CENTRO UNIVERSITÁRIO FEEVALE

INGO JOST

MINERAÇÃO DE DADOS PARA ADOÇÃO DE PRÁTICAS DE MARKETING EM AMBIENTE ACADÊMICO

Novo Hamburgo, junho de 2009.

INGO JOST

MINERAÇÃO DE DADOS PARA ADOÇÃO DE PRÁTICAS DE MARKETING EM AMBIENTE ACADÊMICO

Centro Universitário Feevale
Instituto de Ciências Exatas e Tecnológicas
Curso de Ciência da Computação
Trabalho de Conclusão de Curso

Professor Orientador: Juliano Varella de Carvalho

Novo Hamburgo, junho de 2009.

AGRADECIMENTOS

Gostaria de agradecer a todos os que, de alguma maneira, contribuíram para a realização desse trabalho de conclusão, em especial:

Aos amigos que sempre pude contar, à minha família, à minha namorada Marilaine, minha gratidão, pelo apoio emocional. .

Agradeço também ao professor e amigo Juliano pela orientação e dedicação.

RESUMO

Este trabalho apresenta o projeto que visa utilizar a área de *Data Mining*, disseminada nos mais diversos segmentos, utilizada por empresas que buscam descobrir o conhecimento que se encontra oculto em suas próprias bases de dados a fim de possibilitar um diferencial a seus concorrentes, criando, desta forma, uma aplicação destinada a uma instituição de Ensino Superior. A base de dados desta instituição é composta por informações de vestibulandos, como município de residência, renda familiar, sexo, idade, entre outros. Dentre as diversas técnicas de *Data Mining*, a que será utilizada é a de Classificação, que possui diferentes algoritmos, em que os dados serão tratados e após a aplicação desses algoritmos, serão classificados em perfis pré-definidos. O objetivo é encontrar relação entre os perfis de vestibulandos com a demanda por cursos. Assim, espera-se que, com o conhecimento adquirido, seja possível ao departamento de Marketing adotar práticas direcionadas a áreas ou cursos, com o objetivo de atrair mais alunos à instituição, o que seria um diferencial em um mercado cada dia mais disputado.

Palavras-chave: *Data Mining*. Técnicas de Classificação. Marketing de precisão. Mapeamento de Perfis.

ABSTRACT

This work presents the project that will use Data Mining, spread on many segments, used for companies that aim to know the discovery that has been hidden in their databases to differ from their rivals, to build a application to high school institution. Its database is composed for students that have made vestibular's information, like address, familiar rent, gender, age and others. Among the Data Mining's techniques, the classification technique that will be used. Its has some algorithms. The data will be prepared to begin of application these algorithms and it will be classified into pre-defined profiles. The goal is to find relationship between the profiles and demand for courses. Thus, it is expected that with the knowledge, it is possible for the department of Marketing adopt practices directed to areas or courses, aiming to attract more students to the institution, which would be a gap in the market each day over disputed

Key words: Data Mining. Classification Technique. Precision Marketing. Discovery Profiles.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	Application Programming Interface
BD	Banco de Dados
BI	Business Intelligence
CART	Classification and Regression Tree
CHAID	CHi-squared Automatic Interaction Detector
CIM	Customer Interaction Management
CRM	Customer Relationship Management
FISEM	Final de Semana
GUI	Graphical User Interface
IA	Inteligência Artificial
IDE	Integrated Development Environment
JAR	Java Archive
JDBC	Java Database Connectivity
JDK	Java Standard Edition Development Kit
KDD	Knowledge Database Discovery
OSBI	Open Source Business Intelligence
RNA	Redes Neurais Artificiais
SGBD	Sistema Gerenciador de Banco de Dados
SQL	Structured Query Language
TDIDT	Top-Down Induction of Decision Trees
TI	Tecnologia em Informação
XML	eXtensible Markup Language

SUMÁRIO

INTRODUÇÃO

As corporações têm buscado, no decorrer das últimas décadas, digitalizar os seus dados e adquirir ou desenvolver sistemas de informação que manipulam e alimentam essas bases. No entanto, este gerenciamento das informações é utilizado normalmente apenas para a gestão e otimização de processos, não sendo aproveitado para uma análise mais detalhada, capaz de propiciar novas oportunidades às instituições.

Com esse objetivo, surge a área de Descoberta de Conhecimento em Bases de Dados, cujo nome é originado do inglês *Knowledge Discovey in Databases* (KDD). Segundo Goldschmidt e Passos (2005), é um processo composto pelas seguintes etapas operacionais:

- Pré-processamento: etapa em que os dados recebidos passam por uma preparação para a próxima etapa. Esta preparação consiste na seleção, excluindo os dados irrelevantes e os que possuem inconsistências e transformação destes, sendo realizadas conversões e adaptações;

- Mineração de Dados: quando é realizada a busca propriamente dita pelo conhecimento. Existem diversos algoritmos que implementam diferentes técnicas para a obtenção do conhecimento, destacando-se Associação, Classificação e Agrupamento;

- Pós-processamento: aproveitamento das informações adquiridas, sendo realizada a interpretação e avaliação da importância do conhecimento descoberto, se houver.

A Mineração de Dados é utilizada nas mais diversas áreas, desde a descoberta de pesos de atributos em um sistema de Raciocínio Baseado em Casos, conforme Silveira (2003), até, de acordo com Neves (2003), a definição de padrões em pacientes de Diabetes.

Independente do foco, são utilizadas as técnicas já citadas, inclusive os mesmos algoritmos. Dentre estas técnicas, destacam-se:

- Agrupamento: procura separar em grupos dados similares;
- Classificação: consiste em classificar os registros em categorias (classes) pré-definidas;
- Associação: busca associações entre atributos em diferentes transações.

Existem diversas ferramentas que implementam as técnicas de KDD, destacando-se Sipina, desenvolvida pela Universidade de Lyon e Weka, sendo ambas utilizadas neste projeto. Através das técnicas e tecnologias apresentadas, tem-se como motivação deste trabalho a criação de uma ferramenta que extraia conhecimento da base de dados de um centro universitário. Esta instituição, que conta com milhares de alunos, procura uma solução para melhor distribuir seus esforços em relação ao Marketing, área que vem sendo bastante explorada por aplicações de *Data Mining*, como verificado em *KDNuggets*[1], portal referência em KDD.

Dentre os exemplos na área, destacam-se processos de relacionamento de clientes, conforme Almeida (2005), ou de mapeamento de perfis, como a empresa *Sigma*[2] que possui uma ferramenta com diversos canais de comunicação para que as empresas conheçam as características de seus consumidores e o *Talisma Knowledgebase Software*[3], que permite direcionar o contato da empresa com o cliente, a partir de transações já realizadas, evitando gastos com comunicação.

Este projeto manipulará informações pessoais de alunos, procurando peculiaridades comuns em, por exemplo: escolha de curso, município de residência, idade, escolas freqüentadas. Serão mapeados perfis de vestibulandos para que se encontre associações entre as características dos alunos e a escolha por cursos, possibilitando ao departamento de Marketing adotar práticas direcionadas para institutos, como o de Ciências Exatas, ou especificamente para cursos.

[1] http:// www.kdnuggets.com/
[2] http://www.sigmamarketing.com/
[3] http://www.talisma.com/tal_products/knowledgebase.aspx

Fica a total critério da equipe de Marketing a forma de aproveitamento do conhecimento adquirido, juntamente com as práticas, que podem ser desde o direcionamento de propagandas até o agendamento de visitas de apresentação conforme a escola ou região de maior procura pelo curso (ou menor).

No presente trabalho, será apresentada no Capítulo 1 a área de Descoberta do Conhecimento e algumas de suas técnicas, das quais a de Classificação é detalhada no Capítulo 2. No Capítulo 3 é explicado o Estudo de Caso adotado para o desenvolvimento da Aplicação, sendo esta abordada no Capítulo 4. O último capítulo é referente aos Resultados encontrados após o processo de KDD para finalmente serem destacadas as conclusões e trabalhos futuros.

1 DESCOBERTA DE CONHECIMENTO

A área de Descoberta de Conhecimento em Bases de Dados, cujo nome é originado do inglês *Knowledge Discovey in Databases* (KDD), busca através de dados já cadastrados, aplicar determinadas técnicas que extraem conhecimento. Este conhecimento pode ser utilizado na tomada de decisões ou posicionamento estratégico das instituições.

Dada a grande distinção entre as técnicas, identifica-se na Descoberta de Conhecimento, de acordo com Freitas (1998), uma grande interdisciplinaridade, envolvendo pelo menos três grandes áreas: Estatística, Inteligência Artificial e Banco de Dados. O processo de KDD é composto por diversas etapas, sendo ilustradas na Figura 1.1

Figura 1.1: Etapas do processo de KDD
Fonte: Romani , 2008.

Segundo Goldschmidt e Passos (2005), a Descoberta do Conhecimento é composta pelas etapas operacionais: Pré Processamento, Mineração de Dados e Pós-Processamento.

1.1 Pré Processamento

Dentre as etapas da Figura 1.1, Seleção, Processamento e Transformação se enquadram na fase de Pré-Processamento. A seleção consiste na filtragem dos dados, excluindo aqueles irrelevantes, por exemplo, no mapeamento de perfis de clientes, provavelmente nome e CPF são atributos irrelevantes.

O processamento consiste em executar operações como as de limpeza de dados, evitando inconsistências decorrentes de erros de cadastro (salário ou idade com valores negativos) e tratamento de campos com grande ocorrência de valores nulos, por serem dados não disponíveis ou não informados. É necessário avaliar se devem ser desconsiderados ou receber um valor padrão.

Em relação à transformação, conforme Mongiovi (1998), destaca-se a conversão de dados (para aqueles dados com o mesmo significado, porém em formatos diferentes, por exemplo, o campo *sexo* sendo representado ora por números 1 e 2, ora por caractere ‘M’ e ‘F’), criar categorias para variáveis contínuas (categorias de faixas etárias, ao invés de idade), converter variáveis nominais em numéricas, usar escalas de redução / ampliação (normatizar valores com unidades de medida diferentes) e criar novas variáveis. Estes processos são necessários em maior intensidade quando os dados são oriundos de diferentes bases ou até de diferentes formatos, como banco de dados relacional e documentos em *eXtensible Markup Language* (XML).

Além disso, a transformação dos dados é necessária para que estes fiquem no formato de entrada dos algoritmos de *Data Mining*, como por exemplo, utilizar consultas *Structured Query Language* (SQL) e a partir do resultado executar procedimentos que criam arquivos neste formato. A ferramenta Weka, que será apresentada posteriormente, tem como entrada de dados, arquivos no formato .ARFF (*Attribute-Relation File Format*).

```
@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```

Figura 1.2: Exemplo de Arquivo .arff
Fonte: Santos, 2005

1.2 Mineração de Dados

Etapa mais importante da Descoberta de Conhecimento, segundo Goldschmidt e Passos (2005). Sua relevância é tamanha, que o termo Mineração de Dados confunde-se com KDD. É nesta etapa que se realiza a busca propriamente dita pelo conhecimento, extraindo padrões dos dados. Existem diversos algoritmos que implementam diferentes técnicas para a obtenção do conhecimento. No entanto, somente a utilização das técnicas não é suficiente para atingir os resultados esperados. Para utilização de KDD, é necessário o envolvimento de um especialista na área desde o momento da identificação do problema, passando por todas as etapas descritas, juntamente com especialistas do domínio da aplicação.

Entre as principais técnicas de Mineração de Dados, destacam-se: Associação, Classificação e Agrupamento.

1.2.1 Associação

Procura por associações entre os atributos de um conjunto de dados, conforme Witten e Frank (2000). O exemplo mais comum para ilustrar esta técnica é o de transações de compras: neste caso obtém-se regras de associação com a ocorrência de produtos em diferentes transações. Por exemplo, um mercado pode auferir que 80% dos compradores de

pão, também compram leite. A partir destes conhecimentos descobertos, diversas ações podem ser tomadas, como por exemplo, deixar esses produtos próximos (ou distantes), promover a venda conjunta dos produtos, reduzir o preço de um deles e aumentar o preço do outro, dentre outras medidas mercadológicas cabíveis.

Os algoritmos que descobrem regras de associação, segundo Agrawal e Srikant (1994), devem gerá-las de forma a atender parâmetros que são informados pelo usuário: suporte e confiança. Conforme Gonçalves (2008), suporte é o percentual de incidência da regra no conjunto de transações e confiança indica a validade da regra. Isto pode ser constatado, por exemplo, em um conjunto de transações de compra de produtos, com Suporte 60 e Confiança 30, para que a regra {*pão}→{café}* seja verdadeira, é necessário que em pelo menos 60% das transações os dois produtos estejam presentes e que no mínimo 30% das transações que tenham comprado *pão*, tenham adquirido *café* também.

As regras de associações podem ser ainda: “multidimensionais”, em que atributos de diferentes tipos aparecem na regra, como o exemplo a seguir:

(Sexo = ‘M’) ^ (20<Idade<30) → {cerveja}

Esta regra indica que homens com idade entre 20 e 30 anos compram cerveja.

As regras também podem ser “híbridas”, em que atributos de mesma dimensão aparecem várias vezes na regra:

(Sexo = ‘M’) ^ (20<Idade<30) ^ {café} → {leite}

No caso acima verifica-se que homens com idade entre 20 e 30 anos, que compram café, também adquirem leite.

Por fim, também existem as regras negativas, auferindo que o consumidor que compra os produtos *A* e *B*, não compra o produto *C*, por exemplo.

Há casos em que as associações podem ser descobertas a partir de generalizações, ou seja, os dados estarem em forma de hierarquia, como por exemplo, *cereal* é uma generalização de *arroz* e *feijão*, assim como *roupa* é de *camisa* e *calça*. Alguns algoritmos têm suporte para mineração de dados multi-nível. O algoritmo mais utilizado para a técnica de associação, de acordo com Agrawal e Srikant (1994) é o *Apriori* e suas variantes.

Para ilustração da técnica de exemplo, será usada Tabela 1.1.

Tabela 1.1: Conjunto de Transações

Transação	Café	Leite	Manteiga
1	1	1	0
2	0	1	1
3	1	0	0
4	1	0	0
5	1	1	0
6	1	1	1
7	1	1	1
8	0	1	0
9	1	1	1
10	0	1	0

Fonte: Do autor

Os algoritmos de associação partem do total de transações formando os possíveis conjuntos a estarem presentes nas regras. Seguindo o exemplo da Tabela 1.1, o algoritmo identificaria os seguintes conjuntos com dois elementos: {café, leite}, {leite, manteiga} e {café, manteiga}. Não são levados em consideração conjuntos com um elemento porque não se conseguiria elaborar uma regra.

A seguir, o algoritmo verifica se os conjuntos atendem ao suporte, por exemplo, 40%. O conjunto {café, leite} ocorreu nas transações 1, 5, 6, 7 e 8, representando 50% das transações. O conjunto {leite, manteiga} ocorreu em 40 % dos casos (transações 2, 6, 7 e 9), o conjunto {café, manteiga} ocorreu nas transações 6, 7 e 8 (30 %). Sendo assim, apenas o primeiro e segundo conjuntos atendem ao suporte, sendo então verificado o fator confiança. Como o conjunto {café, manteiga} não atende ao suporte, conseqüentemente o algoritmo não identifica o conjunto {café, leite, manteiga}, por este possuir o subconjunto {café, manteiga}.

Dados os conjuntos {café, leite} e {leite, manteiga}, o algoritmo testará a confiança das seguintes possíveis regras: {café} → {leite}, {leite} → {café}, {leite} → {manteiga} e {manteiga} → {leite}. O teste consiste em verificar em quantas das ocorrências de café ocorre leite, das ocorrências que ocorrem leite ocorrem café e assim por diante. Das transações em que há café (1, 3, 4, 5, 6, 7, 9), em cinco ocorrem leite (1, 5, 6, 7 e 9) representando 5/7 ≈ 71,4 %. Seguindo o mesmo procedimento, as demais regras apresentam a confiança 62,5%, 50% e 100% (em todas as ocorrências de manteiga, há simultaneamente leite). Infere-se então, dada a Tabela 1, suporte de 40% e confiança 70%, as regras {café} → {leite} e {manteiga} → {leite}.

1.2.2 Agrupamento

Técnica que, de acordo com Goldschmidt e Passos (2005), procura separar em grupos, registros com características as mais homogêneas possíveis, ou seja, que possuam propriedades comuns. Desta forma, a técnica de agrupamento também busca a maior distinção entre os grupos definidos.

Conforme Grégio (2007), os algoritmos de agrupamento varrem os dados, identificando grupos e associando os dados de entrada a estes. A implementação desses algoritmos pode ser na técnica de hierarquia, em que a base de dados é dividida em subconjuntos menores, até que os dados se encontrem em um único grupo (nodo raiz).

A outra técnica de agrupamento é o particionamento, em que o número de grupos é pré-definido, após estes serem criados os dados são agrupados de acordo com a similaridade em relação aos grupos. Entre os algoritmos de agrupamento, destacam-se o *K-Modes e K-Means*. Este algoritmo traça aleatoriamente valores a serem adotados como o centro dos *clusters* (grupos), calculando a distância de cada registro aos centróides, o associando ao grupo com menor distância.

1.2.3 Classificação

A técnica de classificação consiste em, conforme Goldschmidt e Passos (2005), classificar os registros em categorias (classes) pré-definidas, possibilitando, quando da inserção de novos registros, que estes já sejam classificados automaticamente. É necessária a construção de um modelo, a partir de um *conjunto de treinamento*, que é composto por registros utilizados como exemplo, para a posterior classificação das tuplas da base de dados. De acordo com Carvalho (2000), O modelo criado para a classificação pode ser representado por regras de classificação, árvores de decisão fórmula matemática ou redes neurais artificiais (RNA).

Segundo Freitas (2000), a principal diferença entre a técnica de classificação e a de associação está no nível sintático (regras de classificação possuem somente um atributo na sua conclusão, enquanto as de associação permitem vários). Já em relação ao agrupamento, a principal diferença é a pré-definição das classes, enquanto os grupos são gerados em tempo de execução. A técnica de classificação será detalhada no Capítulo 2.

1.3 Pós-Processamento

Fase em que ocorre o aproveitamento das informações adquiridas, sendo realizada a interpretação e avaliação da importância do conhecimento descoberto, se houver. É verificada entre o especialista em KDD e o especialista da área da aplicação, a necessidade de repetir o ciclo de etapas.

Na etapa de pós-processamento, são tratados os resultados gerados pelos algoritmos de KDD, que podem estar em formatos não tão claros para o usuário, como informações de natureza estatística, conforme Carvalho (2003).

1.4 Ferramentas

Existem diversas ferramentas *freeware* que permitem a utilização das técnicas de KDD, como RapidMiner[4], Pentaho[5], Weka[6] (ferramenta *open-source* que pode ser acoplada a outras, como o próprio Pentaho) e Sipina[7], especializada em árvores de classificação.

1.4.1 Weka

O pacote Weka (*Waikato Environment for Knowledge Analysis*) é composto por bibliotecas que implementam diversos algoritmos das técnicas de KDD. Inerente ao *software*, está a interface gráfica que permite a escolha da forma de apresentação dos resultados (gráficos, árvores). Não obstante, podem ser desenvolvidas aplicações que utilizam apenas a camada dos algoritmos de Mineração de Dados, conforme Witten e Frank (2000).

Como visto na Figura 1.2 o formato de entrada da ferramenta Weka é arquivo .arff, sendo este composto de Relação, Atributos e Dados (SANTOS, 2005). Na ilustração, *relation* indica o nome da relação, já *attribute* indica um atributo com os seus possíveis valores entre "{}". Após a descrição dos atributos, há o conjunto dos dados (*data*), em que a cada linha possui os valores associados a cada atributo separados por ",".

[4] http://www.rapidminer.com
[5] http://www.pentaho.com
[6] http:// www.cs.waikato.ac.nz/ml/weka/
[7] http:// eric.univ-lyon2.fr/~ricco/sipina.html

O Weka oferece um grande conjunto de algoritmos a serem aplicados nesses arquivos, como demonstra a Figura 1.3, possibilitando a visualização dos resultados em diferentes formatos, como árvores de decisão e gráficos (Figura 1.4).

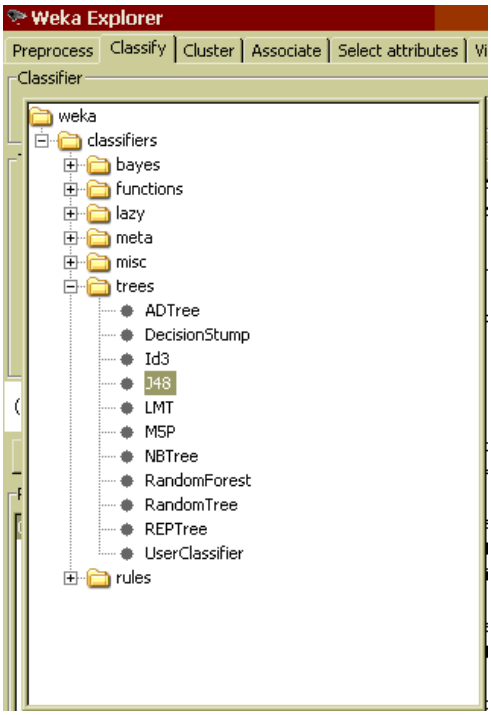

Figura 1.3: Opções do Weka
Fonte: Do autor

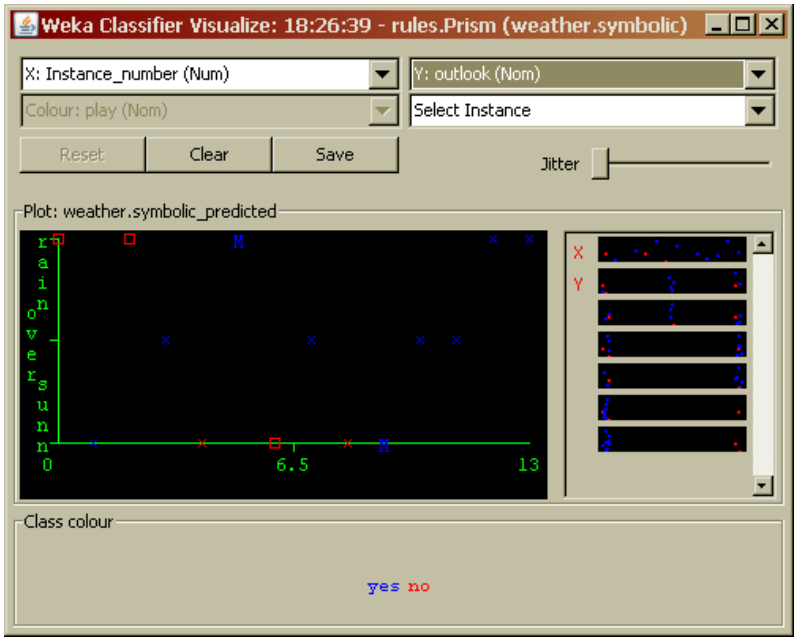

Figura 1.4: Gráfico com as ocorrências
Fonte: Do autor

1.4.2 Sipina

Ferramenta desenvolvida na Universidade de *Lyon*, é especializada na técnica de classificação, possuindo algoritmos próprios. Juntamente com a ferramenta, é disponibilizado um Suplemento para o Microsoft Excel[8]. Realizada a instalação, dados editados em planilhas eletrônicas nesta ferramenta podem servir de entrada de dados para o Sipina, bastando selecionar a opção *Execute Sipina*, no menu Sipina, informando o intervalo das células, conforme a Figura 1.5.

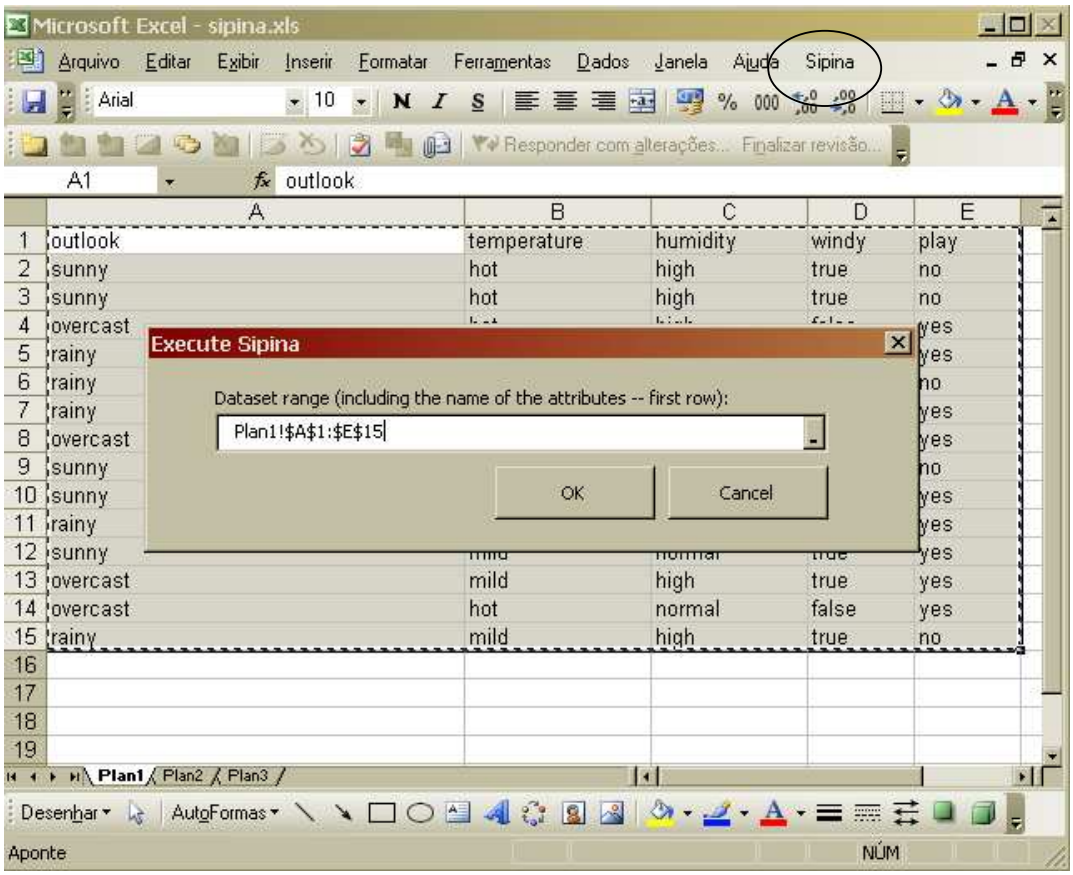

Figura 1.5: Chamada da ferramenta Sipina a partir do Microsoft Excel
Fonte: Do autor

A ferramenta também possui a implementação de diversos algoritmos, conforme na Figura 1.6. A Figura 1.7 apresenta uma árvore de decisão gerada após execução do algoritmo C4.5.

[8] http://office.microsoft.com/pt-br/excel/FX100487621046.aspx

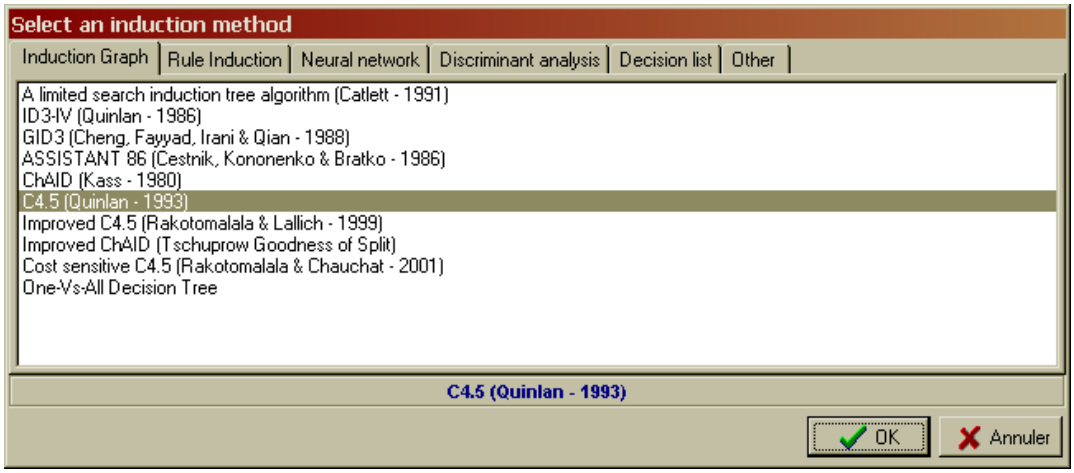

Figura 1.6: Opções de técnicas / algoritmos da ferramenta Sipina
Fonte: Do autor

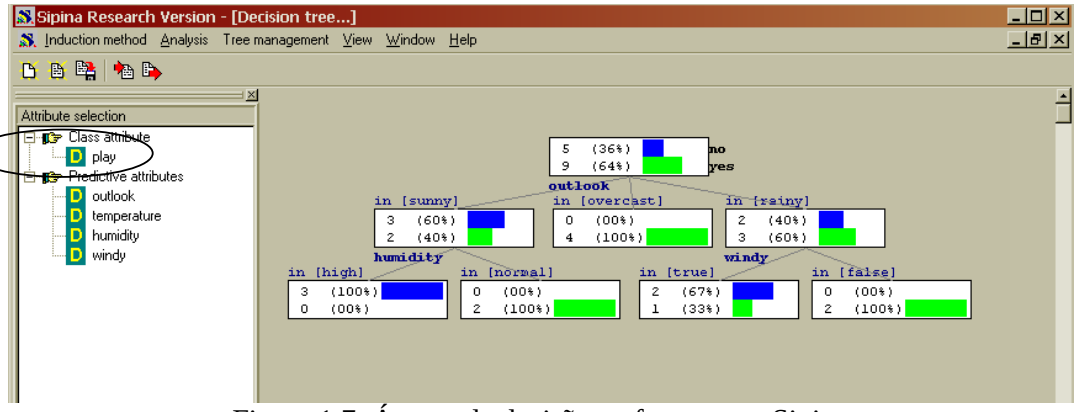

Figura 1.7: Árvore de decisão na ferramenta Sipina
Fonte: Do autor

1.4.3 RapidMiner

Ferramenta *open-source* criada na Alemanha, anteriormente chamada *Yale*. Conforme RapidMiner (2008), possui interface gráfica ao usuário (GUI) e *scripts* baseados em XML, tornando esta uma *Integrated Development Environment* (IDE) e um interpretador para KDD. É desenvolvida sob a plataforma Java, o que facilita integração com outras aplicações sob esta arquitetura.

Um exemplo desta integração, é o RapidMiner possuir incorporado toda a biblioteca Weka. Além disso, pode ser integrada com bibliotecas *Java Database Connectivity* (JDBC), possibilitando a conexão diretamente ao banco de dados, para aplicação dos algoritmos de Mineração de Dados. Na Figura 1.8 é exibida uma aplicação do RapidMiner.

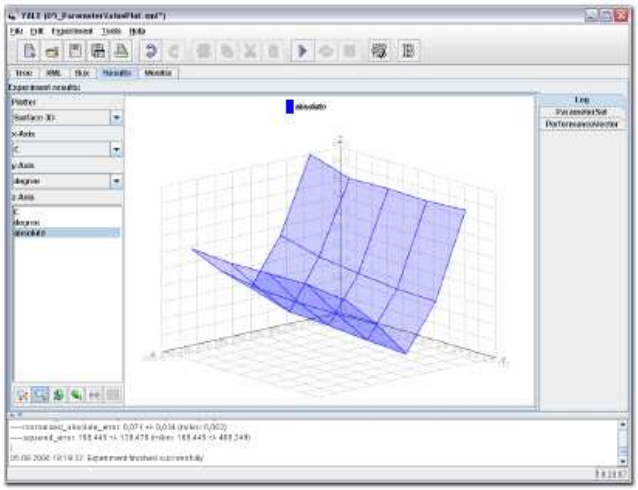

Figura 1.8: *Screenshot* da ferramenta RapidMiner
Fonte: Rapidminer, 2008

1.4.4 Pentaho

Pentaho BI é uma *Open Source Business Intelligence* (OSBI), com módulos de *Data Mining*, Análise, *Dashboards* (painel de indicadores) e Relatório. O módulo de *Data Mining* é a biblioteca *Weka*, que foi incorporada ao ambiente da OSBI. As figuras 1.9 e 1.10 mostram *screensshots* do Pentaho BI.

A ferramenta possui versão para *download* gratuito e versão *enterprise* proprietária, tendo como casos de sucesso de sua utilização empresas de grande porte, como Sun[9] e Mozilla[10].

[9] http://java.sun.com
[10] http://www.mozilla.org

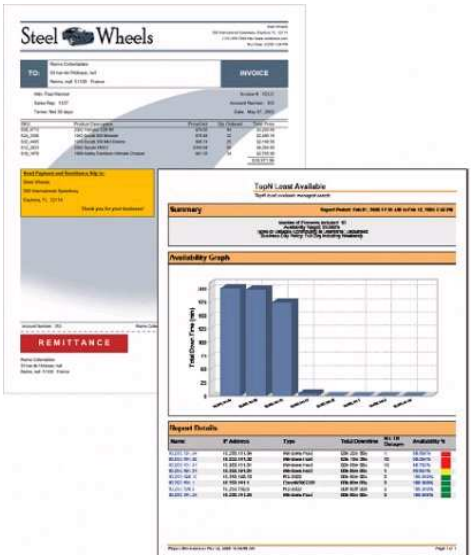

Figura 1.9: Módulo Relatório do Pentaho
Fonte: Pentaho, 2008

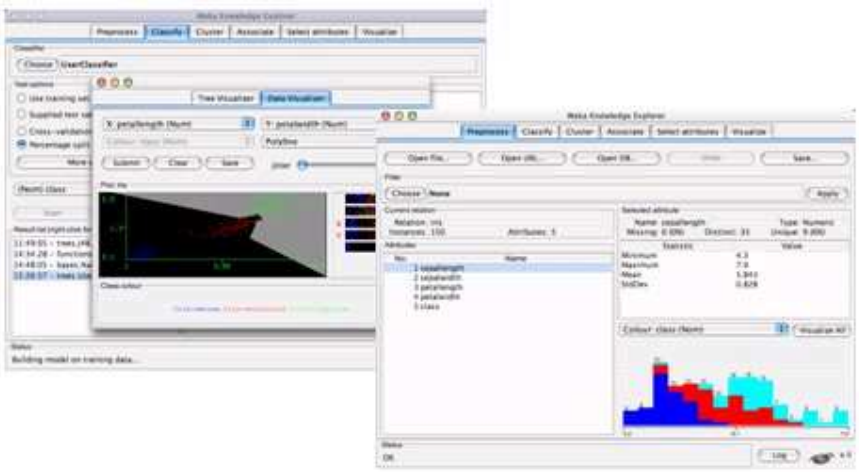

Figura 1.10: Módulo *Data Mining* do Pentaho
Fonte: Pentaho, 2008

Neste capítulo foi abordada a área de Descoberta de Conhecimento, suas etapas e diferentes técnicas. Além disso, foram apresentadas algumas das ferramentas que implementam essas técnicas. No próximo capítulo, será detalhada a técnica de classificação e alguns algoritmos que a utilizam. Esta foi escolhida por trabalhar com classes pré-definidas, o que melhor se adéqua a aplicação de mapeamento de perfis a ser desenvolvida.

2 TÉCNICA DE CLASSIFICAÇÃO

Dentre os algoritmos de classificação, encontram-se diferentes métodos para a indução de conhecimento, destacando-se: redes neurais artificiais (RNA), regras a partir de IF <*condição*> THEN <*classe*>, como os algoritmos PRISM (VASCONCELOS, 2002) e PART (OLIVEIRA, 2002), árvores de decisão, árvores de decisão com regras e método estatístico, como classificadores bayesianos (MCCALLUM, 1998).

2.1 Redes Neurais Artificiais

Área da Inteligência Artificial (IA) assim batizada por ser inspirada no funcionamento do cérebro humano, com as redes de células nervosas e o próprio comportamento. Segundo Carvalho (2000), as RNA são formadas por diversas redes em que existem unidades de processamento chamadas neurônios artificiais. O modelo de um neurônio artificial pode ser visto na Figura 2.1.

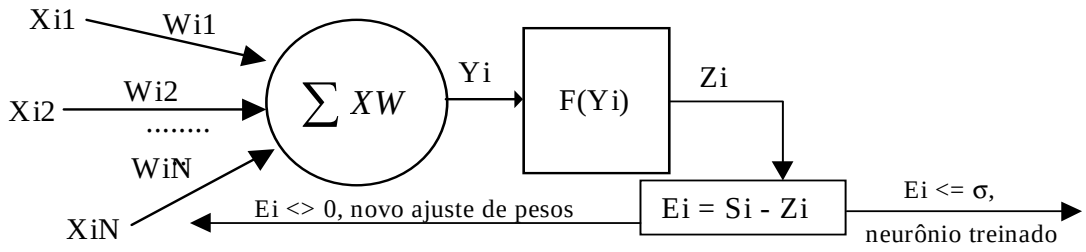

Figura 2.1: Modelo de um neurônio artificial
Fonte: Carvalho (2000)

Conforme Goldschmidt e Passos (2005), as redes neurais artificiais possuem as seguintes características semelhantes ao cérebro humano:

- Busca paralela e endereçamento pelo conteúdo: o conhecimento fica distribuído pelas redes, não existindo endereços de memória.

- Aprendizado por experiência: busca a identificação de padrões através de repetidas apresentações dos dados às redes.

- Generalização: as RNAs conseguem generalizar a partir de exemplos anteriores, facilitando a manipulação de dados com impurezas.

- Associação: as RNAs possuem a capacidade de identificar relações entre padrões de natureza distinta.

- Abstração: é a possibilidade de abstrair a essência de um conjunto de dados de entrada.

- Robustez e Degradação Gradual: a perda de neurônios artificiais não significa prejuízo no desempenho, pois as informações estão distribuídas pela rede.

Já segundo Azevedo (1999), o aprendizado das RNAs pode ser:

- Por independência de quem aprende: o aprendizado ocorre por memorização, contato, exemplos, analogia, exploração e descoberta;

- Por retroação do mundo: o aprendizado pode ser supervisionado ou não-supervisionado, de acordo com a presença ou ausência de realimentação explícita, onde são assinalados erros ou acertos;

- Por finalidade do aprendizado: o aprendizado pode ser por um auto-associador, em que a rede memoriza exemplos, conseguindo reproduzi-los caso sejam apresentados deteriorados posteriormente; hetero-associador, em que os exemplos são apresentados aos pares, e o segundo elemento pode ser reproduzido mesmo que o primeiro seja alterado; e detector de regularidades, em que são identificados padrões pela própria RNA, não sendo estes definidos anteriormente.

As RNA possuem, segundo Barreto (2002), diferentes topologias: diretas, com ciclos e simétricas.

2.1.1 Redes Diretas

É o tipo de rede mais utilizado, haja vista os métodos de aprendizado serem dos mais difundidos e fáceis de usar. Normalmente é utilizado em camadas, em que neurônios que recebem estímulos são chamados de camada de entrada e os que têm sua saída sendo o final da rede camada de saída.

Na Figura 2.2, tem-se um exemplo de uma rede em camada. Note-se que as redes diretas não possuem ciclo.

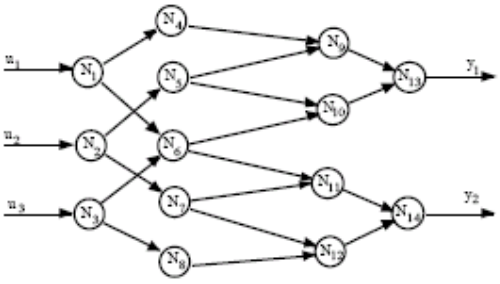

Figura 2.2: Exemplo de RNA direta
Fonte: Barreto, 2002

2.1.2 Redes Com Ciclo

São redes em que o grafo de conectividade possui pelo menos um ciclo. Quando há ocorrência de neurônios dinâmicos, são chamadas de redes recorrentes. Como os neurônios completam ciclos, podem realimentar outros neurônios.

Das redes com ciclo, destacam-se as redes propostas por Hopfield (1984) e as redes bi-direcionais (Kosko, 1988), que podem ser usadas por sistemas especialistas em um de seus principais paradigmas: treinamento com exemplos de uma rede direta e representação do conhecimento de modo localizado pelo uso de rede com ciclos (Azevedo, 1999).

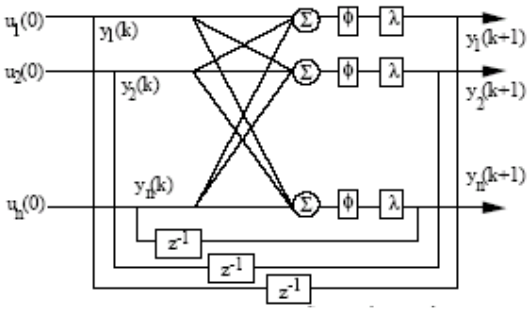

Figura 2.3: Exemplo de RNA com ciclo e neurônios dinâmicos
Fonte: Azevedo, 1999

Na Figura 2.3 é exibido um exemplo de redes com ciclo. Segundo Barreto (2002), existe ainda um tipo específico de redes com ciclo, as redes simétricas, em que a matriz de conectividade é simétrica.

2.2 Árvores de decisão

São uma forma simples de representação das regras, composta de nodos, ligações e folhas, significando, respectivamente, os atributos, seus possíveis valores e as diferentes classes. Na Figura 2.4, é exibida uma árvore de decisão, gerada a partir da aplicação do algoritmo J48 sobre os dados da Figura 1.2.

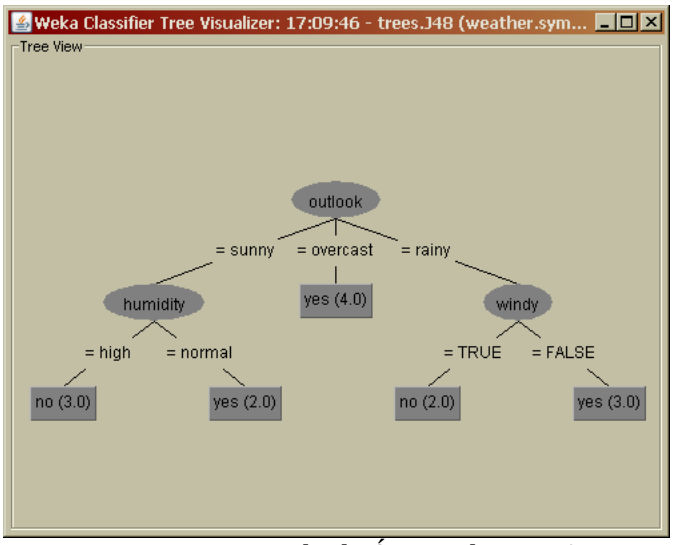

Figura 2.4: Exemplo de Árvore de Decisão
Fonte: Santos, 2005

As árvores de decisão são de fácil entendimento, sendo possível a interpretação inclusive pelos usuários, mesmo quando da representação de grandes bases de dados. No entanto, os algoritmos família TDIDT (*Top-Down Induction of Decision Trees*) não são totalmente eficazes, pois existem regras que não são possíveis de se representar. Esse problema é decorrente, muitas vezes, do fato de o nodo raiz obrigatoriamente constar nas regras, sendo chamado de problema sintático, conforme Mongiovi (1998).

De acordo com Goldschmidt e Passos (2005), os algoritmos dividem o conjunto de treinamento em duas ou mais partes, sendo um processo recursivo repetido até que todos os itens do conjunto pertençam a uma classe. O algoritmo baseado em árvore de decisão mais utilizado é o C4.5 (Quinlan, 1993), que tem sua origem no ID3 (Quinlan, 1979).

2.2.1 ID3

O algoritmo ID3, cujo nome significa *Iterative Dichotomizer Tree*, inicialmente seleciona um atributo para o nodo raiz, gerando ligações para todos os diferentes valores; se todos os sob sobre um nodo pertencem a uma mesma classe, o nodo passa a ser uma folha que recebe o nome da classe; enquanto existem nodos sem classe, o nodo recebe um atributo ainda não utilizado pela árvore com ligações criadas para todos os valores. A escolha dos atributos a serem utilizados pela árvore se dá a partir de informações de entropia e ganho de informação.

O valor da entropia corresponde à impureza do atributo, a falta de homogeneidade, sendo calculada para cada atributo. O ganho de informação é a variação da impureza. O que possuir menor valor de entropia ou maior ganho de informação, segundo Ascenso (2004), é escolhido o nó raiz da árvore. O cálculo de entropia é dado na Figura 2.5, em que cada "p" corresponde ao número de instâncias sobre o de exemplos.

$$\text{Entropia}(S) = -(p1*\log_2 p1 + p2*\log_2 p2 + \ldots + pn*\log_2 pn)$$

Figura 2.5: Fórmula de cálculo de entropia
Fonte: OSÓRIO, 2001

Para ilustrar o algoritmo, serão tabulados os dados do arquivo da Figura 1.2 na Tabela 2.1.

Tabela 2.1: Conjunto para exemplo do ID3

	tempo	temperatura	umidade	vento	jogar
1	ensolarado	quente	alta	não	não
2	ensolarado	quente	alta	sim	não
3	nublado	quente	alta	não	sim
4	chuvoso	amena	alta	não	sim
5	chuvoso	fria	normal	não	sim
6	chuvoso	fria	normal	sim	não
7	nublado	fria	normal	sim	sim
8	ensolarado	amena	alta	não	não
9	ensolarado	fria	normal	não	sim
10	chuvoso	amena	normal	não	sim
11	ensolarado	amena	normal	sim	sim
12	nublado	amena	alta	sim	sim
13	nublado	quente	normal	não	sim
14	chuvoso	amena	alta	sim	não

Fonte: Do Autor

A seguir são realizados os cálculos de entropia para os atributos do conjunto: tempo, temperatura, umidade e vento (jogar é atributo de classe, não sendo calculada sua entropia). Inicialmente são calculadas as entropias dos possíveis valores do atributo.

No cálculo abaixo, “2/9” corresponde ao número de ocorrências de *ensolarado* (duas ocorrências) nas transações em que *jogar = sim* (nove transações). De forma análoga, há três ocorrências de *ensolarado* nas transações em que *jogar = não* (cinco transações).

Entropia (tempo=ensolarado) = - (2/9) * log2(2/9) – (3/5) * log2(3/5) = 0,924

Quando todas as ocorrências de um valor de atributo correspondem a uma mesma classe, a entropia é 0 (em todas as ocorrências de *nublado*, *jogar = sim*).

Entropia (tempo=nublado) = 0

Entropia (tempo=chuvoso) = - (3/9) * log2(3/9) – (2/5) * log2(2/5) = 1,057

Deve-se multiplicar a entropia de cada valor pela divisão entre sua ocorrência e o total de transações. Este cálculo deve ser realizado para todos os possíveis valores do atributo, sendo a soma destes resultados a entropia do atributo.

No exemplo abaixo, 14 é o número de transações; 5, 4 e 5 são as respectivas ocorrências de *ensolarado*, *nublado* e *chuvoso*, sendo 0,924, 0 e 1,057 suas entropias.

Entropia (tempo) = (5/14) 0,924 + (4/14)0 + (5/14) 1,057= 0,70

Entropia (vento=sim) = - (3/9) * log2(3/9) – (3/5) * log2(3/5) = 0,97

Entropia (vento=não) = - (6/9) * log2(6/9) - (2/5) *log2(2/5) = 0,918

Entropia (vento) = (6/14)0,97 + (8/14)0,918 = 0,94

Entropia (temperatura=fria) = - (3/9) * log2(3/9) – (1/5)* log2(1/5) = 0,992

Entropia (temperatura =amena) = - (4/9) * log2(4/9) – (2/5) * log2(2/5) = 1,048

Entropia (temperatura =quente) = - (2/9) * log2(2/9) – (2/5) * log2(2/5) = 1,01

Entropia (temperatura) = (4/14)0,992 + (6/14)1,048 + (4/14)1,01= 1,021

Entropia (umidade=normal) = - (6/9) * log2(6/9) – (1/5) * log2(1/5)= 0,854

Entropia (umidade=alta) = - (3/9) * log2(3/9) - (4/5) * log2(4/5) = 0,785

Entropia (umidade) = (7/14)0,854 + 7/14(0,785)= 0,822

Percebe-se que o atributo de menor entropia é *tempo*, sendo este o nó raiz da árvore a ser gerada. O algoritmo ID3 não necessariamente produz as melhores regras, sendo criadas em alguns casos, árvores muito grandes. Surgiu então, uma evolução do ID3, o C4.5, pelo mesmo criador, Quinlan, em 1993.

2.2.2 C4.5

O algoritmo C4.5 consegue, conforme Coello (2002), manipular registros com valores desconhecidos para alguns atributos, tratar atributos com valores contínuos, e inferir regras a partir da árvore. Para evitar o problema de tamanho da árvore, o C4.5 trabalha com a poda.

A poda é baseada no grau de incerteza de um atributo, quando, por exemplo, a entropia é maior que um determinado valor. A operação consiste em substituir uma sub-árvore, com grande taxa de erro, por um nodo ou por um ramo e é realizada antes ou depois de a árvore ser elaborada. Assim a árvore é simplificada, o que facilita a inserção de novos nodos e sub-árvores, necessidade que vai surgindo conforme o conjunto de dados cresce. No entanto, existem casos em que a poda é difícil de ser aplicada, como em árvores que não se encontram equilibradas.

Para ilustrar um exemplo de poda, utilizou-se o conjunto de dados da tabela 2.1, inserindo-se mais duas transações, Tabela 2.2, com o intuito de tornar a árvore mais complexa.

Tabela 2.2: Novas transações no conjunto

	tempo	temperatura	umidade	vento	Jogar
15	nublado	amena	alta	não	não
16	nublado	fria	alta	sim	sim

Fonte: Do Autor

Utilizando-se da ferramenta Weka, aplicou-se os algoritmos ID3 e J48 (WITTEN; FRANK, 2000) para que esses gerassem suas respectivas árvores. O J48 é a implementação em Java do algoritmo C4.5. Na Figura 2.6 e 2.7 são exibidas as árvores geradas.

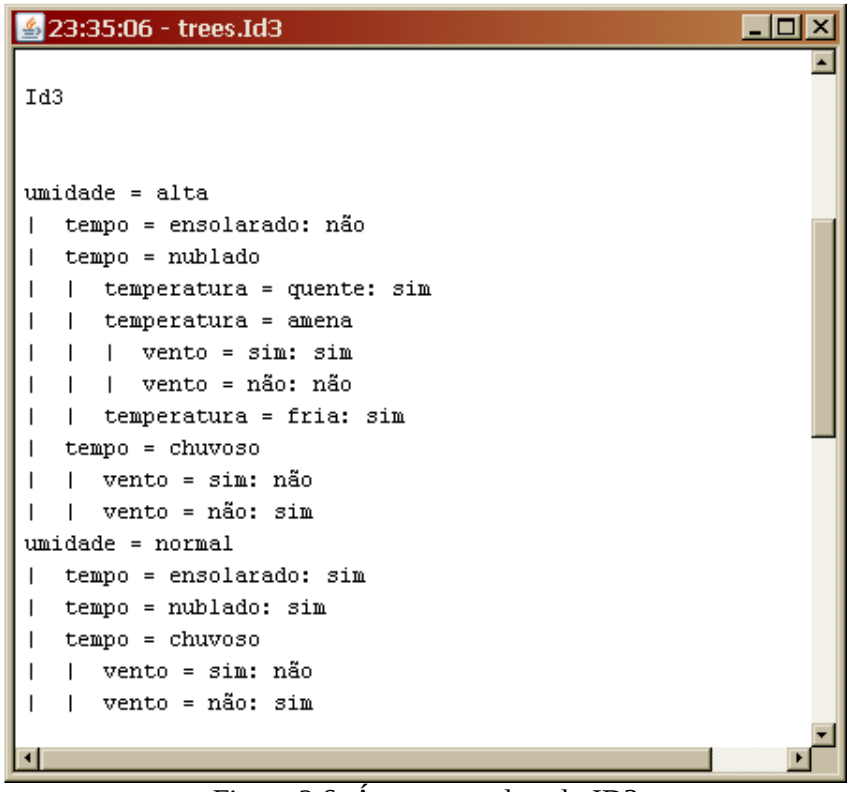

Figura 2.6: Árvore gerada pelo ID3
Fonte: Do autor

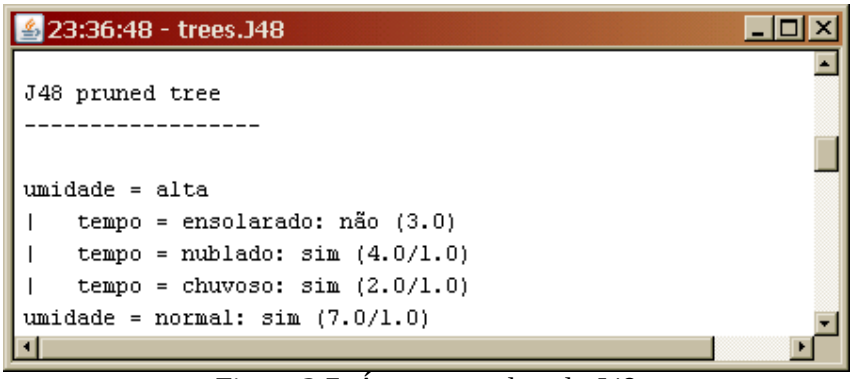

Figura 2.7: Árvore gerada pelo J48
Fonte: Do autor

Ainda com o objetivo de reduzir as árvores, segundo Quinlan (1993), o algoritmo é capaz de agrupar valores de atributos. Conseqüentemente, em uma situação em que um atributo possua *n* valores pertencentes a uma mesma classe, será criado um ramo para o grupo de valores ao invés de *n* ramos. Nos casos em que ocorrem valores desconhecidos para um determinado atributo, o C4.5 trabalha com estatística baseada no conjunto dos dados até então conhecido.

O C4.5 permite ainda gerar regras a partir da árvore de decisão, em que condições irrelevantes são excluídas. Isto permite que as regras se tornem mais simples do que se fossem simplesmente elaboradas a partir de toda a extensão da árvore. Estas regras podem ser no formato convencional (a), com condições combinadas (b) e ainda de dissociação (c), conforme Tabela 2.3.

Tabela 2.3: Formato das regras

	Formato
a	IF <condicao> THEN <classe>
b	IF <condicao1> (and | or)* <condicao2> THEN <classe>
c	IF (NOT)+ <*condicao1*> (and | or)* (NOT)+ <*condicao2*> THEN <*classe*>

Fonte: Do autor

2.3 C&RT

Classification & Regression Trees, também chamado de C&RT, foi proposto por Breiman (1984). É um algoritmo robusto que possui excelente performance, inclusive quando trabalha com imensa quantidade de dados, sendo, segundo Lewis (2000), um dos mais utilizados algoritmos para construção de árvores de decisão.

O algoritmo testa todas as possibilidades de divisão para as variáveis, posteriormente realizando a análise para particionar o nodo. Essa partição é feita através de perguntas com respostas binárias (Sim, Não), sendo possível a manipulação de variáveis com valores contínuos ou categorizados.

2.4 CHAID

Outro algoritmo que possui a característica de construir árvores a partir de regressão. É um acrônimo para *CHi-squared Automatic Interaction Detector*. Diferentemente do C&RT, a partição dos nodos não é binária, acarretando em árvores mais esparsas, o que pode ser uma vantagem para problemas como o de produtos em um supermercado, por exemplo.

O CHAID, segundo Kass (1980), apresenta melhor desempenho para grandes conjuntos de dados. No entanto, apesar deste desempenho, a árvore gerada pelo CHAID é maior que a obtida pelo C4.5. O algoritmo foi executado a partir da ferramenta Sipina, utilizando-se os mesmos dados (Tabelas 2.1 e 2.2).

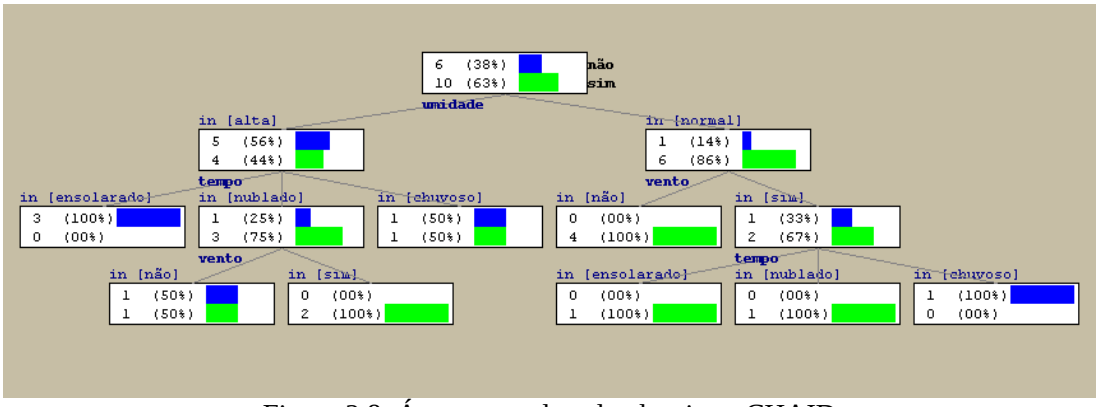

Figura 2.8: Árvore gerada pelo algoritmo CHAID
Fonte: do Autor

O algoritmo recebe como parâmetros *sppliting nodes*, relativo ao número de divisões dos nodos, quanto mais alto, menor a performance e maior a árvore a ser gerada e *merging nodes* em que os nodos são unidos se possuem diferença pouco significativa, sendo este critério relativo ao parâmetro informado.

2.5 Regras

Algoritmos de regras de classificação não pertencem à família TDIDT, não possuindo o problema sintático (vide seção 2.2). A seguir são apresentados Prism e PART.

2.5.1 Prism

O Prism infere regras no formato de IF <*condição*> THEN <*classe*>, em que, conforme Vasconcelos (2002), uma *condição* é um conjunto de termos <atributo_c_t q valor>, onde *atributo_*c_t corresponde a um atributo do conjunto de treinamento; q = {.=., .<., .>., .£., .^{3}., .^{1}.} e *valor* é um possível valor do atributo. Se a condição for verdadeira, o dado é agrupado à *classe*.

Em comparação com as árvores de decisão, são geradas um número menor de regras, sendo estas ainda menos complexas. No entanto, utilizando o Prism podem ocorrer sobreposição de regras.

2.5.2 PART

O algoritmo é uma variação do J48 que apresenta, ao invés de uma árvore de decisão, regras. Estas são geradas em duas etapas: inicialmente são auferidas da árvore de decisão para um posterior refinamento. Sendo as regras criadas, todas as instâncias do conjunto são testadas para descobrir a cobertura das regras. Segundo Entriel (2008), as que possuírem menor cobertura e a árvore de decisão são descartadas.

Aplicando o PART no mesmo conjunto de dados anteriormente utilizado pelo ID3 e pelo C4.5, o conjunto de regras gerado é exibido na Figura 2.9, sendo explícita a ausência do problema sintático, haja vista o atributo “umidade” estar presente em todas regras nas árvores geradas nas Figuras 2.5 e 2.6.

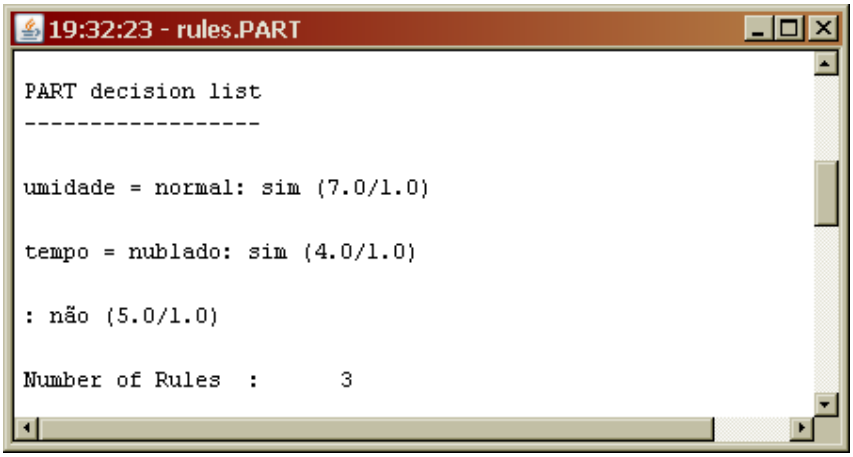

Figura 2.9: Conjunto de regras gerado pelo algoritmo PART
Fonte: Do autor

2.6 Redes Bayesianas

Assim como RNA, é uma área de IA, muito utilizada em problemas para diagnóstico e decisão. As redes bayesianas, conforme Mello (2007), utilizam o Teorema de *Bayes*, decorrente do Teorema da probabilidade, juntamente com grafos, que representam as relações entre os conjuntos das probabilidades, buscando atribuição de níveis de confiabilidade.

Os grafos utilizados nas redes bayesianas são direcionados e acíclicos. Uma ligação de um nodo a outro representa a influência direta entre um e outro. Cada um destes nodos armazena os possíveis estados da variável juntamente com uma tabela de probabilidades que quantificam a influência de outro nodo ligado. Conforme Marques e Dutra (2003), para cada variável I que possui como pais *B1*,..., *Bn*, existe uma tabela *P(A| B1, ..., Bn)*.

Para que a rede seja criada, são passados como entrada dados e informações para que após o processamento, a saída seja a rede propriamente dita. Segundo Guinzani (2006), a aprendizagem consiste inicialmente em gerar a rede que apresenta as relações entre as variáveis, sendo em seguida determinada a representação da distribuição das probabilidades de cada nodo, para assim estimar as probabilidades condicionais com a base de dados, considerando apenas os valores de variáveis relevantes.

As redes bayesianas são amplamente utilizadas na área de diagnósticos médicos. A Figura 2.10 mostra um exemplo de uma rede bayesiana para diagnóstico de doenças cardíacas (SAHESKI, 2005).

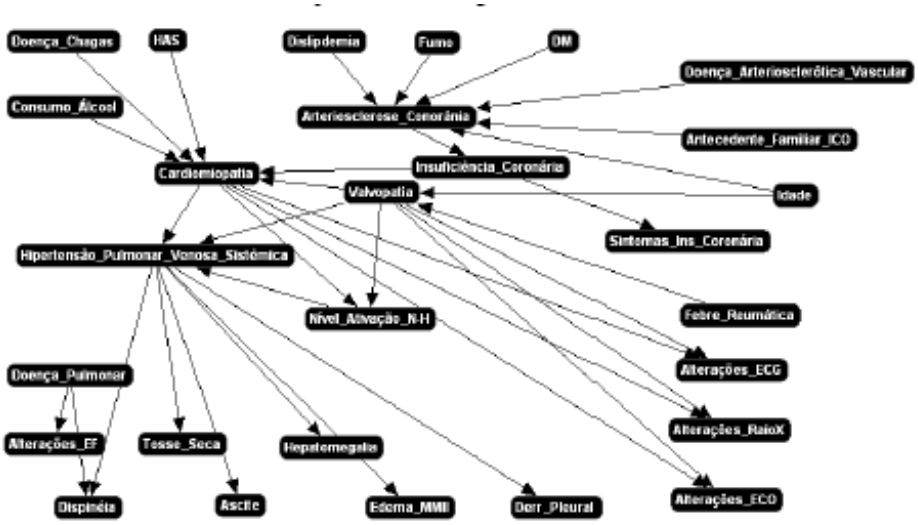

Figura 2.10: Exemplo de Rede Bayesiana
Fonte: Saheski, 2005

2.6.1 Classificadores

Existem vários classificadores bayesianos, sendo, segundo Pinto (2005), o *Naive Bayes* o mais simples. Este funciona através de apresentação de conjuntos com uma classificação associada, sendo que novos elementos são classificados a partir do reconhecimento por parte do algoritmo, técnica muito utilizada em algoritmos anti-spam.

Entre os classificadores ainda, destaca-se o TAN, extensão ao *Naive Bayes* que trata dependência entre os valores de variáveis. As figuras 2.11 e 2.12 ilustram esta diferença, considerando os possíveis valores X1, X2, X3 e X4 para o atributo C.

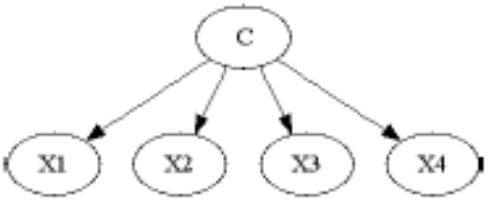

Figura 2.11: Modelo de dependência *Naive Bayes*
Fonte: Pinto (2005)

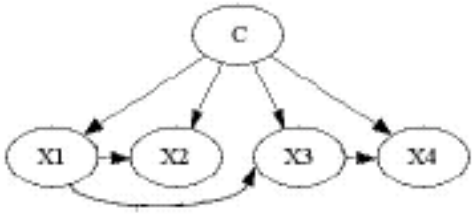

Figura 2.12: Modelo de dependência TAN
Fonte: Pinto (2005)

2.7 Meta-classificação

Os algoritmos de classificação normalmente trabalham com todos os dados do conjunto de treinamento em memória. No entanto, isto pode ser um grande problema quando se trata de grandes bases de dados. Para o devido tratamento deste problema destacam-se algoritmos incrementais, técnicas de amostragem, processamento paralelo e técnicas de partição e combinação de resultados parciais. Segundo Mongiovi (1998), se adotando esta medida, os dados são divididos em subconjuntos e a cada um destes são aplicados os algoritmos tradicionais e os resultados então são combinados. Os classificadores que aplicam algoritmos aos resultados de outros classificadores são chamados de *meta-classificadores.*

Uma das ferramentas utilizadas neste projeto, Weka, possui meta-classificadores, destacando-se *Bagging* e *Boosting*. De acordo com Oliveira (2002), o método *Bagging* constrói os classificadores a partir de conjuntos independentes de amostras de dados, sendo estes gerados a partir do conjunto de treinamento. São extraídas instâncias conforme estas são repetidas nas amostras.

Ainda segundo Oliveira (2002), no método *Boosting* as instâncias possuem um peso associado. Inicialmente, o peso é o mesmo para todas as instâncias, sendo este alterado conforme as instâncias de treinamento são classificadas incorretamente, baseadas nos classificadores gerados anteriormente.

Neste capítulo foram apresentadas metodologias de utilização da técnica de classificação, como RNAs e árvores de decisão e algoritmos que implementam estas. No Estudo de Caso, será abordado como KDD vem sendo utilizado na área de Marketing e as etapas que se sucederão após o recebimento dos dados da instituição para posterior aplicação dos algoritmos.

3 ESTUDO DE CASO

A crescente procura do mercado de trabalho por empregados cada vez mais qualificados, juntamente com a perspectiva de crescimento profissional, acarretou em uma maior procura por cursos de graduação. Conseqüentemente, acabaram por surgir diversas universidades e centro universitários, tornando-se um setor com acirrada concorrência.

Assim como em qualquer segmento, no de Ensino Superior também é imprescindível um diferencial em relação aos concorrentes para a atração dos clientes, no caso, os alunos. Buscando esta qualidade, serão aplicadas as técnicas de KDD para que assim identifiquem-se perfis de alunos e relações entre esses perfis e o interesse por cada curso ou área.

3.1 Aplicações na área de Marketing

A identificação de perfis de clientes vem sendo utilizada cada vez mais por aplicações de *Data Mining* com enfoque na área de Marketing, objetivando o estreitamento na relação com os consumidores, visando melhor atendimento, fidelização e até a conquista de novos clientes.

A área que associa conceitos de Marketing de Relacionamento ao uso de TI é a *Customer Relationship Manager* (CRM), que torna possível às empresas que adotam sistemas da área um serviço personalizado, antecipando as necessidades dos clientes e suprindo-as totalmente. Uma área derivada do CRM é a *Customer Interaction Management* (CIM), em que a empresa norte-americana Talismã possui ferramentas específicas como a *Talisma Knowledgebase Software*, que permite direcionar o contato da empresa com o cliente, a partir de transações já realizadas, evitando gastos com comunicação.

O mapeamento de perfis também é utilizado para análise de crédito, em que características dos clientes são associadas ao histórico de pagamento. As instituições

financeiras utilizam ferramentas específicas para este fim, como o Banco do Brasil, conforme Lemos (2000) possui um *software* desenvolvido internamente chamado Análise de Crédito (ANC).

Um caso de sucesso é o Grupo RBS, que possui a RBS Direct[11], uma empresa específica para o Marketing de Precisão, utilizando ferramenta de CRM da Oracle[12] e conta com uma equipe de desenvolvimento, que entre outras atividades, utiliza *Data Mining* para criar perfis de clientes.

Podem ser mapeadas características de consumidores através do acesso a internet, sendo rastreadas as páginas de maior incidência de acessos para os usuários cadastrados de um *site*, por exemplo. Isto acaba caracterizando uma sub-área de KDD, a *Web Mining* (BOULLOSA, 2002).

3.2 Dados dos vestibulandos

Com o objetivo de encontrar relações entre a demanda dos cursos da instituição e os perfis de vestibulandos a serem traçados, foram selecionados os dados de vestibulares de 2006 a 2008, sendo dois concursos por ano: inverno e verão. Nessas bases de dados, encontram-se informações de endereço, nascimento, renda familiar, escola em que o aluno cursou o Ensino Médio, entre outras, contidas nas respostas do questionário aplicado aos candidatos no momento da inscrição para o concurso (vide Anexo I). O formato dos dados recebidos foi de planilha eletrônica do *Microsoft Excel*.

Os alunos poderão estar distribuídos em vários perfis simultaneamente, pois estes podem possuir informações não mutuamente exclusivas, como por exemplo, informações de sexo e idade em um e município de residência e escola de Ensino Médio em outro. Como esses perfis serão pré-definidos, será utilizada a técnica de classificação.

Será desenvolvida uma aplicação para a leitura das planilhas eletrônicas. Esta ferramenta permitirá a geração de arquivos .arff, a serem utilizados pelo Weka. Evidentemente, não será simplesmente realizada a conversão de um formato para outro, pois as planilhas contêm dezenas de atributos, o que acarretaria em árvores muito extensas. O

[11] http://www.rbsdirect.com.br
[12] http://www.oracle.com

software a ser criado possibilita a escolha dos atributos que serão utilizados para determinadas classificações, sendo então aplicados os algoritmos.

Os resultados serão apresentados para o departamento de Marketing da instituição, para que este, a seu critério, adote medidas a fim de atrair mais alunos à instituição, que podem ser desde o direcionamento de propagandas até o agendamento de visitas de apresentação conforme a escola ou região de maior procura pelo curso (ou menor).

Neste capítulo foi exposto o problema que necessita a aplicação das técnicas de KDD. No seguinte serão apresentados detalhes da implementação do projeto.

4 APLICAÇÃO

De posse dos dados em que serão executados os algoritmos de KDD, partiu-se então para etapas já descritas de Pré-Processamento e Mineração de Dados, sendo descritas nas próximas seções, juntamente com o desenvolvimento da ferramenta

4.1 Pré-Processamento

Conforme explicitado no capítulo anterior, os dados foram recebidos em planilhas eletrônicas. Para que estes fossem transformados em um dos formatos lidos pelo Weka, optou-se pelo desenvolvimento de uma aplicação Java que lê as planilhas e gera os arquivos .arff. A escolha desta tecnologia se deu pela sua portabilidade, por ser a mesma plataforma do Weka e possuir grande diversidade de bibliotecas para leitura de planilhas eletrônicas, como a JExcelApi[13] e POI[14], sendo esta a escolhida devido a experiências anteriores deste autor.

POI é uma API (*Application Programming Interface*) específica para acesso (leitura ou escrita) a arquivos no formato do *Microsoft Office*[15]. É uma ferramenta desenvolvida e mantida pela *Apache Software Foundation*[16].

4.1.1 Leitura das Planilhas

A aplicação criada recebe como entrada o nome do arquivo da planilha (juntamente com o seu diretório), sendo que os procedimentos que utilizam as classes e métodos da API POI abrem o arquivo e o interpretam no seu formato. Em todos os arquivos recebidos, a primeira linha é o cabeçalho, em que cada célula corresponde a um atributo. As demais linhas

[13] jexcelapi.sourceforge.net/
[14] http://poi.apache.org/
[15] http://office.microsoft.com/pt-pt/default.aspx
[16] http://apache.org/

são os registros, conforme exemplo da Figura 4.1, em que o cabeçalho com “QSE_2”, “QSE_3”, “QSE_4” e “QSE_5” correspondem às questões “Qual o tipo de curso de Ensino Médio (2º Grau) que você concluiu?”, “Em que ano concluiu o Ensino Médio?” e “Se você não concluiu o Ensino Médio, em que etapa se encontra?”, respectivamente.

S	T	U	V
QSE_2	**QSE_3**	**QSE_4**	**QSE_5**
EDUCAÇÃO DE JOVENS E ADULTOS	2007	JÁ CONCLUI	TODO EM ESCOLA PU
ENSINO MÉDIO (2º GRAU)	2007	JÁ CONCLUI	TODO EM ESCOLA PA
NÃO CONCLUI	NÃO CONCLUI	3º ANO OU 3ª ETAPA	TODO EM ESCOLA PA
SUPLETIVO REGULAR	2007	JÁ CONCLUI	TODO EM ESCOLA PU
EXAME DA SEC/SUPLETIVO	1999	JÁ CONCLUI	MAIOR PARTE EM ES
ENSINO MÉDIO (2º GRAU)	2001	JÁ CONCLUI	TODO EM ESCOLA PU
ENSINO MÉDIO (2º GRAU)	2007	JÁ CONCLUI	TODO EM ESCOLA PA
CURSO NORMAL OU MAGISTÉRIO	2007	JÁ CONCLUI	TODO EM ESCOLA PA
ENSINO MÉDIO (2º GRAU)	2003	JÁ CONCLUI	TODO EM ESCOLA PU
ENSINO MÉDIO (2º GRAU)	1996	JÁ CONCLUI	TODO EM ESCOLA PU
ENSINO MÉDIO (2º GRAU)	2003	JÁ CONCLUI	TODO EM ESCOLA PU
CURSO TÉCNICO DE NÍVEL MÉDIO	1994	JÁ CONCLUI	TODO EM ESCOLA PU
SUPLETIVO REGULAR	2006	JÁ CONCLUI	MAIOR PARTE EM ES
ENSINO MÉDIO (2º GRAU)	2004	JÁ CONCLUI	TODO EM ESCOLA PU

Figura 4.1: Exemplo de planilha com dados dos vestibulandos
Fonte: Do Autor

A identificação dos atributos (cada pergunta do questionário é um atributo) ocorreu sem empecilhos. No entanto, segundo Heuser (2001), os registros foram exportados de um banco de dados não normalizado. Isto foi percebido quando da constatação de diferentes representações para dados com o mesmo significado, como ausência de acentuação, abreviaturas, erros de ortografia ou omissão de preposições. Por exemplo, os seguintes conjuntos são formados por elementos com o mesmo significado: {sao sebastiao do cai, são sebastião do caí}, {sto a patrulha, santo antônio da patrulha}, {montenegro, mentenegro}.

Com o objetivo de solucionar esses problemas, foi necessária a implementação de métodos que substituíssem caracteres eliminando todos os sinais ortográficos (substituir “Ê” por “E”, “Á” por “A”, e assim por diante) e transformassem os termos lidos para maiúsculas, evitando assim diferenciação de caixa alta e baixa. Além disso, foram incluídos nesses métodos termos específicos para eliminação dos erros ortográficos, como substituir o termo “mentenegro” por “montenegro”, no conjunto supra citado. Corrigidas as inconsistências dos dados, partiu-se então para a preparação dos dados para o Weka.

Os arquivos no formato .arff, para que possam ser interpretados pelo Weka, não podem possuir os caracteres “ ” e “,” nos nomes dos atributos ou nos possíveis valores a serem designados a estes. Por isto, estes caracteres foram incorporados aos conjuntos de

termos a serem substituídos. Este problema ocorreu com o curso com a descrição “(4205) engenharia industrial, bacharelado, habilitação em engenharia industrial mecânica”, que após as referidas modificações, passou para “(4205)_ENG_IND-_BCH-HAB_EM_ENG_IND_MECANICA”.

Como constatado no próprio exemplo anterior, alguns dos atributos possuem valores com descrição muito extensa, a ponto de após o arquivo .arff ser gerado não ser reconhecido. Neste caso, destacam-se, além de curso, o atributo escola, com nomes como “'Veranópolis - Centro Tecnológico Universidade de Caxias do Sul - Unidade de Veranópolis'”. Aproveitando-se dos já implementados métodos que substituem termos por outros, foram realizadas trocas de palavras muito utilizadas por abreviações, por “Escola Estadual de Ensino Médio” passou a ser representado por “EEEM”, “Porto Alegre” por “POA” e assim por diante. O conjunto com todos os termos e suas respectivas substituições podem ser visualizados no Anexo II.

Além destas substituições, dos dados referentes à data de nascimento do vestibulando (lidos pelo POI no formato numérico, sendo necessária a conversão para data) foi aproveitado somente o ano, para uma maior uniformização dos dados. Neste ponto, os procedimentos para leitura dos dados e geração dos arquivos .arff estavam prontos.

4.1.2 Interface Gráfica

Buscando facilitar a utilização dos procedimentos criados, foi elaborada uma interface gráfica utilizando *swing*, conjunto de biblioteca nativo do JDK (*Java Standard Edition Development Kit*) que permite a criação de GUIs. Facilidade esta foi verificada na tela que permite, ao invés de informar o nome do arquivo com a planilha eletrônica no código-fonte da aplicação, selecionar o arquivo através da interface criada, conforme a Figura 4.4. A aplicação desenvolvida pode ser compactada em um arquivo JAR (*Java Archive*) e segue o fluxo descrito na Figura 4.2.

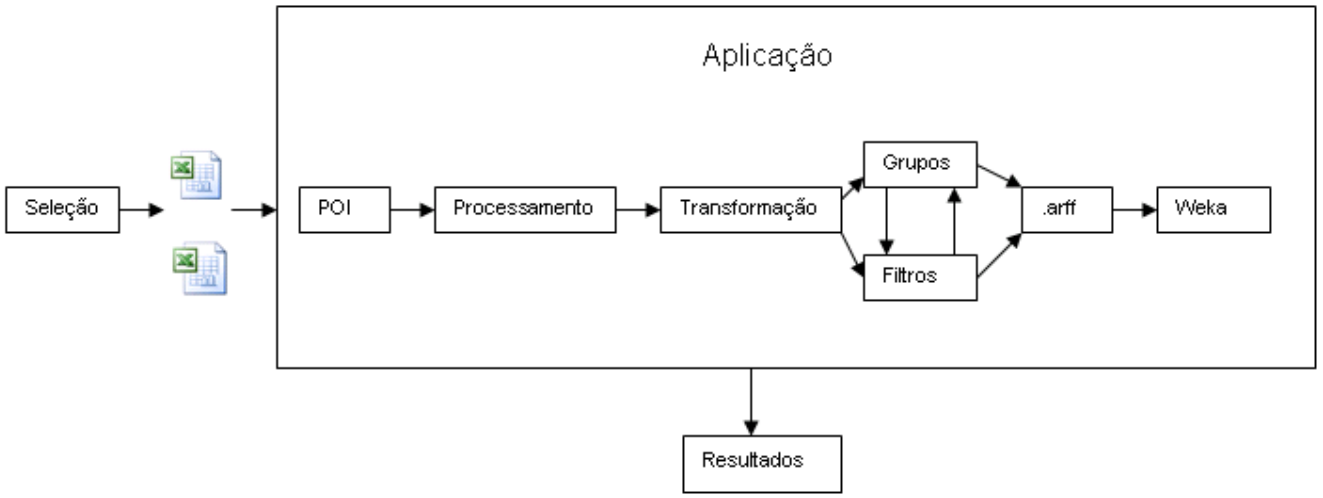

Figura 4.2: Fluxo da Aplicação
Fonte: Do Autor

Ao executar o arquivo .jar é exibida a tela inicial (Figura 4.3).

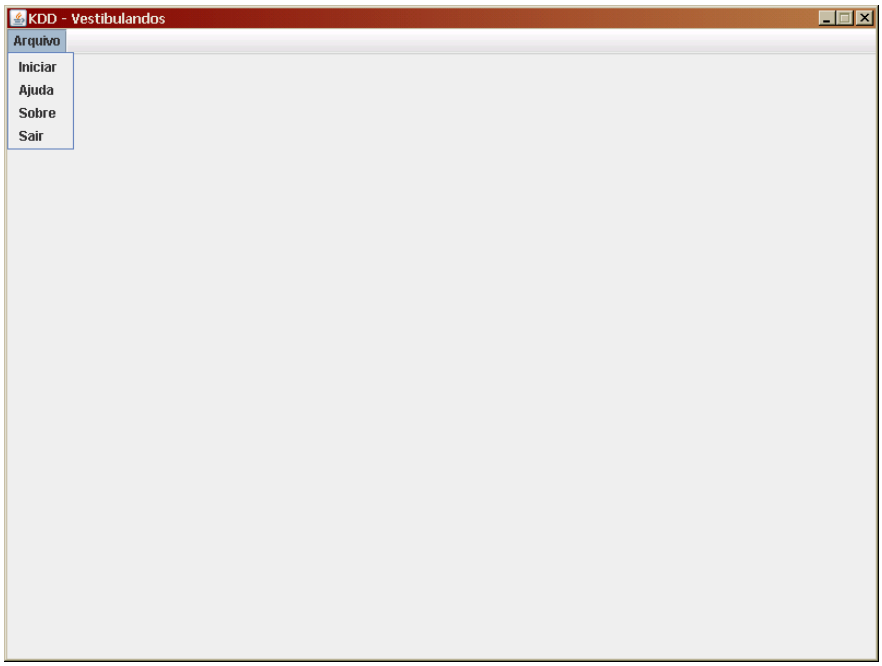

Figura 4.3: Tela inicial da aplicação
Fonte: Do autor

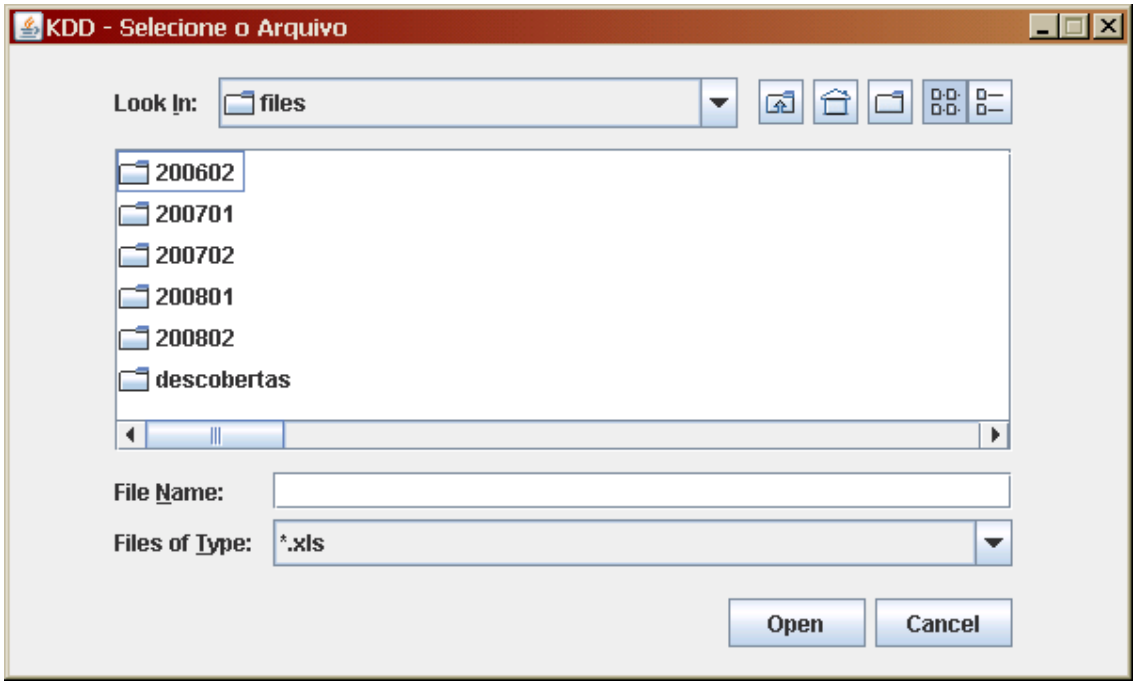

Figura 4.4: Tela de seleção de arquivo de dados
Fonte: Do Autor

Escolhendo a opção “Iniciar”, abre-se a janela de seleção do arquivo com a planilha dos registros dos vestibulandos. Com o arquivo já carregado, a aplicação possibilita ao usuário a criação de Grupos, Filtro e Seleção de Atributos.

4.1.3 Grupos

Mesmo com uma posterior redução do número de atributos, surgiu a necessidade de diminuir para cada um destes, o número de valores possíveis. Isto porque foi identificada a semelhança entre alguns, por exemplo, no vestibular de 2008/02 há ocorrência dos valores: “(4401) Bacharelado em Direito” e “(4501) Bacharelado em Direito” para o atributo relativo ao curso escolhido.

Um grupo consiste em definir um nome (descrição do grupo) que substitua um conjunto de valores. Por exemplo, pode-se criar um grupo chamado “VALE_DO_PARANHANA” para o atributo cidade, que substituiria todos os registros que contenham os valores “IGREJINHA”, “PAROBE”, “RIOZINHO”, “ROLANTE”, “TAQUARA” ou “TRÊS_COROAS”. Assim, a cada ocorrência desses municípios, no arquivo .arff gerado, ela será substituída por “VALE_DO_PARANHANA”. Da mesma forma, pode-se utilizar o ano de nascimento do vestibulando, agrupando-se os valores “1985”,

“1986”, “1987”, “1988” e “1989” em “20_a_25_anos” (os grupos não podem possuir espaços em branco no nome), por exemplo.

Por conseguinte, mais registros serão classificados no grupo do que se cada cidade (ou ano de nascimento) continuasse individualizada. Assim, é possível o mapeamento de regras mais consistentes, devido a esta maior freqüência. A tela de Criação de Grupos, exibida na Figura 4.5, é exibida logo após a planilha ter sido carregada.

Inicialmente são exibidos todos os atributos encontrados. Ao lado esquerdo de cada atributo há um botão que ao passar o *mouse* sobre este é exibida a descrição da questão associada ao atributo. Isto porque nas planilhas recebidas, há questões com o título “QSE_2”, “QSE_3” (que correspondem a “Qual o tipo de curso de Ensino Médio (2º Grau) que você concluiu?” e “Em que ano concluiu o Ensino Médio?”, respectivamente) e assim por diante.

Ao serem pressionados, os botões abrem uma tela com todos os valores possíveis ao atributo selecionado (Figura 4.6). Cada um destes valores possui um *checkbox* associado e, desta forma, para a criação do grupo deve-se selecionar os valores que o comporão, informar um nome e clicar em “Criar”.

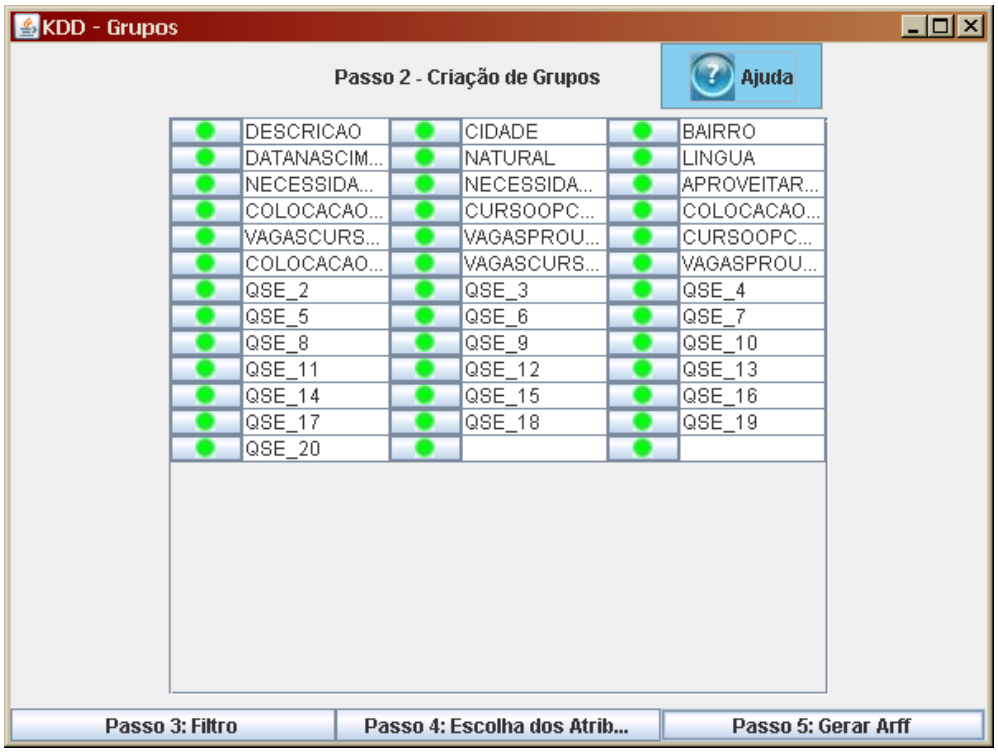

Figura 4.5: Tela de Criação de Grupos
Fonte: Do autor

Figura 4.6: Tela com lista de valores para criação de grupos
Fonte: Do autor

4.1.4 Filtros

Em muitas situações, o objetivo será de descobrir padrões em um subconjunto dos dados de vestibulandos, como apenas os candidatos de determinado(s) curso(s), ou de cidade ou região específica. Com o intuito de suprir essa necessidade foi implementada a tela Criação de Filtros, a qual já é acessível a partir da tela de Grupos. O leiaute e o processo são muito semelhantes ao da tela anterior. Também são listados os atributos encontrados e ao pressionar o botão que os acompanha é aberta a tela para seleção dos valores a serem filtrados, sendo ambas as telas exibidas nas figuras 4.7 e 4.8.

Figura 4.7: Tela de Criação de Filtros
Fonte: Do Autor

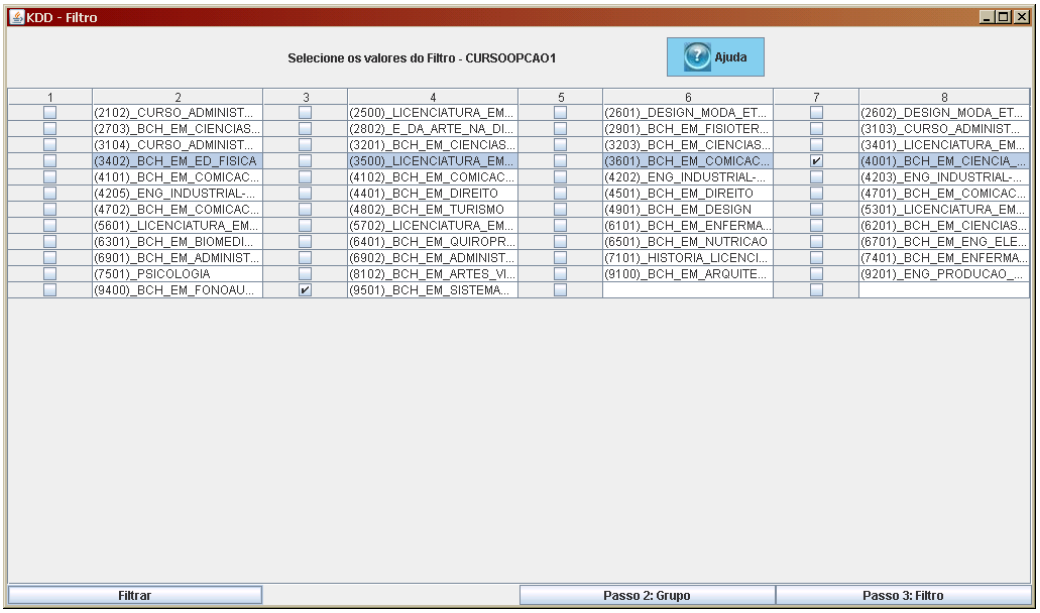

Figura 4.8: Tela com lista de valores para filtro
Fonte: Do autor

Tanto os filtros selecionados, como os grupos cadastrados, pode ser visualizados ou cancelados, sendo sua exibição na tela inicial da aplicação, conforme Figura 4.9.

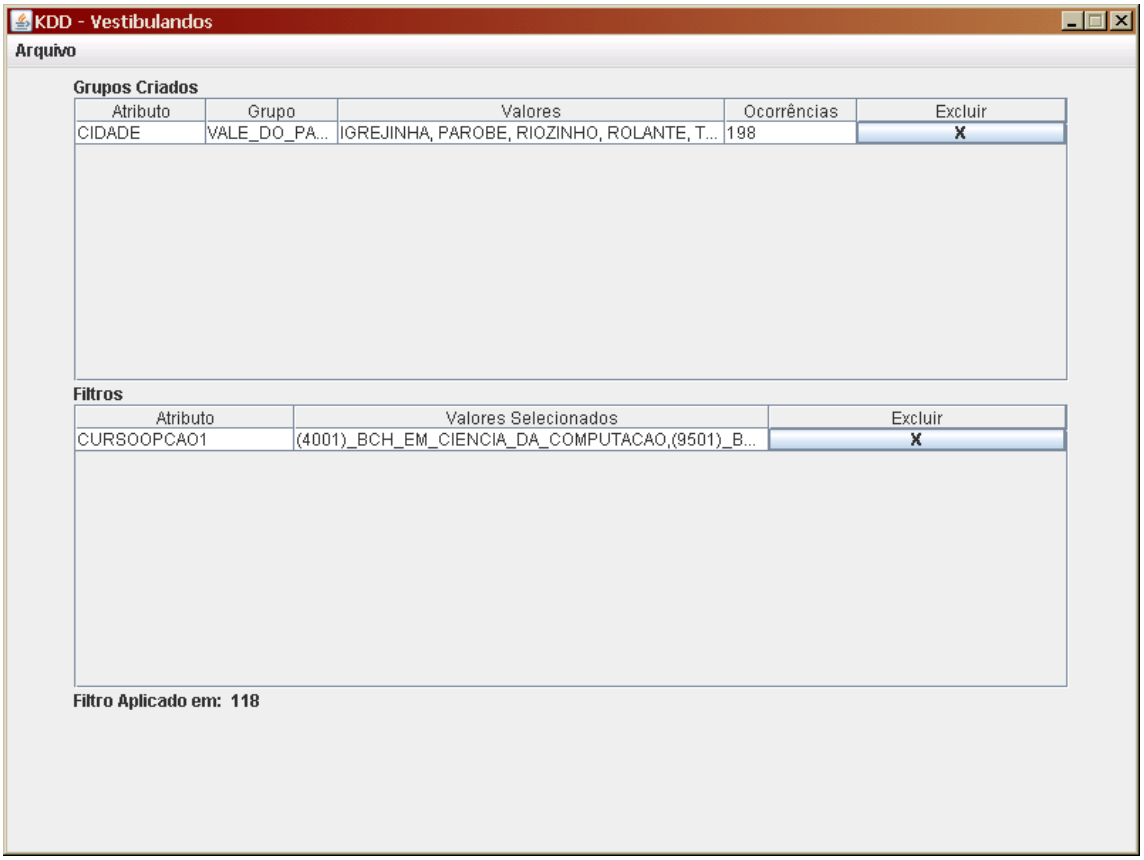

Figura 4.9: Relação de Grupos e Filtros criados
Fonte: Do autor

4.1.5 Seleção de Atributos

Gerando o arquivo .arff com todos os atributos, as regras tendem a ser grandes, expandindo a árvore. Tornou-se fundamental então, a possibilidade de o usuário escolher os atributos, para que na situação de procurar padrões em, por exemplo, cidade, opção de curso e renda familiar, sejam apenas selecionados apenas estes atributos.

Então, com o intuito de evitar o número elevado de regras e seu tamanho excessivo, essa funcionalidade de seleção de atributos passou a ser obrigatória, para que possa ser gerado o arquivo arff. A tela de Seleção de Atributos é exibida na Figura 4.10. Nela são apresentados os atributos com um *checkbox* ao lado, o qual deve ser selecionado caso o atributo deva constar no arquivo .arff.

Realizada a seleção dos atributos, pode-se partir para a criação do arquivo a ser utilizado pelo Weka, clicando em “Passo 5: Gerar Arff”. Buscando construir os conjuntos de teste e treinamento, ao pressionar o botão de geração de arquivo é aberta uma janela com a

possibilidade de definir o percentual a ser utilizado pelo conjunto de teste. Assim, informando o valor "30", este corresponde ao percentual aproximado do conjunto de teste enquanto os 70% restantes serão utilizados pelo conjunto de treinamento, para elaboração do modelo por parte do Weka. A geração ocorrendo com sucesso, são exibidos os números de registros em cada conjunto.

Figura 4.10: Tela de Seleção de Atributos
Fonte: Do Autor

4.2 KDD

Neste ponto, o arquivo já pode ser utilizado pelo Weka. Abrindo o arquivo com esta ferramenta, já são visualizados os atributos encontrados, conforme verificado na Figura 4.11.

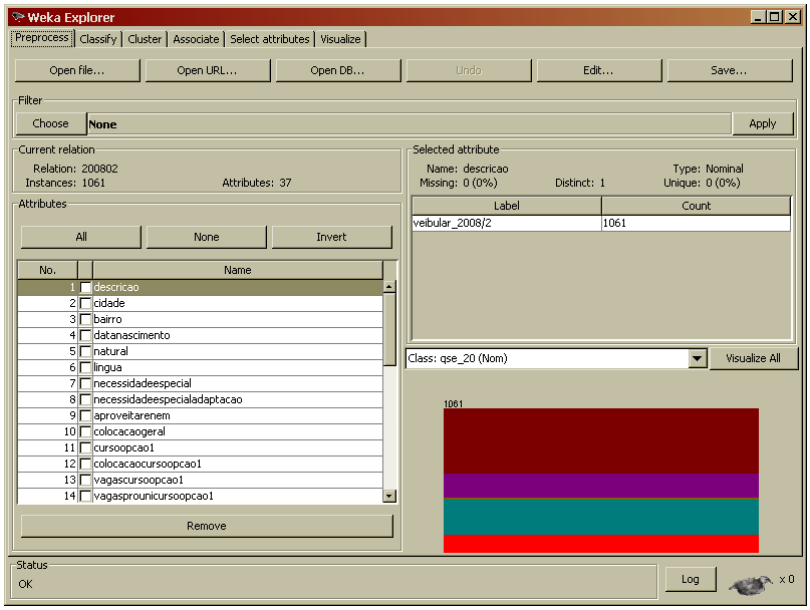

Figura 4.11: Arquivo .arff gerado acessado pelo Weka.
Fonte: Do Autor

Os atributos são apresentados na aba "*Preprocess*", sendo possível a seleção dos que serão utilizados para a aplicação de KDD. Nas demais abas há destaque para diferentes tipos de técnicas: uma específica para classificação, associação, agrupamento. Além disso, há uma para algoritmos a fim de selecionar os melhores atributos e outra para visualização de resultados.

Conforme visto na seção 4.1.5, a aplicação desenvolvida permite a escolha dos atributos a serem utilizados para a geração do arquivo .arff. Esta é uma funcionalidade presente no Weka. No entanto, os conjuntos de teste e treinamento devem possuir exatamente os mesmos atributos. Assim, ao excluir atributos no arquivo de treinamento lido pelo Weka, ocorreria conflito ao informar o de teste (pois a exclusão de atributos não refletiria no conjunto de teste). Sendo assim, a seleção de atributos deve ser anterior a geração do.arff.

Criados os arquivos de entrada para o Weka, optou-se inicialmente pela utilização do algoritmo J48, haja vista a sua utilização por um elevado número de trabalhos voltados para a técnica de classificação, e também por ser do segmento de árvores de decisão, com resultados que podem ser facilmente interpretados inclusive pelo usuário final. Os primeiros atributos escolhidos foram: cidade, curso, tipo de curso no Ensino Médio (Ensino Médio Regular, Técnico, Educação de Jovens e Adultos, Supletivo Regular, Magistério ou Exame da SEC) e forma de estudo no Ensino Médio (todo em escola pública, todo em escola particular, maior

parte em escola pública ou maior parte em escola particular), sendo Curso, o atributo escolhido como Classe.

Infelizmente, utilizando-se os conjuntos de teste e treinamento para todos os períodos, de 2006/2 a 2008/2, foi elevado o número de registros classificados erroneamente, conforme a Matriz de Confusão gerada pelo modelo. A Matriz de Confusão é um confronto dos dados do conjunto de teste com o modelo criado a partir do treinamento, informando quantos registros foram classificados corretamente.

Mais satisfatório foi o resultado quando da aplicação do algoritmo em atributos relativos ao grupo familiar dos vestibulandos, utilizando-se questões referentes à participação econômica do vestibulando na família, quem pagará os estudos, número de pessoas componentes do grupo familiar e se a atividade profissional está relacionada ao curso escolhido. Para estes atributos, a Matriz de Confusão indicou que a maioria dos registros foram classificados corretamente. No entanto, os melhores resultados não ocorreram para a classe Curso, como o exemplo na Figura 4.12, em que a classificação é decorrente da participação econômica na família.

A Matriz de Confusão abaixo deve ser lida da seguinte maneira: o valor na coluna correspondente a classe foi classificado corretamente, destacando-se 414 registros para a classe *a* (“nao_trabalho”) e 164 registros para a classe *c* (“trabalho-respondo_por_meu_sustento_e_ajudo_a_familia”) e assim por diante.

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 414   0  35   0   0 |   a = nao_trabalho
   7  23  29   0   0 |   b = sou_economicamente_independente_da_familia
  19  13 164   0   0 |   c = trabalho-_respondo_por_meu_sustento_e_ajudo_a_familia
   8   6  69   0   0 |   d = trabalho-_respondo_por_meu_sustento_e_da_familia
 136   5  92   0   0 |   e = trabalho_e_recebo_ajuda_da_familia
```

Figura 4.12: Exemplo de Matriz de Confusão
Fonte: Do Autor

Constatou-se que uma das principais causas dos resultados até então insatisfatórios é a dispersão dos dados. Por exemplo, no período com maior número de registros, 2007/01, dos 3.480 candidatos, aproximadamente 50% (1.622) inscreveu-se em Administração, Educação Física, Direito, Design, Enfermagem ou Biomedicina, sendo os demais nos 33 cursos restantes.

Buscando um melhor resultado, foram agrupadas as informações dos vestibulares de mesmo período (2006/02, 2007/02 e 2008/02 em uma planilha e 2007/01 e 2008/01 em outra). As regras geradas tiveram um maior número de registros as atendendo, todavia, o índice de registros classificados incorretamente continuou elevado. Utilizando-se do mesmo critério, foram solicitados à instituição os dados referentes aos demais períodos: 2003/01, 2003/02, 2004/01, 2004/02, 2005/01 e 2009/01.

4.3 Integração Aplicação x Weka

Com o objetivo de obter melhores resultados, buscou-se através de reuniões com funcionários do departamento de Marketing estabelecer os melhores parâmetros a serem utilizados pelos algoritmos de KDD, tanto na criação de grupos, como filtro e seleção de atributos. Além desses encontros, a ferramenta desenvolvida foi disponibilizada a esses funcionários, para que com a freqüente utilização, surgissem novas opções de parâmetros.

Conseqüentemente, careceu-se da integração entre o Weka (que pode ser acoplado a outros *software*, como mencionado na seção 1.4.1) e a ferramenta desenvolvida, para que o pré-processamento e a aplicação dos algoritmos de KDD ocorram no mesmo ambiente. A tela que utiliza procedimentos do Weka, é exibida logo após a geração do arquivo .arrf e pode ser visualizada na Figura 4.13

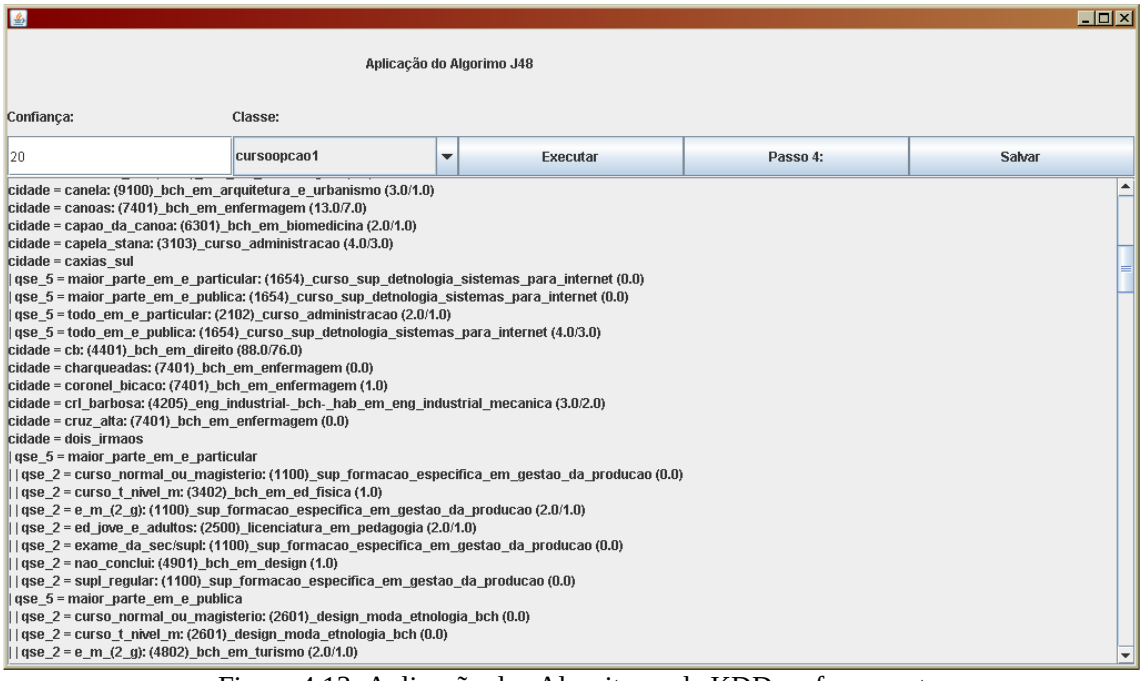

Figura 4.13: Aplicação dos Algoritmos de KDD na ferramenta
Fonte: Do Autor

O primeiro campo da tela é o fator de confiança, que corresponde ao percentual de incidências de registros que devam atender a regra a ser gerada. O segundo campo é uma caixa de seleção onde deve ser indicado o atributo que será utilizado como classe. O botão “Executar” aciona o algoritmo J48, sendo o resultado exibido no *frame* da tela.

A aplicação permite que seja exportado o resultado obtido, juntamente com a configuração do teste. O arquivo salvo possui todos os grupos criados, os filtros selecionados, os atributos escolhidos, o nome do arquivo da planilha eletrônica com os dados e a árvore gerada. Um exemplo de um arquivo gerado pode ser visto na Figura 4.14.

```
Arquivo: C:\Ingo\Projetos\tc\files\060708__2.xls

Grupos:

CIDADE
      PARANHANA: IGREJINHA, PAROBE, RIOZINHO, ROLANTE, TAQUARA

Filtros:

CURSOOPCAO1: (2102)_CURSO_ADMINISTRACAO, (3103)_CURSO_ADMINISTRACAO, (3104)_CURSO_ADMINISTRACAO,
(6901)_BCH_EM_ADMINISTRACAO_HAB_EM_NEGOCIOS_INTERNACIONAIS, (6902)_BCH_EM_ADMINISTRACAO_HAB_EM_NEGOCIOS_INTERNACIONAIS

Atributos Selecionados:CIDADE,QSE_8,QSE_11

Árvore:


Options: -C 0.20

J48 pruned tree
------------------

QSE_11 = ACIMA_9_SALARIOS_MINIMOS: NAO_TRABALHO (61.0/36.0) 41.00%
QSE_11 = ATE_3_SALARIOS_MINIMOS: TRABALHO-_RESPONDO_POR_MEU_SUSTENTO_E_AJUDO_A_FAMILIA (164.0/108.0) 34.00%
QSE_11 = ENTRE_3_E_6_SALARIOS_MINIMOS: TRABALHO-_RESPONDO_POR_MEU_SUSTENTO_E_AJUDO_A_FAMILIA (180.0/125.0) 31.00%
QSE_11 = ENTRE_6_E_9_SALARIOS_MINIMOS: TRABALHO_E_RECEBO_AJUDA_DA_FAMILIA (99.0/65.0) 34.00%
QSE_11 = QUESTAO_NAO_APLICADA_NO_CONCURSO: TRABALHO-_RESPONDO_POR_MEU_SUSTENTO_E_AJUDO_A_FAMILIA (291.0/211.0) 27.00%

Number of Leaves  :     5
```

Figura 4.14: Formato do arquivo com o teste salvo
Fonte: Do Autor

Neste capítulo foi apresentada a aplicação juntamente com as etapas de Pré-processamento e Mineração de Dados. No próximo são relatados os resultados, na etapa de Pós-processamento.

5 RESULTADOS

Neste capítulo são expostos os resultados obtidos a partir da aplicação dos algoritmos J48 e *Naive Bayes*, sendo cada seção correspondente a um conjunto de atributos em que foi executado um algoritmo ou ambos. Por conseguinte são destacadas as regras que podem vir a ter importância ao departamento de Marketing da instituição, das quais algumas foram salientadas inclusive por funcionários deste.

Após reuniões com estes funcionários e a disponibilização da ferramenta desenvolvida, constatou-se que resultados obtidos com o algoritmo J48 que até então eram descartados por gerar apenas uma regra, tem grande valia, como pode ser verificado na próxima seção.

5.1 Meio de Comunicação

Selecionando os dados do vestibular de inverno de 2006 a 2008, filtrando apenas por vestibulandos que para a questão: “Além da internet, qual o outro canal de comunicação que você obteve informações sobre este vestibular?” responderam “Através do Jornal NH”, obteve-se a seguinte regra para a classificação para a questão “Cidade”:

- NOVO_HAMBURGO (635.0/354.0) 44.00%

Esta regra indica que 635 vestibulandos tiveram a resposta selecionada, e que destes 44 % são residentes na cidade de Novo Hamburgo. O mais relevante, contudo, é a informação que está implícita: mais que a metade dos vestibulandos que receberam informações do vestibular a partir do Jornal NH, principal jornal do município de Novo Hamburgo, são provenientes de outras cidades, constatando que o investimento em publicidade neste jornal tem grande repercussão na região.

5.2 Relação Tipo de ensino médio X curso

Selecionando apenas os atributos referentes ao curso e ao tipo de Ensino Médio, foram encontradas semelhanças nesta relação nos vestibulares de inverno de 2007 e 2008, sendo as regras exibidas nas Figuras 5.1 e 5.2. Verificou-se que independente do tipo de Ensino Médio, o curso de enfermagem é o de maior freqüência.

```
J48 pruned tree
------------------

QSE_2 = CURSO_NORMAL_OU_MAGISTERIO: (7401)_BCH_EM_ENFERMAGEM (70.0/58.0) 17.00%
QSE_2 = CURSO_T_NIVEL_M: (7401)_BCH_EM_ENFERMAGEM (217.0/165.0) 24.00%
QSE_2 = ED_JOVE_E_ADULTOS: (3103)_CURSO_ADMINISTRACAO (135.0/116.0) 14.00%
QSE_2 = EXAME_DA_SEC_SUPL: (7401)_BCH_EM_ENFERMAGEM (41.0/35.0) 15.00%
QSE_2 = E_M_(2_G): (3103)_CURSO_ADMINISTRACAO (1285.0/1137.0) 12.00%
QSE_2 = SUPL_REGULAR: (7401)_BCH_EM_ENFERMAGEM (135.0/110.0) 19.00%

Number of Leaves  :     6
```

Figura 5.1: Tipo de ensino médio X curso em 2007/02
Fonte: Do autor

No destaque da figura acima, aufere-se que dos vestibulandos que responderam “Curso Normal ou Magistério” para a pergunta “Qual o tipo de curso de Ensino Médio (2º Grau) que você concluiu” (“QSE_2”), 17% inscreveram-se em Bacharelado em Enfermagem. Este percentual é o resultado aproximado de (70-58)/70, onde 70 é o total de concorrentes que responderam à alternativa em questão e 58 é o número de registros descartados para elaboração da regra. Já para a execução do algoritmo no concurso de 2008/02, é destacada a ocorrência do maior percentual entre os dois períodos apurados, 33%.

```
J48 pruned tree
------------------

QSE_2 = CURSO_NORMAL_OU_MAGISTERIO: (3500)_LICENCIATURA_EM_PEDAGOGIA (70.0/58.0) 17.00%
QSE_2 = CURSO_T_NIVEL_M: (7401)_BCH_EM_ENFERMAGEM (148.0/126.0) 15.00%
QSE_2 = ED_JOVE_E_ADULTOS: (7401)_BCH_EM_ENFERMAGEM (104.0/96.0) 8.00%
QSE_2 = EXAME_DA_SEC_SUPL: (7401)_BCH_EM_ENFERMAGEM (21.0/18.0) 14.00%
QSE_2 = E_M_(2_G): (7401)_BCH_EM_ENFERMAGEM (1003.0/897.0) 11.00%
QSE_2 = NAO_CONCLUI: (6301)_BCH_EM_BIOMEDICINA (231.0/209.0) 10.00%
QSE_2 = SUPL_REGULAR: (7401)_BCH_EM_ENFERMAGEM (84.0/56.0) 33.00%

Number of Leaves  :     7
```

Figura 5.2: Tipo de ensino médio X curso em 2008/02
Fonte: Do Autor

Utilizando como base os mesmos concursos (2007/2 e 2008/2) foi executado o J48 somente em registros de vestibulandos para o curso de Ciência da Computação e em ambos gerou-se somente uma regra correspondente ao Ensino Médio regular (“E_M_(2_G)”), 71% (41/58) e 62% (25/40) respectivamente.

A implementação do *Naive Bayes* para o Weka apresenta a probabilidade para todos os valores possíveis dos atributos. Aplicando o algoritmo com a mesma configuração das figuras 5.1 e 5.2, para cada tipo de Ensino Médio é exibida a ocorrência de todos os cursos.

As tabelas 5.1 e 5.2 exibem a classificação dos cursos com maior probabilidade de vestibulandos com curso técnico de nível médio. Assim como com o algoritmo J48, o curso com maior incidência foi o de Bacharelado em Enfermagem.

Como vantagem desses resultados, tem-se a avaliação dos demais cursos, não somente o de maior incidência. É verificada a ocorrência de Ciência da Computação, Bacharelado em Direito, Engenharia Industrial – Bacharelado, Habilitação em Engenharia Industrial Mecânica, Administração e Formação Específica em Gestão da Produção nas duas relações.

No entanto, essa forma de apresentação pode ser muito extensa. Ressalta-se que os resultados a seguir representam a classificação para apenas um dos valores possíveis para o tipo de Ensino Médio (técnico de nível médio). Se necessária a análise dos resultados na íntegra, seria produzida uma tabela com o número de células próximo ao produto do número de diferentes tipos de Ensino Médio pela quantidade de cursos.

Tabela 5.1: Tipo de ensino médio X curso em 2007/02 – *Naive Bayes*

Curso	Valor
(7401)_BCH_EM_ENFERMAGEM	0,2086
(1100)_SUP_FORMACAO_ESPECIFICA_EM_GESTAO_DA_PRODUCAO	0,0708
(4401)_BCH_EM_DIREITO	0,0629
(3103)_CURSO_ADMINISTRACAO	0,0511
(4001)_BCH_EM_CIENCIA_DA_COMPUTACAO	0,0472
(4205)_ENG_INDUSTRIAL-_BCH-_HAB_EM_ENG_INDUSTRIAL_MECANICA	0,0393
(3201)_BCH_EM_CIENCIAS_CONTABEIS	0,0354
(3203)_BCH_EM_CIENCIAS_CONTABEIS	0,0354
(6701)_BCH_EM_ENG_ELETRONICA	0,0354

Fonte: Do autor

Salienta-se que ambas as tabelas apresentam apenas os cursos com maior possibilidade de um vestibulando com curso técnico de nível médio estar inscrito, sendo os

dados ordenados por probabilidade decrescente, destacando-se a grande diferença entre Bacharelado em Enfermagem das demais, sendo mais que o dobro que o segundo curso.

Tabela 5.2: Tipo de ensino médio X curso em 2008/02 – *Naive Bayes*

Curso	Valor
(7401)_BCH_EM_ENFERMAGEM	0,1216
(1100)_SUP_FORMACAO_ESPECIFICA_EM_GESTAO_DA_PRODUCAO	0,0740
(3104)_CURSO_ADMINISTRACAO	0,0740
(1654)_CURSO_SUP_DETNOLOGIA_SISTEMAS_PARA_INTERNET	0,0476
(4205)_ENG_INDUSTRIAL-_BCH-_HAB_EM_ENG_INDUSTRIAL_MECANICA	0,0423
(4205)_ENG_INDUSTRIAL-_BCH-_HAB_EM_ENG_INDUSTRIAL_MECANICA	0,0423
(6301)_BCH_EM_BIOMEDICINA	0,0370
(4501)_BCH_EM_DIREITO	0,0370
(2602)_DESIGN_MODA_ETNOLOGIA_BCH	0,0317

Fonte: Do autor

5.3 Relação Escola X Curso

Uma das relações mais interessantes para o departamento de Marketing é a descoberta de qual curso tem maior incidência de candidatos para cada escola de Ensino Médio. Contudo, filtros devem ser aplicados, pois são centenas de diferentes escolas de origem para cada concurso.

Buscando conhecimento dessa relação nos vestibulares de inverno de 2006, 2007 e 2008, aplicou-se o filtro pela cidade de Novo Hamburgo. Os atributos escolhidos foram curso e escola de Ensino Médio. Executado o J48, alternando os atributos de classe, foram geradas as seguintes árvores:

```
J48 pruned tree
------------------
: (3103)_CURSO_ADMINISTRACAO (1484.0/1297.0) 13.00%

Number of Leaves  :     1
```

Figura 5.3: Escola X Curso – Classificado por Curso
Fonte: Do Autor

```
(1100)_SUP_FORMACAO_ESPECIFICA_EM_GESTAO_DA_PRODUCAO: NOVO_HAMBURGO_FUND_ESC_T_LIBERATO_SALZANO_V_DA_CUNHA (32.0/27.0) 16.00%
(2102)_CURSO_ADMINISTRACAO: NOVO_HAMBURGO_C_EST_SENADOR_ALBERTO_PASQUALINI (26.0/23.0) 12.00%
(2601)_DESIGN_MODA_ETNOLOGIA_BCH: NOVO_HAMBURGO_C_MARISTA_PIO_XII (13.0/9.0) 31.00%
(2602)_DESIGN_MODA_ETNOLOGIA_BCH: NOVO_HAMBURGO_UNID_E_FUND_EVANG (35.0/29.0) 17.00%
(2901)_BCH_EM_FISIOTERAPIA: NOVO_HAMBURGO_C_EST_25_JUL (35.0/30.0) 14.00%
(3103)_CURSO_ADMINISTRACAO: NOVO_HAMBURGO_C_MARISTA_PIO_XII (187.0/169.0) 10.00%
(3104)_CURSO_ADMINISTRACAO: NOVO_HAMBURGO_C_MARISTA_PIO_XII (77.0/70.0) 9.00%
(3201)_BCH_EM_CIENCIAS_CONTABEIS: NOVO_HAMBURGO_C_ES (42.0/38.0) 10.00%
(3203)_BCH_EM_CIENCIAS_CONTABEIS: NOVO_HAMBURGO_C_EST_25_JUL (35.0/30.0) 14.00%
(3401)_LICENCIATURA_EM_ED_FISICA: NOVO_HAMBURGO_C_EST_25_JUL (37.0/30.0) 19.00%
(3500)_LICENCIATURA_EM_PEDAGOGIA: NOVO_HAMBURGO_C_ES (24.0/19.0) 21.00%
(3601)_BCH_EM_COMICACAO_SOCIAL-_HAB_EM_RELACOES_PUBLICAS: NOVO_HAMBURGO_E_ED_BASICA_ESC_APLICACAO_FEEVALE (23.0/19.0) 17.00%
(4001)_BCH_EM_CIENCIA_DA_COMPUTACAO: NOVO_HAMBURGO_C_E_25_JUL (56.0/51.0) 9.00%
(4101)_BCH_EM_COMICACAO_SOCIAL_HAB_EM_PUBLICIDADE_E_PROPAGANDA: NOVO_HAMBURGO_E_ED_BASICA_ESC_APLICACAO_FEEVALE (42.0/34.0) 19.
(4203)_ENG_INDUSTRIAL-_BCH-_HAB_EM_ENG_INDUSTRIAL_QUIMICA: NOVO_HAMBURGO_FUND_ESC_T_LIBERATO_SALZANO_V_DA_CUNHA (20.0/16.0) 20.
(4205)_ENG_INDUSTRIAL-_BCH-_HAB_EM_ENG_INDUSTRIAL_MECANICA: NOVO_HAMBURGO_FUND_ESC._TEC._LIBERATO_SALZANO_V._DA_CUNHA (35.0/31.
(4401)_BCH_EM_DIREITO: NOVO_HAMBURGO_C_ES (108.0/92.0) 15.00%
(4501)_BCH_EM_DIREITO: NOVO_HAMBURGO_C_ES (37.0/31.0) 16.00%
(4702)_BCH_EM_COMICACAO_SOCIAL_HAB_EM_JORNALISMO: NOVO_HAMBURGO_C_ES (33.0/28.0) 15.00%
(4901)_BCH_EM_DESIGN: NOVO_HAMBURGO_C_STA_CATARINA (54.0/47.0) 13.00%
(5601)_LICENCIATURA_EM_LETRAS_HAB_PORTUGUES_INGLES_E_RESPECTIVAS_LITERATURAS: NOVO_HAMBURGO_C_EST_25_JUL (20.0/13.0) 35.00%
(6201)_BCH_EM_CIENCIAS_FARMACEUTICAS: NOVO_HAMBURGO_C_EST_25_JUL (22.0/18.0) 18.00%
(6301)_BCH_EM_BIOMEDICINA: NOVO_HAMBURGO_C_MARISTA_PIO_XII (59.0/49.0) 17.00%
(6401)_BCH_EM_QUIROPRAXIA: NOVO_HAMBURGO_C_STA_CATARINA (16.0/12.0) 25.00%
(6501)_BCH_EM_NUTRICAO: NOVO_HAMBURGO_C_ES (21.0/18.0) 14.00%
(7401)_BCH_EM_ENFERMAGEM: NOVO_HAMBURGO_C_STA_CATARINA (106.0/84.0) 21.00%
(7501)_PSICOLOGIA: NOVO_HAMBURGO_UNID_E_FUND_EVANG (40.0/34.0) 15.00%
(9100)_BCH_EM_ARQUITETURA_E_URBANISMO: NOVO_HAMBURGO_C_STA_CATARINA (57.0/51.0) 11.00%
```

Figura 5.4: Escola X Curso – Classificado por Escola
Fonte: Do autor

Ocorrendo a classificação por escolas, é verificado um excessivo número de regras, sendo várias destas pouco relevantes (algumas foram retiradas). No entanto, acredita-se ser esta uma característica da técnica de classificação: a cada aplicação de algoritmo, das várias regras geradas, apenas algumas são aproveitadas. Neste exemplo, há várias informativas, como por exemplo, a que indica que 21% dos vestibulandos para o curso de Bacharelado em Enfermagem do município de Novo Hamburgo são provenientes do Colégio Santa Catarina.

5.4 Meio de Comunicação - Jornais

O classificador *Naive Bayes* demonstrou-se muito eficaz para atributos com poucos valores, por exemplo, para a pergunta “Além da Internet, qual o outro canal de comunicação que você obteve informações sobre este vestibular?”, filtrando-se pelos Jornais em que houve divulgação do Vestibular, o J48 retornou apenas a regra:

```
J48 pruned tree
------------------
: ATRAVES_JORNAL_NH (800.0/165.0) 79.00%

Number of Leaves  :     1
```

Figura 5.5: Regra gerada pelo J48 - Meios de Comunicação
Fonte: Do Autor

Já o *Naive Bayes* indica as probabilidades de todos os valores possíveis, exibidas na Tabela 5.3.

Tabela 5.3: Meios de Comunicação – *Naive Bayes*

Jornal	Probabilidade
JORNAL_COMERCIO	0.01234568
JORNAL_CORREIO_POVO	0.03209877
JORNAL_NH	0.78518519
ZERO_HORA	0.15555556
DIARIO	0.00246914
GRAMADO	0.00246914
O_DIARIO	0.00246914
PIONEIRO	0.00246914

Fonte: Do autor

5.5 Faixa Salarial – Grupo Familiar

Algumas questões eram referentes à situação econômica do vestibulando. É aberta então a possibilidade de identificar a faixa salarial predominante nos vestibulandos de determinado curso (ou área). Selecionando apenas os cursos Sistemas de Informação e Ciência da Computação, obtiveram-se as regras da Figura 5.6. Com a execução do J48 sobre os dados filtrados por candidatos da área da Saúde, obteve-se resultado (Figura 5.7) semelhante ao curso de Sistemas de Informação.

```
Options: -C 0.20

J48 pruned tree
------------------

CURSOOPCAO1 = (4001)_BCH_EM_CIENCIA_DA_COMPUTACAO: ATE_3_SALARIOS_MINIMOS (98.0/59.0) 40.00%
CURSOOPCAO1 = (9501)_BCH_EM_SISTEMAS_INFORMACAO: ENTRE_3_E_6_SALARIOS_MINIMOS (46.0/28.0) 39.00%

Number of Leaves  :     2
```

Figura 5.6: Faixa Salarial X Ciência da Computação e Sistemas de Informação
Fonte: Do Autor

```
Options: -C 0.20

J48 pruned tree
------------------
: ENTRE_3_E_6_SALARIOS_MINIMOS (1022.0/599.0) 41.00%

Number of Leaves  :     1
```

Figura 5.7: Faixa Salarial X Ciência da Computação e Sistemas de Informação
Fonte: Do Autor

Se tratando de mais uma aplicação com poucos atributos, o resultado obtido com o *Naive Bayes* é exibido na íntegra na Figura 5.8, contatando-se mais uma vez que os

algoritmos produziram resultados semelhantes, haja vista a classe com maior probabilidade (ENTRE_3_E_6_SALARIOS_MINIMOS, com aproximadamente 41%) ser a mesma apontada pelo algoritmo J48. Salienta-se a diferença de probabilidade entre esta e as demais: ATE_3_SALARIOS_MINIMOS, com 32%; ENTRE_6_E_9_SALARIOS_MINIMOS, com 16% e ACIMA_9_SALARIOS_MINIMOS, com 9%.

```
Naive Bayes (simple)

Class ACIMA_9_SALARIOS_MINIMOS: P(C) = 0.09356725 100%

Class ATE_3_SALARIOS_MINIMOS: P(C) = 0.32358674 100%

Class ENTRE_3_E_6_SALARIOS_MINIMOS: P(C) = 0.41325536 100%

Class ENTRE_6_E_9_SALARIOS_MINIMOS: P(C) = 0.16959064 100%
```

Figura 5.8: Faixa Salarial X Cursos da área da Saúde – *Naive Bayes*
Fonte: Do Autor

5.6 Instituições concorrentes – Ciência da Computação

Foi realizado o filtro por candidatos para o curso de Ciência da Computação para os vestibulares de verão de 2008 e 2009, sendo selecionado o atributo referente à questão “Em qual destas instituições você também está prestando vestibular?” com o objetivo de identificar o concorrente mais procurado pelos candidatos. Assim como o teste da seção 5.3.2, o algoritmo J48 gerou apenas uma regra, exibida na Figura 5.9.

```
Options: -C 0.20

J48 pruned tree
------------------
: NENHUMA (162.0/83.0) 49.00%

Number of Leaves  :     1
```

Figura 5.9: Instituições concorrentes - Ciência da Computação - J48
Fonte: Do Autor

Apesar de ser interessante para a instituição a informação de que a maior parte dos candidatos não presta vestibular em um concorrente, a aplicação do J48 neste caso foi de pouca valia. Isto porque não retornou, por menor que seja, uma posição relativa às outras instituições. Conforme ocorrido outrora, o resultado gerado pelo *Naive Bayes* foi mais informativo, sendo tabulado e ordenado por probabilidade na Tabela 5.4.

Tabela 5.4: Instituições concorrentes – Ciência da Computação - *Naive Bayes*

Instituição	Probabilidade
Nenhuma	0.33057851
UNISINOS	0.21900826
UFRGS	0.06198347
FACCAT	0.02066116
UCS	0.01652893
ULBRA	0.01652893
PUC	0.01239669
UERGS	0.01239669
UFV-MG	0.00826446
UNIVALI	0.00826446
UTFPR	0.00826446

Fonte: Do Autor

5.7 Motivo de Escolha

Aplicando-se o filtro por vestibulandos residentes em cidades vizinhas, tem-se o objetivo de investigar o principal motivo que levou a escolha da instituição em questão por parte dos candidatos.

A pergunta em questão foi “Qual o principal motivo que o levou a optar pela Feevale?” e as cidades selecionadas foram Campo Bom, Canoas, Estância Velha, Esteio, Porto Alegre, Portão, São Leopoldo, Sapiranga, Sapucaia, Taquara, Viamão. O algoritmo J48 gerou a árvore da Figura 5.10, utilizando o atributo referente à pergunta como classe.

```
Options: -C 0.20

J48 pruned tree
------------------

CIDADE = CAMPO_BOM: LOCALIZACAO (261.0/153.0) 41.00%
CIDADE = CANOAS: OFERECE_O_CURSO_DESEJADO (49.0/36.0) 27.00%
CIDADE = ESTANCIA_VELHA: LOCALIZACAO (180.0/92.0) 49.00%
CIDADE = ESTEIO: OFERECE_O_CURSO_DESEJADO (34.0/21.0) 38.00%
CIDADE = POA: OFERECE_O_CURSO_DESEJADO (196.0/130.0) 34.00%
CIDADE = PORTAO: QUALIDADE_ACADEMICA (47.0/33.0) 30.00%
CIDADE = SAO_LEOPOLDO: OFERECE_O_CURSO_DESEJADO (251.0/152.0) 39.00%
CIDADE = SAPIRANGA: LOCALIZACAO (241.0/152.0) 37.00%
CIDADE = SAPUCAIA: OFERECE_O_CURSO_DESEJADO (38.0/26.0) 32.00%
CIDADE = TAQUARA: OFERECE_O_CURSO_DESEJADO (78.0/44.0) 44.00%
CIDADE = VIAMAO: DISPONIBILIDADE_CURSOS_NO_TURNO_FISEM (10.0/6.0) 40.00%

Number of Leaves  :     11
```

Figura 5.10: Motivo de Escolha X Cidade
Fonte: Do autor

O resultado encontrado é muito satisfatório para a instituição, haja vista que em apenas três das cidades selecionadas (Campo Bom, Estância Velha e Sapiranga) o principal motivo foi a localização, sendo os demais referentes a atribuições do centro universitário: qualidade, disponibilidade de cursos em finais de semana (FISEM) ou de curso específico.

O algoritmo *Naive Bayes* apresentou o resultado da Tabela 5.5. Note-se que os resultados de ambos os algoritmos foi idêntico: cada motivo classificado para as cidades pelo J48 foi o de maior probabilidade calculada pelo *Naive Bayes*

Tabela 5.5: Motivo Escolha X Cidade - *Naive Bayes*

Cidade	Classe	Probabilidade
Campo Bom	Localização	0,34713376
Canoas	Oferece o curso desejado	0,1372549
Estância Velha	Localização	0,38197425
Esteio	Oferece o curso desejado	0,16091954
Porto Alegre	Oferece o curso desejado	0,26907631
Portão	Qualidade Acadêmica	0,15
São Leopoldo	Oferece o curso desejado	0,32894737
Sapiranga	Localização	0,30612245
Sapucaia do Sul	Oferece o curso desejado	0,14285714
Taquara	Oferece o curso desejado	0,26717557
Viamão	Disponibilidade FISEM	0,07936508

Fonte: Do autor

5.8 Candidatos com Ensino Médio concluído X não concluído

Em todos os concursos, conforme verificado nas planilhas, há ocorrência de candidatos que ainda não estão aptos para o ingresso no Ensino Superior, pois prestam o exame para testar seus conhecimentos antes de concluir o Ensino Médio. Devido a este fato, passa a ser interessante para instituição a descoberta de relações entre cursos e concluintes e não concluintes, para um maior controle entre oferta de vagas e vestibulandos com possibilidade de iniciar o curso desejado.

Selecionando os registros dos concursos de inverno de 2007 e 2008, e verificando o atributo referente a situação do candidato em relação ao Ensino Médio, encontrou-se a regra exibida na Figura 5.11 após a execução do algoritmo J48.

```
J48 pruned tree
------------------
: JA_CONCLUI (3544.0/735.0) 79.00%

Number of Leaves  :     1
```

Figura 5.11: Situação em relação ao Ensino Médio
Fonte: Do autor

Constatado que aproximadamente 79% dos candidatos já concluíram o Ensino Médio, partiu-se então com testes utilizando os 735 registros de candidatos que ainda não o concluíram. Buscando uma relação entre os não concluintes, município e curso, obteve-se a árvore apresentada na Figura 5.12.

Apesar das informações encontrarem-se dispersas, com várias regras com baixa freqüência, destacam-se que a procura dos não concluintes dos municípios de Ivoti, Novo Hamburgo, Porto Alegre e São Leopoldo: Publicidade e Propaganda (21%), Administração (10%), Design (29%) e Administração (16%), respectivamente.

```
CIDADE = BOM_PRINCIPIO: (4901)_BCH_EM_DESIGN (1.0) 100%
CIDADE = CAC: (6401)_BCH_EM_QUIROPRAXIA (1.0) 100%
CIDADE = CAMAQUA: (2901)_BCH_EM_FISIOTERAPIA (1.0) 100%
CIDADE = CAMPO_BOM: (4401)_BCH_EM_DIREITO (69.0/61.0) 12.00%
CIDADE = CANELA: (6301)_BCH_EM_BIOMEDICINA (4.0/1.0) 75.00%
CIDADE = CANOAS: (6301)_BCH_EM_BIOMEDICINA (8.0/5.0) 38.00%
CIDADE = CARAZINHO: (2601)_DESIGN_MODA_ETNOLOGIA_BCH (1.0) 100%
CIDADE = CAXIAS_SUL: (6401)_BCH_EM_QUIROPRAXIA (6.0/4.0) 33.00%
CIDADE = CURITIBA: (7401)_BCH_EM_ENFERMAGEM (1.0) 100%
CIDADE = DOIS_IRMAOS: (2703)_BCH_EM_CIENCIAS_BIOLOGICAS (21.0/19.0) 10.00%
CIDADE = ESTANCIA_VELHA: (3103)_CURSO_ADMINISTRACAO (29.0/26.0) 10.00%
CIDADE = ESTEIO: (3103)_CURSO_ADMINISTRACAO (1.0) 100%
CIDADE = FARROUPILHA: (6301)_BCH_EM_BIOMEDICINA (2.0) 100%
CIDADE = GRAVATAI: (2901)_BCH_EM_FISIOTERAPIA (1.0) 100%
CIDADE = IGREJINHA: (3103)_CURSO_ADMINISTRACAO (21.0/18.0) 14.00%
CIDADE = IVO: (4101)_BCH_EM_COMICACAO_SOCIAL_HAB_EM_PUBLICIDADE_E_PROPAGANDA (14.0/11.0) 21.00%
CIDADE = LINDOLFO_COLLOR: (6201)_BCH_EM_CIENCIAS_FARMACEUTICAS (1.0) 100%
CIDADE = LINHA_NOVA: (3201)_BCH_EM_CIENCIAS_CONTABEIS (1.0) 100%
CIDADE = MONTENEGRO: (2901)_BCH_EM_FISIOTERAPIA (9.0/7.0) 22.00%
CIDADE = NH: (6301)_BCH_EM_BIOMEDICINA (1.0) 100%
CIDADE = NOVA_HARTZ: (4501)_BCH_EM_DIREITO (13.0/10.0) 23.00%
CIDADE = NOVA_PRATA: (6301)_BCH_EM_BIOMEDICINA (1.0) 100%
CIDADE = NOVO_HAMBURGO: (3103)_CURSO_ADMINISTRACAO (255.0/229.0) 10.00%
CIDADE = OSORIO: (2601)_DESIGN_MODA_ETNOLOGIA_BCH (1.0) 100%
CIDADE = PARAI: (2601)_DESIGN_MODA_ETNOLOGIA_BCH (1.0) 100%
CIDADE = PAROBE: (4501)_BCH_EM_DIREITO (21.0/18.0) 14.00%
CIDADE = PELOTAS: (9100)_BCH_EM_ARQUITETURA_E_URBANISMO (1.0) 100%
CIDADE = POA: (2601)_DESIGN_MODA_ETNOLOGIA_BCH (31.0/22.0) 29.00%
CIDADE = ROLANTE: (6401)_BCH_EM_QUIROPRAXIA (11.0/9.0) 18.00%
CIDADE = SAO_LEOPOLDO: (3103)_CURSO_ADMINISTRACAO (37.0/31.0) 16.00%
```

Figura 5.12: Curso X Município X não concluintes
Fonte: Do autor

5.9 Conhecimento Prévio do Curso

Com o objetivo de comparar o conhecimento anterior que os candidatos possuem do curso desejado, foram realizados testes com todos os registros, ainda nos concursos de

inverno de 2007 e 2008, e também especificamente para o curso de Ciência da Computação. O atributo selecionado foi referente à pergunta “Qual o grau de conhecimento que você possui sobre o curso pretendido?”. As figuras 5.13 e 5.14 apresentam os resultados após aplicação do algoritmo *Naive Bayes*.

```
Naive Bayes (simple)

Class ELEVADO_(ACIMA_DA_MEDIA): P(C) = 0.16032685 100%
Class NENHUM: P(C) = 0.05494505 100%
Class POUCO_(ABAIXO_DA_MEDIA): P(C) = 0.18258664 100%
Class RAZOAVEL_(NA_MEDIA): P(C) = 0.57311919 100%
Class TOTAL: P(C) = 0.02902226 100%
```

Figura 5.13: Conhecimento do Curso - todos
Fonte: Do autor

```
Naive Bayes (simple)

Class ELEVADO_(ACIMA_DA_MEDIA): P(C) = 0.19417476 100%
Class NENHUM: P(C) = 0.03883495 100%
Class POUCO_(ABAIXO_DA_MEDIA): P(C) = 0.15533981 100%
Class RAZOAVEL_(NA_MEDIA): P(C) = 0.58252427 100%
Class TOTAL: P(C) = 0.02912621 100%
```

Figura 5.14: Conhecimento do Curso – Ciência da Computação
Fonte: Do autor

Constatou-se que os candidatos de Ciência da Computação têm perfil semelhante ao perfil do conjunto todo, haja vista a classificação com maior probabilidade ter sido RAZOAVEL_(NA_MEDIA) em ambos os testes. Todavia, as classificações não foram idênticas devido ao fato de, após ordenar as classes por probabilidades, ocorrer a inversão na ordem das classes POUCO_(ABAIXO_DA_MEDIA) e ELEVADO_(ACIMA_DA_MEDIA), ou seja, os vestibulandos de Ciência da Computação possuem maior probabilidade de conhecimento elevado do curso do que pouco conhecimento, o oposto do perfil geral dos vestibulandos.

5.10 Curso X Atividade Profissional

Assim como no teste anterior, foi realizada uma comparação entre candidatos do conjunto todo e específicos do curso de Ciência da Computação. Foi selecionado o atributo referente à questão “Sua atividade profissional está relacionada com o curso pretendido?”, com as alternativas “Sim”, “Não” e “Não Estou Trabalhando”, nos concursos de verão de 2008 e 2009. A figura 5.15 indica que em 35% dos vestibulandos já atuam profissionalmente na área do curso escolhido. Já na Figura 5.16 é indicado o resultado apresentado pela

aplicação do J48 somente em candidatos do curso de Ciência da Computação, ocorrendo nesta relação divergência entre os perfis, já que para este curso a classe destacada foi NAO_ESTOU_TRABALHANDO, com 35%.

```
J48 pruned tree
------------------
: SIM (6733.0/4382.0) 35.00%

Number of Leaves  :     1
```

Figura 5.15: Relação de vestibulandos com atividade profissional
Fonte: Do autor

```
J48 pruned tree
------------------
: NAO_ESTOU_TRABALHANDO (162.0/106.0) 35.00%

Number of Leaves  :     1
```

Figura 5.16: Ciência da Computação X atividade profissional
Fonte: Do autor

No entanto, buscando por uma classificação para todas as alternativas da pergunta, aplicou-se o *Naive Bayes* ainda sobre os candidatos de Ciência da Computação, verificando uma mínima diferença de probabilidade entre cada alternativa, como pode ser visto na Figura 5.17.

```
Naive Bayes (simple)

Class NAO: P(C) = 0.32727273 100%
Class NAO_ESTOU_TRABALHANDO: P(C) = 0.34545455 100%
Class SIM: P(C) = 0.32727273 100%
```

Figura 5.17: Ciência da Computação X atividade profissional – *Naive Bayes*
Fonte: Do autor

Através da aplicação do J48 sobre todos os vestibulandos, mas desta vez classificando com relação aos cursos obteve-se uma árvore com praticamente todas as regras possuindo valores bastante significativos. A árvore em questão é exibida nas figuras 5.18 e 5.19.

Das regras geradas, destacam-se as elevadas freqüências da classe SIM para os cursos Gestão da Produção (88%), Gestão em Recursos Humanos (62%), Enfermagem (66%) e Administração (68%); NÃO para os cursos Ciências Biológicas (52%), Turismo (58%) e Educação Física (Licenciatura e Bacharelado com mais de 50%) e NAO_ESTOU_TRABALHANDO, salientando-se Publicidade e Propaganda e cursos na área da Saúde, como Quiropraxia, Fisioterapia e Nutrição, com mais de 50%.

```
J48 pruned tree
------------------

CURSOOPCAO1 = (1000)_SUP_FORMACAO_ESPECIFICA_EM_GESTAO_DA_PRODUCAO: SIM (70.0/15.0) 79.00%
CURSOOPCAO1 = (1100)_SUP_FORMACAO_ESPECIFICA_EM_GESTAO_DA_PRODUCAO: SIM (96.0/12.0) 88.00%
CURSOOPCAO1 = (1648)_CURSO_SUP_DETNOLOGIA_EM_GESTAO_RECURSOS_HUMANOS: SIM (69.0/26.0) 62.00%
CURSOOPCAO1 = (1649)_CURSO_SUP_DETNOLOGIA_EM_COMERCIO_EXTERIOR: NAO_ESTOU_TRABALHANDO (193.0/128.0) 34.00%
CURSOOPCAO1 = (1650)_CURSO_SUP_DETNOLOGIA_EM_JOGOS_DIGITAIS: NAO_ESTOU_TRABALHANDO (84.0/43.0) 49.00%
CURSOOPCAO1 = (1654)_CURSO_SUP_DETNOLOGIA_SISTEMAS_PARA_INTERNET: SIM (70.0/29.0) 59.00%
CURSOOPCAO1 = (2102)_CURSO_ADMINISTRACAO: SIM (95.0/58.0) 39.00%
CURSOOPCAO1 = (2194)_CURSO_SUP_DETNOLOGIA_EM_GESTAO_FINANCEIRA: SIM (36.0/14.0) 61.00%
CURSOOPCAO1 = (2195)_CURSO_SUP_DETNOLOGIA_EM_CONSTRUCAO_EDIFICIOS: SIM (17.0/7.0) 59.00%
CURSOOPCAO1 = (2196)_CURSO_SUP_DETNOLOGIA_EM_COMICACAO_ASSISTIVA: SIM (3.0/1.0) 67.00%
CURSOOPCAO1 = (2500)_LICENCIATURA_EM_PEDAGOGIA: NAO (26.0/10.0) 62.00%
CURSOOPCAO1 = (2601)_DESIGN_MODA_ETNOLOGIA_BCH: NAO_ESTOU_TRABALHANDO (167.0/58.0) 65.00%
CURSOOPCAO1 = (2602)_DESIGN_MODA_ETNOLOGIA_BCH: NAO_ESTOU_TRABALHANDO (169.0/108.0) 36.00%
CURSOOPCAO1 = (2703)_BCH_EM_CIENCIAS_BIOLOGICAS: NAO (137.0/66.0) 52.00%
CURSOOPCAO1 = (2802)_E_DA_ARTE_NA_DIVERSIDADE_LICENCIATURA: NAO (3.0/1.0) 67.00%
CURSOOPCAO1 = (2901)_BCH_EM_FISIOTERAPIA: NAO_ESTOU_TRABALHANDO (215.0/102.0) 53.00%
CURSOOPCAO1 = (3103)_CURSO_ADMINISTRACAO: SIM (546.0/274.0) 50.00%
CURSOOPCAO1 = (3104)_CURSO_ADMINISTRACAO: SIM (168.0/54.0) 68.00%
CURSOOPCAO1 = (3201)_BCH_EM_CIENCIAS_CONTABEIS: SIM (155.0/95.0) 39.00%
CURSOOPCAO1 = (3203)_BCH_EM_CIENCIAS_CONTABEIS: SIM (96.0/43.0) 55.00%
CURSOOPCAO1 = (3401)_LICENCIATURA_EM_ED_FISICA: NAO (192.0/85.0) 56.00%
CURSOOPCAO1 = (3402)_BCH_EM_ED_FISICA: NAO (117.0/57.0) 51.00%
CURSOOPCAO1 = (3500)_LICENCIATURA_EM_PEDAGOGIA: NAO (98.0/53.0) 46.00%
CURSOOPCAO1 = (3601)_BCH_EM_COMICACAO_SOCIAL-_HAB_EM_RELACOES_PUBLICAS: SIM (75.0/48.0) 36.00%
CURSOOPCAO1 = (4001)_BCH_EM_CIENCIA_DA_COMPUTACAO: NAO_ESTOU_TRABALHANDO (162.0/106.0) 35.00%
CURSOOPCAO1 = (4101)_BCH_EM_COMICACAO_SOCIAL_HAB_EM_PUBLICIDADE_E_PROPAGANDA: NAO_ESTOU_TRABALHANDO (168.0/86.0) 49.00%
CURSOOPCAO1 = (4102)_BCH_EM_COMICACAO_SOCIAL_HAB_EM_PUBLICIDADE_E_PROPAGANDA: NAO_ESTOU_TRABALHANDO (20.0/3.0) 85.00%
```

Figura 5.18: Relação de vestibulandos com atividade profissional X Curso
Fonte: Do Autor

```
CURSOOPCAO1 = (4202)_ENG_INDUSTRIAL-_BCH-_HAB_EM_ENG_INDUSTRIAL_QUIMICA: NAO_ESTOU_TRABALHANDO (76.0/47.0) 38.00%
CURSOOPCAO1 = (4203)_ENG_INDUSTRIAL-_BCH-_HAB_EM_ENG_INDUSTRIAL_QUIMICA: SIM (134.0/76.0) 43.00%
CURSOOPCAO1 = (4205)_ENG_INDUSTRIAL-_BCH-_HAB_EM_ENG_INDUSTRIAL_MECANICA: SIM (187.0/94.0) 50.00%
CURSOOPCAO1 = (4401)_BCH_EM_DIREITO: NAO (385.0/252.0) 35.00%
CURSOOPCAO1 = (4501)_BCH_EM_DIREITO: NAO_ESTOU_TRABALHANDO (134.0/84.0) 37.00%
CURSOOPCAO1 = (4701)_BCH_EM_COMICACAO_SOCIAL_HAB_EM_JORNALISMO: NAO_ESTOU_TRABALHANDO (14.0/6.0) 57.00%
CURSOOPCAO1 = (4702)_BCH_EM_COMICACAO_SOCIAL_HAB_EM_JORNALISMO: NAO (151.0/85.0) 44.00%
CURSOOPCAO1 = (4802)_BCH_EM_TURISMO: NAO (90.0/38.0) 58.00%
CURSOOPCAO1 = (4901)_BCH_EM_DESIGN: NAO_ESTOU_TRABALHANDO (235.0/138.0) 41.00%
CURSOOPCAO1 = (5301)_COMPUTACAO_LICENCIATURA: SIM (20.0/8.0) 60.00%
CURSOOPCAO1 = (5301)_LICENCIATURA_EM_COMPUTACAO: NAO_ESTOU_TRABALHANDO (3.0/1.0) 67.00%
CURSOOPCAO1 = (5601)_LICENCIATURA_EM_LETRAS_HAB_PORTUGUES_INGLES_E_RESPECTIVAS_LITERATURAS: NAO (86.0/38.0) 56.00%
CURSOOPCAO1 = (5701)_LICENCIATURA_EM_LETRAS_HAB_PORTUGUES_INGLES_E_RESPECTIVAS_LITERATURAS: NAO (4.0/1.0) 75.00%
CURSOOPCAO1 = (5702)_LICENCIATURA_EM_LETRAS_HAB_PORTUGUES_ESPANHOL_E_RESPECTIVAS_LITERATURAS: NAO (2.0/1.0) 50.00%
CURSOOPCAO1 = (6101)_BCH_EM_ENFERMAGEM: NAO_ESTOU_TRABALHANDO (93.0/53.0) 43.00%
CURSOOPCAO1 = (6201)_BCH_EM_CIENCIAS_FARMACEUTICAS: NAO_ESTOU_TRABALHANDO (142.0/81.0) 43.00%
CURSOOPCAO1 = (6301)_BCH_EM_BIOMEDICINA: NAO_ESTOU_TRABALHANDO (247.0/95.0) 62.00%
CURSOOPCAO1 = (6401)_BCH_EM_QUIROPRAXIA: NAO_ESTOU_TRABALHANDO (146.0/65.0) 55.00%
CURSOOPCAO1 = (6501)_BCH_EM_NUTRICAO: NAO_ESTOU_TRABALHANDO (124.0/60.0) 52.00%
CURSOOPCAO1 = (6701)_BCH_EM_ENG_ELETRONICA: SIM (97.0/53.0) 45.00%
CURSOOPCAO1 = (7101)_HISTORIA_LICENCIATURA: NAO (65.0/26.0) 60.00%
CURSOOPCAO1 = (7401)_BCH_EM_ENFERMAGEM: SIM (402.0/136.0) 66.00%
CURSOOPCAO1 = (7501)_PSICOLOGIA: NAO_ESTOU_TRABALHANDO (144.0/80.0) 44.00%
CURSOOPCAO1 = (8102)_BCH_EM_ARTES_VISUAIS: NAO (41.0/24.0) 41.00%
CURSOOPCAO1 = (9100)_BCH_EM_ARQUITETURA_E_URBANISMO: NAO_ESTOU_TRABALHANDO (297.0/167.0) 44.00%
CURSOOPCAO1 = (9201)_ENG_PRODUCAO_HAB_CALCADOS_E_COMPONENTES: SIM (55.0/21.0) 62.00%
CURSOOPCAO1 = (9401)_BCH_EM_FONOAUDIOLOGIA_SERIADO: NAO_ESTOU_TRABALHANDO (21.0/12.0) 43.00%
CURSOOPCAO1 = (9500)_BCH_EM_SISTEMAS_INFORMACAO: SIM (69.0/37.0) 46.00%
CURSOOPCAO1 = (9501)_BCH_EM_SISTEMAS_INFORMACAO: SIM (22.0/11.0) 50.00%

Number of Leaves  :     56
```

Figura 5.19: Relação de vestibulandos com atividade profissional X Curso – continuação
Fonte: Do autor

Neste capítulo foram expostos parte dos resultados dos testes realizados com os algoritmos J48 e *Naive Bayes* nos dados de vestibulandos, sendo estes algoritmos comparados. A ferramenta desenvolvida foi disponibilizada ao departamento de marketing, possibilitando a continuidade de execuções dos algoritmos com diferentes combinações de atributos e classes buscando a descoberta do conhecimento.

CONCLUSÃO

A Mineração de Dados, quando bem aplicada, é capaz de propiciar muitos benefícios às corporações, ajudando na tomada de decisões ou na descoberta de conhecimento que pode ser utilizado em posicionamento estratégico. Em determinados segmentos, como no de instituições de ensino superior, em que a concorrência está cada vez mais acirrada, as técnicas de *Data Mining* estão se tornando essenciais.

A área de Marketing se tornou o foco de diversas aplicações de Descoberta de Conhecimento, tanto acadêmica como comercialmente. No entanto, ainda faz-se necessário um estudo das técnicas que facilitem a identificação de perfis, um dos principais objetivos de *softwares* na área.

Além dessas técnicas, carece-se da pesquisa dos algoritmos que as implementam, inclusive com diferentes formas de indução de conhecimento. No desenvolvimento da aplicação deste projeto, foi dada ênfase à técnica de classificação, procurando abordar diferenças e semelhanças entre algoritmos que geram árvores de decisão, auferem conhecimento em regras e geram redes *bayesianas*.

Espera-se com este projeto, que o conhecimento já descoberto, juntamente com o que ainda poderá ser através da utilização da aplicação, ajude o trabalho do departamento de Marketing da instituição com o objetivo de direcionar seus esforços para, além de um melhor aproveitamento dos recursos, propiciar a atração de novos alunos. No entanto, as medidas a serem tomadas e o seu foco (se perfis com alta ou baixa demanda por cursos), ficam a total critério deste setor.

Com as metas sendo atingidas, proporcionando bons resultados à instituição, seria um grande benefício integrar a ferramenta desenvolvida com sistemas de informação já presentes na instituição, facilitando posteriores manutenções.

Finalizando, é pretendida a obtenção de bases de dados não somente de vestibulandos, mas de alunos desde o início do curso até o final da graduação para que também sejam aplicados algoritmos de KDD (não necessariamente de classificação) em possíveis trabalhos futuros. Com isto, poderão ser tomadas medidas a fim de evitar que matrículas sejam trancadas ou a própria evasão.

Além disso, em relação à aplicação, melhorias poderiam ser realizadas, tais como utilizar banco de dados para o armazenamento das informações dos vestibulandos, disponibilização em um ambiente web, implementação da funcionalidade de salvar as configurações da aplicação (grupos e filtros criados e seleção dos atributos) para que sirvam de entrada para a ferramenta e estudo da viabilidade da utilização de outras bibliotecas para manipulação das planilhas.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, Rakesh; SRIKANT, Ramakrishnan. **Fast Algorithms for Mining Association Rules, In: 20th International Conference on Very Large Data Bases, 1994**. **Procedings.** San Francisco, 1994. p. 487-499

ALMEIDA, Fernando C. et al. **Data Mining no contexto de Customer Relationship Management.** Caderno de Pesquisas em Administração, São Paulo, v. 12, n. 2, p. 85-97, 2005.

ASCENSO, João. **Reconhecimento de Padrões.** Escola Superior de Tecnologia – Engenharia Informática. Setúbal, 2004

AZEVEDO, Fernando Mendes de. **Algoritmos Genéticos em Redes Neurais Artificiais.** In: V Escola de Redes Neurais, São José dos Campos, 1999

BARRETO, Jorge M.. **Introdução às Redes Neurais.** Laboratório de Conexionismo e Ciências Cognitivas – UFSC, Florianópolis, 2002

BOULLOSA, José Roberto de Freitas. **Um Ambiente para Mineração de Utilização da Web.** Tese - Universidade Federal do Rio de Janeiro, 2002

BREIMAN, L; FRIEDMAN, J. H.; OLSHEN, R. A. **Classication and regression trees.** Belmont: Wadsworth Statistical Press, 1984

BROWN, Myra; SAHIN, Bilge. **Using Data Mining for Competitive Advantage in Direct Marketing Machine Learning Algorithms - Decision Trees.** Itom 6032, Spring 2002

CARVALHO, Deborah Ribeiro; BUENO, Marcos; NETO, Wilson Alves. **Ferramenta de Pré e Pós-processamento para Data Mining.** In: XII Seminário de Computação. Blumenau, 2003.

CARVALHO, Juliano V. **Reconhecimento de Caracteres Manuscritos Utilizando Regras de Associação.** Campina Grande: 2000. Dissertação (Mestrado) - UNCG, 2000.

COELLO, Adán; MANUEL, Juan. **Aprendizado de heurísticas para o escalonamento de sistemas de tempo real.** In: Anais do IV Workshop Brasileiro sobre Sistemas de Tempo-Real, evento integrante do 20o. Simpósio Brasileiro de Redes de Computadores- SBRC'2002. Sociedade Brasileira de Computação (SBC), pp. 3-10, Búzios, RJ, 20 a 24 de maio de 2002.

ENTRIEL, Aparecida Laino. **Uma Proposta para o Entendimento do Comportamento do Consumidor e Gestão de Marketing de Serviços.** Rio de Janeiro: 2008. Tese (Doutorado) - Universidade Federal do Rio de Janeiro, 2008.

FREITAS, Alex A. **Data Mining**, In: XIII Simpósio Brasileiro de Banco de Dados. Maringá/Brasil, 1998.

FREITAS, Alex A. **Understanding the Crucial Differences Between Classification and Discovery of Association Rules** – A Position Paper. SIGKDD Explorations, New York, 2000.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. ***Data Mining* :** Um guia prático. Rio de Janeiro: Elsevier, 2005.

GONÇALVES, Eduardo Corrêa. Data Mining de Regras de Associação. Disponível em < http://www.devmedia.com.br/articles/viewcomp.asp?comp=7065> . Acessado em 27/09/2008

GRÉGIO, André Ricardo Abed. **Aplicação das Técnicas de *Data Mining* para a Análise de *Logs* de Tráfego TCP/IP.** São José dos Campos: 2007. Dissertação (Mestrado) - INPE, 2007

GUINZANI, Jhonas Bonfante; SIMÕES, Priscyla Waleska Targino de Azevedo; MATTOS, Merisandra Côrtes de; BETTIOL, Jane. **Mineração de Dados em Redes Bayesianas Utilizando a API da Shell Belief Network Power Constructor (BNPC).** In: II Congresso Sul Catarinense de Computação, Criciúma, 2006.

HEUSER, Carlos Alberto. **Projeto de Banco de Dados.** Porto Alegre, RS: Editora Sagra Luzzatto, 2001.

HOPFIELD, J. **Neurons with Graded Response Have Collective Computational Properties Like Those of Two-State Neurons.** In: Proceedings of the National Academy of Sciences, vol.81, 1984, pp.3088-3092

KASS, G. V. **An Exploratory Technique for Investigating Large Quantities of Categorical Data.** Journal of Applied Statistics, Vol. 29, No. 2 (1980), pp. 119-127.

KOSKO, B. **Bidirectional associative memories.** IEEE Transactions on Systems, Man, and Cybernetics, vol.18, no.1, pp.49-60, 1988.

LEMOS, Eliane Prezepiorski; STEINER, Maria Teresinha Ams; NIEVOLA, Júlio César. **Análise de Crédito Bancário por Meio de Redes Neurais e Árvores de Decisão:** uma aplicação simples de Data Mining. Revista de Administração USP. Ed jul/ago/set 2005.

LEWIS, Roger J. **An Introduction to Classification and Regression Tree (CART) Analysis.** In: Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, 2000.

LOH, Stanley**; Data Mining.** Disponível em <http://atlas.ucpel.tche.br/~loh/dm-ppt.pdf>. Acessado em 02/10/2008.

MARQUES, Roberto Ligeiro, DULTRA, Inês. **Redes Bayesianas**: o que são, para que servem, algoritmos e exemplos de aplicações. Disponível em:

http://www.cos.ufrj.br/~ines/courses/cos740/leila/cos740/Bayesianas.pdf. Acessado em: 10/05/2009.

MCCALLUM, Andrew; NIGAM, Kamal**. A Comparison of Event Models for Naive Bayes Text Classication.** In: AAAI-98 Workshop on Learning for Text Categorization. Stanford, 1998.

MELLO, Márcio Pupin de et al. **Redes Bayesianas no Delineamento de Culturas Agrícolas Usando Informações Contextuais.** In: XXIII Congresso Brasileiro de Cartografia, Rio de Janeiro, 2007

MONGIOVI, Giuseppe. **T.E.I. Data Mining.** Notas de aula. Campina Grande. 1998.

NEVES, Cledjalma Ferreira. **Descobertas de Padrões Usando Técnicas de Extração de Conhecimento.** Palmas: Centro Universitário Luterano de Palmas - Práticas de Estágio, 2003

OLIVEIRA, Fernando Luiz et al. **Utilização de Algoritmos Simbólicos para a Identificação do Número de Caroços do Fruto Pequi.** In: II Encontro de Informática do Tocantins, Palmas, 2002

OSÓRIO, Fernando. **Machine Learning: Aprendizado simbólico a partir de exemplos.** Disciplina de Redes Neurais 2001/2, Unisinos, São Leopoldo

PENTAHO, Sourceforge. Disponível em <http://sourceforge.net/projects/pentaho/> acessado em 26/10/2008

PINTO, Carlos Manuel S. **Algoritmos Incrementais para Aprendizagem Bayesiana.** Porto: 2005. Tese (Mestrado) - Faculdade de Economia da Universidade do Porto, 2005.

QUINLAN, J.R. ***Discovering rules by induction from large collection of examples.*** Expert Sysmtes in the Micro Electronic Age. Edinburgh, UK: Edinburgh University Press, 1979.

QUINLAN, J.R. **C4.5: *Programs for Machine Learning.*** São Francisco: Morgan Kaufmann, 1993.

RAMOS, Túlio Terra. Estudo em mineração de dados aplicado nos Parâmetros de Controle do Processo de Produção de Agentes Tanantes. Trabalho de Conclusão, ULBRA, Canoas, 2004.

RAPIDMINER, Dortmund – Alemanha. **RapidMiner 4.1 User Guide.** Dortmund, 2008.

ROMANI, Daniel David; KNUPPE, Gustavo; SARAIVA, Marco Antônio Barbosa. ***Data Mining.*** Disponível em

<http://paginas.terra.com.br/informatica/arruda/Downloads/Artigos/artigo07/index.htm>. Acessado em 24/09/2008

SAHEKI, André Hideaki. **Construção de uma rede Bayesiana aplicada ao diagnóstico de doenças cardíacas.** São Paulo: 2005. Dissertação (Mestrado) - Escola Politécnica, USP, 2005.

SANTOS, Rafael. **Weka na Munheca:** Um guia para uso do Weka em scripts e integração com aplicações em Java. Apostila Princípios e Aplicações de Mineração de Dados. S.l., 2005.

SILVEIRA, Rosemari de Freitas. Mineração de Dados Aplicada à Definição de Índices em Sistemas de Raciocínio Baseado em Casos. Porto Alegre: CPGCC da UFRGS, 2003

VASCONCELOS, Benitz; SAMPAIO, Marcus Costa. **Mineração Eficiente de Regras de Classificação com Sistemas de Banco de Dados Objeto-Relacional.** In: XVII Simpósio Brasileiro de Banco de Dados, Gramado, 2002.

WITTEN, Ian H.; FRANK, Eibe. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Diego: Morgan Kaufmann Publishers, 2000

ANEXO I

1 - Em qual município você reside?

2 - Qual o tipo de curso de Ensino Médio (2º Grau) que você concluiu?

3 - Em que ano concluiu o Ensino Médio?

4 - Se você não concluiu o Ensino Médio, em que etapa se encontra?

5 - Como você fez seus estudos de Ensino Médio (2º Grau)?

6 - Em que turno você fez os seus estudos de Ensino Médio (2º Grau)?

7 - Em que escola você concluiu o Ensino Médio (2º Grau)?

8 - Qual a sua participação na vida econômica familiar?

9 - Quem pagará seus estudos?

10 - Quantas pessoas compõem o seu grupo familiar? Considera-se grupo familiar, o conjunto de pessoas residindo na mesma moradia, que sejam relacionadas ao candidato pelos seguintes graus de parentesco: pai, padrasto, mãe, madrasta, cônjuge, companheiro(a), filho(a), enteado(a), irmão(ã) e avô(ó).(Fonte: art. 6° da portaria MEC n° 4 , de 18/05/2006)

11 - Em qual faixa de renda situa-se o seu grupo familiar? Entende-se como renda bruta mensal familiar a soma de todos os rendimentos auferidos por todos os membros do grupo familiar. (Fonte: parágrafo único do art. 6º da portaria MEC n° 4, de 18/05/2006)

12 - Sua atividade profissional está relacionada com o curso pretendido?

13 - Em qual destas instituições você também está prestando vestibular?

14 - Qual o grau de conhecimento que você possui sobre o curso pretendido?

15 - Qual o principal motivo que o levou a ingressar em um curso superior?

ANEXO II

_	ADVENTISTA ADV
%	BACHARELADO BCH
__ _	CARLOS CRL
/ _	CACHOEIRINHA CAC
-RS	CMAQUA CAMAQUA
/RS	CENECISTA CNC
Â A	CIENTIFICO CIE
Ã A	EVANGELICA EVANG
Ã A	EVANGELICO EVANG
Á A	EDUCACAO ED
Â A	EDUCAAO ED
Ê E	EDUCACIONAL ED
É E	ELDORADO_SUL ES
É E	ELDORADO_SUL ES
Ã A	EDUC ED
Í I	E_TEC ET
'I I	MEDIO M
Í I	_DO_ _
Î I	_-_ _
Í I	_DE_ _
Õ O	FEDERAL FED
Ó O	FUNDACAO FUND
Ô O	FUNDAMENTAL FUND
Ú U	FREDERICO FRED
, -	GRAMADO GRAM
Ç C	GRAU G
º	HABILITACAO HAB
°	JANEIRO JAN
ENGENHARIA ENG	JULHO JUL
ESTADUAL E	IVOTI IVO
ESTAD E	IGREJINA IGREJINHA
ESCOLA E	LA_SALLE LS
ENSINO E	LUTERANO LUT
INSTITUTO I	NOVEMBRO NOV
INSTUICAO I	NOVO_HAMBURGO-RS NOVO_HAMBURGO
INTITUTO I	NOVO_HAMBURGO/RS
INST I	NOVO_HAMBURGONOVA_HAMBURGO
COLEGIO C	NOVO_HAMBURGO
COLEGO C	NOVO_HAMBUGO NOVO_HAMBURGO
TECNICA T	NOVO_HAMBURPO NOVO_HAMBURGO
MUNICIPAL M	NOVO_HMABURGO NOVO_HAMBURGO
MUN M	NOVO_HANBURGO NOVO_HAMBURGO
TECNICO T	NOVO_HAMURGO NOVO_HAMBURGO
ED_JOVENS_E_ADULTOS EJA	NOVA_HAMBURO NOVO_HAMBURGO
ED_JOVE_E_ADULTOS EJA	NOVO_HAMBURHO NOVO_HAMBURGO
ENS E	NOVO_HAMBRUGO NOVO_HAMBURGO

```
NOVO_AMBURGO NOVO_HAMBURGO
NOVA_HAMBURGO NOVO_HAMBURGO
NOVO_HAMBUGO NOVO_HAMBURGO
NOVO_HAMBURPO NOVO_HAMBURGO
NOVO_HMABURGO NOVO_HAMBURGO
NOVO_HANBURGO NOVO_HAMBURGO
NOVO_HAMURGO NOVO_HAMBURGO
NOVA_HAMBURO NOVO_HAMBURGO
NOVO_HAMBURHO NOVO_HAMBURGO
NOVO_HAMBRUGO NOVO_HAMBURGO
NONO_HAMBURGO NOVO_HAMBURGO
NOVO_AMBURGO NOVO_HAMBURGO
OBJETIVO OBJ
OBJETIVA OBJ
OUTUBRO OUT
PELOTENSE PEL
PROFESSOR PROF
PRESIDENTE PRES
'
PORTO_ALEGRE POA
SAO_LEPOLDO SAO_LEOPOLDO
SANTA STA
SANTO STO
SAO_JOSE SJE
SAO_JOAO SJO
SAPUCAIA_DO_SUL SAPUCAIA
SAPUCAIA_SUL SAPUCAIA
SENHORA SRA
SETEMBRO SET
SOCIEDADE SOC
SUPERIOR SUP
SUPLETIVO SUPL
UNIDADE UNID
UNIVERSITARIO UNIV
UNIVERSITARIO UNIV
UNIVERSIDADE UNIV
SUPLETIVO SUPL
```