurusamy, R. (1994). Kdd-93: Progress and challenges in knowledge discovery in databases. *AI magazine*, *15*(3), 77.

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). Machine learning, neural and statistical classification.

Ngai, E. W., Xiu, L., Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert system with applications*, 36(2), 2592-2602

Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning.*

Rezende, S. O., Marcacini, R. M., & Moura, M. F. (2011). O uso da Mineração de Textos para extração e organização não supervisionada de conhecimento. *Revista de Sistemas de Informação da FSMA*, *7*, 7-21.

Rygielski, C., Wang, J. C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in society*, *24*(4), 483-502.

Seifert, J. W. (2004). Data mining: An overview. *National security issues*, 201-217.

Pallotta, V., Delmonte, R., Vrieling, L., Walker, D. (2013). Interaction Mining: The New Frontier of Customer Interaction Analytics. *New Challenges in Distributed Information*, (439):91-111.

Tan, A. H. (1999, April). Text mining: The state of the art and the challenges. In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases (Vol. 8, pp. 65-70).

Tan, P. N., Steinback, M., Kumar, V. (2009). *Introdução ao datamining: Mineração de Dados.* Ciência Moderna.

Zaman, F., Hogan, G., Der Meer, S., Keeney, J., Robitzsch, S., Muntean, G. M. (2015). A recommender system architecture for predictive telecom network management. *Communications Magazine*, *IEEE,* 53(1), 286-293.

Zhang, N., & Lu, W. F. (2007, June). An Efficient Data Preprocessing Method for Mining Customer Survey Data. In *Industrial Informatics, 2007 5th IEEE International Conference on* (Vol. 1, pp. 573-578). IEEE.

Wazlawick, Raul. *Metodologia de pesquisa para ciência da computação, 2a edição*. Vol. 2. Elsevier Brasil, 2014.

Weiss, G. M. (2005). Data Mining in Telecommunications. *Data Mining and Knowledge Discovery Handbook*, pages 1189-1201. Springer.

Witten, I. H., Moffat, A., Bell, T. C. (1999). *Managing gigabytes: compressing and indexing documents and images.* Morgan Kaufmann.

Wu, S., Kang, N., Yang, L. (2014). Fraudulent Behavior Forecast in Telecom Industry Based on Data Mining Technology. *Communications of the IIMA*, 7(4), 1.

Ye, L., Qiu-ru, C., Hai-xu, X., Yi-jun, L. Zhi-min, Y. (2012). Telecom customer segmentation with K-means clustering. *In Computer Science & Education (ICCSE), 2012 7th International Conference on* (pp. 648-651). IEEE.