UNIVERSIDADE REGIONAL DE BLUMENAU

CENTRO DE CIÊNCIAS EXATAS E NATURAIS

CURSO DE CIÊNCIA DA COMPUTAÇÃO – BACHARELADO

FERRAMENTA PARA PREDIÇÃO DE DADOS PROEMINENTES DE SISTEMAS RH

RAFAEL SEMANN

BLUMENAU
2015

2015/2-19

RAFAEL SEMANN

FERRAMENTA PARA PREDIÇÃO DE DADOS PROEMINENTES DE SISTEMAS RH

Trabalho de Conclusão de Curso apresentado ao curso de graduação em Ciência da Computação do Centro de Ciências Exatas e Naturais da Universidade Regional de Blumenau como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação.

Prof. Cláudio Ratke, Mestre – Orientador

BLUMENAU
2015

2015/2-19

FERRAMENTA PARA PREDIÇÃO DE DADOS PROEMINENTES DE SISTEMAS RH

Por

RAFAEL SEMANN

Trabalho de Conclusão de Curso aprovado para obtenção dos créditos na disciplina de Trabalho de Conclusão de Curso II pela banca examinadora formada por:

Presidente:	Prof. Cláudio Ratke Orientador, Mestre – Orientador, FURB
Membro:	Prof. Everaldo Grahl, Mestre – FURB
Membro:	Prof. Roberto Heinzle, Doutor – FURB

Blumenau, 04 de dezembro de 2015

AGRADECIMENTOS

Aos meus pais Márcio e Simone por todo o apoio que me deram durante está longa jornada, nunca me deixando desistir.

A minha namorada Jéssica, por todo o suporte, ajuda e compressão pelo tempo que estive ausente em função deste trabalho.

Ao meu orientador Cláudio Ratke, por ter aceitado a missão de me orientar e por todo o engajamento demonstrado ao longo do trabalho.

Um agradecimento especial para todas as pessoas que disponibilizaram seu tempo para responder a avaliação referente a ferramenta.

Por último, mas de forma alguma menos importante, agradecimento a empresa Benner Sistemas por ter fornecido a base de dados de RH, para que fosse possível realizar a mineração de dados pela ferramenta.

I don’t want to believe. I want to know.

Carl Sagan

RESUMO

Este trabalho apresenta o desenvolvimento de uma ferramenta *World Wide Web* (Web) para mineração de dados sobre sistema Recursos Humanos (RH), que tem por objetivo detectar padrões na rotatividade de pessoal de um sistema especifico de RH. De forma mais detalhada objetiva-se a identificar grupos que possuem tendência a serem demitidos ou se demitirem, modificar o algoritmo C4.5 para permitir atributos providos de mineração de texto, demonstrar a consulta de predição sobre árvore de decisão e avaliar a importância e legitimidade das informações geradas para o RH. Para realizar a mineração de dados foi utilizado o método de árvore de decisão através do algoritmo C4.5, onde foi introduzido a mineração de texto através de um atributo do mesmo tipo. Por meio da árvore de decisão gerada, foi utilizado o *Predictive Model Markup Language* (PMML) para realizar consulta sobre os dados demissionais. Ao final demonstrando que a ferramenta é útil para auxiliar na tomada de decisão referente a rotatividade de pessoal.

Palavras-chave: Mineração de dados. Recursos humanos. Predição. Tomada de decisão.

ABSTRACT

This work shows the development of a World Wide Web (Web) tool to data mining over a Human Resources (HR), which aims to detect patterns in the turnover of a specific HR. In more detail aim to identify groups that have tendency of been fired or to resign, modify the C4.5 algorithm to allow attributes provides by the text mining, demonstrate a forecasting search on the decision tree, and measure the importance and legitimacy of the generated information to the HR. For the data mining was used the decision tree method over the C4.5 algorithm, in this algorithm was introduced the text mining through an attribute of the same type. Throughout the generated decision tree was used the Predicted Model Markup Language (PMML) to do a search in the dismissal data. Finally demonstrating that the tool is useful to help in the decision-making relative to turnover.

Key-words: Data mining. Human resources. Forecasting. Decision-making.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

PMML - *Predictive Model Markup Language*

RH – Recursos Humanos

SI – Sistemas de Informação

SQL - Structured Query Language

TI – Tecnologia da Informação

WEB – *World Wide Web*

XML - *eXtensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Os Sistemas de Informação (SI) passaram por uma grande evolução ao longo dos anos, de meros sistemas de controle transacionais se tornando ferramentas indispensáveis no crescimento e desenvolvimento de empresas. De acordo com Quintella e Soares Junior (2009), com o avanço tecnológico, a difusão dos computadores foi aumentando, bem como sua capacidade de coleta e armazenamento de dados. Essa difusão citada pelos autores propiciou a evolução e aperfeiçoamento dos SI.

Atualmente, no Brasil, os SI compõem um mercado que movimenta um grande volume de investimentos, ultrapassando o montante de 12 bilhões de dólares, somente em 2014, com crescimento de 12,6 % com relação ao ano anterior, dados da Associação Brasileira das Empresas de Software (2015). Com tanto potencial, ainda existe pouca exploração dos dados contidos em um sistema, visando auxiliar o usuário nas tomadas de decisão. A afirmação de Quintella e Soares Junior (2009) justifica tal fato, pois, de acordo com os autores, não houve aproveitamento total da capacidade de utilização dos dados. Sendo assim, a necessidade por sistemas que façam melhor uso dessas informações torna-se importante.

Os sistemas de RH são um exemplo de sistemas que possuem uma grande quantidade de informações armazenadas com potencial para exploração, que, contudo, são pouco utilizadas. Verificando os números dos sistemas de RH se percebe que esse é um segmento na qual se deve investir para um avanço dos mesmos. Somente em 2013 no Brasil, o mercado de sistemas de RH obteve uma receita de mais de 518 milhões de reais, dados da Série estudos.

Sistemas de RH podem ser definidos como sistemas integrados usados para coletar, armazenar e analisar informações sobre recursos humanos de uma organização (HENDRICKSON, 2003). Atualmente os sistemas de RH apresentam uma grande quantidade de informações, porém estas são pouco utilizadas pelos gestores para auxiliar na tomada de decisões, assim, a decisão realizada possui baixa qualidade e precisão, isto porque, segundo Chien e Chen (2008), a aplicação de mineração de dados não tem atraído muita atenção das pessoas no campo de Recursos Humanos.

Por outro lado, a mineração de dados através de suas técnicas que se propõem a explorar e encontrar padrões dentro de grandes quantidades de informações, poderia auxiliar no processo de tomada de decisão. Isto porque, “A mineração de dados é o processo de descoberta automática de informações úteis em grandes depósitos de dados.” (TAN; STEINBACH; KUMAR, 2009, p.3). Portanto, o uso da mineração de dados torna possível

encontrar relacionamentos entre os dados de um sistema de RH, bem como criar modelos de previsões para mostrar as saídas voluntárias, por exemplo.

Para realizar a descoberta e processamento das informações dentro da mineração de dados existe uma série de técnicas que podem ser utilizadas. Dentro destas variadas técnicas, será utilizada a técnica de árvore geradora, pois de acordo com Maimon e Rokach (2007) a mesma visa encontrar a melhorar estratégia para alcançar o objetivo desejado, ou seja, privilegia os dados mais relevantes em níveis superiores.

Dentro de uma base de dados de um sistema RH as informações podem encontrar-se nos mais variados formatos. Sendo assim, existe uma grande quantidade de informações que devem ser introduzidas na árvore de decisão, tais como, idade do funcionário, escolaridade, entre outros. A utilização do algoritmo C4.5 possibilita que o dado inserido na árvore seja classificado de acordo com sua importância, com relação ao objetivo desejado. Para que essa classificação seja possível. Quinlan (1993), sugeriu a utilização da entropia, sendo esta uma medida que indica o grau de aleatoriedade do atributo, permitindo assim que os dados mais consistentes fiquem nos níveis superiores quando na indução da árvore.

Contudo o algoritmo C4.5 não consegue utilizar um atributo de texto descritivo para a indução da árvore geradora, uma vez que dentro deste existem várias informações dispersas. Os atributos com textos descritivos, em geral nos sistemas de RH, armazenam entrevistas demissionais, avaliações de desempenho, entre outros, informações essas que são muito relevantes no contexto. Para evitar que este dado não seja desprezado como atributo, a mineração de texto é utilizada antes da indução da árvore.

“O objetivo principal da mineração de texto é a análise e descobrimento de padrões interessantes, incluindo tendências e valores discrepantes [...]” (AGGARWAL; ZHAI, 2012, p.2). Para realizar essa busca de tendências e categorização de palavras dentro de textos, devem ser realizadas técnicas como a redução adverbial, retirada de sufixos, singularização das palavras, entre outros. Além disto, um passo muito importante é a criação de uma tabela de sinônimos, visando assim agrupar um conjunto de palavras que possuem significado semelhante. Por exemplo, as palavras “gestor”, “líder”, “gerente” podem ser generalizadas para “chefe”, assim gerando uma melhor precisão quando utilizado na indução da árvore geradora.

Dentro desse contexto, será desenvolvida uma ferramenta que utilizará a mineração de dados, através da técnica de indução de árvore de decisão e com isso encontrar padrões nos dados provenientes de um sistema de RH. Esta ferramenta também fará uso do algoritmo C4.5 para realizar a indução da árvore geradora, permitindo assim o fornecimento de um melhor

agrupamento dos dados, bem como o uso da mineração de texto para permitir que os textos de entrevistas demissionais possam ser aproveitados como atributos na árvore geradora. Todas essas técnicas em conjunto serão aplicadas com o objetivo de mostrar os dados encontrados na melhor forma possível para auxiliar na tomada de decisão.

1.1 OBJETIVOS

O objetivo deste trabalho é desenvolver uma ferramenta para detectar padrões na rotatividade de pessoal de um sistema especifico de RH.

Os objetivos específicos são:

a) identificar grupos que possuem tendência a serem demitidos ou se demitirem;
b) modificar o algoritmo C4.5 para permitir atributos providos de mineração de texto;
c) demonstrar a consulta de predição sobre árvore de decisão;
d) avaliar a importância e legitimidade das informações geradas para o RH.

1.2 ESTRUTURA

Este trabalho é dividido em quatro capítulos. No primeiro capitulo é descrita as justificativas para o desenvolvimento deste trabalho, bem como seus objetivos e sua estrutura.

O segundo capítulo abrange a fundamentação teórica, na qual apresenta a evolução dos sistemas de informação, a história dos sistemas de RH, além disto, explica conceitos gerais sobre a mineração de dados e mineração de texto que devem ser considerados para o desenvolvimento da aplicação. Ao final, sendo apresentados os trabalhos correlatos.

O terceiro capítulo descreve o desenvolvimento da aplicação, no qual são listados os requisitos, bem como sua especificação através dos diagramas de caso de uso, atividade, sequência e classe. Além disto, é descrito a implementação, técnicas e ferramentas utilizadas, operacionalidade e ao final é apresentado a discussão dos resultados obtidos.

No quarto capitulo são descritas as conclusões e as extensões sugeridas para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são abordados os aspectos teóricos relacionados ao trabalho. Inicialmente conceituando sobre Sistemas de Informação, depois, apresenta-se uma fundamentação sobre Sistema RH e sua evolução histórica. Em seguida, é apresentado a mineração de dados e as estruturas que compõem o mesmo, após isto, é realizada uma abordagem sobre *forecasting*. Ao final, são apresentados os trabalhos correlatos.

2.1 SISTEMAS DE INFORMAÇÃO

Um Sistema de informação é definido por O'Brien (2004) como "[...] um conjunto organizado de pessoas, hardware, software, redes de comunicações e recursos de dados que coleta, transforma e dissemina informações em uma organização." (O'BRIEN, 2004, p.6).

Inicialmente os SI eram simples no seu armazenamento de dados, técnicas e métodos utilizados, além de sua ineficiência e limitação. Sendo simples armazenadores de informações, como cadastro de funcionários, faltas, itens, entre outros. Contudo, ao longo dos anos passaram a ser parte significante dentro de uma empresa, onde "[...] a efetiva utilização pelas organizações tem sido considerada crucial para a sobrevivência e a estratégia competitiva." (AUDY; BRODBECK, 2009, p.15).

Atualmente os SI estão em um processo de evolução exponencial, com estes sendo utilizados para tomada de decisões estratégicas dentro de uma corporação. Grandes empresas utilizam os SI para realizar predições, um exemplo é a Cisco que realiza predições sobre as vendas, fazendo uma análise sobre vendas antigas e pedidos atuais (Russel, 2013).

A Figura 1 demostra os três papéis principais do SI dentro de uma empresa elencados por O'Brien (2004), no qual cada nível da pirâmide representa o grau de maturidade de uma empresa dentro do potencial dos SI.

Segundo Riley (2012), os SI podem ser divididos em seis grupos principais:

a) transacional: sistemas simples que realizam apenas cadastro de informações;
b) controle gerencial: geralmente utilizam os dados do sistema transacional e evidenciam estas informações através de relatórios;
c) apoio a decisões: são desenvolvidos para auxiliar na tomada de decisão em situações onde não existe certeza sobre os possíveis retornos da decisão. Geralmente possuem ferramentas e técnicas para auxiliar o acumulo de informações relevantes, para análise de opções e alternativas;
d) planejamento estratégico: auxilia a gerência na tomada de decisões estratégicas, o mesmo acumula, análise e sumariza as informações;

e) gestão do conhecimento: existem para ajudar as empresas a criar e compartilhar informações, são geralmente utilizados em empresas onde funcionários criam novos conhecimentos e *expertise*. No qual, estes podem ser compartilhados com o restante da empresa podendo gerar futuras oportunidades comerciais;

f) automação de escritório: sistemas que visam melhorar a produtividade de funcionários que precisam processar dados e informações, um exemplo deste sistema é o Microsoft Office.

Figura 1 - Papéis SI conforme nível de maturidade da empresa

Fonte: O'Brien (2004).

Dentro desses grupos estão os mais diversos tipos de sistemas, que atendem diversas áreas, sistemas *Enterprise Resource Planning* (ERP), RH, *Customer Relationship Management* (CRM), entre outros, todos estes sistemas fazendo parte do ecossistema dos SI.

2.2 SISTEMA RH

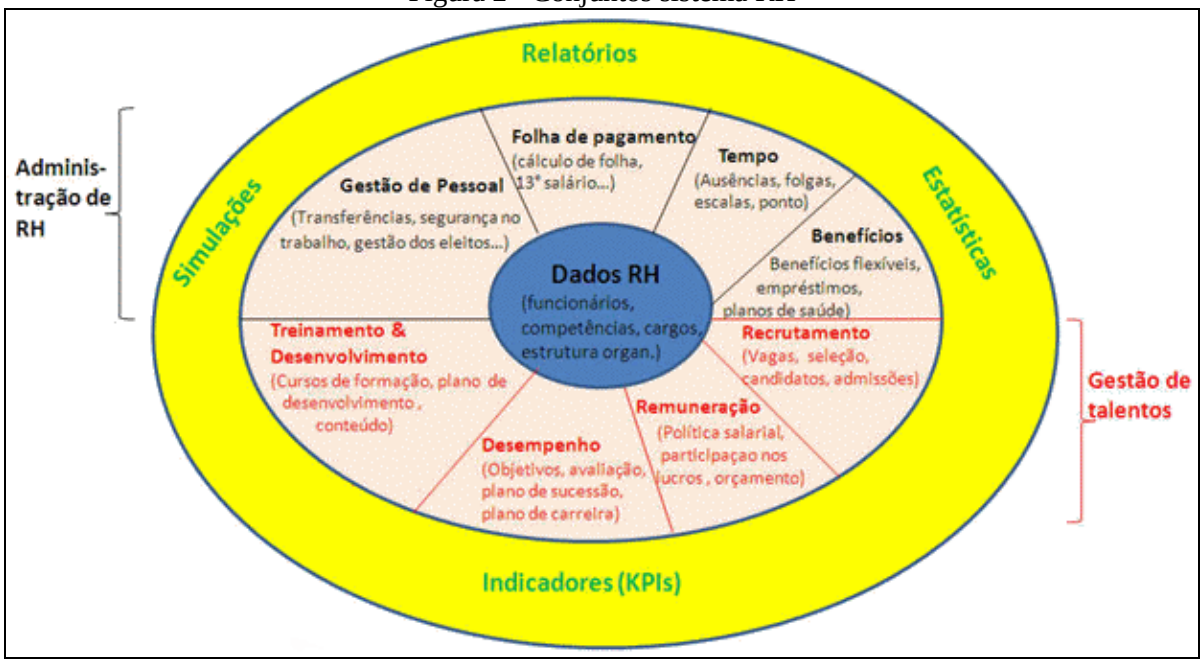
O sistema RH é responsável por armazenar o histórico de um funcionário na empresa, bem como permitir a seleção de candidatos, pesquisas organizacionais, etc. Conforme Binotto, Nakayama e Pilla (2006):

> Um sistema de informação de recursos humanos utiliza como fonte de dados os elementos fornecidos por recrutamento e seleção de pessoal, treinamento e desenvolvimento de pessoal, avaliação de desempenho, administração de salários, registros e controles de pessoal, estatística de pessoal, higiene e segurança, respectivas chefias, entre outros. (BINOTTO; NAKAYAMA; PILLA, 2006, p.305).

A Figura 2 apresenta o sistema RH dividido em três grandes conjuntos, sendo o mais interno à administração do RH, ou seja, cadastro de funcionários, cargos, funções, entre outros. O grupo intermediário compreende a parte onde são realizadas as operações de pagamentos, benefícios, tributos, feedback, gestão de talentos dos funcionários, entre outros, no qual este grupo se utiliza dos dados do grupo inferior. A parte mais externa compreende a

parte analítica, onde se encontram os relatórios, cubos, indicadores, sendo estas informações providas através de dados dos outros grupos descritos.

Figura 2 - Conjuntos sistema RH



Fonte: Liman (2011).

Dentro dos conjuntos exibidos na Figura 2 pode-se destacar os módulos:

a) folha de pagamento: existente desde o surgimento dos sistemas RH, módulo chave que cuida dos pagamentos realizados ao funcionário;
b) treinamento e desenvolvimento: área muito nova dentro dos sistemas, visa a evolução dos funcionários dentro das necessidades da empresa;
c) desempenho: realiza avaliações, determina objetivos e projeções para com os funcionários, visando o prolongamento do mesmo dentro da empresa;
d) estatísticas: fornece para os gestores uma visão analítica sobre os departamentos, unidades, setores da empresa visando a tomada de decisões.

Segundo Parry (2007) a implantação de um sistema RH dentro de uma empresa traz uma série de benefícios, como redução de custos, melhor comunicação entre as áreas e acabam por vezes acarretando na reestruturação do departamento de RH, oferecendo assim uma contribuição estratégica para o setor.

2.2.1 Evolução histórica

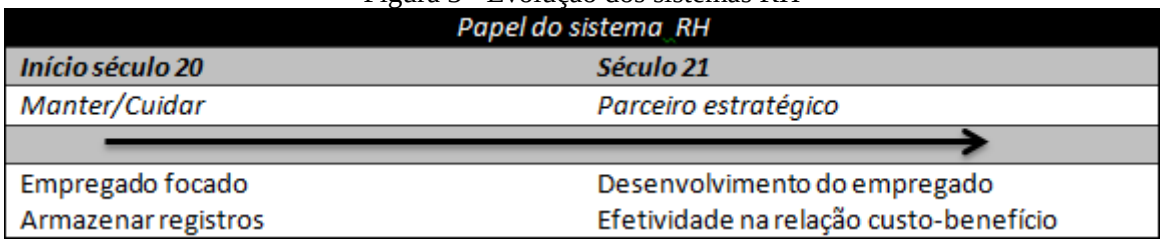
Sistemas de Informação de Empregados surgiram na década de 60 com a tarefa de automatizar o registro dos empregados (DESANCTIS, 1986). Inicialmente eram utilizados

apenas como um simples sistema para armazenamento dos dados dos funcionários, limitados a capacidade da Tecnologia da Informação (TI) na época.

Ao longo dos anos foi evoluindo e assumiu um papel de importância no desenvolvimento e crescimento das empresas. Segundo Bhuiyan, Chowdhury e Ferdous (2014), Recursos Humanos obtiveram melhora da performance nas tarefas tradicionais com a TI. Desde então, os funcionários deixaram de ser somente uma mão-de-obra de produção, se tornando uma das principais ferramentas competitivas do século 21.

A Figura 3 ilustra as características inicias dos sistemas de RH, sendo apenas de registro, evoluindo para um sistema estratégico, com foco no desenvolvimento do funcionário e na efetividade de custos. Como se pode observar, Kavanagh e Thite (2015) fazem uma comparação do começo do século 20 com o século 21.

Figura 3 - Evolução dos sistemas RH



Fonte: Baseado em Kavanagh e Thite (2015).

Atualmente os sistemas RH vêm evoluindo muito em suas estratégias quanto ao funcionário, sendo incorporado aos sistemas ferramentas que auxiliam na capacitação e feedback do colaborador. Um exemplo é a avaliação de seus funcionários no modelo 360º, processo no qual o funcionário é avaliado por múltiplas fontes (DE ALMADA et al., 2008).

2.3 MINERAÇÃO DE DADOS

A mineração de dados é o processo que consiste na descoberta de informações relevantes em grandes bases de dados. As técnicas de mineração de dados são realizadas sobre depósitos de dados de modo a encontrar padrões úteis e recentes, que poderiam passar despercebidos. Além disto, fornecem a capacidade de se prever resultados de uma observação futura, como, por exemplo, a previsão de quantos desligamentos voluntários futuros irão ocorrer (TAN; STEINBACH; KUMAR, 2009).

2.3.1 Etapas da mineração de dados

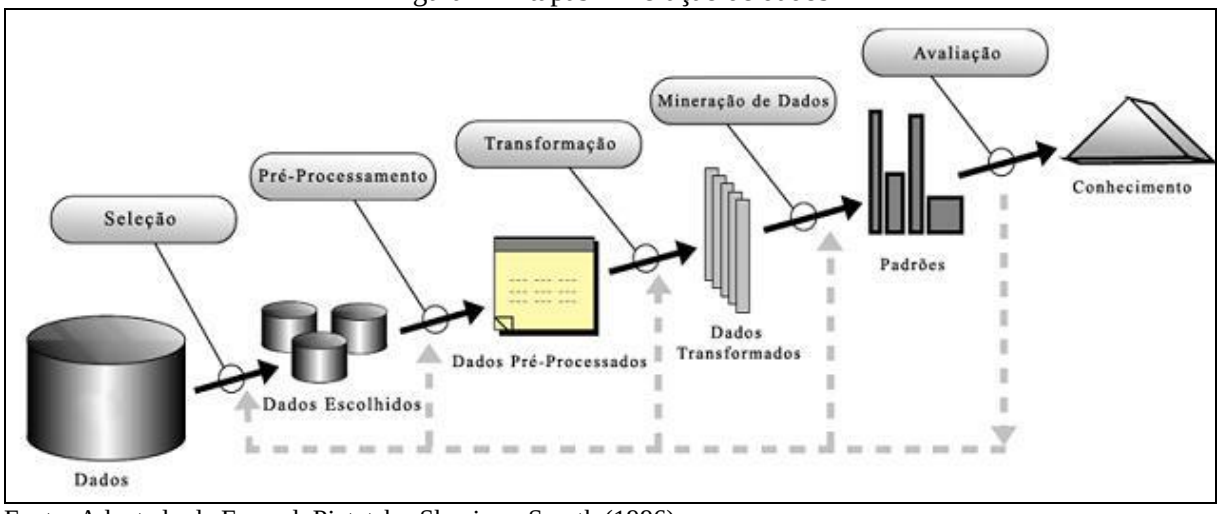
O processo de mineração de dados é formado por cinco etapas, sendo que em cada uma delas é realizada uma operação visando ao final se obter conhecimento sobre uma base

de dados. De acordo com Kantardzic (2011), as etapas a seguir formam o ciclo do processo de mineração de dados:

a) selecionar os dados: buscar as informações que são utilizadas na descoberta de conhecimento, podendo estas serem randômicas ou padronizadas;
b) pré-processamento do dado: verificar quais informações são consistentes para o processo, evitando assim dados que possuam erros de mensuração, de código ou de registro, ou até mesmo erro humano. Esse processo é repetido até que haja o dado mais filtrado possível;
c) transformação: localizar características úteis para a representação dos dados de acordo com o objetivo da tarefa, visando a diminuição de instâncias e variáveis a serem consideradas para o conjunto de dados, além de enriquecer semanticamente as informações;
d) estimar o modelo (minerar o dado): selecionar e implementar a técnica apropriada de mineração de dados, de modo a extrair o conhecimento;
e) interpretação do modelo e conclusões: visualização do conhecimento gerado, podendo assim se tirar conclusões e tomar decisões.

A Figura 4 demonstra de forma macro as cinco etapas da mineração e a maneira com que as mesmas se interagem.

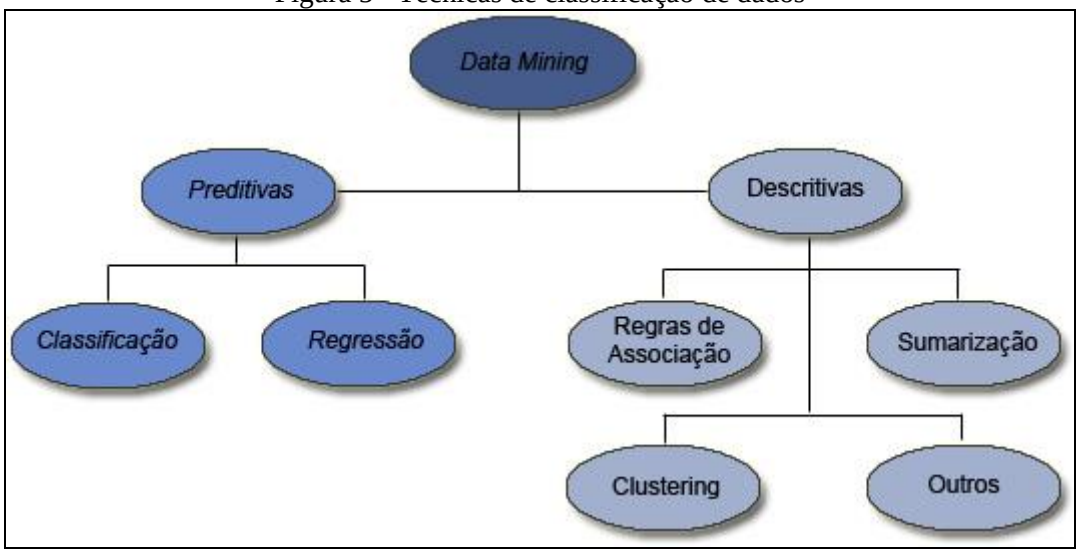
Figura 4 - Etapas mineração de dados



Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Para realizar a classificação dos dados existem diversas técnicas que podem ser aplicadas. A Figura 5 demonstra algumas dessas técnicas e a forma com que as mesmas se relacionam e agrupam, tendo como origem a mineração de dados em si.

Figura 5 - Técnicas de classificação de dados



Fonte: Rezende (2003).

Dentro da etapa de estimar conforme a Figura 4 o modelo (minerar o dado), a classificação é o processo responsável por mapear um conjunto de atributos de modo a classificar os mesmos, desta forma gerando a classe de resultado. “Classificação é a tarefa de aprender uma função alvo f que mapeie cada conjunto de atributos x para um dos rótulos de classes y pré-determinados” (TAN; STEINBACH; KUMAR, 2009, p.172).

De acordo com Tan, Steinbach e Kumar (2009), as técnicas de classificação da mineração de dados abordam uma sistemática visando à construção de modelos de classificação derivados de um conjunto de entrada de dados. Nesta sistemática cada técnica emprega um algoritmo de aprendizagem de modo a identificar um modelo que seja mais ajustado para a relação entre o conjunto de atributos e o rótulo da classe dos dados de entrada.

Existem uma variedade de técnicas entre as quais se destacam:

a) árvore de decisão: simula a estrutura de uma árvore para realizar a classificação dos dados, no qual uma informação grande é partida em várias partes visando alcançar o objetivo desejado (TAN; STEINBACH; KUMAR, 2009);
b) redes neurais: são inspiradas nas redes neurais biológicas, sendo a técnica compostas de nós (neurônios) onde cada nó se conecta ao outro por ligações, onde cada ligação possui um peso associado (MELO, 2010);
c) genético: visão simular o processo evolucionário natural, realizando assim buscas e otimizando a descoberta de padrões. Com o passar do tempo o algoritmo tende a aprender e a se melhorar, de tal forma que somente soluções com grande poder de acerto nas predições são aceitos (FRACALANZA, 2009);
d) regras de associação: consiste no estabelecimento de um conjunto de argumentos

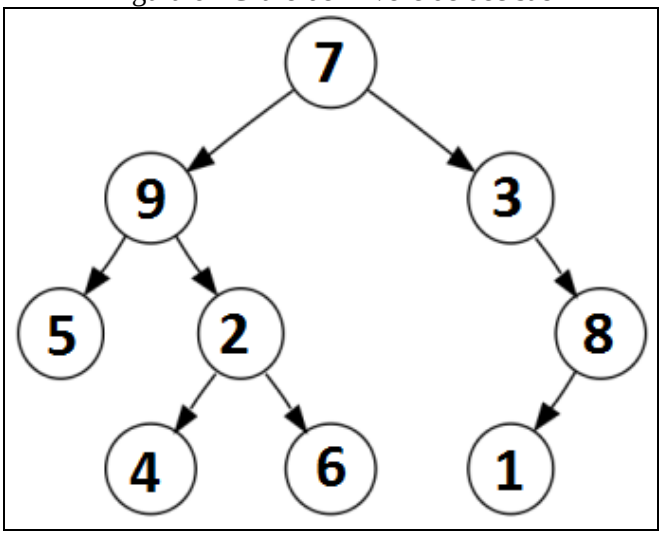
estáticos para um critério macro sobre o dado e um conjunto de argumentos subjetivos como critério de modo a revelar relações inesperadas. É uma técnica com potencial para geração uma grande quantidade de padrões, onde realizar a seleção dos padrões mais interessantes é uma tarefa difícil, na qual se deve estabelecer critérios bem claros para avaliação (TAN; STEINBACH; KUMAR, 2009).

2.3.2 Árvore de decisão

A árvore de decisão é um processo que consiste na divisão de um problema em várias partes, criando assim uma sequência de decisões para o problema. Segundo Magerman (1995), uma árvore de decisão é um dispositivo de tomada de decisão que atribui uma probabilidade para cada uma das escolhas possíveis com base no contexto da decisão. Além disto, a árvore de decisão é poderosa e popular tanto para classificação quanto predição (BERRY; LINOFF, 2004).

A Figura 6 ilustra uma árvore representada através de um grafo, contendo apenas um nó raiz, no qual os demais nós são conectados a este ou a um dos nós conectados a ele, sendo que os números nos nós não precisam possuir qualquer relação de ordem.

Figura 6 - Grafo de Árvore de decisão



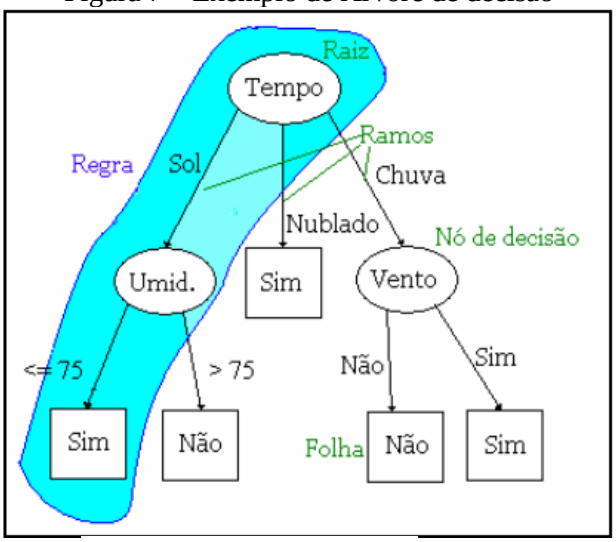
Segundo Tan, Steibach e Kumar (2009), a estrutura de uma árvore é composta de:

a) nodo raiz: este representa o nó principal, no qual todos os demais nós descenderam;
b) nodos internos/decisão: realiza um teste sobre o atributo, onde para cada valor do atributo existe um ramo para uma outra subárvore ou folha;
c) nodos folha ou terminais: é o último nível da árvore, representando o final da

busca na árvore.

A seguir à Figura 7 mostra um exemplo de árvore de decisão para avaliar a probabilidade de se jogar tênis, no qual existe apenas um caminho da raiz até a folha.

Figura 7 – Exemplo de Árvore de decisão



Fonte: Borges, Justino e Ratke (2003).

Através da Figura 7, uma forma de demonstração do processo de regras de produção seria, se Tempo igual a “Sol” e Umidade menor ou igual a 75 então joga.

2.3.2.1 Indução da Árvore de Decisão

Consiste no processo de classificação dos dados e construção da árvore de decisão. De acordo com Tan, Steinbach e Kumar (2009), a partir de um conjunto de atributos é possível construir uma infinidade de árvores de decisão.

Existe uma série de algoritmos sendo desenvolvidos para a indução da árvore de decisão, estes algoritmos em geral utilizam uma série de decisões locais sobre qual atributo utilizar para particionar os dados realizando a estratégia que cresce a árvore de decisão (TAN; STEINBACH; KUMAR, 2009). Dentre esses algoritmos podem ser citados:

a) ID3: a ideia básica deste algoritmo é construir a árvore através do emprego da descendência (*top-down*) da mesma, a cada nodo seleciona-se o atributo que melhor classifica os exemplos de treinamento locais, sendo este processo repetido até o término da classificação (CHEN; PENG; ZHOU, 2009);
b) C4.5: é uma melhoria do algoritmo ID3, no qual esse passou a permitir atributos contínuos, discretização de atributos, poda de ramos visando uma diminuição da árvore, entre outros (QUINLAN, 1993);
c) Hunt: neste algoritmo a árvore cresce de forma recursivo, no qual os registros

treinos são partidos em sucessivos subconjuntos mais puros, o algoritmo somente funcionará se cada combinação possuir apenas um único rótulo de classe e toda combinação de valores estiver presente nos registros treinos (TAN; STEINBACH; KUMAR, 2009);

d) CART: *Classification and Regression Trees* é um método que utiliza dados históricos para construir a árvore de decisão. Este algoritmo apenas responde sim/não para questões, no qual será buscado sempre todas as possíveis variáveis e valores com o objetivo de encontrar a divisão mais homogênea (TIMOFEEV, 2004).

Dentre esses algoritmos o C4.5 utiliza a tática de crescimento de árvore para realizar a indução. O mesmo foi proposto por Quinlan, 1993, se baseando no algoritmo ID3 proposto pelo mesmo, no qual este gera uma boa árvore de decisão através de uma grande quantidade de atributos e objetos sem muita computação (QUINLAN, 1986). O algoritmo C4.5 realiza uma classificação no qual os atributos mais relevantes se localizem nos nós superiores, realizando este processo até que não haja mais atributos relevantes para a classe meta.

Para que essa classificação seja possível, em 1993, Quinlan sugeriu a utilização da entropia, sendo esta uma medida que indica o grau de aleatoriedade do atributo, permitindo assim que os dados mais consistentes fiquem nos níveis superiores quando na indução da árvore. A expressão de entropia proposta por Quinlan em 1993, no qual p(D, j) é a quantidade de casos em D que pertencem à classe j e C é o número total de classes, é demonstrada abaixo:

$$Info(D) = -\sum_{j=1}^{C} p(D,j) \times \log 2 \ (p(D,j))$$

Segundo Quinlan (1993), a capacidade de previsão de um atributo é inversamente proporcional a entropia, onde quanto menor a entropia maior a capacidade de previsão, isto porque quanto menor a entropia menor é o grau de aleatoriedade do atributo. O ganho de informação visa definir o teste que possui a divisão com maior ganho sobre a entropia (QUINLAN, 1996). Desta forma, o ganho de informação é obtido através da expressão:

$$Gain\ (D,T) = Info(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} \times Info(D_i)$$

O k determina o número de classes existentes e D_i são os casos em cada classe.

O Quadro 1 exibe o pseudocódigo do algoritmo C4.5, no qual é possível visualizar a aplicação do cálculo da entropia e de ganho descritos acima, de forma a gerar a indução da árvore de decisão.

Quadro 1 – Pseudocódigo algoritmo C4.5

```
Função C4.5(A atributos, E exemplos);
Inicio
      Se E está vazio, retorna falha
      Se E contém exemplos da mesma classe
            Retorna folha com a classe

      Para cada atributo em A faça
            Se atributo[A] contínuo então
                  Discretiza A[i]
            Calcula valor de Ganho de A[i]
      Fim para

      A[x] = Atributo com maior informação de ganho
      Adicionar nó para atributo A[x]
      Adicionar ramos
      Para cada ramo
            Chamar função C4.5(A, Ex)
      Fim para
Fim.
```

Fonte: Adaptado de Borges, Justino e Ratke (2003).

Conforme o Quadro 1, se um atributo for do tipo contínuo é realizada a discretização, que envolve a transformação de um atributo contínuo em um categórico, onde são determinados pontos de divisão que possibilitam mapear os valores para suas respectivas categorias, podendo esta divisão ser feita através da média, mediana, entre outras (TAN; STEINBACH; KUMAR, 2009).

Na indução se pode determinar a profundidade no qual a árvore atingirá, visando torna-la mais compacta e assertiva. Segundo Quilici-Gonzalez e Zampirolli (2015, p.71), a poda tem por objetivo retirar os nós que refletem um super ajuste (*overfitting*), mitigando ruídos ou *outliers*, sendo este processo realizada durante ou após a conclusão da indução da árvore.

2.3.3 Mineração de texto

A Mineração de texto é utilizada para se encontrar padrões dentro de textos que parecem não possuir sentindo ou ligação alguma entre si. Indo além de acesso a informação, mas para um facilitador aos usuários auxiliando na análise e simplificação das informações visando a tomada de decisão (AGGARWAL; ZHAI, 2012).

Atualmente a mineração de texto vem sendo muito utilizada após o advento das mídias sociais, no qual empresas se utilizam de técnicas de mineração de texto para captar a percepção dos usuários com relação a sua marca. Aggarwal e Zhai (2012), citam que a

mineração de texto obteve um rápido crescimento em diferentes contextos de aplicações Web como as redes sociais, no qual frequentemente ocorrem em contextos de multimídia ou outros domínios de dados heterogêneos.

A mineração de texto possui uma série de algoritmos para extração de conhecimento dentro de textos para os mais diferentes objetivos. O algoritmo de mineração de opinião em um texto trabalha sobre o contexto de opiniões e entrevistas de usuários, no qual o mesmo minera sobre as opiniões para revelar e sumarizar opiniões sobre o tópico mais discutido, desta forma otimizando decisões e *business inteligence* (AGGARWAL; ZHAI, 2012).

Em geral uma opinião pode ser sobre qualquer coisa, um indivíduo, organização, entre outros, sendo a opinião chamada de entidade. Segundo Liu e Zhang (2012), existe dois tipos principais de opiniões: as regulares, que se referem frequentemente a opiniões simples, e opiniões comparativas no qual existe uma relação de similaridade ou diferença entre duas ou mais entidades.

Para que haja uma melhor classificação e análise sobre as opiniões a realização de estruturas preparatórias se torna importante, como a retirada de preposições, singularização das palavras, criação de uma tabela de sinônimos que visa agrupar uma série de palavras em um mesmo grupo. Conforme Marcacini, Moura e Rezende (2011):

> Para a extração e organização não supervisionada de conhecimento a partir de dados textuais, o diferencial está na etapa de extração de padrões, na qual são utilizados métodos de agrupamento de textos para organizar coleções de documentos em grupos. Em seguida, são aplicadas algumas técnicas de seleção de descritores para os agrupamentos formados, ou seja, palavras e expressões que auxiliam a interpretação dos grupos. (MARCACINI; MOURA; REZENDE, 2011, p.8).

Segundo Wives (2002), as fases para realizar a busca automática de palavras relevantes e semelhanças são a identificação de termos, a remoção de *stop words*, a normalização morfológica e seleção de termos. A Figura 8 mostra as etapas em ordem de aplicação, sendo que dependendo da situação a ordem pode ser alterada e algumas etapas não serem utilizadas.

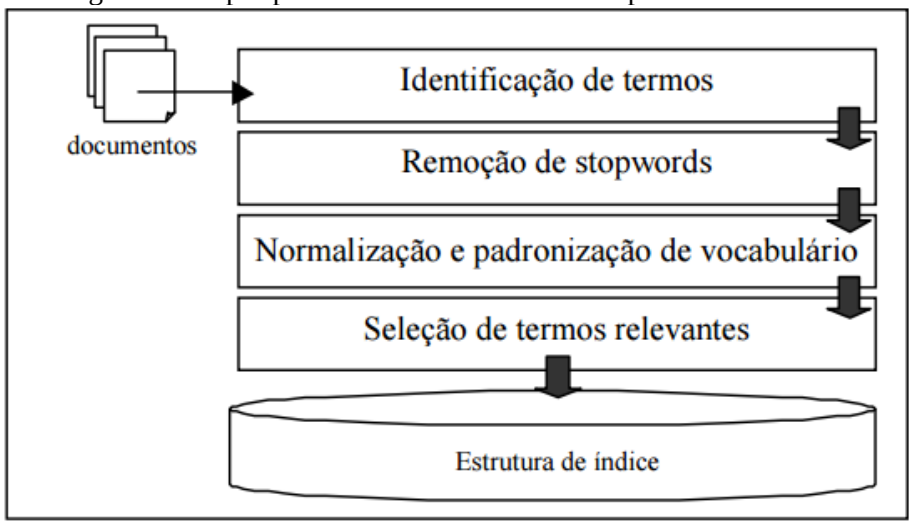
As características de cada etapa propostas por Wives (2002) são:

a) identificação de termos: nesta fase é aplicado um analisador léxico que identifica as palavras e ignorados símbolos, caracteres de controle ou de formatação;
b) remoção de *stop words*: consiste no processo de eliminar palavras que funcionam apenas para realizar a ligação entre frases, sendo que estas não necessitam ser incluídas. Por exemplo: retirada de palavras como “nas”, “das”, “ou”, “seja”, entre outras;
c) normalização e padronização de vocabulário: este processo visa eliminar as variações morfológicas de uma palavra, através da identificação do radical livre

desta palavra, onde os prefixos e sufixos são eliminados e os radicais resultantes são utilizados. “Assim, uma ideia, independentemente de ter tendo sido escrita através de seu substantivo, adjetivo ou verbo, é identificada por um mesmo (e único) radical.” (WIVES, 2002, p.53);

d) seleção de termos relevantes: essa etapa consiste na exclusão dos termos com menor importância, existe uma série de técnicas para a seleção de termos que podem se basear na posição dos termos ou na sua posição quanto a sintaxe.

Figura 8 - Etapas processo busca automática de palavras relevantes



Fonte: Wives (2002).

Após a aplicação das etapas o resultado será o conjunto de palavras que possuíram a maior importância dentro do contexto analisado. Sendo que, através destas palavras pode-se detectar pontos negativos e positivos, permitindo assim, que seja tomada decisões a partir do resultado gerado (WIVES, 2002).

Para exibir as palavras encontradas no final do processo, pode-se utilizar a nuvem de palavras ou *Wordcloud* conforme demonstra a Figura 9, no qual as palavras de acordo com a sua importância dentro das informações mineradas possuem tamanhos distintos.

Figura 9 - Nuvem de palavras

A nuvem de palavras possibilita uma análise rápida e superficial das palavras encontradas, além disto, permite acompanhar a evolução de uma palavra dentro do contexto analisado (GRAHAM; MILLIGAN; WEINGART, 2013).

2.4 FORECASTING

Historicamente, o termo *forecasting* é ligado ao termo predição. Os recentes trabalhos em métodos para determinar valores futuros baseados em dados coletados utilizam o termo predição, ao invés de *forecasting* (PALIT; POPIVIC, 2005). O *forecasting* é predominantemente associado com o problema de análise de séries temporais de modo a se encontrar relações futuras.

Os modelos de *forecasting* são classificados em:

a) objetivos: realizados utilizando o simples julgamento, intuição, conhecimento comercial e outras informações relevantes;
b) univariados: utilizam do modelo unidimensional para coletar dados e na exploração do padrão de séries temporais;
c) multivariados: baseados no modelo simultâneo de observação de duas ou mais variáveis e nos modelos de séries temporais multivariadas.

As abordagens de *forecasting* normalmente são combinadas entre si, de modo a realizar uma previsão mais apurada. Por exemplo, a predição de demissões voluntárias em um sistema RH baseada em vários dados estáticos do passado que podem ser combinados com a experiência ou conhecimento de uma pessoa altamente envolvida no RH (PALIT; POPIVIC, 2005).

Conforme Rey, Kordon e Wells (2012), a razão para a integração de mineração de dados e *forecasting* é fornecer as previsões da mais alta qualidade possível. Assim, pode-se encontrar de forma mais qualitativa tendências que afetam a rotatividade de pessoal, aposentadoria, ociosidade, atrasos, entre outros de modo a permitir o planejamento e a tomada de decisões.

A integração da mineração de dados com *forecasting* permitem entender de que forma fatores como a matéria prima, logística, mão-de-obra, interagem com o processo de produção de uma empresa por exemplo (REY; WELLS, 2013).

Uma forma de utilizar a mineração de dados juntamente com o *forecasting* é através do PMML, onde, o Data Mining Group (2014), demonstra que uma árvore gerada pode ser utilizada como uma estrutura preditiva, onde cada nó contém uma expressão preditiva lógica que define a regra para escolha de um ou outro nó.

2.4.1 PMML

O PMML é uma linguagem baseada em *eXtensible Markup Language* (XML) com o objetivo de representar predição e descrição de modelos de mineração de dados, permite a troca de modelos entre diferentes ferramentas e ambientes, em geral evitando incompatilidades (Guazzelli et al., 2009). Grandes empresas como IBM e SAS utilizam do PMML para suas soluções visando levar ao usuário toda a capacidade da mineração de dados de uma forma demonstrativa e simples.

Segundo Guazzelli et al. (2009), as estruturas padrões de um XML gerado a partir do PMML são:

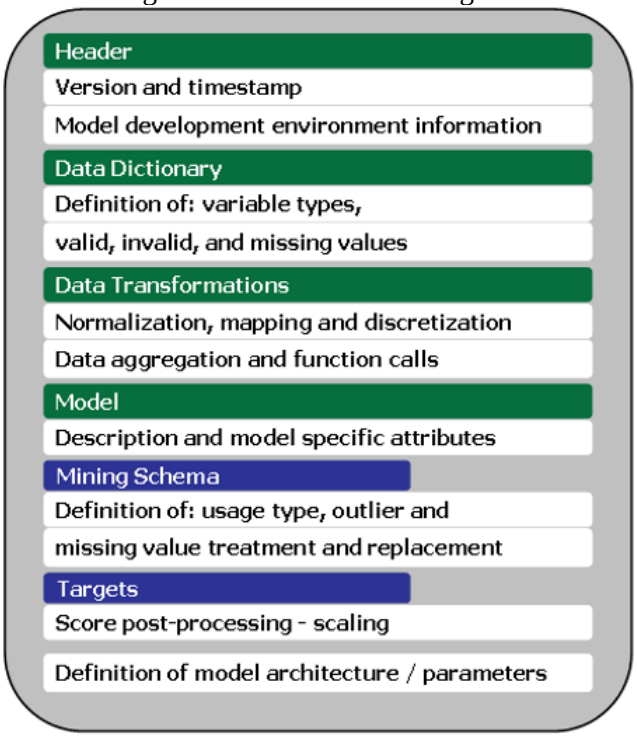
a) cabeçalho: contém atributos informativos referente a informações da versão do PMML utilizado, data em que o mesmo foi gerado, dados do ambiente de geração do modelo, entre outros;
b) dicionário de dados: área onde são declarados os atributos que serão futuramente utilizados pelo modelo, caso o atributo não esteja declarado e se faça utilização do mesmo um erro acontecerá. Um atributo deve ser caracterizado entre contínuo, categórico ou ordinal, cada um possuindo informações extras a serem declaradas;
c) transformação dos dados: permite que sejam realizadas modificações nos dados através de funções como discretização, normalização, entre outros. Esta habilidade de representar transformações de dados em conjunto com os parâmetros que definem os modelos é um dos conceitos chaves do PMML;
d) modelo: nesta seção onde é especificada a técnica que será utilizada pelo modelo e

toda a estrutura de funções necessárias para realizar a mesma;

- esquema de mineração: lista de todos os campos que serão utilizados, contendo informações específicas dos campos como nome e uso do mesmo, sendo nesta informação de uso onde no caso da árvore de decisão se informa a classe meta;
- metas: permite o dimensionamento de variáveis previstas.

A Figura 10 demonstra de forma macro a estrutura de um modelo PMML, mostrando a sequência da estrutura de cima para baixo.

Figura 10 - PMML estrutura geral



Fonte: Guazzelli et al. (2009).

O PMML possuí uma série de modelos para utilização sobre os métodos de mineração de dados como árvore de decisão, redes neurais, regressão, etc. Todos os modelos utilizam a estrutura descrita acima, possuindo apenas algumas partes específicas para atender suas peculiaridades de execução.

2.5 TRABALHOS CORRELATOS

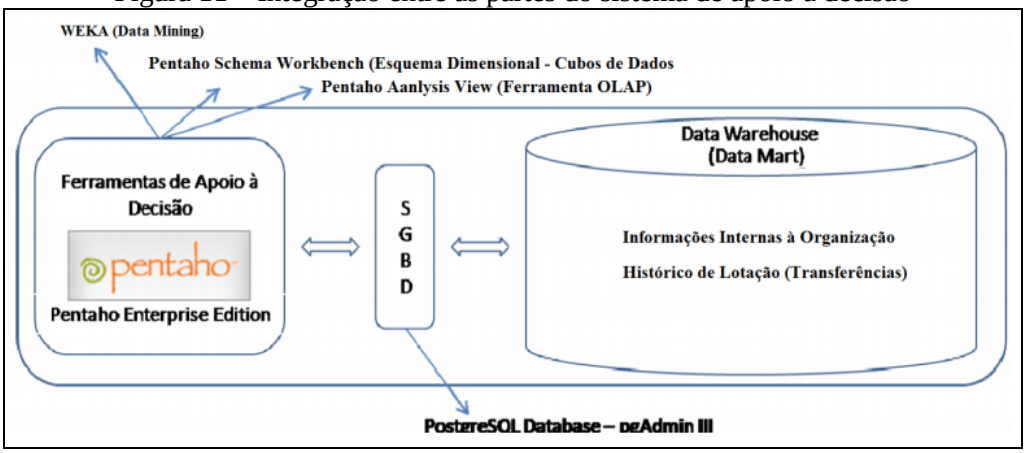
Como trabalho correlato foi encontrado uma tese de mestrado de Mendes (2013) onde o objetivo é caracterizar *turnover* interno através da mineração de dados. A ferramenta comercial Universal RH Explorer (2015) objetiva a mineração de dados e relacionamento de informações nas mais variadas bases de dados RH.

2.5.1 Aplicação de técnicas de *data mining* na caracterização de *turnover* interno para o suporte à gestão de pessoas

A tese de Mendes (2013) teve por objetivo descobrir e elucidar informações dentro de uma base de dados de Gestão de Pessoas com relação à rotatividade interna de pessoal e seus impactos na empresa. Para realizar a validação do estudo, foi utilizada uma base de dados de uma instituição financeira que possuía um histórico de rotatividade de pessoas.

Na Figura 11 como se pode observar, Mendes (2013) mostra como os componentes do sistema se interagem, onde o Sistema de Gerenciamento de Banco de Dados (SGBD) faz a ponte entre a ferramenta de apoio a decisão (mineração de dados, cubos de dados) e o *Data Warehouse* (informações internas à organização e históricos).

Figura 11 – Integração entre as partes do sistema de apoio a decisão



Fonte: Mendes (2013).

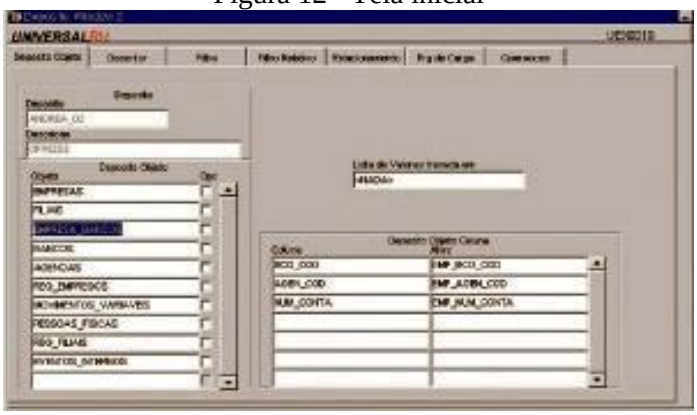
Foram desenvolvidos algoritmos de mineração de dados como modelos de árvores de decisão, agrupamento e regras de associação para poder descrever o perfil dos funcionários que se movimentam na empresa. Para a complementação dos resultados encontrados pelas técnicas de mineração de dados, foi criado um protótipo multidimensional de dados de modo a encontrar conhecimento sobre a rotatividade de pessoal.

2.5.2 Universal RH Explorer

O Universal RH Explorer (2015), trata-se de uma ferramenta que utiliza técnicas de mineração de dados, como árvore de decisão e regras de associação para realizar extração de informações de uma base de dados de Gestão de Pessoas, de modo que o usuário possa realizar combinações e cruzamentos dos dados extraídos para encontrar padrões nos desligamentos, afastamentos, etc.

Na Figura 12 encontra-se a tela principal da ferramenta, onde na barra esquerda são demonstrados os objetos que serão minerados dentro no sistema de Gestão de pessoas e no centro, as tabelas e campos que compõem este objeto.

Figura 12 - Tela inicial



Fonte: Universal RH Explorer (2015).

Através das informações coletadas é possível gerar relatórios, cubos e gráficos. A Figura 13 demostra um exemplo de cubo gerado através dos dados buscados.

Figura 13 - Exemplo de cubo

Fonte: Universal RH Explorer (2015).

A ferramenta tem como base a plataforma Web, sendo desenvolvida na linguagem Java e tendo como banco de dados o Oracle. Para realizar extrações mais complexas através de Structured Query Language (SQL) é utilizado o Procedural Language/Structured Query Language (PL/SQL), porém, existe a limitação de somente poder extrair informações de base de dados de Gestão de Pessoas que sejam baseadas em Oracle.

3 DESENVOLVIMENTO

Nesta seção são apresentas as etapas do desenvolvimento da ferramenta. São apresentados os requisitos, a especificação e a implementação, demonstrando as técnicas e ferramentas utilizadas. Ao final é apresentada a operacionalidade da ferramenta e os resultados obtidos.

3.1 REQUISITOS

A seguir são mostrados os Requisitos Funcionais (RF) e Não Funcionais (RNF) atendidos pela ferramenta, apresentados respectivamente no Quadro 2 e no Quadro 3, que para cada RF foi relacionado o seu caso de uso.

Quadro 2 - Requisitos funcionais

Requisitos funcionais (RF)	**Caso de uso (UC)**
RF01: permitir cadastro de conexão para busca no banco de dados do sistema de RH.	UC01
RF02: permitir o cadastro das configurações para indução da árvore.	UC02
RF03: permitir o cadastro das configurações dos atributos da árvore geradora.	UC03
RF04: permitir o cadastro do agendamento da árvore geradora para determinar a frequência de busca das informações.	UC04
RF05: permitir o cadastro da tabela de sinônimos.	UC05
RF06: permitir o cadastro da tabela de *stop words*.	UC06
RF07: permitir realizar consulta PMML.	UC07
RF08: disponibilizar uma tela com gráficos demonstrando a previsão de rotatividade de pessoal.	UC08
RF09: permitir o cadastro de usuário.	UC09
RF10: permitir o cadastro de grupos de usuário.	UC10
RF11: permitir cadastrar grupo de usuário para o usuário.	UC11

Quadro 3 - Requisitos não funcionais

Requisitos não funcionais (RNF)
RNF01: utilizar o banco de dados SQL Server.
RNF02: utilizar o ambiente Visual Studio e a linguagem C#.
RNF03: utilizar o *framework* PMML4Net para execução do PMML.
RNF04: utilizar o *framework* Bootstrap para o design das páginas Web.
RNF05: utilizar um serviço do Windows para execução dos agendamentos.
RNF06: utilizar o componente D3.js para gerar os gráficos.

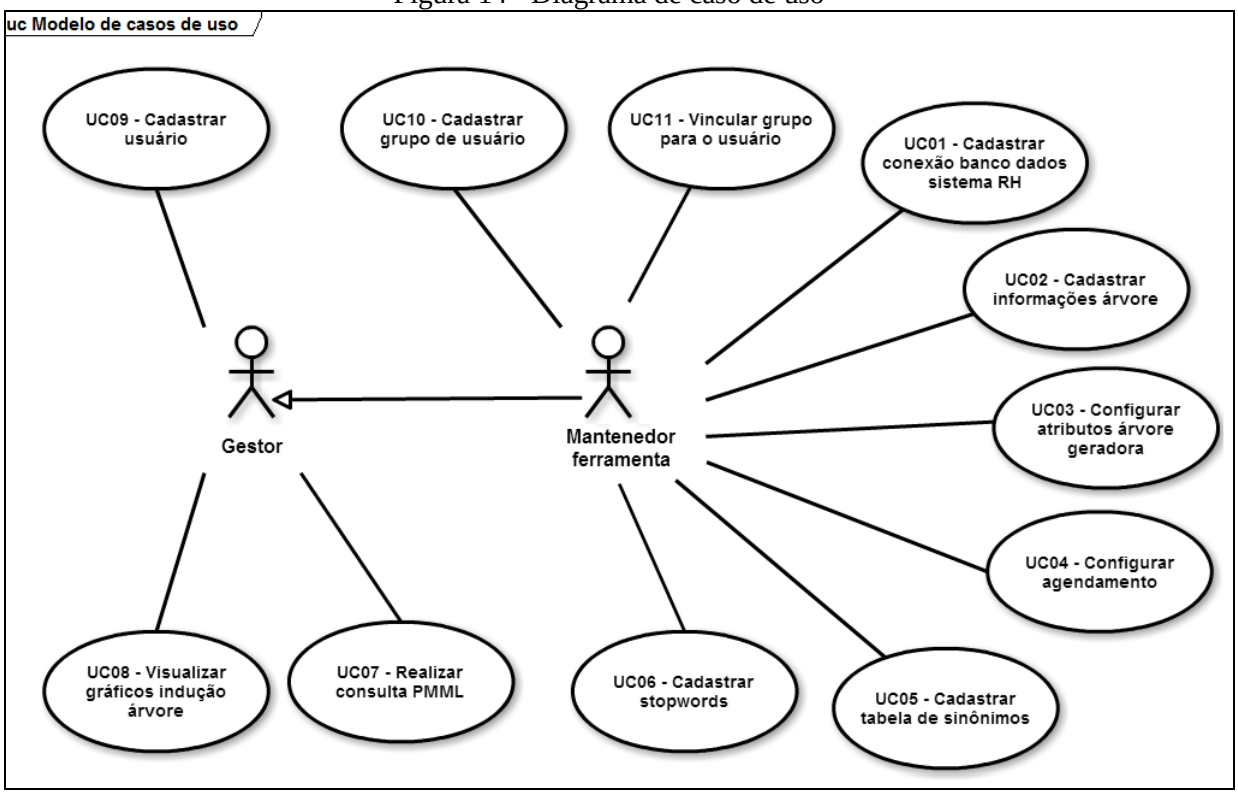
3.2 ESPECIFICAÇÃO

Nesta seção é apresentado a especificação da ferramenta. A especificação foi desenvolvida seguindo o *Unified Modeling Language* (UML), sendo os diagramas desenvolvidos através das ferramentas Astah Community 7.0, Gliffy e Cacco.

3.2.1 Diagrama de Caso de uso

A Figura 14 demonstra o diagrama de caso de uso da ferramenta, na qual o papel de mantenedor da ferramenta é responsável por manter a ferramenta e gerar as informações baseado (Árvore, *stopwords*, sinônimos, etc.) nas demandas requisitadas pelo papel do gestor, sendo este responsável por analisar as informações geradas pelas árvores e realizar a consulta PMML. A descrição detalhada dos principais casos de uso é apresentada no Apêndice A.

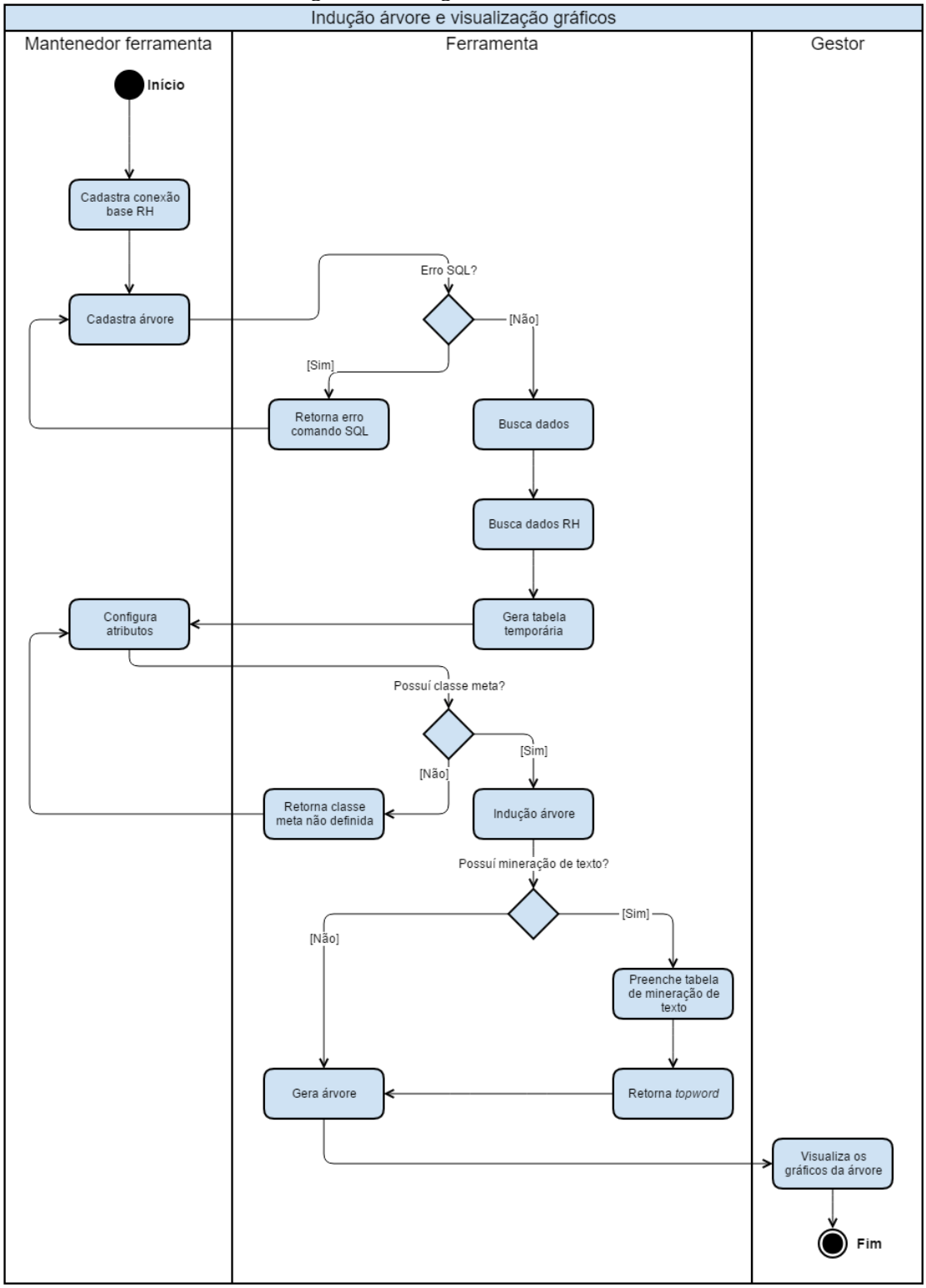
Figura 14 - Diagrama de caso de uso



3.2.2 Diagrama de atividades

Nesta seção é apresentado o diagrama de atividades. A Figura 15 demonstra o processo para indução da árvore e visualização dos gráficos gerados. O usuário deve cadastrar a conexão do banco de dados da base RH, as configurações da árvore (Comando SQL, nível poda, etc.), configurar os atributos da árvore (definir classe meta, tipo atributo, etc.). Após isto, o mesmo poderá visualizar os gráficos gerados a partir da indução da árvore.

Figura 15 - Diagrama de atividades

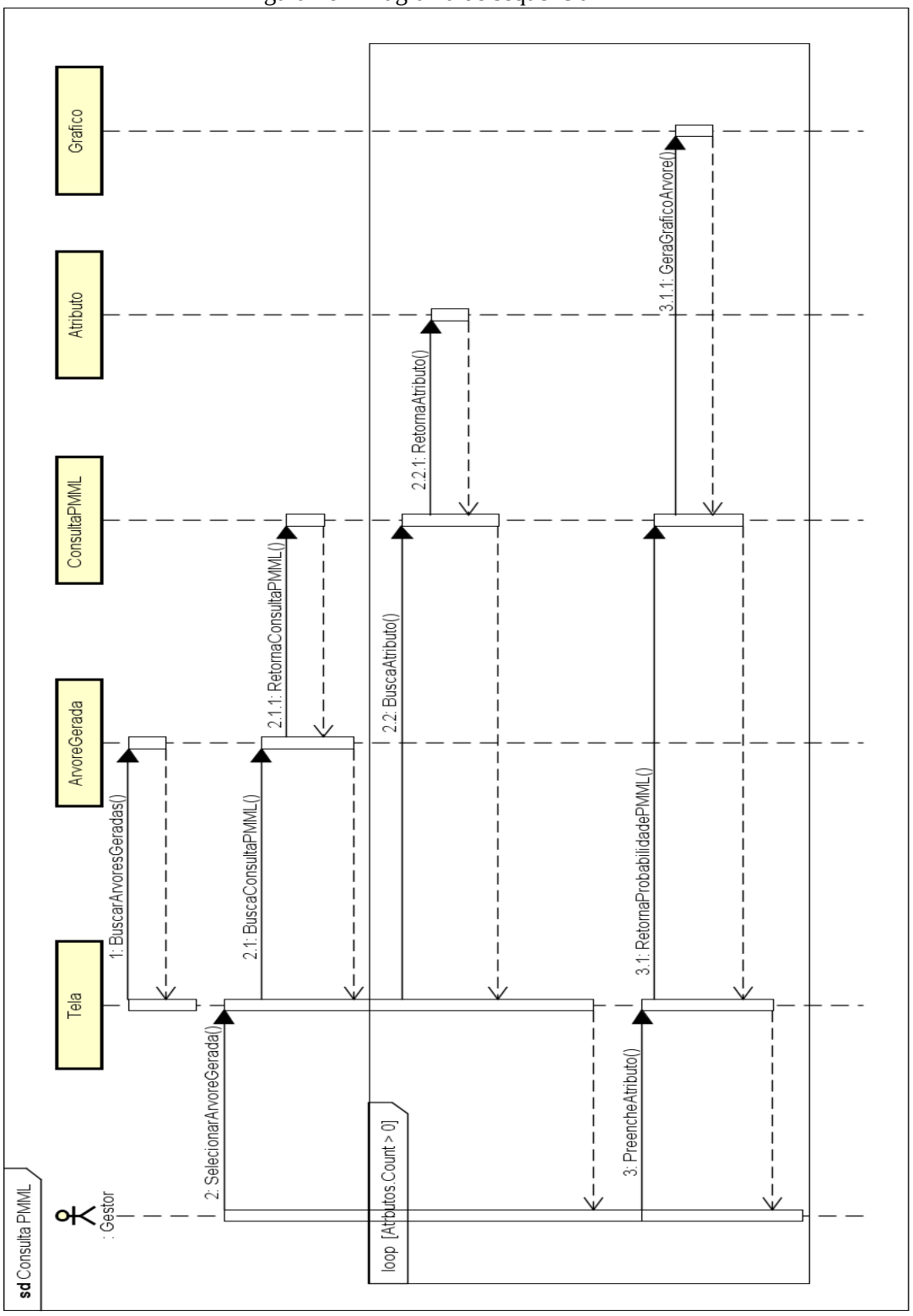


3.2.3 Diagrama de sequência

O diagrama de sequência da ferramenta apresentado na Figura 16 apresenta o processo de consulta PMML. O gestor entra na tela de arvores geradas, sendo então retornadas todas as árvores geradas até o momento. O mesmo deve selecionar a árvore que deseja consultar, ao

realizar a seleção será buscada a consulta PMML relativa a árvore selecionada, através disto será disponibilizado um atributo a ser preenchido. Uma vez preenchido o atributo é realizado a probabilidade através do PMML e retornado o caminho gerado na árvore, este processo de preenchimento de atributo e retorna da árvore gerada se repete até que não haja mais atributos.

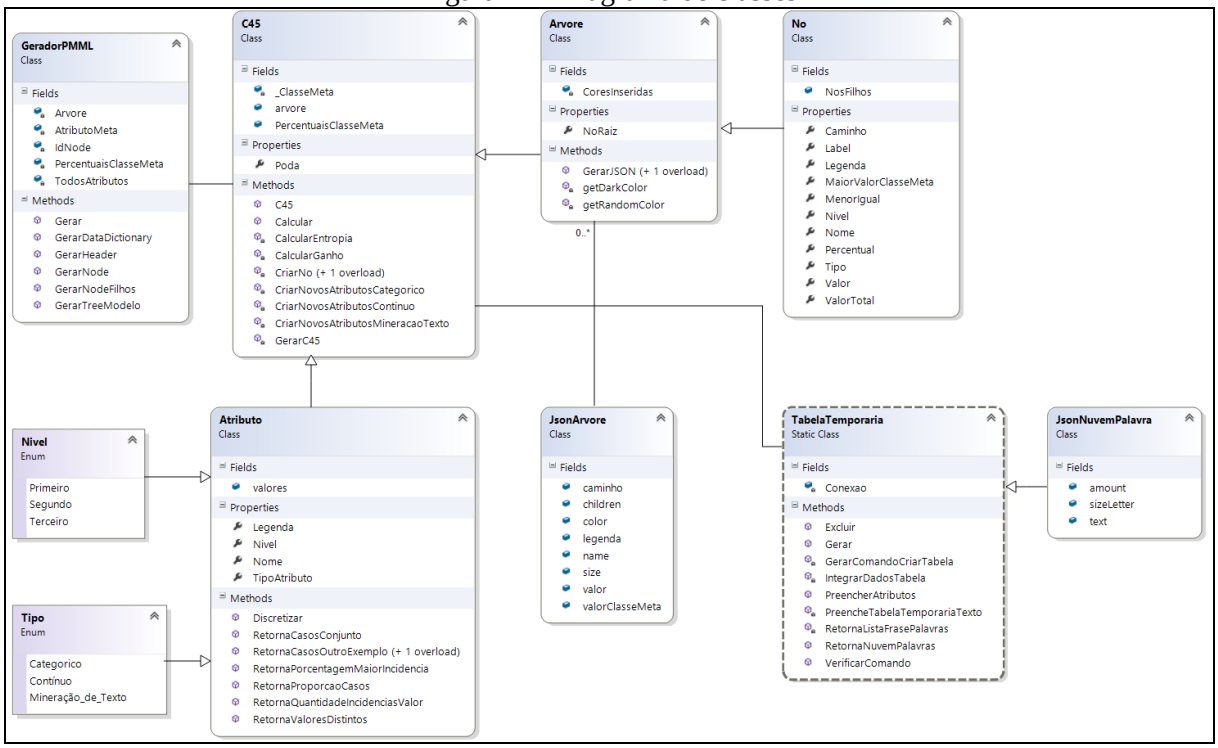
Figura 16 - Diagrama de sequência



3.2.4 Diagrama de classes

Na Figura 17 é representado o diagrama de classes da ferramenta com as principais classes. Através da classe `C45` é realizado a indução da árvore de decisão (Construção da árvore composta pela classe `Árvore` e `No`), onde inicialmente são buscados os dados da classe `TabelaTemporario`, em seguida sendo utilizados para determinar a entropia da classe meta (`_ClasseMeta`), o ganho de cada atributo (`Atributo`) de acordo com o nível atual da árvore e nível de poda (`Poda`). A partir da árvore gerada, são chamados os métodos da classe `GeradorPMML` e `JsonArvore` para que sejam gerados a consulta PMML e os arquivos *JavaScript Object Notation* (JSON) que serão utilizados pelos gráficos (Interativo, árvore e nuvem).

Figura 17 - Diagrama de classes



3.3 IMPLEMENTAÇÃO

A seguir são mostradas as técnicas e ferramentas utilizadas e a operacionalidade da implementação.

3.3.1 Técnicas e ferramentas utilizadas

A ferramenta foi implementada na linguagem C# para as operações a serem realizadas no servidor, para a parte do cliente foi utilizada a linguagem JavaScript visando uma resposta mais rápida para o usuário e para as operações de mineração de texto foi utilizada a

linguagem nativa do banco de dados SQL Server desta forma trabalhando direto na fonte com os dados.

Utilizado para codificação o Visual Studio 2013, pois o mesmo possui uma excelente integração e ferramentas para desenvolvimento em C# e JavaScript. Visando uma melhor integração e controle das fontes, foi utilizado o Microsoft Team Fundation.

Para as operações de banco de dados foi utilizado o Entity Framework que realiza as operações no banco através de objetos, além deste trabalhar muito bem em conjunto com o SQL Server 2012 por serem feitos pela mesma empresa.

Foi utilizada a biblioteca PMML4Net (Carol, 2013) para desenvolvimento do processador PMML, onde está prove classes para leitura e consumo de modelo de árvores em formato PMML. Esta biblioteca foi selecionada por ser *open source* e ter sua implementação feita em C#, possuindo assim, alta compatibilidade com a linguagem de servidor no qual a ferramenta foi desenvolvida.

A geração dos gráficos foi realizada utilizando a biblioteca Data-Driven Documents (Bostock, Heer e Ogievetsky, 2011) sendo *open source* e escrita em JavaScript, como trabalha no lado do cliente proporciona melhor performance para o usuário.

3.3.2 Desenvolvimento da ferramenta

Nesta seção são apresentados os detalhes da implementação dos processos que compõem a ferramenta.

3.3.2.1 Indução árvore

No Quadro 4 pode-se visualizar o trecho de código responsável por implementar o algoritmo C4.5 para a indução da árvore. Incialmente é verificado o nível de poda da árvore, em seguida é realizado o cálculo do ganho de informação sobre a classe meta e o ganho sobre cada atributo, determinando assim o atributo que irá compor o nível. O Quadro 5 mostra um trecho fragmento de código do método `GerarC45` onde este faz uma chamada recursiva para realizar a geração dos nós da árvore.

Quadro 4 – Fragmento da Rotina algoritmo C4.5

```
private void GerarC45(string atributoMeta, List<Atributo> atributos,
                      No no)
{
if (Poda.HasValue && no.Nivel >= Poda.Value)
    return;

var valoresClasseMeta = atributos.FirstOrDefault(atr =>
                                      atr.Nome.Equals(_ClasseMeta.Nome));
double valorEntropia = CalcularEntropia(valoresClasseMeta);

Atributo atributoMaiorGanho = new Atributo();
double maiorGanho = 0;

foreach (var atributo in atributos.Where(atr =>
                                     !atr.Nome.Equals(_ClasseMeta.Nome)))
{
    var ganhoAtr = CalcularGanho(valorEntropia, valoresClasseMeta,
                                 atributo);
    if (ganhoAtr > maiorGanho)
    {
        maiorGanho = ganhoAtr;
        atributoMaiorGanho = atributo;
    }
}
```

Quadro 5 - Chamada recursiva algoritmo C4.5

```
No noAtual;
if (atributoMaiorGanho.TipoAtributo == Tipo.Categorico)
{
  var valoresDistintos = atributoMaiorGanho.RetornaValoresDistintos();
  foreach (var valor in valoresDistintos)
  {
    var novosAtributos = CriarNovosAtributosCategorico(valor.ToString(),
                                        atributoMaiorGanho, atributos);
    noAtual = CriarNo(atributoMaiorGanho.Nome,
                      atributoMaiorGanho.Legenda, Tipo.Categorico,
                      valor.ToString(), novosAtributos[0].valores.Count,
                      no.ValorTotal, novosAtributos.FirstOrDefault(atr =>
       atr.Nome == _ClasseMeta.Nome).RetornaPorcentagemMaiorIncidencia(),
                      no.Nivel + 1);
    no.NosFilhos.Add(noAtual);
    GerarC45(valor.ToString(), novosAtributos, noAtual);
  }
}
```

3.3.2.2 Cálculo da Entropia

O Quadro 6 mostra o código fonte relativo ao cálculo de entropia realizado para a classe meta de acordo com nível atual, sendo esse valor utilizado para determinar o atributo que possuem maior ganho com relação a classe meta.

Quadro 6 - Função que calcula a entropia da classe meta

```
private double CalcularEntropia(Atributo valoresClasseMeta)
{
var valoresDistintos = valoresClasseMeta.RetornaValoresDistintos();
float totalInformacoes = valoresClasseMeta.valores.Count;
double entropia = 0;

foreach (var valorDistinto in valoresDistintos)
{
  float divisaoCasosTotal= valoresClasseMeta
       .RetornaProporcaoCasos(valorDistinto.ToString())/totalInformacoes;
  entropia += (divisaoCasosTotal) * (Math.Log(divisaoCasosTotal, 2));
}

if (entropia < 0)
  entropia *= -1;

return entropia;
}
```

3.3.2.3 Cálculo do Ganho

No Quadro 7 pode-se visualizar o fragmento de código que executa o cálculo de ganho de informação para o atributo do tipo categórico com relação à classe meta. São retornados todos os valores distintos dentro do atributo, em seguida sobre cada atributo é realizado o cálculo de frequência sobre a classe meta e ao final o valor total dos atributos é subtraído da entropia da classe meta para encontrar o valor de ganho.

Quadro 7 - Rotina para cálculo do ganho do atributo

```
double ganho = 0;
List<double> listaCalculoD = new List<double>();
if (valoresAtributoCalculo.TipoAtributo == Tipo.Categorico)
{
  var valoresDistAtrCalculo =
           valoresAtributoCalculo.RetornaValoresDistintos();
 foreach (var valor in valoresDistAtrCalculo)
 {
   float totalCasos =
          valoresAtributoCalculo.RetornaProporcaoCasos(valor.ToString());
   var casosClasseMeta =
        valoresAtributoCalculo.RetornaCasosOutroExemplo(
                              valor.ToString(),valoresClasseMeta.valores);
   double casosSomados = 0;
   foreach (var caso in casosClasseMeta)
   {
     double divisao = caso.Value / totalCasos;
     casosSomados += divisao * (Math.Log(divisao, 2));
   }
   double calculoD = (totalCasos / valoresAtributoCalculo.valores.Count)
                     * casosSomados;
   listaCalculoD.Add(calculoD < 0 ? calculoD *= -1 : calculoD);
 }
}
ganho = entropia - (listaCalculoD.Sum());
```

3.3.2.4 Discretização

No Quadro 8 é possível visualizar o código no qual é realizada a discretização dos atributos contínuos, visando transformar o mesmo em um atributo categórico. Sendo que para realizar a partição dos valores foi utilizado o cálculo da média, onde são somados todos os valores e dividido pela quantidade. Desta forma, dividindo os valores em dois grupos os de menores ou iguais à média e maiores que a média.

Quadro 8 - Método discretização atributo contínuo

```
public double Discretizar()
{
double pontoReferencia = 0;
foreach (var valor in valores)
{
   double ret = 0;
   bool converte = double.TryParse(valor.ToString(), out ret);
   pontoReferencia += ret;
}

pontoReferencia = pontoReferencia / valores.Count;
return pontoReferencia;
}
```

3.3.2.5 Mineração de texto

Para realizar a mineração de texto foram criadas diversas funções no banco de dados, visando velocidade para a mineração. No Quadro 9 é possível visualizar a função principal responsável por buscar a *top word* dentro do atributo de mineração de texto determinado para a árvore que será gerada.

Quadro 9 - Fragmento da função para mineração de texto

```
FETCH NEXT FROM cPALAVRAS INTO @PALAVRA1
WHILE @@FETCH_STATUS = 0
BEGIN
   SET @PALAVRA2 = @PALAVRA1

INSERT INTO @RESULTADO
 SELECT @PALAVRA2,
  DBO.FN_TIRAACENTO(DBO.RETORNAMASCULINO(
                    DBO.RETORNASINGULAR(RTRIM(VALOR)))),
                    DBO.RETORNADICIONARIOPADRAO(
                     DBO.FN_TIRAACENTO(
                      DBO.RETORNAMASCULINO(
                       DBO.RETORNASINGULAR(
                        RTRIM(VALOR)))))
 FROM SPLIT(@PALAVRA1, ' ')
  WHERE VALOR NOT IN (SELECT PALAVRA  FROM PALAVRASEMSIGNIFICADO)
    AND VALOR NOT IN (SELECT PALAVRA  FROM STOPWORD)
    AND VALOR NOT IN (',','.',';','?','!','-',' ');

FETCH NEXT FROM cPALAVRAS INTO @PALAVRA1
END
```

No Quadro 10 é apresentada a função responsável pela radicalização da palavra, um dos passos descritos por Wives (2002) para realizar a mineração de texto.

Quadro 10 - Função para radicalização das palavras

```
FUNCTION [DBO].[RETORNARADICAL] (
      @ENTRADA VARCHAR(80) )
      RETURNS VARCHAR(80)
BEGIN
DECLARE @SAIDA VARCHAR (40) ;
DECLARE @AUX VARCHAR (40) ;

SET @ENTRADA = UPPER (LTRIM(RTRIM(@ENTRADA)))
SET @SAIDA = @ENTRADA

IF (@ENTRADA LIKE '%s')
BEGIN
   SET @AUX= DBO.RETORNASINGULAR(@ENTRADA)
   SET @SAIDA = @AUX ;
END
ELSE BEGIN
   IF (@ENTRADA LIKE '%a')
   BEGIN
      SET @AUX= DBO.RETORNAMASCULINO(@SAIDA)
      SET @SAIDA = @AUX;
      SELECT @AUX= DBO.RETORNASEMSUPERLATIVO (@SAIDA )
      SET @SAIDA = @AUX;
      SELECT @AUX= DBO.RETORNAREDUCAOADVERBIAL(@SAIDA )
      SET @SAIDA = @AUX ;

      SELECT @AUX= DBO.RETORNASEMSUFIXO(@SAIDA )
      IF (@SAIDA <> @AUX)
      BEGIN
         SELECT @AUX= DBO.RETORNASEMSUFIXOVERBAL (@SAIDA )
         SET @SAIDA = @AUX ;
      END
   END
END
```

O Quadro 11 apresenta um fragmento de código da função responsável pela singularização da palavra. Transformando a palavra da sua forma pluralizada para a singularizada, por exemplo, a palavra “sermões” submetida a está função será convertida para “sermão”.

Quadro 11 - Fragmento da função para singularização da palavra

```
SET @ENTRADA = UPPER(LTRIM(RTRIM(@ENTRADA)))
SET @SAIDA = @ENTRADA

IF (dbo.CHARINDEX2('NS', @SAIDA)+ 1 = LEN(@SAIDA)) AND (LEN(@SAIDA) > 3)
   SET @SAIDA = dbo.REPLACE2(@SAIDA , 'NS' , 'M' )

IF (dbo.CHARINDEX2('ÕES',@SAIDA)+ 2 = LEN(@SAIDA)) AND (LEN(@SAIDA) > 6)
   SET @SAIDA = dbo.REPLACE2 ( @SAIDA , 'ÕES' , 'ÃO' ) ;

IF (dbo.CHARINDEX2('ÃES', @SAIDA)+ 2 = LEN(@SAIDA)) AND (LEN(@SAIDA) > 3)
   IF ( @SAIDA = 'MÃES' OR @SAIDA = 'MAMÃES')  SET @SAIDA = @SAIDA ;
      ELSE SET @SAIDA = dbo.REPLACE2 ( @SAIDA , 'ÃES' , 'ÃO' ) ;
   ELSE
      SET @SAIDA = @SAIDA;

IF (dbo.CHARINDEX2('AIS', @SAIDA)+ 2 = LEN(@SAIDA)) AND (LEN(@SAIDA) > 3)
   IF ( @SAIDA IN ( 'CAIS' , 'MAIS' ) )  SET @SAIDA = @SAIDA ;
      ELSE SET @SAIDA = dbo.REPLACE2 ( @SAIDA , 'AIS' , 'AL' ) ;
   ELSE
      SET @SAIDA = @SAIDA;
```

3.3.2.6 Gerador Arquivo PMML

Para permitir que seja realizada a consulta PMML sobre a árvore, é necessário transformar a mesma para o formato de arquivo que o PMML é capaz de utilizar. No Quadro 12 é exibido parte do código responsável por criar o arquivo XML interpretado pelo PMML. A árvore gerada pelo método `GerarC45` é utilizada para determinar os atributos a serem utilizados, o atributo a ser realizada a predição, possíveis caminhos a serem percorridos e resultados que os caminhos retornam.

Quadro 12 - Método para geração do arquivo PMML

```
public XmlDocument Gerar(List<Atributo> todosAtributos,
                         string atributoMeta, Arvore arvore,
                         HashSet<string> percentuaisClasseMeta)
{
this.TodosAtributos = todosAtributos;
this.AtributoMeta = atributoMeta;
this.Arvore = arvore;
this.PercentuaisClasseMeta = percentuaisClasseMeta;

XmlDocument doc = new XmlDocument();
XmlDeclaration xmlDeclaration = doc.CreateXmlDeclaration("1.0", "UTF-8",
                                                         null);

XmlElement root = doc.DocumentElement;
doc.InsertBefore(xmlDeclaration, root);

var atrVersion = doc.CreateAttribute("version");
atrVersion.Value = "4.1";
var atrXmlns = doc.CreateAttribute("xmlns");
atrXmlns.Value = "http://www.dmg.org/PMML-4_1";

XmlElement elementPMML = doc.CreateElement(string.Empty, "PMML",
                                           string.Empty);
elementPMML.Attributes.Append(atrVersion);
elementPMML.Attributes.Append(atrXmlns);
doc.AppendChild(elementPMML);
GerarHeader(doc, elementPMML);
GerarDataDictionary(doc, elementPMML);
GerarTreeModelo(doc, elementPMML);

return doc;
}
```

3.3.2.7 Executor consulta PMML

O Quadro 13 demonstra o fragmento de código que realiza a consulta PMML, no qual é utilizado o arquivo previamente gerado pela função `Gerar` para obter as informações da árvore que será consultada. A partir deste arquivo e dos dados inseridos pelo usuário na tela de consulta PMML a função realiza a consulta, no qual cada atributo inserido pelo usuário permite que a busca avance para níveis inferiores da árvore.

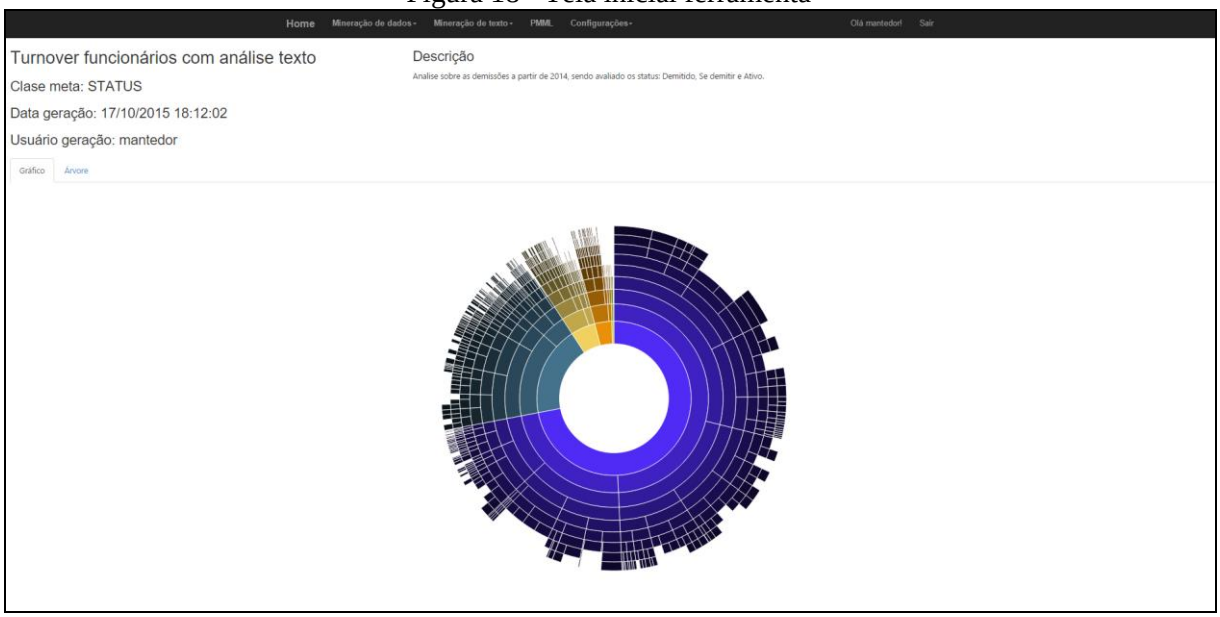
Quadro 13 - Método para execução da consulta PMML

```
public ActionResult Consultar(List<ConsultaPmml> atributos)
{
ViewBag.PossuiNosFilhos = true;
var arvoreGeradaId = atributos[0].ArvoreGeradaId;
var arvoreGerada = db.ArvoreGerada.FirstOrDefault(arvore => arvore.ID ==
                                                    arvoreGeradaId);
XmlDocument doc;
var xml = new
            System.Xml.Serialization.XmlSerializer(typeof(XmlDocument));
using (TextReader reader = new StringReader(arvoreGerada.XmlPmml))
{
  doc = (XmlDocument)xml.Deserialize(reader);
}

doc.Save(HostingEnvironment.MapPath("/Content/pmml.xml"));
Pmml pmml =
      Pmml.loadModels(HostingEnvironment.MapPath("/Content/pmml.xml"));
ModelElement model = pmml.Models[0];
Dictionary<string, object> dict = new Dictionary<string, object>();
foreach (var atr in atributos)
{
  if (!atr.ClasseMeta)
  {
    if (atr.Valor != null)
      dict.Add(atr.Nome, atr.Valor);
    else
    {
      ModelState.AddModelError("", "Campo "+atr.Label+" não foi
                                    preenchido.");
      return View(atributos);
     }
  }
}

ScoreResult result = model.Score(dict);
if (result.Value != null)
{
  Session.Clear();
  Session.Add("ConsultaPMML", atributos);
  ViewBag.Resultado = result.Value;
}
else
  ModelState.AddModelError("", "Consulta PMML não obteve resultado. Favor
                                verificar as informações preenchidas.");
}

return View(atributos);
}
```

3.3.3 Operacionalidade da implementação

A seguir são demonstradas as principais funcionalidades existentes na ferramenta. Na Figura 18 é apresentada a tela inicial da ferramenta, esta possui o comportamento de sempre exibir a última árvore gerada.

Figura 18 - Tela inicial ferramenta



A estrutura das telas do sistema é constituída por um menu superior, que permite a navegação entre as funcionalidades da ferramenta e na parte central são exibidas as informações referentes ao item selecionado do menu. Cada opção em geral apresenta uma listagem dos registros com opções para realizar inclusão, alteração, exclusão e visualização, conforme Figura 19.

Figura 19 - Tela de manutenção de palavras irrelevantes (*Stop words*)

3.3.3.1 Cadastros

Através do menu principal é possível acessar as informações de cadastros e as consultas existentes na ferramenta conforme visualizado na Figura 20. Foram criados 3 itens com subdivisões visando facilitar a navegação na ferramenta, sendo eles:

a) `mineração de dados`: neste item é possível realizar o cadastro ou manutenção da `estrutura da árvore` e visualizar as `árvores geradas` (Figura 21);

b) `mineração de texto`: através deste menu, é possível realizar o cadastro ou manutenção de `stop words` e `agrupadores de texto` (Figura 22);

c) `configurações`: neste menu está disponível o cadastro ou manutenção dos `grupos de usuários` e a configuração da `conexão base RH` (Figura 23).

Figura 20 - Menu principal da ferramenta

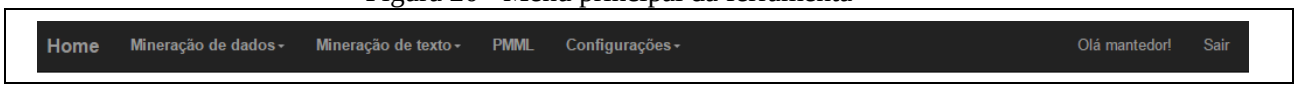


Figura 21 - Menu mineração de dados e subitens

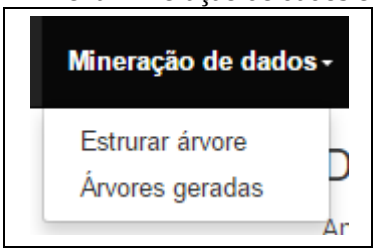


Figura 22 - Menu mineração de texto e subitens

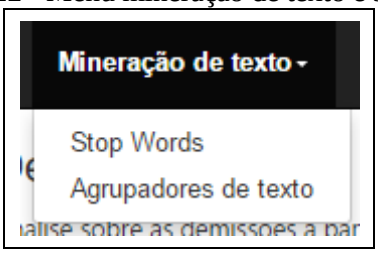
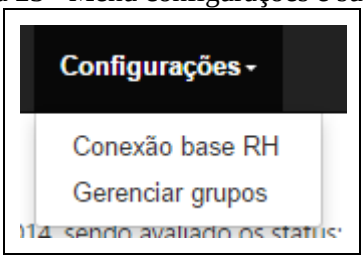


Figura 23 - Menu configurações e subitens



A Figura 24 demonstra o cadastramento das informações de conexão para busca no banco de dados da base RH, sendo que esse cadastro sempre possuirá apenas um registro.

Figura 24 - Cadastro conexão base RH

Conforme a Figura 25 no cadastro de estrutura da árvore são informados o `nome`, `data da criação` e `descrição`. No campo `SQL` deve ser inserido a consulta, em linguagem SQL, a

ser realizada na base RH para geração da árvore. O campo `poda` é opcional, caso não informado no momento da geração da árvore a ferramenta irá assumir dez níveis.

Figura 25 - Cadastro da estrutura de árvore

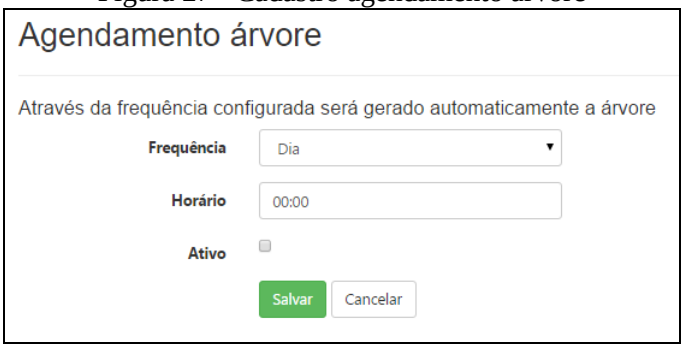
Para que seja possível a geração da árvore se faz necessário realizar a configuração dos atributos da árvore (Figura 26). Os atributos são carregados a partir da consulta `SQL` informada no cadastro da árvore. Deve-se informar o atributo que será `classe meta`, o `tipo` de cada atributo, sendo que para o tipo mineração de texto deve-se informar o `nível de busca`: primeiro (Agrupador de texto), segundo (Palavra agrupadora) e terceiro (Palavra). Opcionalmente se pode informar uma `legenda` para o atributo visando uma melhor apresentação deste nos gráficos.

Figura 26 - Configuração atributos da árvore de decisão

Na Figura 27 é apresentado o cadastro de agendamento de uma árvore e este cadastro permite a geração da árvore automaticamente, havendo assim sempre uma atualização dos

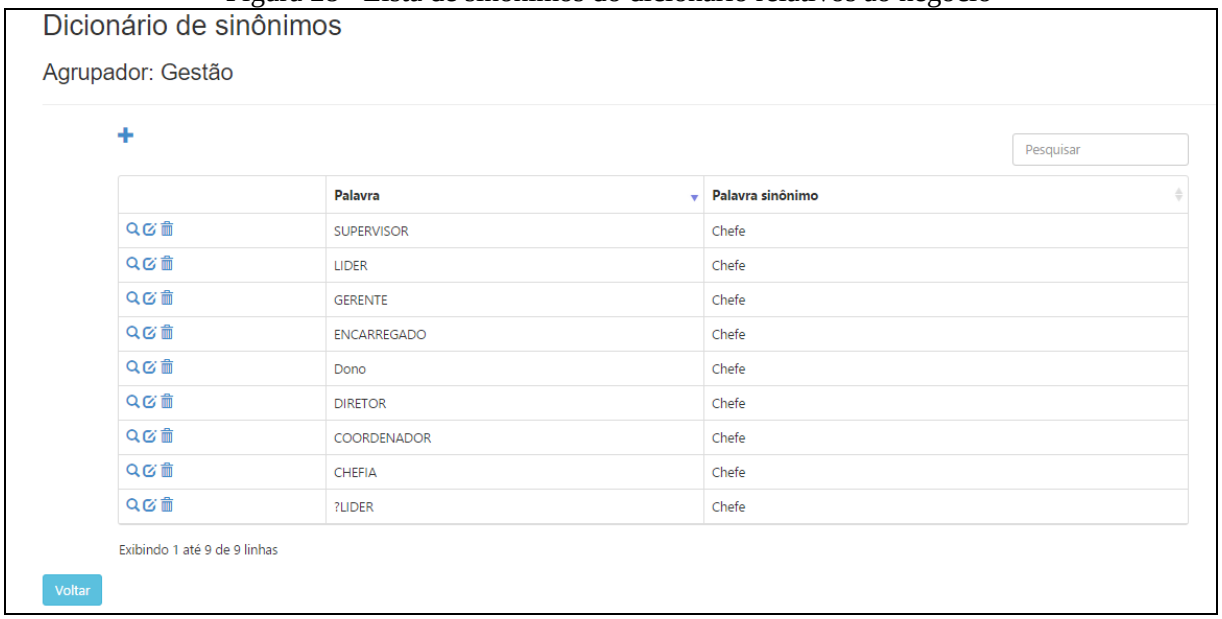
dados da árvore, sendo necessário informar uma `frequência` (Dia, semana ou mês), `horário` e uma opção para tornar o agendamento `ativo`.

Figura 27 - Cadastro agendamento árvore



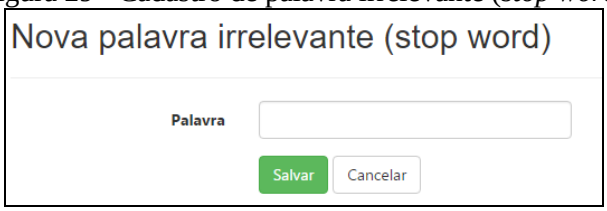
A Figura 28 apresenta o dicionário de sinônimos ligados ao agrupador de texto e estes cadastros são utilizados para o nível de busca na mineração de texto, visando uma melhor busca da *top word*.

Figura 28 - Lista de sinônimos do dicionário relativos ao negócio



Além das *stop words* internas da ferramenta é possível realizar o cadastro de palavra irrelevante (Figura 29) para permitir tratar casos específicos que possam ser encontrados dentro da mineração de texto realizada sobre a base RH.

Figura 29 - Cadastro de palavra irrelevante (*stop word*)

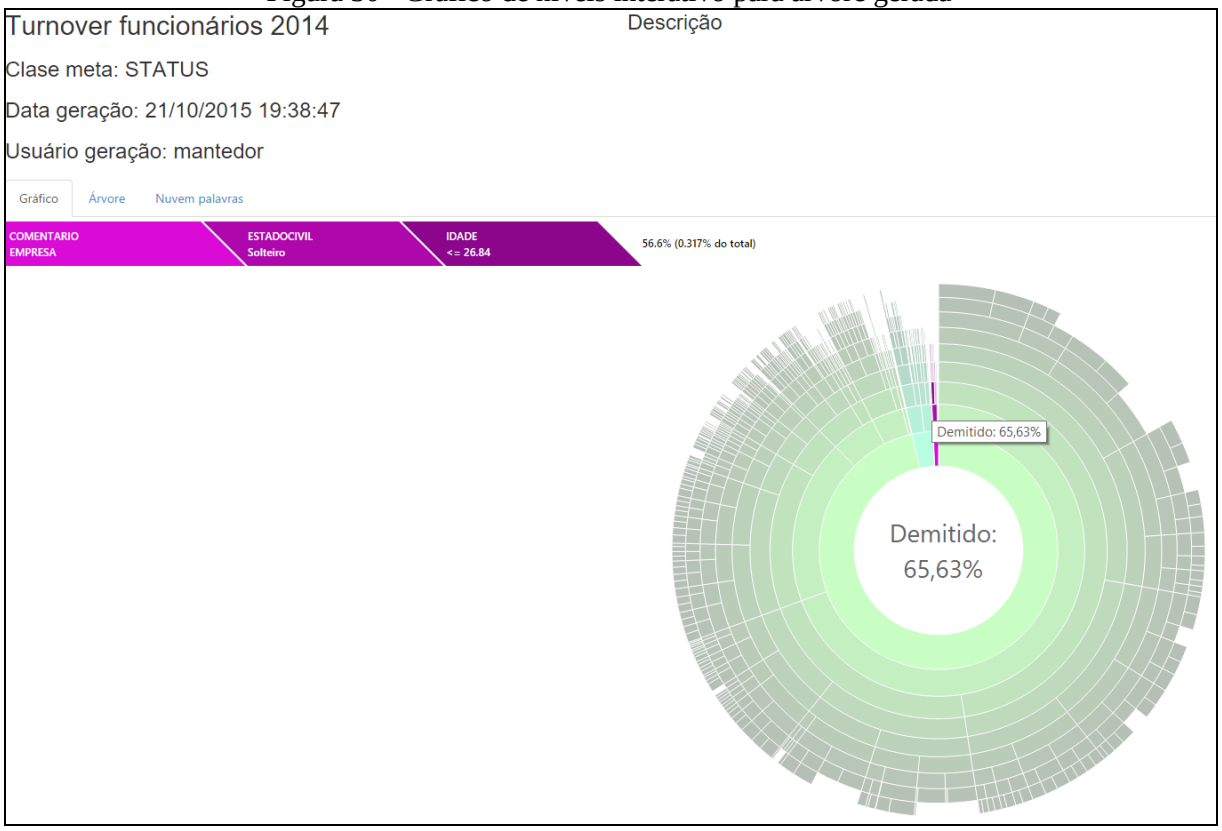


3.3.3.2 Gerar Árvore Decisão

Na tela de configuração de atributos (Figura 26) ao clicar no botão `Salvar/Gerar gráfico` será realizado a indução da árvore. Após realizada a geração é apresentada uma tela com os gráficos gerados a partir da árvore, sendo inicialmente apresentando o gráfico de níveis interativo (Figura 30).

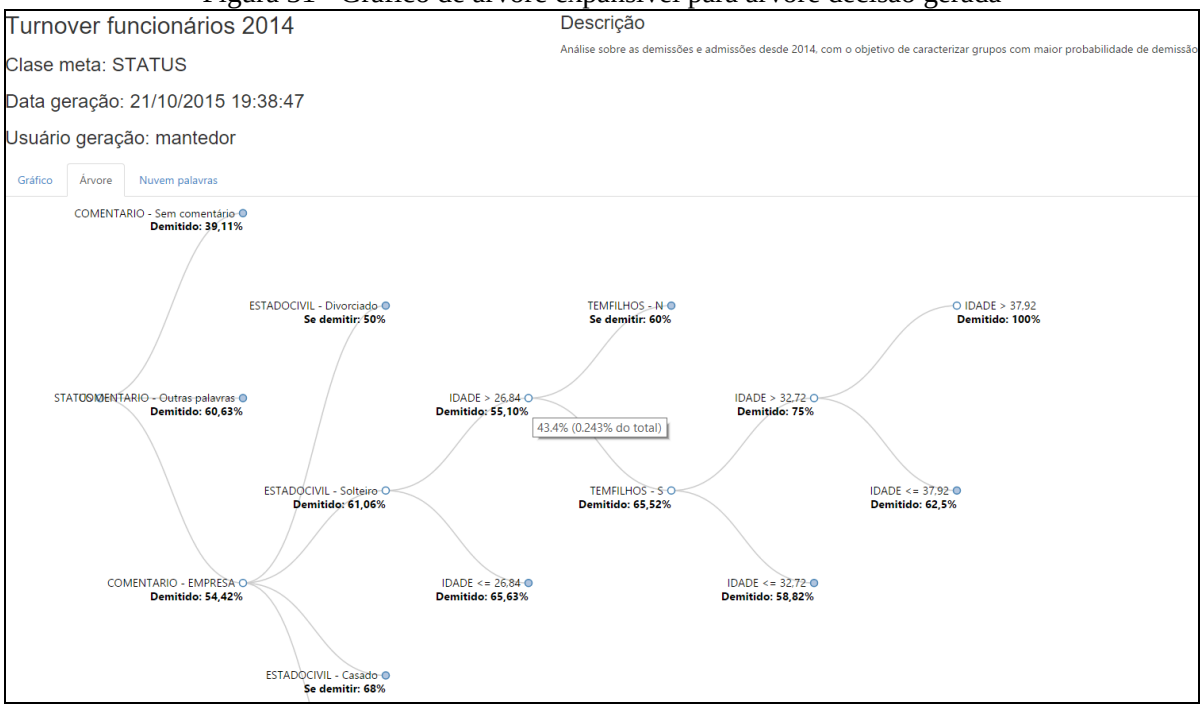
No gráfico da Figura 30 é possível ver o atributo com a maior chance baseada na classe meta por nível, na barra de caminhos é possível visualizar a confiabilidade do nível e percentual de informações com relação ao total. Além disto, é possível visualizar o atributo do tipo `Mineração de texto` no primeiro nível com legenda `Comentário`, no qual a palavra `Empresa` foi gerada como a *top word*.

Figura 30 - Gráfico de níveis interativo para árvore gerada



Nesta mesma tela é possível visualizar o gráfico de árvore (Figura 31), sendo possível fechar e expandir os níveis da árvore, além de visualizar os mesmos dados apresentados no gráfico de níveis interativo. Podendo se verificar a utilização do atributo do tipo `Mineração de texto` no primeiro nível da árvore, com legenda de `Comentário`.

Figura 31 - Gráfico de árvore expansível para árvore decisão gerada



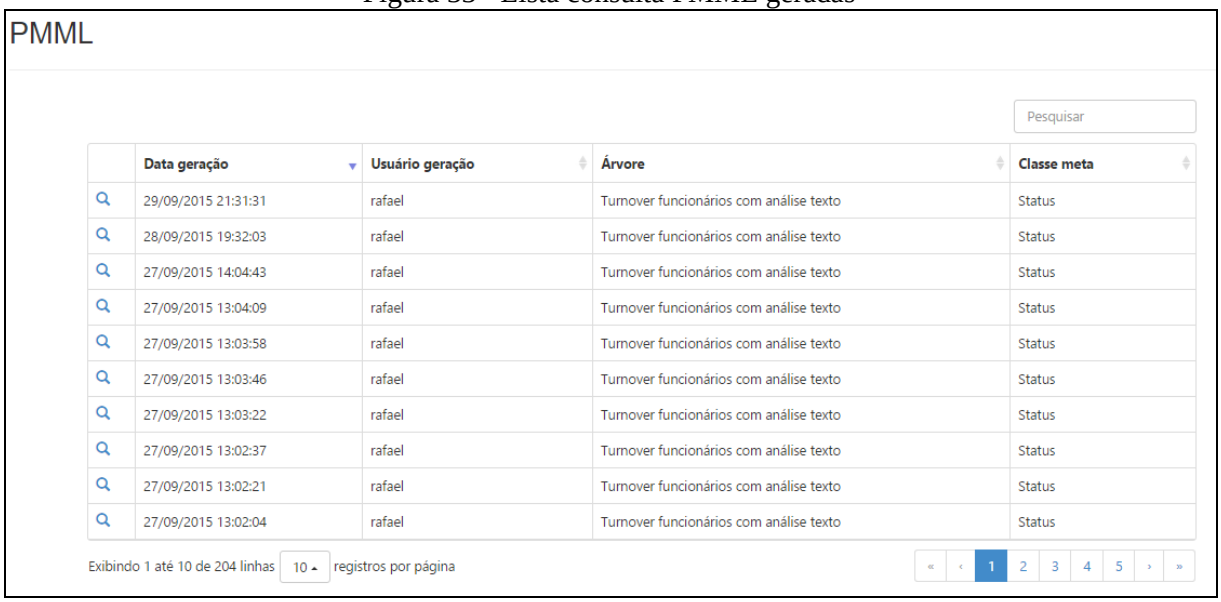
Caso a árvore gerada possua atributo do tipo `Mineração de texto` é possível visualizar o gráfico de nuvem de palavras (Figura 32). Este gráfico exibe as palavras de acordo com o nível de consulta selecionado, passando-se o mouse sobre a palavra é apresentada a quantidade de vezes que a palavra foi encontrada dentro das informações analisadas.

Figura 32 - Gráfico de nuvem de palavras

3.3.3.3 Consulta PMML

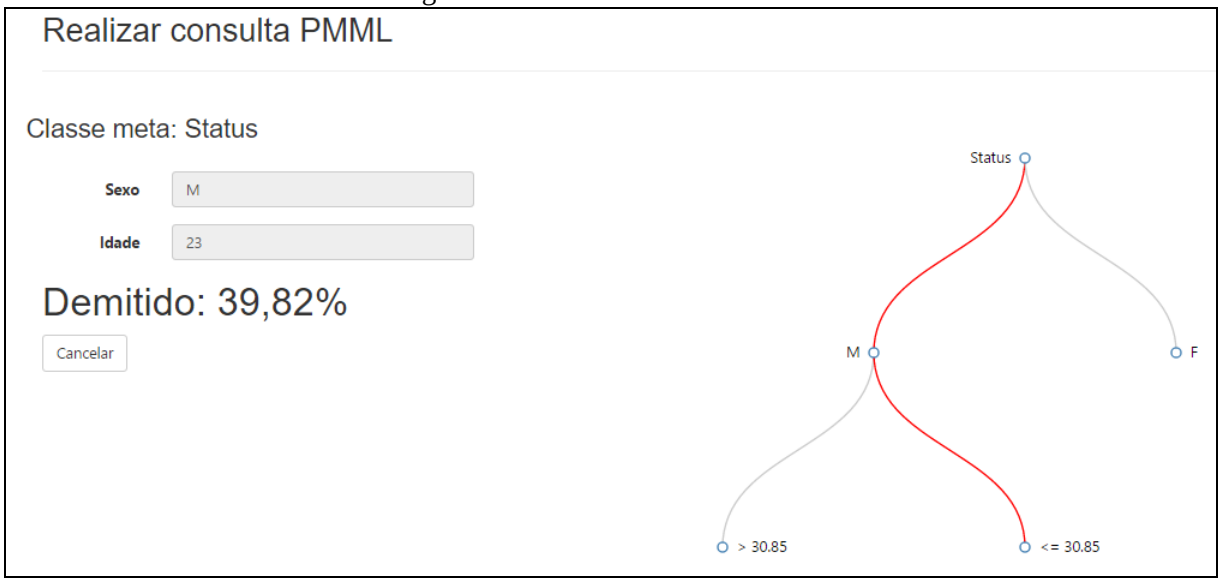
Durante a geração da árvore de decisão as informações utilizadas pelo processo PMML são salvas em um arquivo XML. Através do menu (Figura 20) é possível acessar a tela com a lista das consultas PMML disponíveis. Na Figura 33 se pode visualizar a tela.

Figura 33 - Lista consulta PMML geradas



Ao selecionar uma das consultas é apresentada a tela para consulta PMML (Figura 34), onde é apresentado um campo para preenchimento do atributo e o gráfico de uma árvore. Conforme preenchido o atributo é retornada a maior probabilidade sobre a classe meta e é atualizado o gráfico da árvore.

Figura 34 - Tela de consulta PMML



3.4 RESULTADOS E DISCUSSÕES

Em comparação com os trabalhos correlatos, a ferramenta desenvolvida possui algumas diferenças significantes. Com relação a Mendes (2013), o sistema desenvolvido possui sua plataforma toda Web permitindo uma maior mobilidade no acesso, além de implementar o algoritmo de mineração de dados utilizado, não necessitando de ferramenta de terceiro para analisar os dados. Comparando-se com Universal RH Explorer (2015), se tem como principal diferença que a ferramenta desenvolvida possibilita realizar mineração de texto sobre os dados e utiliza estes para a árvore de decisão, conforme visualizado na Figura 32 e Figura 33. Além disto, a ferramenta desenvolvida é a única que utiliza o PMML para permitir predição sobre os dados gerados pela mineração de dados.

Através do Quadro 14 é apresentada uma análise geral das funcionalidades disponibilizadas na ferramenta desenvolvida com relação aos trabalhos correlatos. No qual é possível visualizar que cada trabalho utilizou de diferentes técnicas de mineração de dados e formas de visualização para os dados gerados. Sendo que, os trabalhos correlatos têm seu software desenvolvido na plataforma Desktop enquanto a ferramenta deste trabalho é desenvolvida na Web. Além disto, a principal diferença entre o trabalho desenvolvido e os correlatos é que o mesmo possuí mineração de texto e utilização de consulta de predição (PMML).

Quadro 14 - Comparação com trabalhos correlatos

Funcionalidade	Semann (2015)	Mendes (2013)	Universal RH Explorer (2015)
Plataforma	Web	Desktop	Desktop
Disponibiliza conexão com outras bases de dados	Sim	Não	Sim
Possui mineração de texto	Sim	Não	Não
Aplicação do PMML sobre dado gerado pela mineração de dados	Sim	Não	Não
Formas de visualização	Gráficos e nuvem de	Tabelas	Gráficos e cubos

dos dados	palavras		
Método	Árvore de decisão	Associação e agrupamento	Associação e árvore de decisão

Com o objetivo de avaliar o resultado final deste trabalho, foi elaborado um formulário de avalição através do Google Docs. A ferramenta foi demonstrada de forma individual para 8 pessoas que trabalham na empresa Benner Sistemas, para que os mesmos pudessem avaliá-la. A avaliação foi realizada com pessoas das mais diferentes profissões (Figura 35) que tem contato com o RH, sendo que entre essas pessoas 75% delas possui 10 ou mais anos de experiência com Recursos Humanos (Figura 36).

Figura 35 – Gráfico da questão 1 – Profissão exercida dos respondentes

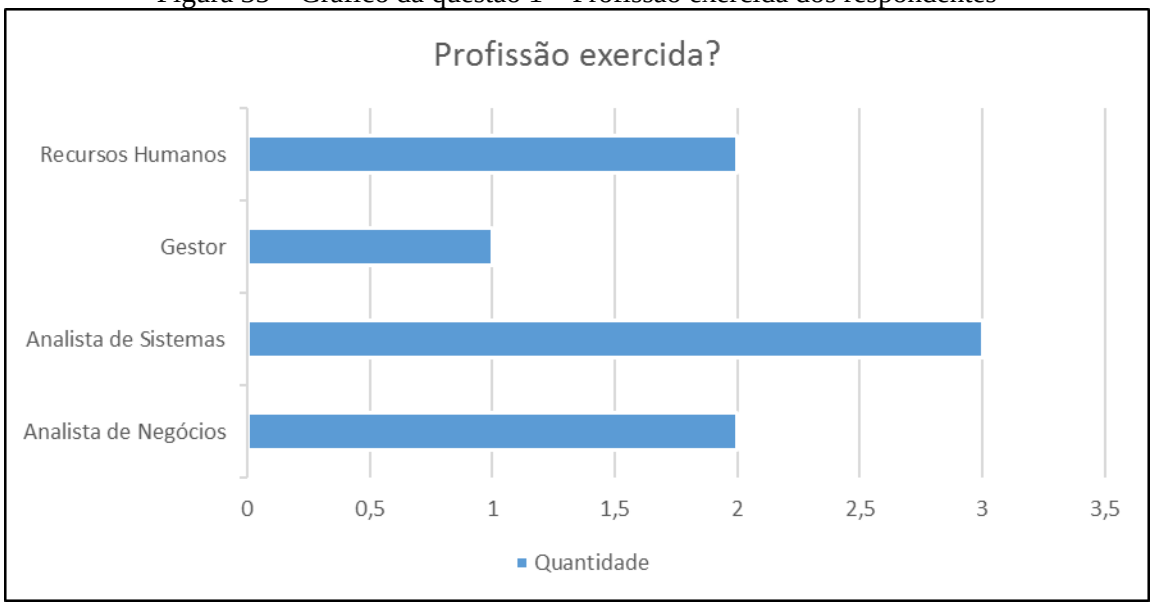
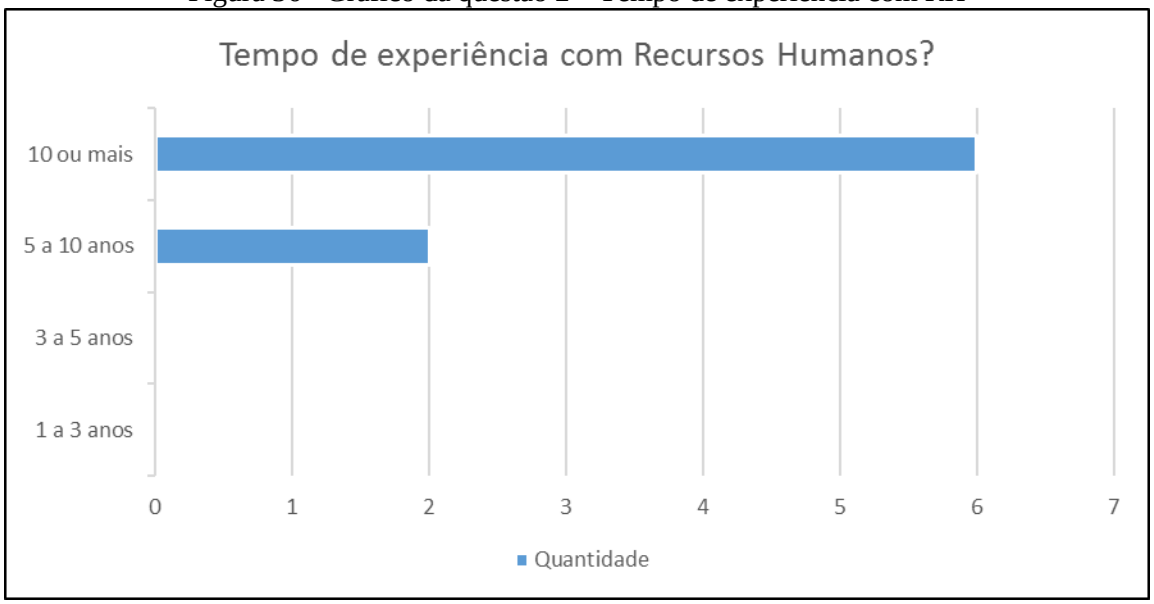
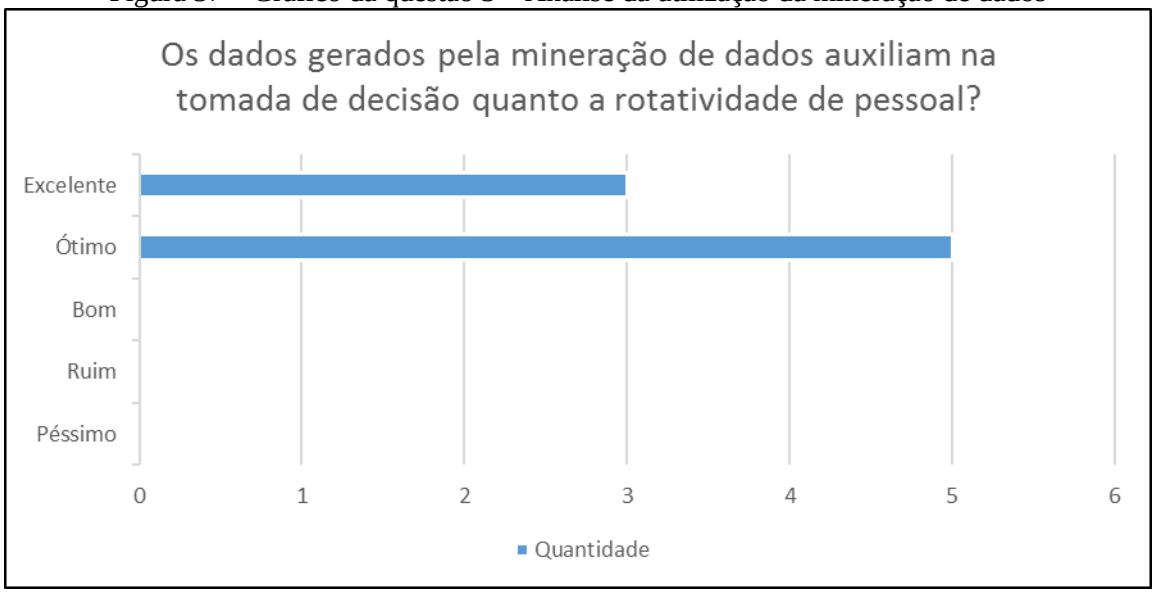


Figura 36 - Gráfico da questão 2 – Tempo de experiência com RH



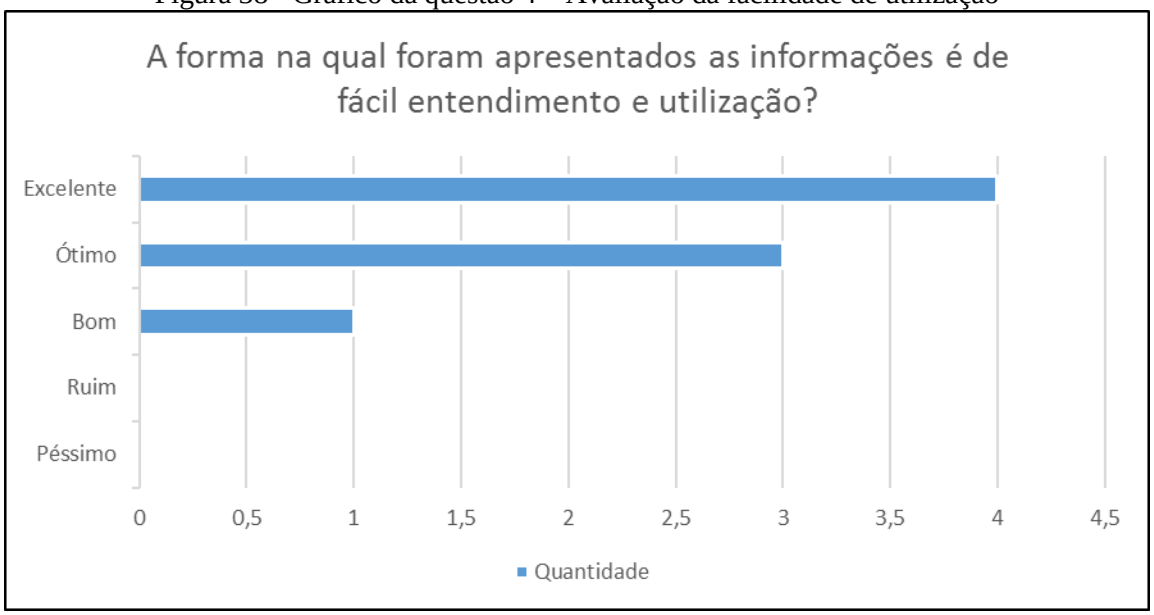
A Figura 37 demostra o gráfico gerado a partir das respostas da questão 3, sendo possível verificar que o processo de mineração de dados realizado pela ferramenta sobre sistema RH oferece informações que auxiliam na tomada de decisão no âmbito da rotatividade de pessoal.

Figura 37 – Gráfico da questão 3 – Análise da utilização da mineração de dados



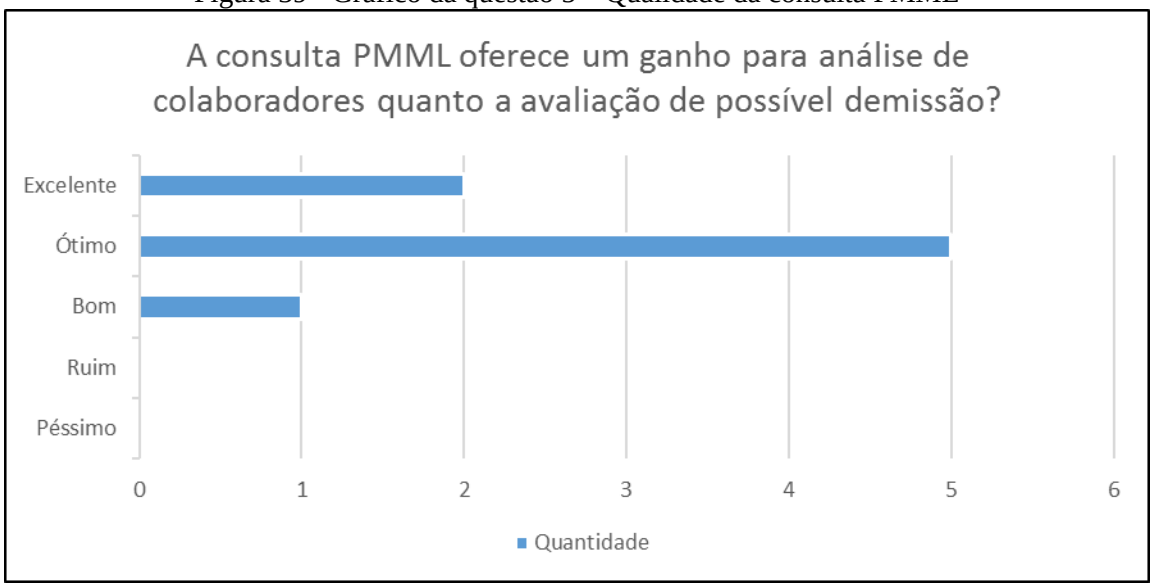
A questão 4 tinha o intuito de avaliar se as informações mineiradas na ferramenta são apresentadas de maneira que propiciam fácil compreensão, seja pelo gráfico de níveis interativo, gráfico de árvore decisão ou nuvem de palavras. Conforme pode ser observado na Figura 38 os respondentes consideraram que a forma no qual os dados são apresentados prove uma qualidade no entendimento das informações.

Figura 38 - Gráfico da questão 4 – Avaliação da facilidade de utilização



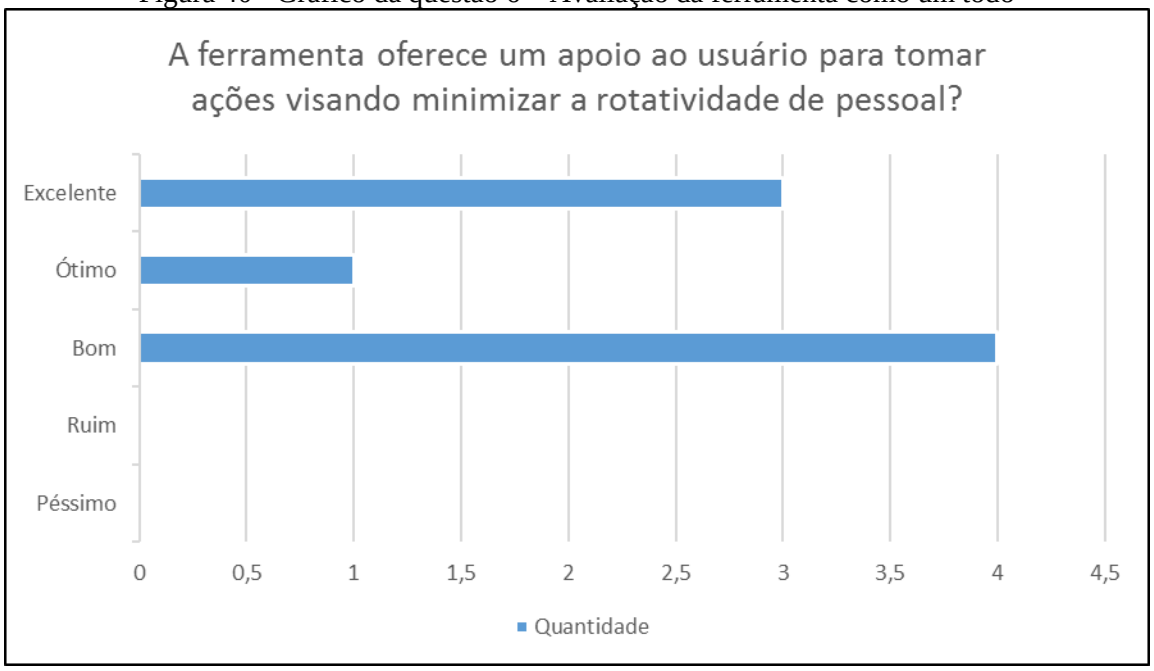
O objetivo da questão 5 era avaliar a efetividade da consulta PMML como suporte a análise de possível demissão, seja sobre a consulta de um candidato para uma nova vaga na empresa ou sobre colaboradores que trabalham na empresa. Através das respostas obtidas conforme Figura 39, fica evidenciado que a consulta alcançou seu objetivo de auxiliar na tomada de decisão quanto a possíveis casos de demissão, visando minimiza-los.

Figura 39 - Gráfico da questão 5 – Qualidade da consulta PMML



A questão 6 visava avaliar a ferramenta como um todo (Telas, gráficos, cadastros, entre outros) quanto a sua efetividade para auxiliar na tomada decisões sobre a rotatividade de pessoal. Como pode ser visto na Figura 40 a ferramenta foi avaliada positivamente por todos os questionados, demonstrando que a mesma alcançou o seu objetivo de auxiliar sobre análise de rotatividade de pessoal.

Figura 40 - Gráfico da questão 6 – Avaliação da ferramenta como um todo



4 CONCLUSÕES

O trabalho conseguiu alcançar o objetivo proposto. Os resultados obtidos se mostraram válidos e condizentes com o cenário real dos Recursos Humanos, sendo capaz de analisar os mais diferentes tipos de dados contidos em um sistema RH. Através desta análise gerando uma árvore de decisão que auxilia na tomada de decisão referente a rotatividade de pessoal dentro da empresa.

Com o processo de árvore de decisão foi possível encontrar grupos que possuem maior tendência de requerer a demissão em uma empresa, desta forma, concretizando o primeiro objetivo específico. Pois, com a capacidade de se encontrar grupos de risco, permite-se que sejam realizadas ações. Seja para evitar a contratação de pessoas que se encaixam neste padrão ou para melhorar a condição deste grupo, visando equalizar a probabilidade de demissão com relação aos demais grupos da empresa.

A ferramenta inclui um novo conceito, a inclusão de atributo obtido na mineração de texto, adicionado ao o algoritmo C4.5 (Figura 25), permitindo uma análise mais profunda sobre os dados existentes na base de dados, desta forma, legitimando o segundo objetivo específico. Pois com isto, foi possível incluir informações existentes nas entrevistas demissionais na geração da árvore de decisão, conforme pode ser verificado na Figura 29. Permitindo assim, se encontrar padrões não possíveis pelo algoritmo C4.5 padrão. Em adição a mineração de texto, a ferramenta permitiu agrupar as palavras em 3 níveis (Figura 25), permitindo um maior controle e lógica sobre as palavras encontradas.

Outra funcionalidade importante foi a geração do PMML sobre a indução de árvore. Permitindo assim, percorrer os níveis da árvore de modo a predizer um status conforme informações previamente preenchidas. Desta forma a permitir que possa ser realizado, por exemplo, análise sobre currículos, verificando a probabilidade que o candidato possuí de demissão dentro da empresa, podendo utilizar este fator como corte prévio de candidatos. Através desta funcionalidade foi possível concretizar o terceiro objetivo especifico do trabalho.

Foi realizada uma avaliação sobre a ferramenta com pessoas experientes em RH (75% possuíam 10 ou mais anos de experiência com RH) e dos mais diferentes ramos de atuação. Sendo que através da análise sobre os resultados da avaliação foi possível verificar que 50% responderam como “Bom”, 12% “Ótimo” e 37,5% “Excelente” o apoio que a ferramenta oferece para a tomada de decisão visando minimizar a rotatividade de pessoal, com isto o último objetivo especifico foi alcançado. Pois, ficou evidenciado a efetividade que a

ferramenta possui para auxiliar na tomada de decisão. No qual, através dos gráficos apresentados pela ferramenta, foi possível realizar uma análise de forma simples e clara quanto aos grupos com maior chance de demissão dentro do sistema de RH.

Destaca-se como limitação o problema da utilização de mineração de texto como atributo na árvore de decisão, quando os registros analisados possuem muitos desligamentos sem a entrevista de desligamento informada, acarretando com isso em muita informação nula. Desta forma, existe um problema para se encontrar a *top word* e dificultando no cálculo de ganho do atributo no algoritmo C4.5, pois a informação nula acaba gerando uma informação de ganho incorreta.

4.1 EXTENSÕES

A ferramenta pode ser melhorada, ampliada ou incrementada em futuros trabalhos sendo agregado novas funcionalidades ou novas técnicas como:

a) utilizar árvore de decisão difusa para melhorar a acurácia em atributos contínuos;
b) adicionar funcionalidade para determinar a quantidade mínima de nós por nível, para o mesmo ser considerado relevante;
c) criar *WebService* para permitir que o sistema RH possa consumir as funções existentes na ferramenta;
d) aplicar a ferramenta em sistemas de outras áreas;
e) permitir mais de um atributo do tipo mineração de texto em uma mesma árvore.

REFERÊNCIAS

AGGARWAL, Charu C.; ZHAI, ChengXiang. **Mining Text Data**. Nova York: Springer Science & Business Media, 2012. 524 p, il.

Associação Brasileira das Empresas de Software. **Mercado Brasileiro de Software**: panorama e tendências. São Paulo: Associação Brasileira das Empresas de Software, v. 1, jun. 2015. Disponível em: <http://central.abessoftware.com.br/Content/UploadedFiles/Arquivos/Dados 2011/ABES-Publicacao-Mercado-2015-digital.pdf>. Acesso em: 2 set. 2015.

AUDY, Jorge Luis Nicolas; BRODBECK, Ângela Freitag. **Sistemas de Informação**: Planejamento e alinhamento estratégico nas organizações. 2. ed. Porto Alegre: Bookman Editora, 2009. 160 p.

BERRY, Michael J A; LINOFF, Gordon S. **Data Mining Techniques**: For Marketing, Sales, and Customer Relationship Management. Indianapolis: John Wiley & Sons, 2004. 643 p, il.

BHUIYAN, Faruk; CHOWDHURY, Mustafa Manir; FERDOUS, Farzana. Historical Evolution of Human Resource Information System (HRIS): An Interface between HR and Computer Technology. **Human Resource Management Research**, v. 4, n. 4, p. 75-80, 2014.

BINOTTO, Erlaine; NAKAYAMA, Marina Keiko; PILLA, Bianca Smith. **e-RH**: conceitos e práticas de RH eletrônico. Passo Fundo: Ed. UPF, 2006. 320 p, il.

BORGES, Paulo Sérgio da Silva; JUSTINO, Gilvan; RATKE, Cláudio. Uso da Função Sigmóide para Pertinência em Árvores de Decisão Difusas In: II ECTEC - Encontro de Ciência e tecnologia, 2003, Lages. **II ECTEC - Encontro de Ciência e tecnologia**, 2003.

BOSTOCK, Michael; HEER, JEFFREY; OGIEVETSKY, Vadim. **D3**: Data-Driven Documents. Stanford: Computer Science Department of Stanford University, 2011. Disponível em: <http://vis.stanford.edu/files/2011-D3-InfoVis.pdf>. Acesso em: 11 out. 2015.

CAROL, Damien. **PMML4Net**. [S.l.], 2013. Disponível em: <http://staging.nuget.org/packages/pmml4net/>. Acesso em: 28 ago. 2015.

CHEN, Juhua; PENG, Wei; ZHOU, Haiping. **An Implementation of ID3 --- Decision Tree Learning Algorithm**. Sydney, 2009. Disponível em: <http://web.arch.usyd.edu.au/~wpeng/DecisionTree2.pdf>. Acesso em: 14 ago. 2015.

CHIEN, Chen-Fu; CHEN, Li-Fei. **Data mining to improve personnel selection and enhance human capital**: A case study in high-technology industry. [S.l.], 2008. Disponível em: <http://www.researchgate.net/publication/223190807_Data_mining_to_improve_personnel_selection_and_enhance_human_capital_A_case_study_in_high-technology_industry>. Acesso em: 15 ago. 2015.

Data Mining Group. **PMML 4.2 Tree Models**. [S.l.], 2014. Disponível em: <http://dmg.org/pmml/v4-2-1/GeneralStructure.html>. Acesso em: 01 set. 2015.

DE ALMADA, Valéria Ferreira et al. Gestão de desempenho por competências: integrando a gestão por competências, o balanced scorecard e a avaliação 360 graus. **Revista de Administração Pública**, v. 42, n. 5, p. 875-898, 2008.

DESANCTIS, Gerardine. **Human Resource Information Systems**: A Current Assessment. [S.l.], 1986. Disponível em: < http://aisel.aisnet.org/misq/vol10/iss1/1/>. Acesso em: 28 ago. 2015.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, Brasília, v. 17, n. 1, 1996. Disponível em: <http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf>. Acesso em: 27 ago. 2015.

FRACALANZA, Livia Fonseca. **Mineração de Dados voltada para Recomendação no Âmbito de Marketing de Relacionamento**. 2009. 59 f. Dissertação de Mestrado (Mestrado em Ciência da Computação) – Pontifícia Universidade Católica do Rio de Janeiro, Faculdade de Tecnologia. Departamento de Ciências da Computação. Rio de Janeiro, 2009.

GRAHAM, Shawn; MILLIGAN, Ian; WEINGART, Scott. **Exploring Big Historical Data**: The Historian's Macroscope. Londres: Imperial College Press, 2015. 308p. Disponível em: <http://www.themacroscope.org/?page_id=362>. Acesso em: 30 set. 2015.

GUAZZELLI, Alex et al. PMML: An Open Standard for Sharing Models. **The R Journal**, Youngstown, v. 1, n. 1, 2009. Disponível em: < http://journal.r-project.org/archive/2009-1/RJournal_2009-1_Guazzelli+et+al.pdf>. Acesso em: 27 ago. 2015.

HENDRICKSON, Anthony R. Human resource information systems: Backbone technology of contemporary human resources. **Journal Of Labor Research.** Ames, p. 381-394. set. 2013. Disponível em: <http://link.springer.com/article/10.1007/s12122-003-1002-5>. Acesso em: 26 ago. 2015.

KANTARDZIC, Mehmed. **Data Mining**: Concepts, Models, Methods, and Algorithms. Nova Jersey: John Wiley & Sons, 2011. 520 p, il.

KAVANAGH, Michael J; THITE, Mohan. Evolution of Human Resource Management and Human Resource Information Systems: The Role of Information Technology. In **Human Resource Information Systems**; KAVANAGH, Michael J.; THITE, Mohan; JOHNSON, Richard D. (orgs.). SAGE Publications, Inc. 2015. p. 1-24.

LIMAN, Ahmed. **Como a Tecnologia em Recursos Humanos Pode Ajudar as Empresas Brasileiras a Terem Mais Sucesso**. [S. l.], 2011. Disponível em: <http://www.rhevistarh.com.br/portal/?p=2754>. Acesso em: 26 ago. 2015.

LIU, Bing; ZHANG, Lei. A Survey of opinion mining and sentiment analysis. In: AGGARWAL, Charu C.; ZHAI, Chengxiang. **Mining Text Data**. Nova York: Springer Us, 2012. p. 415-463. Disponível em: <http://www.cs.unibo.it/~montesi/CBD/Articoli/SurveyOpinionMining.pdf>. Acesso em: 5 set. 2015.

MAGERMAN, David M. Statistical Decision-Tree Models for Parsing. In: ANNUAL MEETING OF THE ASSOCATION FOR COMPUTATIONAL LINGUISTICS, 33., 1995, Cambridge. **Proceedings of the 33rd Annual Meeting of the Assocation for Computational Linguistics**. Stroudsburg: Association For Computational Linguistics, 1995. p. 276 - 283. Disponível em: <http://www3.nd.edu/~dchiang/papers/acl00-1.pdf>. Acesso em: 7 set. 2015.

MAIMON, Oded; ROKACH, Lior. **Data Mining with Decision Trees – Theory and Applications**. Singapura: World Scientific Publishing Company, 2007. 264 p.

MARCACINI, Ricardo M.; MOURA, Maria F.; REZENDE, Solange O. O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento. **Revista de Sistemas de Informação da FSMA**, Macaé, v. 1, n. 7, 2011. Disponível em: <http://www.fsma.edu.br/si/edicao7/FSMA_SI_2011_1_Principal_3.pdf>. Acesso em: 10 set. 2015.

MELO, Marcelo Damasceno. **Um processo de mineração de dados para predição de níveis criminais de áreas geográficas**. 2010. 129 f. Dissertação de Mestrado (Mestrado em Ciência da Computação) – Universidade Estadual do Ceará, Faculdade de Tecnologia. Centro de Ciências Científicas. Ceará, 2010.

MENDES, Alessandro de Souza. **Aplicação de técnicas de data mining na caracterização de turnover interno para o suporte à gestão de pessoas**. 2013. 103 f. Dissertação de Mestrado (Mestrado em Engenharia Elétrica) – Universidade de Brasília, Faculdade de Tecnologia. Departamento de Engenharia Elétrica. Brasília, 2013.

O'BRIEN, James A; MOREIRA, Célio Knipel. **Sistemas de informação:** e as decisões gerenciais na era da internet. 2. ed. São Paulo : Saraiva, 2004. xxiii, 431p, il. Tradução de: Introduction to information systems.

PALIT, Ajoy K.; POPOVIC, Dobrivoje. **Advances in Industrial Control** (Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications). Londres: Springer, 2005. 381 p, il.

PARRY, Emma. **HR and Technology**: Impact and Advantages. Londres: Chartered Institute of Personnel, 2007. 43 p, il.

QUILICI-GONZALEZ, José Artur; ZAMPIROLLI, Francisco de Assis. **Sistemas Inteligentes e Mineração de Dados**. Santo André: Triunfal Gráfica e Editora, 2015. 150 p. Disponível em: <https://books.google.com.br/books?id=X76VBgAAQBAJ&dq=poda+minera%C3%A7%C3%A3o+de+dados&hl=pt-PT&source=gbs_navlinks_s>. Acesso em: 30 set. 2015.

QUINLAN, J. R. Improved Use of Continuous Attributes in C4.5. **Journal of Artificial Intelligence Research 4**, El Segundo, v. 4, n. 1, 1996. Disponível em: <https://www.jair.org/media/279/live-279-1538-jair.pdf>. Acesso em: 13 set. 2015.

QUINLAN, J. R. Induction of Decision Trees. **Machine Learning**. 1986. v. 1, 1 ed., p. 81-106. Disponível em: <http://hunch.net/~coms-4771/quinlan.pdf>. Acesso em: 18 ago. 2015.

QUINLAN, J. R. **C4.5 Programs for Machine Learning**. San Francisco: Morgan-Kauffman, 1993. 302 p, il.

QUINTELLA, Rogério Hermida; SOARES JUNIOR, Jair Sampaio. Descoberta de conhecimento em bases de dados públicas: uma proposta de estruturação metodológica. In **Avaliação e sociedade**: a negociação como caminho [online]; Tenório, Robinson Moreira; Vieira, Marcos Antônio (orgs.). Salvador: EDUFBA, 2009. p. 165-201.

REY, Tim; KORDON, Arthur; WELLS, Chip. **Applied Data Mining for Forecasting Using SAS**. Cary : SAS Institute, 2012. 336 p, il.

REZENDE, Solange Oliveira. **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Editora Manole Ltda, 2003. 525 p, il.

RILEY, Jim. **ICT – types of information system**. [S. l.], 2012. Disponível em: <http://tutor2u.net/business/ict/intro_information_system_types.htm>. Acesso em: 11 set. 2015.

RUSSEL, Hasan Shahriar. **Case Studies**. [S.l.], 2013. Disponível em: <http://www.scribd.com/doc/173623865/Case-Studies>. Acesso em: 28 ago. 2015.

Série estudos. **Série Estudos Tecnologia - 2013**. Rio de Janeiro: Série estudos, v. 13, set. 2013. Disponível em: <http://data.axmag.com/data/201310/20131027/U93356_F247019/FLASH/index.html>. Acesso em: 3 set. 2015.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Datamining**: mineração de dados. Rio de Janeiro: Ciência Moderna, 2009. xxi, 900 p, il.

TIMOFEEV, Roman. **Classification and Regression Trees (CART) Theory and Applications**. 2004. 40 f. Dissertação (Pós-graduação em Ciência da Computação) –Center of Applied Statistics and Economics, Universidade de Humbold. Berlim.

UNIVERSAL RH EXPLORER. **Universal RH Explorer**. [S.l.], 2015. Disponível em: <http://www.universalrh.com.br/solucoes/explorer/index.html>. Acesso em: 15 ago. 2015.

WIVES, Leandro Krug. **Tecnologias de descoberta de conhecimento em textos aplicados à inteligência competitiva**. 2002. 166 f. Dissertação (Pós-graduação em Computação) - Curso de Pós-Graduação em Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre.

APÊNDICE A – Descrição dos Casos de Uso

Este apêndice apresenta a descrição dos principais casos de uso descritos na seção de especificação deste trabalho. No Quadro 15 estão descritos os casos de uso UC02, UC03 e UC07.

Quadro 15 - Descrição dos Casos de Uso

UC02 - Cadastrar informações árvore
Permite ao usuário cadastrar informações para indução da árvore geradora. Possibilitando a inserção do nome, data de criação, nível da poda, descrição e o comando SQL a ser utilizado para buscar informações na base RH.

Constraints
Pré-condição. Usuário deve ter realizado o cadastro de conexão para o banco de dados RH.
Pós-condição. Uma nova configuração de árvore foi incluída, alterada ou excluída na ferramenta.

Cenários
Incluir informação árvore {Principal}.
1. Usuário solicita a inserção de uma árvore;
2. Sistema apresenta tela para inserção de uma árvore;
3. Usuário efetua o cadastramento das devidas informações;
4. Sistema grava a configuração da árvore;

Alterar configuração árvore {Alternativo}.
No passo 1, o usuário opta por alterar uma árvore existente.
1.1. Usuário seleciona uma árvore que deseja ser alterar.
1.2. Usuário efetua a alteração na árvore.
1.3. Ferramenta grava a árvore.

Excluir configuração árvore {Alternativo}.
No passo 1, o usuário seleciona uma árvore.
1.1. Ferramenta apresenta os dados para exclusão.
1.2. Usuário confirma a exclusão.
1.3. Ferramenta exclui a árvore.

Obrigatoriedade de campos no bloco {Exceção}.
No passo 3, o usuário não informa todas as informações obrigatórias a serem cadastrados. A ferramenta informa a mensagem de inconsistência, sendo necessário informar tais informações.

Comando SQL incorreto {Exceção}.
No passo 3, caso o comando SQL esteja incorreto. A ferramenta informa a mensagem de inconsistência, sendo necessário corrigir o comando SQL.

Existe gráfico gerado para a árvore {Exceção}.
No passo 1.3, caso exista gráfico gerado para a árvore. A ferramenta retorna uma inconsistência informando que não é possível realizar a alteração/exclusão.

UC03 – Configurar atributos árvore geradora
Permite ao usuário configurar os atributos da árvore para indução da mesma, determinando qual é a classe meta, tipo (categórico, contínuo ou mineração de texto) e uma legenda para apresentação do atributo nos gráficos.

Constraints
Pré-condição. Usuário deve ter realizado o cadastro da configuração da árvore.
Pós-condição. Uma nova configuração dos atributos da árvore foi incluída, alterada ou excluída na ferramenta.

Cenários
Incluir configurações atributos {Principal}.
1. Usuário solicita a configuração dos atributos;
2. Ferramenta apresenta tela para configuração dos atributos;
3. Usuário efetua o cadastramento das devidas informações;
4. Ferramenta grava a configuração dos atributos;
5. Ferramenta gera os gráficos;

Alterar configuração atributos {Alternativo}.
No passo 1, o usuário opta por alterar a configuração dos atributos existentes.
1.1. Usuário seleciona uma árvore na qual deseja alterar os atributos.
1.2. Usuário efetua a alteração nos atributos.
1.3. Ferramenta grava os atributos.
1.4. Ferramenta gera os gráficos;

Classe meta não definida {Exceção}.
No passo 3, o usuário não informa um atributo como classe meta. A ferramenta informa a mensagem de inconsistência, sendo necessário informar uma classe meta.

UC07 – Realizar consulta PMML
Permite ao usuário realizar consulta preditivas sobre as árvores geradas, onde a cada atributo preenchido é executado o PMML e exibido um gráfico da árvore percorrida com sua probabilidade.

Constraints
Pré-condição. Usuário deve ter realizado a indução da árvore.
Pós-condição. Exibido para o usuário probabilidade da sua consulta e um gráfico de árvore.

Cenários
Realização da consulta PMML {Principal}.
1. Usuário solicita a árvore que deseja realizar a consulta;
2. Ferramenta apresenta tela para preenchimento do atributo;
3. Para cada atributo novo atributo disponível para preenchimento:
4. 1 - Usuário realiza o preenchimento do atributo;
5. 2 - Usuário clica no comando “Gerar probabilidade”;
6. 3 - Ferramenta apresenta a probabilidade e gráfico de árvore;
7. 4 – Usuário clica no comando “Avançar”;

Não avançar consulta {Alternativo}.
No passo 7, o usuário opta por não avançar a consulta.
1.1. Usuário consulta a probabilidade gerada pela ferramenta;

Cancelar consulta {Alternativo}.
No passo 7, o usuário opta por cancelar a consulta.
1.1. Usuário clica no comando “Cancelar” e para a consulta;

Atributo com valor incorreto {Exceção}.
No passo 5, o usuário informa o atributo de forma incorreta. A ferramenta informa a mensagem de inconsistência, sendo necessário informar o valor do atributo novamente.

Consulta não retornou resultado {Exceção}.
No passo 5, a ferramenta não encontrou resultado na consulta. A ferramenta informa a mensagem de inconsistência, sendo necessário cancelar a consulta.

APÊNDICE B – Formulário de avaliação da ferramenta desenvolvida

Conforme descrito na seção 3.4 foi elaborado um formulário para avaliar a ferramenta desenvolvida. A Figura 41 e Figura 42 mostram o formulário que foi aplicado.

Figura 41 – Primeira parte formulário de avaliação

Ferramenta para predição de dados proeminentes de sistemas RH

Este formulário tem por objetivo avaliar os resultados obtidos através da ferramenta.

* Required

Profissão exercida? *

- Analista de Sistemas
- Analista de Negócios
- Recursos Humanos
- Gestor

Tempo de experiência com Recursos Humanos? *

- 1 a 3 anos
- 3 a 5 anos
- 5 a 10 anos
- 10 ou mais

Os dados gerados pela mineração de dados auxiliam na tomada de decisão quanto a rotatividade de pessoal? *

Análise sobre os dados gerados e apresentados pelos gráficos, avaliando a efetividade destes.

- Péssimo
- Ruim
- Bom
- Ótimo
- Excelente

A forma na qual foram apresentados as informações é de fácil entendimento e utilização? *

Avaliação quanto ao gráfico interativo de níveis, árvore e nuvem de palavras.

- Péssimo
- Ruim
- Bom
- Ótimo
- Excelente

Figura 42 - Segunda parte formulário de avaliação