ÉDEN DE OLIVEIRA PINTO COELHO

DESCOBERTA DE CONHECIMENTO SOBRE O PROCESSO SELETIVO DA UFLA

Monografia de Graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do Curso de Ciência da Computação para a obtenção do título de Bacharel.

LAVRAS
MINAS GERAIS – BRASIL
2007

ÉDEN DE OLIVEIRA PINTO COELHO

DESCOBERTA DE CONHECIMENTO SOBRE O PROCESSO SELETIVO DA UFLA

Monografia de Graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do Curso de Ciência da Computação para a obtenção do título de Bacharel.

Área de Concentração:

Banco de Dados

Orientador / Co-orientador:

Prof. Marcelo Silva de Oliveira

Prof. Ahmed Ali Abdalla Esmin

LAVRAS
MINAS GERAIS – BRASIL
2007

Ficha Catalográfica

Pinto Coelho, Éden de Oliveira.

Descoberta de Conhecimento sobre o Processo Seletivo da / Éden de Oliveira Pinto Coelho. Lavras – Minas Gerais, 2007. 56p : il.

Monografia de Graduação – Universidade Federal de Lavras. Departamento de Ciência da Computação.

1. Descoberta de Conhecimento. 2. Banco de Dados. 3. Mineração de Dados. I. PINTO COELHO, E. O. II. Universidade Federal de Lavras. III. Título.

ÉDEN DE OLIVEIRA PINTO COELHO

DESCOBERTA DE CONHECIMENTO SOBRE O PROCESSO SELETIVO DA UFLA

Monografia de Graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do Curso de Ciência da Computação para a obtenção do título de Bacharel.

Aprovada em (8/8/2007)

_____
Prof. Paulo Henrique de Souza Bermejo

_____
Prof. Wilian Soares Lacerda

_____
Prof. Marcelo Silva de Oliveira
(Orientador)

_____
Prof. Ahmed Ali Abdalla Esmin
(Co-orientador)

LAVRAS
MINAS GERAIS – BRASIL

DESCOBERTA DE CONHECIMENTO SOBRE O PROCESSO SELETIVO DA UFLA

RESUMO

A informação vem desempenhando um papel fundamental no desenvolvimento e sucesso das grandes organizações. Os sistemas de suporte a decisão tornam mais confiáveis as tarefas de coletar, tratar, interpretar e utilizar informações. As empresas tendem, com o passar do tempo, a aumentar consideravelmente seu volume de dados. Entretanto, há uma relação inversa entre o volume de dados existentes e a necessidade de conhecimento estratégico, ou seja, apesar das informações resumidas e significativas para tomada de decisão seja de volume menor, geralmente elas não estão disponíveis e exigem que sejam extraídas a partir de grande quantidade de dados. Descoberta de Conhecimento em Banco de Dados (KDD – *Knowledge Discovery in Databases*) refere-se ao processo de extração de conhecimento a partir de grande base de dados. Mineração de Dados (ou *Data Mining*), refere-se a uma determinada etapa deste processo. Este trabalho apresenta uma aplicação prática do processo de KDD na base de dados sobre os candidatos ao processo seletivo dos vestibulares ocorridos no ano de 2006 da UFLA. Neste trabalho, utilizando-se de uma ferramenta chamada WEKA (*Weikato Enviroment for Knowledge Analysis*), foram aplicadas as técnicas de Mineração Visual de Dados, Árvore de Decisão, Regras de Associação e Redes Neurais. Os resultados obtidos poderão ser usados para traçar perfis dos candidatos ao processo seletivo do vestibular da UFLA, a fim de levantar informações relevantes que tragam subsídios para as instituições de ensino em geral na tomada de decisões.

Palavras-chave: Descoberta de Conhecimento em Banco de Dados, Mineração de Dados.

KNOWLEDGE DISCOVERY ABOUT UFLA SELECTIVE PROCESS

ABSTRACT

The information has been having a fundamental role on companies' growth, development and sucess. The making-decision supporting systems, available at these companies, make the work of collecting, treating and analyzing. There is also a tendency in these companies to increase their data amount. However, there is an inverse relation between the data amount and the need of a strategic knowledge, that is, although the resumed and meaningful information to making- decision are fewer, generally they are not available and demand to be extracted from big data amounts. KDD - Knowledge Discovery in Databases refers to the extration of knowledge from a big database amounts. Data Mining refers to a especific phase of this process. This study demostrates a practical application of KDD Process to the database of 2006 UFLA's Entrance Examination or Selective Process. Coherently to WEKA research tool - Weikato Enviroment for Knowledge Analysis, the Data Visual Mining, Decision Tree, Association Rules and Neural Networks were applied. The results can be used to biuld up the candidates profile, in order to extract important information that offer support to this institution on the making-decision process.

Key-words: Knowledge Discovery in Databeses, Data Mining.

SUMÁRIO

Lista de Figuras ..... **viii**

Lista de Tabelas ..... **ix**

Abreviaturas ..... **x**

1 Introdução **1**
1.1 Tema e Problema da Pesquisa 1
1.2 Objetivos da Pesquisa 2
1.3 Justificativa 2
1.4 Limitações da Pesquisa 2
1.5 Estrutura do Trabalho 3

2 Revisão de Literatura sobre KDD e Mineração de Dados **4**
2.1 Descoberta de Conhecimento e Mineração de Dados 4
2.2 Etapas do Processo de KDD 6
2.2.1 Seleção dos Dados-Alvo 7
2.2.2 Limpeza e Pré-Processamento dos Dados 7
2.2.3 Transformação e Adequação dos Dados 8
2.2.4 Mineração dos Dados 8
2.2.5 Interpretação e Avaliação dos Resultados 9
2.3 Principais Tarefas de Mineração de Dados 9
2.3.1 Classificação 10
2.3.2 Associação 10
2.3.3 Clusterização (Segmentação) 10
2.3.4 Estimativa (Regressão) 11
2.4 Principais Técnicas de Mineração de Dados 11
2.4.1 Mineração Visual de Dados 11
2.4.2 Árvore de Decisão 13
2.4.3 Regras de Associação 15
2.4.4 Redes Neurais Artificiais 16
2.5 Ferramentas de Mineração de Dados 19
2.5.1 Weka 20
2.6 Trabalhos Relacionados 21
2.7 Considerações Finais 22

3 Desenvolvimento do Trabalho **23**
3.1 Considerações Finais 25

4 Testes e Resultados **26**

4.1 Visualização Gráfica dos Dados....26
4.2 Árvore de Decisão....28
4.3 Regras de Associação....30
4.4 Redes Neurais....31
4.5 Considerações Finais....34

5 Conclusões e Sugestões para Trabalhos Futuros....35
5.1 Conclusões....35
5.2 Sugestões para Trabalhos Futuros....36

Bibliografia....37

ANEXO....40

LISTA DE FIGURAS

Figura 2-1: Relação entre KDD e Data Mining. **Fonte:** Carvalho (2002)..... 5

Figura 2-2: Etapas do Processo de KDD. **Fonte:** Fayadd (1999).....6

Figura 2-3: Modelo de Referência de Visualização. **Fonte:** Card (1999)..... 12

Figura 2-4: Um exemplo de Árvore de Decisão. **Fonte:** Correa (2007)..... 14

Figura 2-5: Exemplo de uma Rede Neural. **Fonte:** Thomé (2007)..... 17

Figura 2-6: Interfaces Gráficas da ferramenta WEKA..... 21

Figura 3-1: Banco de Dados Access Referente ao Questionário Sócio-Econômico..... 24

Figura 3-2: Dados no formato ARFF.....25

Figura 4-1: Visualização gráfica dos dados na ferramenta WEKA.....26

Figura 4-2: Árvore de Decisão gerada pela ferramenta WEKA.....28

Figura 4-3: Rede Neural no ambiente WEKA.....32

Figura 4-4: Resultado da Rede Neural treinada no ambiente WEKA..... 33

LISTA DE TABELAS

Tabela 2-1: Principais Ferramentas de Mineração de Dados disponíveis no mercado. **Fonte:** Aracele (2004).....20

ABREVIATURAS

KDD – Knowledge Discovery in Database – Descoberta de Conhecimento em Banco de Dados

MD – Mineração de Dados

UFLA – Universidade Federal de Lavras

WEKA – Waikato Environment for Knowledge Analysis

SQL – Structured Query Language – Linguagem de Consulta Estruturada

1 INTRODUÇÃO

1.1 Tema e Problema da Pesquisa

O Brasil vive uma tendência, que hoje é mundial, de valorização da formação e da educação formal pela sociedade e pelo mercado de trabalho. Esta tendência está relacionada à expansão da visão de comunidade, propiciada pelos avanços tecnológicos e de comunicação que se tornam disponíveis para uma gama cada vez maior das atividades humanas. Tais avanços estão revolucionando o conceito de espaço, tempo e fronteiras nas comunicações entre pessoas, no acesso a informação, na produção e na reconstrução do conhecimento. O fenômeno da globalização, em relação ao qual as tecnologias de informação têm hoje em dia grande responsabilidade, obriga que todos os agentes que intervêm na sociedade, estejam preparados para a mudança, de forma a garantir sua sobrevivência num mercado mais amplo e competitivo.

O crescimento da procura de vagas na educação superior no Brasil deve-se à necessidade de maior qualificação da mão-de-obra, à globalização da economia, às novas tecnologias, aos novos sistemas de gestão, entre outros. No entanto, o crescimento da procura de vagas no ensino superior não significa que a democratização do acesso e permanência, dos jovens brasileiros, a este nível de ensino já se efetivou. Os dados do Instituto Brasileiro de Geografia e Estatística (IBGE) do ano de 2005, revelam que apenas 9,7% dos cerca de 19,6 milhões de jovens entre 18 e 24 anos chagaram à universidade e, somente 1,3% nessa faixa etária concluem uma faculdade.

No entanto, ainda nos dias atuais, percebe-se uma certa deficiência na seleção dos candidatos através do vestibular, tanto que, o processo seletivo, na maioria das instituições de ensino superior, passa por mudanças constantes na tentativa de não só realizar uma seleção mais justa, mas também, de evitar a evasão dos egressos, que apresenta um percentual ainda muito grande.

Na busca por melhores níveis de ensino e como forma de obter informações que possam levar ao conhecimento sobre os candidatos ao processo seletivo de admissão para o ensino superior, a maioria das instituições solicita o preenchimento de uma ficha com o questionário sócio-econômico cultural. Esses dados podem auxiliar os administradores das instituições na tomada de decisões, a fim de melhorar a qualidade de ensino.

1.2 Objetivos da Pesquisa

Esta pesquisa tem por objetivo geral delinear o perfil dos candidatos ao processo seletivo de admissão para o ensino superior da Universidade Federal de Lavras (UFLA), através da aplicação da ferramenta WEKA.

Os objetivos específicos são:

1) Caracterizar as diferenças sócio-econômicas e culturais existentes entre os candidatos, e

2) Estudar e aplicar técnicas e algoritmos de mineração de dados para descobrir padrões de comportamento do vestibulando.

1.3 Justificativa

Diante da ociosidade das vagas nas instituições de ensino superior, o governo federal buscou implementar um novo sistema de financiamento para esse nível de ensino que proporcione a utilização das vagas noturnas no ensino público e das ociosas no privado, através do Programa Universidade para Todos (ProUni). O programa foi lançado no dia 13 de abril de 2004 e permitirá que, em cinco anos, 300 mil estudantes de baixa renda e professores públicos sem formação superior ingressem na universidade (IBGE, 2003).

Tomando como base a preocupação existente em relação ao baixo índice de jovens brasileiros que têm acesso ao ensino superior e as altas taxas de evasão dos pós-egressos, este trabalho visa buscar conhecimentos interessantes sobre as variáveis sócio-econômicas dos vestibulandos da UFLA, e também, servir de base para outros trabalhos na área.

Os resultados obtidos através desses estudos poderão auxiliar os administradores da UFLA na tomada de decisões em relação ao projeto acadêmico a ser desenvolvido junto aos alunos da universidade.

1.4 Limitações da Pesquisa

As limitações para o desenvolvimento dessa pesquisa estão situadas na base de dados. A base de dados fornecida a respeito das variáveis sócio-econômicas dos vestibulandos não contemplava informações sobre a efetivação da matrícula do aluno aprovado no vestibular. Este fato constitui uma limitação porque após a aplicação das técnicas de mineração de dados, poderíamos obter regras relacionadas à condição social do candidato, como

por exemplo, se o candidato não efetivou a matrícula pelo fato de estar trabalhando, bem como outras regras provenientes desse tipo de informação.

1.5 Estrutura do Trabalho

Para a apresentação da pesquisa realizada, estruturou-se este trabalho em cinco capítulos, que estão relacionados a seguir.

Neste capítulo encontra-se a introdução que contempla o tema e problema de pesquisa, os objetivos, as justificativas, as limitações e estrutura do trabalho.

No capítulo 2 está a revisão da literatura sobre o processo de descoberta de conhecimento e mineração de dados, as etapas deste processo, o conceito de mineração de dados, as principais tarefas e técnicas de mineração de dados, a apresentação do software WEKA, e algumas aplicações de mineração de dados.

No capítulo 3, será apresentada a metodologia adotada, bem como os materiais e equipamentos utilizados, permitindo a compreensão e interpretação dos resultados obtidos.

No capítulo 4 são descritos os testes realizados e os resultados da aplicação das técnicas de mineração de dados.

Finalmente, no capitulo 5, são apresentadas as conclusões e as sugestões de trabalhos futuros.

2 REVISÃO DE LITERATURA SOBRE KDD E MINERAÇÃO DE DADOS

A grande quantidade de informação armazenada em meio digital nas atuais organizações é atualmente um dos problemas mais graves trazidos com o advento da tecnologia. A maioria destas informações está armazenada em base de dados, cujo tamanho cresce exponencialmente. Mas, no entanto, para algumas aplicações, muitos destes dados armazenados não têm significado nenhum quando analisados isoladamente e sem nenhuma interpretação (CARVALHO et. al., 2003).

Para se manterem competitivas no mercado, as organizações precisam ter acesso às informações importantes, geralmente "escondidas" entre os dados de seus sistemas transacionais, e, ainda, ter meios de utilizá-las no processo de tomada de decisões. Para tanto, necessitam de técnicas e ferramenta de análise de dados automatizadas. Neste contexto, está o processo de descoberta de conhecimento em banco de dados (*Knowlege Discovery in Databases* – KDD), no qual mineração de dados (*Data Mining*) é a principal etapa.

A descoberta de conhecimento em banco de dados (KDD), é uma área de pesquisa crescente que atrai esforços de pesquisadores. Fundamenta-se no fato de que as grandes bases de dados podem ser uma fonte de conhecimento útil, porém, não explicitamente representado, e cujo objetivo é desenvolver e validar técnicas, metodologias e ferramentas capazes de extrair o conhecimento implícito nesses dados e representá-lo de forma acessível aos usuários (FELDENS, 1996).

Neste capítulo são apresentados conceitos e características do processo de KDD e Mineração de Dados (MD); tais como as etapas deste processo, as principais tarefas e técnicas de MD, ferramentas de MD e exemplos de aplicações práticas de MD.

2.1 Descoberta de Conhecimento e Mineração de Dados

Inicialmente, foram designados vários nomes à noção de achar padrões úteis em dados brutos, tais como mineração de dados, extração de conhecimento, descoberta de informação e processamento de padrões em dados. Apenas em 1989, o termo "Descoberta de Conhecimento em Banco de Dados" foi utilizado para se referir ao processo total de procu-

rar conhecimento em banco de dados, com a aplicação de técnicas de mineração de dados (FAYYAD, et al., 1996).

Segundo Carvalho (2002), muitas vezes os termos "Mineração de Dados" e "Descoberta de Conhecimento em Banco de Dados" são confundidos como sinônimos. Porém, o termo KDD é empregado para descrever todo o processo de extração de conhecimento de um conjunto de dados. O termo Mineração de Dados (*Data Mining*) refere-se a uma das etapas deste processo. A relação existente entre KDD e *Data Mining* pode ser visualizada graficamente através da Figura 2-1.

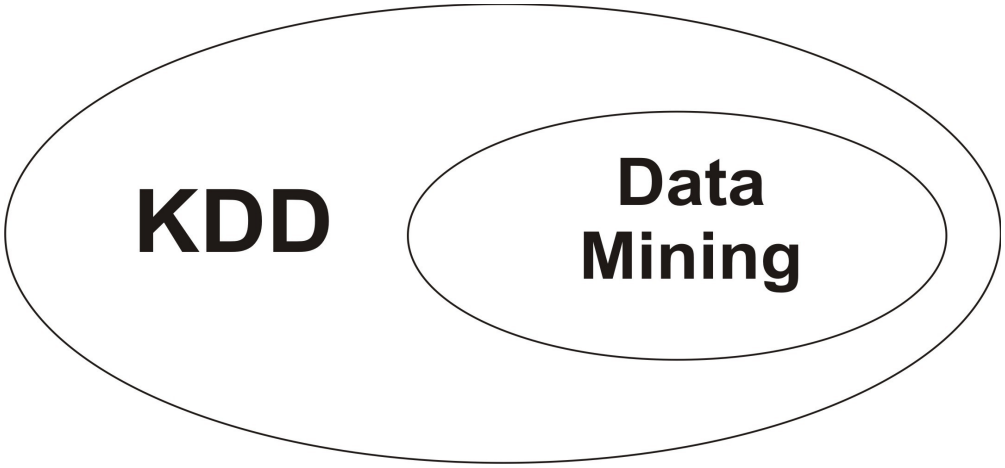

Figura 2-1: Relação entre KDD e Data Mining. **Fonte:** Carvalho (2002).

Uma definição formal, de acordo com Fayyad et al. (1996), é que KDD é o processo não trivial de identificação de padrões. Esse processo deve conter, nas bases de dados, as características de validade, novidade, utilidade e assimilabilidade. A característica de validade se encontra na descoberta de padrões que deve ser válida em novos dados com algum grau de certeza ou probabilidade. A novidade refere-se aos padrões que se destacam por serem novos (pelo menos no contexto da análise). Os padrões devem ser úteis para a tomada de decisões e medidos por alguma função. Em relação à característica assimilável, segundo Fayyad et al. (1996), um dos objetivos do KDD é tornar os padrões assimiláveis ao conhecimento humano.

O processo de KDD é iterativo, uma vez que pode ser executado várias vezes até a obtenção do resultado desejado, e interativo, por permitir a interferência do usuário a qualquer momento e o retorno a passos anteriores e subdividido em várias etapas (FAYYAD et. al., 1996). De uma maneira mais geral, estas etapas podem ser agrupadas em três gran-

des fases: preparação, análise e interpretação. Grande parte das pesquisas nesta área estão centradas nas tarefas de análise, mais especificamente na mineração de dados, que trata da aplicação de métodos sofisticados de análise estatística e de aprendizagem automática a fim de buscar padrões sobre um grande volume de dados. Contudo, estima-se que a mineração de dados propriamente dita consome apenas 15 a 20% de todos os esforços do processo (BRACHMAN, 1996). Na prática, boa parte do processo é centrada na fase de preparação, que inclui a compreensão, seleção, transformação e limpeza dos dados, a fim de enquadrar o problema existente no domínio como um problema de mineração de dados, e definir as variáveis e dados relevantes para o processo de análise.

2.2 Etapas do Processo de KDD

A transformação dos dados em informações que possam auxiliar à tomada de decisões é um processo complexo, conforme afirma Fayyad et. al. (1996). Esse processo pode ser organizado em cinco passos, de acordo com a Figura 2-2.

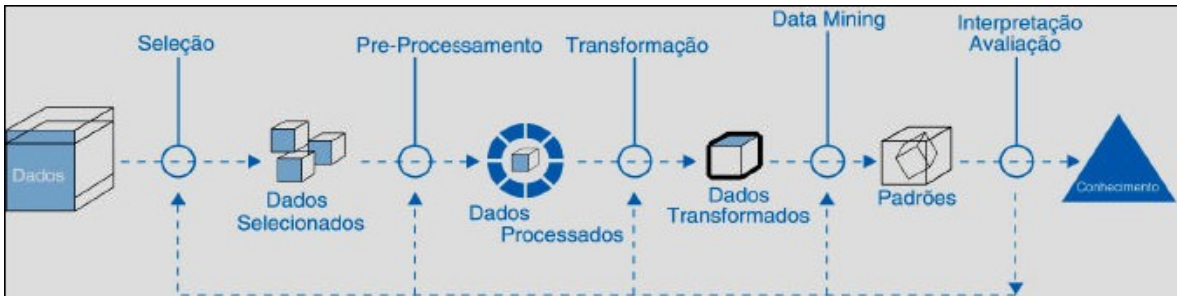

Figura 2-2: Etapas do Processo de KDD. **Fonte:** Fayadd (1999).

O processo de KDD é iniciado através da compreensão do domínio da aplicação e o estabelecimento dos objetivos a serem obtidos. Esta definição geralmente é realizada pelos especialistas do domínio ou conhecedores da base de dados com o apoio do especialista do processo de KDD. Pesquisas e análises devem ser realizadas para que o conhecimento desejado pelo usuário final, ou seja, o tipo de conhecimento que se deseja extrair do banco de dados possa ser reconhecido. Além de definir os resultados e o domínio da aplicação, deve-se levantar um conhecimento prévio relevante sobre a viabilidade e custos da aplicação, duração do projeto, entre outros.

2.2.1 Seleção dos Dados-Alvo

Após a definição do domínio, deve-se localizar e escolher quais as fontes de dados estão relacionadas a este domínio para que um conjunto de dados apropriado possa ser selecionado das mesmas.

Dependendo do caso, esta etapa também consiste em criar uma nova base de dados usando conjuntos de elementos de várias fontes de dados. Isto envolve considerações de homogeneidade, ou seja, esta nova base de dados deve estar coerente com a base de dados original, e identificação das fontes desses dados, as quais podem ser internas ou externas.

As fontes internas normalmente são fontes de dados que já estão incorporadas ao sistema de aplicação do domínio em questão. Por estarem geralmente armazenados em algum tipo de repositório estruturado, requerem menos processamento, pois, eles podem ser recuperados utilizando técnicas, linguagens, ferramentas e comandos convencionais de banco de dados, como SQL.

As fontes externas são compostas por outros tipos de localidades que habitualmente não são incorporados ao sistema de aplicação, como, por exemplo: documentos, livros, internet e informações do especialista do domínio. Estes dados precisam ser formatados e inseridos de forma que possam ser recuperados e preparados para análise.

Ainda nesta etapa poderá haver a escolha do tipo de técnica de mineração a ser adotada. A questão a ser minerada e a própria técnica a ser trabalhada ajuda a definir qual parte da massa de dados inicial vai ser utilizada e, portanto, selecionada.

2.2.2 Limpeza e Pré-Processamento dos Dados

Após reunir dados de determinadas fontes, é muito provável que o conjunto de dados venham a conter registros duplicados, erros e dados ausentes. A geração desses ruídos pode ocorrer devido a problemas em migrações de um sistema para outro, quedas de tensão na hora do processamento, desligamento do computador tendo arquivos abertos, falta de tratamento adequado no armazenamento das entradas de dados vindas de sistemas operados por usuários comuns, dentre outros.

As operações de remoção de ruídos dos dados, tratamento de atributos ausentes e remoção de registros repetidos feitas nesta etapa compreende, entre outros, os seguintes aspectos:

a) Padronização dos valores dos atributos: Em um conjunto de dados constituídos por diversas fontes pode acontecer, por exemplo, que o "sexo" possua diferentes valores e tipos com o mesmo significado como "masculino", "mas", "m" ou 1. Então, deve-se transformar estes valores como sendo um tipo único e iguais para todo o conjunto de dados.

b) Remoção de registros duplicados: Pode ocorrer que os mesmos dados de um cliente estarem contido em diferentes registros e sendo considerado como duas pessoas distintas, pois, por exemplo, o nome de uma pessoa em um determinado registro aparece sem abreviações, e em outro, ele aparece abreviado.

c) Tratamento e eliminação de ruídos: Muitas vezes, os dados coletados podem conter erros ocasionados por diversos fatores no momento da recuperação de diversas fontes, como acidentes, falta de energia elétrica, entre outros. Os campos que contém ruídos nos seus valores devem ser tratados atribuindo o valor correto aos dados ou devem ser eliminados da base de dados, caso não tenham como ser tratadas.

d) Tratamento de valores ausentes: É muito comum encontrar registros cujos campos possuam valores ausentes. Isto pode ocorrer, entre outras coisas, devido a erros na entrada dos dados, por exemplo, se no momento em que um operador estiver cadastrando as informações de pessoas, ele pode esquecer (ou ignorar) alguns campos de dados. Dessa forma, deve-se estabelecer critérios para o tratamento de atributos ausentes.

Alguns estudos mostram que a etapa de limpeza dos dados pode tomar até 80% do tempo necessário para todo o processo de descoberta de conhecimento. Por isso, esta etapa é considerada uma das etapas mais importantes para o sucesso do processo como um todo.

2.2.3 Transformação e Adequação dos Dados

Os dados pré-processados passam ainda por uma transformação com o objetivo de facilitar seu uso pelas técnicas de mineração. É necessário fazer certas adequações no conjunto de dados de acordo com a técnica de Data Mining a ser utilizada, pois existem diversos tipos de algoritmos, e cada um necessita de uma entrada específica, além das conversões de dados, criação de novas variáveis e categorização de variáveis contínuas.

2.2.4 Mineração dos Dados

Consiste na efetiva aplicação das tarefas e das técnicas escolhidas sobre os dados a serem analisados com o objetivo de encontrar os padrões desejados, ou seja, descobrir no-

vas relações, não identificáveis a olho nu. Tarefas são classes de problemas que foram definidas através de estudos na área de mineração de dados. Técnicas são grupos de soluções (algoritmos) para os problemas propostos nas tarefas. Cada tarefa apresenta várias técnicas, e algumas técnicas podem ser utilizadas para solucionar tarefas diferentes.

Data Mining é uma das principais etapas dentro do processo de KDD, que pode ser automático ou, geralmente, semi-automático, sendo que, a qualidade dos resultados dessa etapa depende diretamente da correta realização das etapas anteriores.

Os algoritmos utilizados para se criar modelos a partir de dados, normalmente, provem de áreas como Aprendizado de Máquina, Reconhecimento de Padrões e Estatística. Estas técnicas, muitas vezes, podem ser combinadas para se obter resultados melhores.

2.2.5 Interpretação e Avaliação dos Resultados

Os padrões encontrados na etapa de *Data Mining* devem ser validados a partir da interpretação e avaliação destes. Os usuários envolvidos devem interpretar os padrões extraídos, e para isto, podem lançar mão de ferramentas estatísticas e de visualização que permitam fazer uma "leitura" precisa sobre os resultados obtidos, de forma a possibilitar a verificação da validade e novidade, ou mesmo, a irrelevância dos padrões encontrados.

É importante ressaltar que esta tarefa deve ser realizada em conjunto com os usuários envolvidos no processo de extração de conhecimento, pois somente com o auxílio de especialistas é que um analista terá condições de avaliar se o que foi descoberto é um conhecimento relevante ou não.

Caso o conhecimento não seja validado, então, provavelmente, deve-se retornar às etapas anteriores e tentar refazê-las ou melhorá-las. Esta iteração pode ocorrer até que se obtenha resultados aceitáveis ou concluir-se que não seja possível extrair conhecimento relevante dos dados.

2.3 Principais Tarefas de Mineração de Dados

O desenvolvimento de sistemas de KDD está relacionado com diversos domínios de aplicações em marketing, nas análises corporativas, na astronomia, na medicina, na biologia, entre outros. Existem diversas tarefas de KDD que são, principalmente, dependentes do domínio da aplicação e do interesse do usuário. Cada tarefa de KDD extrai um tipo di-

ferente de conhecimento da base de dados e pode requerer um algoritmo diferente para cada tarefa (Wiley, 2004).

Mineração de dados dispõe de tarefas básicas classificadas nas categorias descritivas, que envolvem a descoberta de padrões interpretáveis por humanos que descrevam os fatos cadastrados na base de dados, e preditivas, que utilizam determinadas variáveis para prever valores desconhecidos de outras variáveis de interesse . A seguir, as principais tarefas de MD são descritas.

2.3.1 Classificação

Consiste em examinar as características de um dado e atribuir a ele uma classe pré-definida. Ou seja, esta tarefa objetiva a construção de modelos que permitam o agrupamento de dados em classes. Esta tarefa é considerada preditiva, pois uma vez que as classes são definidas, ela pode prever automaticamente a classe de um novo dado. Por exemplo, uma população pode ser dividida em categorias para avaliação de concessão de crédito com base em um histórico de transações de créditos anteriores. Em seguida, uma nova pessoa pode ser enquadrada, automaticamente, em uma categoria de crédito específica, de acordo com suas características.

2.3.2 Associação

Estuda um padrão de relacionamento entre itens de dados. Por exemplo, uma análise das transações de compra em um supermercado pode encontrar itens que tendem a ocorrerem juntos em uma mesma compra, como café e leite. Os resultados desta análise podem ser úteis na elaboração de catálogos e *layout* de prateleiras de modo que produtos a serem adquiridos na mesma compra fiquem próximos um do outro. Essa tarefa é considerada descritiva, ou seja, ela é usada para identificar padrões em dados históricos.

2.3.3 Clusterização (Segmentação)

As informações podem ser particionadas em classes de elementos similares. Neste caso, nada é informado ao sistema a respeito das classes existentes. O próprio algoritmo descobre as classes a partir das alternativas encontradas na base de dados, agrupando assim um conjunto de objetos em classes com características semelhantes. Por exemplo, uma população inteira de dados sobre tratamento de uma certa doença pode ser dividida em grupos baseados na semelhança de efeitos colaterais produzidos; acessos a *web* realizados por

um conjunto de usuários em relação a um conjunto de documentos podem ser analisados para revelar clusters ou categorias de usuários. Esta tarefa é considerada descritiva.

2.3.4 Estimativa (Regressão)

Objetiva definir um valor numérico de alguma variável desconhecida a partir dos valores de variáveis conhecidas. Exemplos de aplicações são: estimar a probabilidade de um paciente sobreviver dado o resultado de um conjunto de diagnósticos de exames; predizer quantos carros passam em determinado pedágio, tendo alguns exemplos contendo informações como: cidades mais próximas, preço do pedágio, dia da semana, rodovia em que o pedágio está localizado, entre outros. Essa tarefa é considerada preditiva.

2.4 Principais Técnicas de Mineração de Dados

As técnicas de MD, de acordo com Resende (2003), descrevem um paradigma de extração de conhecimento e vários algoritmos podem seguir este paradigma, ou seja, para uma técnica, pode-se ter vários algoritmos.

Um ponto importante é que cada técnica tipicamente resolve melhor alguns problemas do que outros, não há um método universal. Para as aplicações, grande parte do esforço vai para a formulação do problema, ou seja, a especificação de que tipo de informações o algoritmo de mineração deve procurar no conjunto de dados disponíveis.

A seguir, são descritas as técnicas de mineração de dados que serão usadas neste trabalho.

2.4.1 Mineração Visual de Dados

Representações gráficas têm sido utilizadas, largamente, como instrumentos de comunicação desde os primórdios da humanidade. Com o advento da ciência, as representações gráficas passaram a embutir significados por convenções, como gráficos matemáticos e cartográficos. Estas representações, normalmente, têm o propósito de comunicar uma idéia existente em um conjunto de valores. Contudo, a fim de aproveitar as características da percepção visual humana, uma segunda abordagem consiste em utilizar as representações gráficas para criar ou descobrir uma idéia que esteja embutida nos valores. Esta segunda abordagem tem crescido consideravelmente, devido a evolução dos computadores para a geração de representações significativas.

Visualização é a área em que as representações gráficas produzem um significado que permitam aos usuários desenvolver suas próprias idéias ou de confirmar suas expectativas. Visualização é o processo de mapeamento de dados e informações em um formato gráfico, baseando-se em representações visuais e em mecanismos de interação, fazendo uso de suporte computacional (Card, et. al., 1999).

Na Figura 2-3 pode ser observado um modelo de referência para visualização. A visualização pode ser observada como sendo uma seqüência de mapeamentos "ajustáveis" de dados para uma representação visual, por meio da interação do usuário.

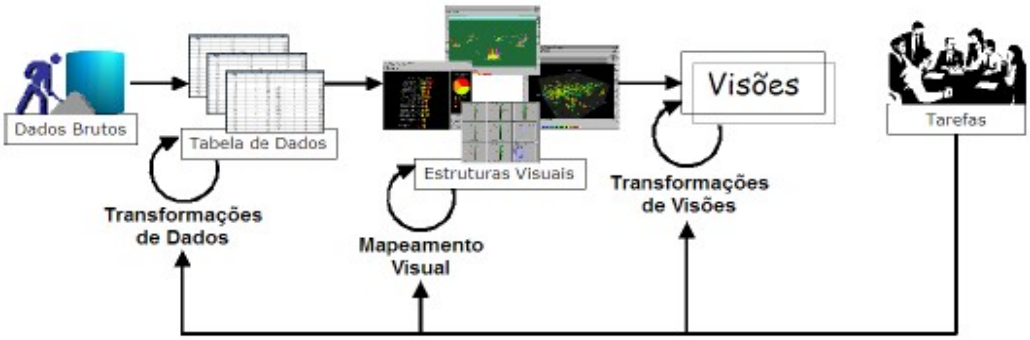

Figura 2-3: Modelo de Referência de Visualização. **Fonte:** Card (1999).

A visualização de informação baseia-se em analisar uma grande quantidade de dados não-físicos, tais como coleções de documentos e dados financeiros. Este tipo de avaliação possui uma complexidade maior devido a tornar visíveis as características inerentes destes dados.

Os dados para a visualização de informação possuem, normalmente, uma grande quantidade de registros, sendo que cada registro consiste em uma observação, medida, transação e um número de variáveis ou dimensões.

A mineração de dados tem se tornado um fenômeno de uso por várias áreas do conhecimento. Esse fenômeno é motivado pela abrangência das técnicas para uso com grandes bases de dados a fim de obter padrões inteligíveis para os usuários, sejam especialistas, analistas ou usuários finais. Porém, somente padrões em forma de árvores, regras, redes ou qualquer linguagem de descrição, normalmente, não são suficientes. A visualização dos padrões encontrados tem-se tornado uma necessidade, pois para uma grande base de dados é gerada uma grande quantidade de padrões. Essa grande quantidade pode ser filtrada, por exemplo, observando-se padrões ou comportamentos gráficos em uma visualização.

A forma de integração também pode auxiliar os usuários do processo a obter um maior ganho de interatividade, pois um acoplamento forte entre a visualização e a mineração permite ao usuário verificar ou analisar hipóteses já sugeridas pelo especialista.

2.4.2 Árvore de Decisão

Possui este nome porque a sua estrutura se assemelha a uma árvore, fácil de entender e assimilar. Dividem os dados em subgrupos, com base nos valores das variáveis. O resultado é uma hierarquia de declarações do tipo "Se...então..." que são utilizadas, principalmente, quando o objetivo da mineração de dados é a classificação dos dados. É conveniente usar árvore de decisão quando o objetivo for categorizar dados.

Na árvore, cada nó especifica o teste de um atributo da instância, e cada ramificação corresponde a um dos possíveis valores do atributo. Uma instância é classificada, começando pela raiz da árvore, testando o atributo especificado, movendo-se para um nível abaixo. Este processo é repetido para a sub-árvore, enraizada pelo novo nó.

Uma árvore de decisão utiliza a estratégia chamada "dividir-para-conquistar" que divide um problema maior em outros menores. Assim, sua capacidade de discriminação dos dados provém da divisão dos espaços definidos pelos atributos em subespaços. Para Witten et. al. (2000), uma característica das árvores de decisão é que cada um dos caminhos desde a raiz até as folhas representa uma conjunção de testes sobre os atributos.

Uma árvore de decisão é formada por um conjunto de regras de classificação. Cada caminho da raiz até uma folha representa uma destas regras. A árvore de decisão deve ser definida de forma que, para cada observação da base de dados, haja um e apenas um caminho da raiz até a folha. As quatro regras de classificação a seguir, compõem a árvore de decisão da Figura 2-4.

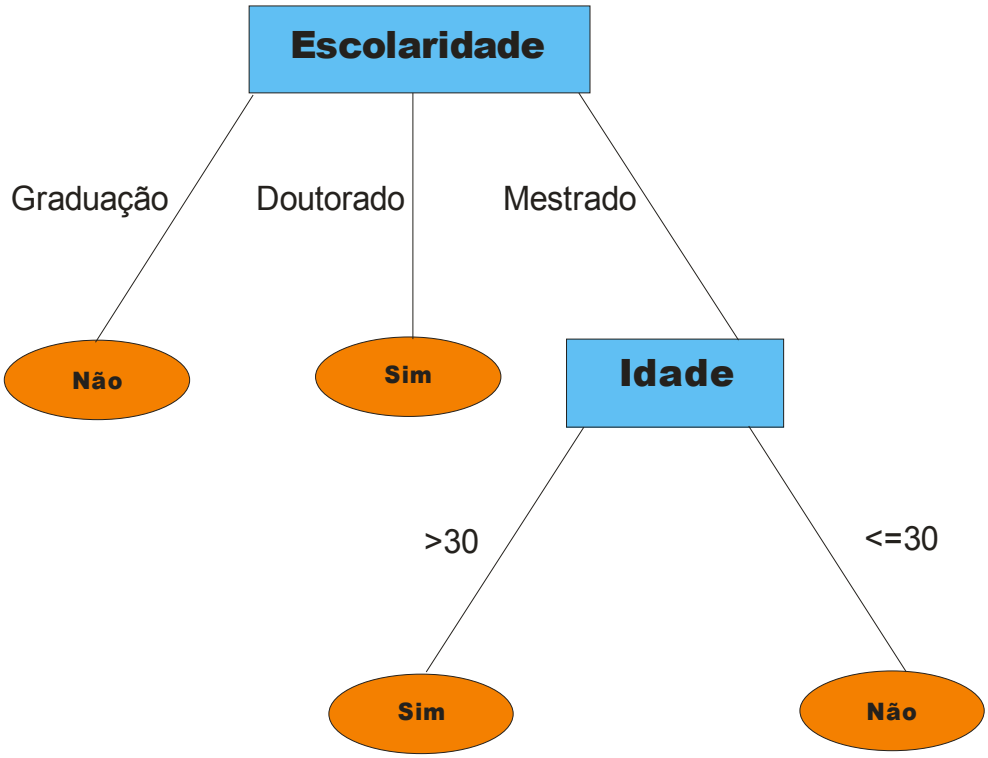

Figura 2-4: Um exemplo de Árvore de Decisão. **Fonte:** Correa (2007).

1. Se (*Escolaridade* = "*Graduação*") então (*Rico* = "*Não*")
2. Se (*Escolaridade* = "*Doutorado*") então (*Rico* = "*Sim*")
3. Se (*Escolaridade* = "*Mestrado*") e (*Idade* = "*>30*") então (*Rico* = "*Sim*")
4. Se (*Escolaridade* = "*Mestrado*") e (*Idade* = "*<=30*") então (*Rico* = "*Não*")

As vantagens das árvores de decisão é que podem ser aplicadas a um grande conjunto de dados, possibilitando uma melhor visão, e o resultado do algoritmo é de fácil compreensão pelo usuário. As desvantagens estão na possibilidade de erros nas classificação quando existem muitas classes.

Dias (2001) destaca alguns exemplos de algoritmos para a construção de um árvore de decisão, que são: CART (BERRY e LINOFF, 1997), CHAID(BERRY e LINOFF, 1997), ID3(QUINLAN, 1983), C4.5(QUINLAN, 1983), SLIQ(METHA et al, 1996) E SPRINT (SHAFER, et al, 1996).

2.4.3 Regras de Associação

A regra de associação é uma expressão representada na forma X => Y (X implica em Y), em que X e Y são conjunto de itens na base de dados e X ∩ Y = Ø. X é o antecedente da regra (lado esquerdo) e Y é o conseqüente da regra (lado direito) e pode envolver qualquer número de itens em cada lado da regra. O significado desta regra é que as transações da base que contém X tendem a conter Y. Um exemplo prático é afirmar que "30% dos registros que contêm X também contêm Y; 2% dos registros contêm ambos" (AGRAWAL et al., 1997).

A regra de associação possui dois parâmetros básicos: o suporte e a confiança. Estes parâmetros limitam a quantidade de regras que serão extraídas e descrevem a qualidade delas.

Considerando que o conjunto de itens X e Y estão sendo analisados, o suporte é definido como a fração de registros que satisfaz a união dos itens no conseqüente (Y) e no antecedente (X), correspondendo ao significado estatístico da regra (AGRAWAL et al., 1997).

A confiança é expressa pelo percentual de registros que satisfaz o antecedente (X) e o conseqüente (Y) em relação ao numero de registros que satisfaz o antecedente, medindo a força da regra ou sua precisão (AGRAWAL et al., 1997). No exemplo anteriormente citado, 30% é o fator de confiança e 2% é o suporte da regra.

Berry et. al. (1997) define a confiança como a freqüência com que o relacionamento mantém-se verdadeiro na amostra de treinamento e o suporte como a freqüência com que a combinação acontece. Assim, uma associação pode se manter 100% do tempo e ter a mais alta confiança, porém pode ser de pouca utilidade se o suporte ocorrer raramente.

Para Agrawal et al. (1997), o problema das regras de associação é encontrar todas as regras que possuam o suporte e a confiança acima de um determinado valor mínimo, pois, na prática, os usuários normalmente estão interessados somente num subconjunto de associações.

É importante destacarmos que a técnica de descoberta de regras de associação é própria da tarefa de associação (DIAS, 2001). A facilidade de interpretação das regras de associação, aliada a uma utilidade prática muito forte, incentivou inúmeros investigadores a desenvolverem algoritmos de descoberta de regras de associação. Os primeiros algorit-

mos a serem utilizados na descoberta de regras de associação foram o AIS (AGRAWAL et al., 1993) e SETM (HOUSTOMA et. al., 1993). Porém, depois desta data, vários algoritmos foram criados. Um dos algoritmos atualmente, mais referenciados para este tarefa é o Apriori (AGRAWAL et.al., 1997).

2.4.4 Redes Neurais Artificiais

Redes Neurais Artificiais são soluções computacionais que envolvem o desenvolvimento de estruturas matemáticas com a habilidade de aprendizagem. As redes neurais têm uma notável habilidade de derivar medidas de dados complicadas ou imprecisas e podem ser utilizadas para extrair padrões e detectar tendências que são muitos complexas para serem percebidas tanto por humanos quanto por outras técnicas computacionais (DWBRASIL, 2004). Devido a este fato, elas têm sido cada vez mais aplicadas em técnicas de *data mining,* pela relativa simplicidade do seu uso quando comparadas às demais tecnologias.

Os modelos neurais foram concebidos com base na estrutura do sistema nervoso, mais especificamente na estrutura do cérebro humano e, assim, sua principal característica está na capacidade de aprender com base na exposição a exemplos. A construção de uma rede neural se constitui, portanto, na configuração da sua arquitetura interna (uma rede interligada de neurônios) e no treinamento desta rede com base em exemplos, até que ela própria consiga aprender como resolver o problema.

Para isto, os pesquisadores tiveram que buscar alternativas para modelar o neurônio biológico, tanto na sua estrutura como na sua funcionalidade, na conectividade, na interatividade dos neurônios e, principalmente, na dinâmica operacional do sistema biológico. Este tipo de rede necessita de arquiteturas paralelas, de algoritmos adequados na fase de aprendizado e alta capacidade de processamento (DWBRASIL, 2004).

De acordo com Silva (2003), o treinamento das redes neurais pode ser do tipo supervisionado (as classes são conhecidas) e não supervisionado (as classes não são conhecidas). No primeiro tipo, utiliza-se algoritmos para construir modelos preditivos que podem capturar interações não lineares entre os atributos. O treinamento não supervisionadas é usado para dividir os dados em agrupamentos de acordo com certas regras pré-definidas.

A rede é geralmente organizada em camadas, isto é, um neurônio situado em uma determinada camada tem sua saída conectada a todos os neurônios da camada seguinte (a

sua direita) e a nenhum outro neurônio de qualquer outra camada, seja ela anterior, posterior ou a sua própria. A Figura 2-5 exemplifica esta topologia.

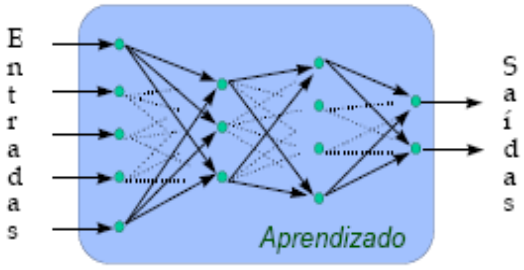

Figura 2-5: Exemplo de uma Rede Neural.
Fonte: Thomé (2007).

Cabe ressaltar que o aumento do número de camadas acarreta o aumento da complexidade e do tempo de processamento da rede. Logo, o objetivo de projeto deve ser o de resolver o problema com a menor topologia possível.

Na fase de treinamento, a rede neural aprende o problema e tenta resolvê-lo auto ajustando seus parâmetros internos. Uma vez que a rede tenha aprendido, isto é, ela tenha chegado a uma condição de erro considerada satisfatória, seus parâmetros são congelados e ela, a partir de então, está pronta para ser usada com dados da situação corrente. Existem várias arquiteturas e várias técnicas de treinamento de rede neural propostas na literatura, cada uma com suas vantagens e desvantagens, dependendo do problema e da aplicação específica a que se destinam.

Na etapa do treinamento, é escolhido o algoritmo de aprendizado juntamente com os parâmetros de aprendizado – taxa de aprendizado (*learning rate*), momento (*momemtum*), condição de parada e dinâmica de treinamento. O aprendizado é o processo pelo qual a rede adapta seus parâmetros (em geral os pesos das conexões entre os neurônios) de forma a satisfazer os requisitos de mapeamento estabelecidos. A dinâmica de treinamento representa a freqüência com que estes parâmetros (pesos) são atualizados.

Nesta dinâmica, a ordem da apresentação dos padrões é importante para a velocidade de aprendizado da rede e, em alguns casos, deve-se reorganizar esta ordem, de forma a acelerar o treinamento.

A taxa de aprendizado é um valor positivo, geralmente menor do que 1.0, que regula a intensidade com que as atualizações dos parâmetros (pesos) serão efetuadas. Taxas muito baixas, próximas de zero, tendem a fazer com que o aprendizado seja bastante lento, porém taxas muito altas, próximas de 1.0, podem fazer com que a rede oscile, como se estivesse aprendendo e desaprendendo, e às vezes, nem consiga chegar a um patamar aceitável de aprendizado.

A taxa de momento é um parâmetro de uso opcional, de valor também positivo e menor do que 1.0, cuja utilização visa imprimir uma dinâmica no treinamento tal que, eventualmente, possibilite o algoritmo buscar o menor erro possível (erro mínimo).

A condição de parada é geralmente estipulada com base na ocorrência de dois eventos: erro mínimo e número de ciclos. A parada pelo erro mínimo ocorre se e quando o algoritmo de treinamento levar a rede a convergir para um erro menor que o mínimo estipulado como critério de término. A parada pelo número de ciclos de treinamento encerra o processo independentemente do nível de aprendizado alcançado pela rede. A parada por este critério deve sempre ser utilizada em conjunto com qualquer outro, com vistas a evitar processos de treino intermináveis.

A vantagem principal da utilização de redes neurais, conforme Adriaans et. al. (1996), é a versatilidade e o resultado satisfatório em áreas complexas. Tem um excelente desempenho em problemas de classificação e reconhecimento de padrões como para o reconhecimento de caracteres, de imagens, de voz, na identificação de impressões digitais, análise de crédito, dentre outros.

Adriaans et. al.(1996), afirma que as desvantagens existentes dizem respeito à solução final, que depende das condições finais estabelecidas na rede, pois os resultados dependem dos valores aprendidos. Outra desvantagem consiste na apresentação de uma "caixa preta" que não contém informações que justifique as conclusões obtidas. As redes neurais não podem provar uma teoria a partir do que aprenderam. Elas são simples "caixa preta" que produzem respostas, mas não demonstram claramente o desenvolvimento de como chegaram aos resultados.

Alguns exemplos de redes neurais são: Perceptron, Rede MLP, Redes de Kohnen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Countepropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB [AZEVEDO, 2000], [BRAGA, 2000], [HAYKIN, 2001].

2.5 Ferramentas de Mineração de Dados

De acordo com Dias (2001), muitas ferramentas atualmente disponíveis são ferramentas genéricas da Inteligência Artificial ou da comunidade estatística. Tais ferramentas geralmente operam separadamente da fonte de dados, requerendo uma quantidade significativa de tempo gasto com exportação e importação de dados, pré e pós-processamentos e transformação de dados. Entretanto, segundo os autores, a conexão rígida entre a ferramenta de descoberta de conhecimento e a base de dados analisada, utilizando o suporte do SGBD (Sistema de Gerenciamento de Banco de Dados) existente, é claramente desejável. Para Goebel et. al. (1999), as características a serem consideradas na escolha de uma ferramenta de descoberta de conhecimento devem ser as seguintes:

1) A habilidade de acesso a uma variedade de fontes de dados, de forma *on-line* e *off-line*;

2) A capacidade de incluir modelos de dados orientados a objeto ou modelos não padronizados (tal como multimídia, espacial ou temporal);

3) A capacidade de processamento com relação ao numero máximo de tabelas/tuplas/atributos;

4) A capacidade de processamento com relação ao tamanho do banco de dados, e

5) Variedade de tipos de atributos que a ferramenta pode manipular.

Existem ferramentas que implementam uma ou mais técnicas de mineração de dados. A tabela 2-1 relaciona alguma dessas ferramentas, fornecendo informações tais como: a empresa fornecedora, as técnicas implementadas de mineração de dados e exemplos de aplicações.

Tabela 2-1: Principais Ferramentas de Mineração de Dados disponíveis no mercado. **Fonte:** Aracele (2004).

Nome	Fabricante	Funções	Destaque	Distribuição
Intelligent Miner	IBM	Algoritmos para regras de associação, classificação, regressão, padrões seqüenciais, clustering	Integrado com o SGBD B2D da IBM. Grande escalabilidade dos algoritmos	Software Proprietário
Enterprise Miner	SAS Institute Inc.	Algoritmos de classificação, regressão, pacotes de análise estatística	Grande variedade de ferramentas estatísticas	Software Proprietário
MineSet	Silicon Graphics Inc.	Algoritmos para regras de associação, classificação, análise estatística	Um robusto conjunto de ferramentas avançadas de visualização	Software Proprietário
Clementine	Integral Solutions Ltd.	Algoritmos para regas de in-dução, redes neurais, clas-sificação e ferramentas de visualização	Interface Orientada-Objeto	Software Proprietário
DBMiner	DBMiner Technology	Algoritmos para regras de associação, classificação e clusterização	Data Mining utilizando OLAP	Software Proprietário
Gemanics Expression	Gemanics Developer	Algoritmos de análises de seqüências	Aplicativo para análises de seqüências de proteínas	Software Proprietário
Weka	Universidade de Waikato	Algoritmos para regras de associação, classificação, regressão, clusterização, redes neurais e ferramentas de visualização	Implementada várias taferas e técnicas de Mineração de Dados	Software Livre de Domínio Público

2.5.1 Weka

A ferramenta WEKA (*Waikato Environment for Knowledge Analysis*), também tem sido bastante utilizada na realização da etapa de mineração de dados. Essa ferramenta foi implementada na linguagem Java e desenvolvida no meio acadêmico da Universidade de Waikato, na Nova Zelândia, em 1999. Foi utilizada esta ferramenta pelo fato de ser de domínio público, estando disponível para donwload em http://www.cs.waikato.ac.nz/weka, onde pode ser melhor compreendida. Esta ferramenta é formada por um conjunto de algoritmos que implementam diversas técnicas para resolver problemas reais de mineração de dados (WITTEN et. al., 2000).

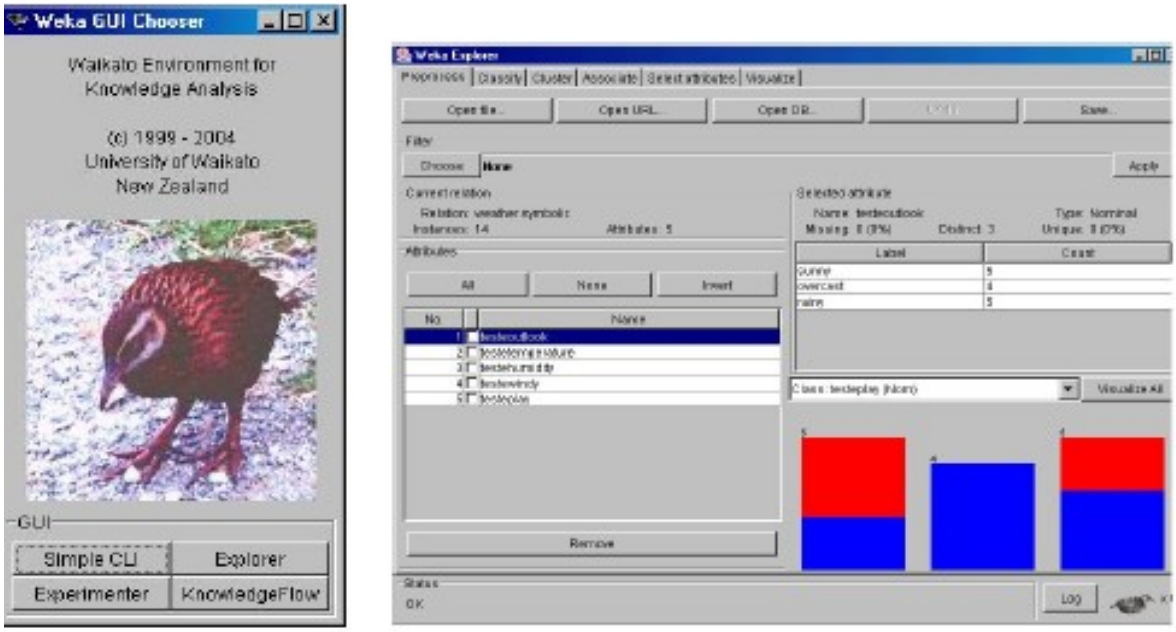

Figura 2-6: Interfaces Gráficas da Ferramenta WEKA.

É composta de dois pacotes que podem ser embutidos em outros programas escritos em Java, permitindo que um desenvolvedor possa criar seu próprio ambiente de mineração de dados. O primeiro pacote possui interfaces para manipulação interativa de algoritmos de MD e o segundo possui classes Java que "encapsulam" esses algoritmos. A ferramenta pode ser utilizada de duas formas: através de linhas de comando ou de uma interface gráfica.

2.6 Trabalhos Relacionados

Na Universidade Federal de Minas Gerais (UFMG), as técnicas de MD foram utilizadas na determinação do perfil dos alunos, com o objetivo de analisar o desempenho dos candidatos ao vestibular desta instituição em 1997. Para isto, foram utilizadas várias características sócio-econômicas coletadas através do questionário respondido pelos candidatos que se inscreveram no vestibular. Os resultados verificam que estes fatos estão, como previsto, fortemente associados com o desempenho dos alunos.

Na Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), um programa de obtenção de conhecimento depois de examinar milhares de alunos, forneceu a seguinte regra: Se o candidato é do sexo feminino, trabalha e teve aprovações com boas notas, então não faz matrícula nesta instituição. Estranho, ninguém havia pensado nisso! Mas uma reflexão justifica a regra fornecida pelo programa: De acordo com os costumes do Rio de Ja-

neiro, uma mulher em idade de vestibular, se trabalha, é porque precisa, e, neste caso, deve ter feito inscrição para ingressar na universidade pública gratuita. Se teve boas notas, provavelmente foi aprovada na universidade pública, onde efetivará matrícula.

No Centro Universitário de Formiga (Unifor/MG), observou-se uma característica interessante obtido através da aplicação da técnica de Mineração de Dados do questionário sócio-econômico deste vestibular: 43% dos candidatos que moram em uma distância máxima de 100 km de Formiga, ficaram sabendo do processo seletivo através de panfletos. Através deste resultado, pode-se tentar melhorar a qualidade da divulgação e do processo seletivo, atingindo, assim, um público-alvo maior e realmente interessado na qualidade de ensino desta instituição.

2.7 Considerações Finais

Atualmente, os sistemas de descoberta de conhecimento são utilizados nas empresas por exercer um papel fundamental na realização de suas atividades relacionadas à tomada de decisões. Com a crescente competitividade, existe uma tendência de permanecer no mercado aquelas empresas que estiverem preparadas e melhor souberem usar as informações disponíveis em seus bancos de dados.

Para tanto, a aplicação das técnicas de mineração de dados em sistemas de descoberta de conhecimento em banco de dados, busca uma fonte de conhecimento útil, porém não explicitamente representada pelo usuário. Para Dias (2001), o usuário de um sistema de descoberta de conhecimento em banco de dados precisa ter um entendimento sólido do negócio da empresa para ser capaz de selecionar corretamente os subconjuntos de dados e as classes de padrões mais interessantes.

Na implementação de mineração de dados, dificilmente haverá uma técnica que resolva todos os problemas de uma empresa ou de uma instituição de ensino. Ter conhecimento das tarefas de mineração de dados, bem como dos algoritmos para a aplicação das mesmas é necessário para proporcionar a melhor abordagem de acordo com os problemas apresentados. A tarefa específica que será executada e, os dados disponíveis para a análise são dois fatores importantes que influenciam na escolha das técnicas de mineração de dados.

3 DESENVOLVIMENTO DO TRABALHO

Para o desenvolvimento deste trabalho, foi utilizada a base de dados dos vestibulares realizados no primeiro e segundo semestres do ano de 2006, que contém os dados coletados do questionário sócio-econômico e cultural preenchido pelos mesmos.

Inicialmente foram comparadas as perguntas do questionário sócio-econômico cultural referentes aos dois semestres do ano de 2006. Com essa análise, notou-se que algumas perguntas sofreram alterações na sua estrutura, como, por exemplo, na pergunta em que se deseja saber o estado civil do vestibulando, notou-se a existência de dois tipos de formulações:

Qual o seu estado civil? (1/2006)

Opções de resposta: 01. Solteiro; 02. Casado

Qual o seu estado civil? (2/2006)

Opções de resposta: 01. Solteiro; 02. Casado; 03. Viúvo; 04. Separado judicialmente ou divorciado; 05. Outro.

Para sanar este problema, todas as perguntas foram ajustadas conforme ao último questionário sócio-econômico elaborado, que se encontra em anexo, pois este era o mais atual e abrangente.

O próximo passo foi enviar a todos os coordenadores dos cursos de graduação desta universidade um documento pedindo-os que elaborassem, com base neste questionário, questões de seu interesse referente ao curso que eles coordenam. Isto teve como meta tentar definir o tipo de informação que seria interessante de ser descoberta na base de dados e iniciar o processo de KDD, através da compreensão do domínio da aplicação e do estabelecimento dos objetivos a serem atingidos.

Logo após, iniciou-se a limpeza dos dados propriamente dita. Originalmente, o questionário estava armazenado na base de dados access, como ilustra a Figura 3.1, que mostra uma parte do banco de dados.

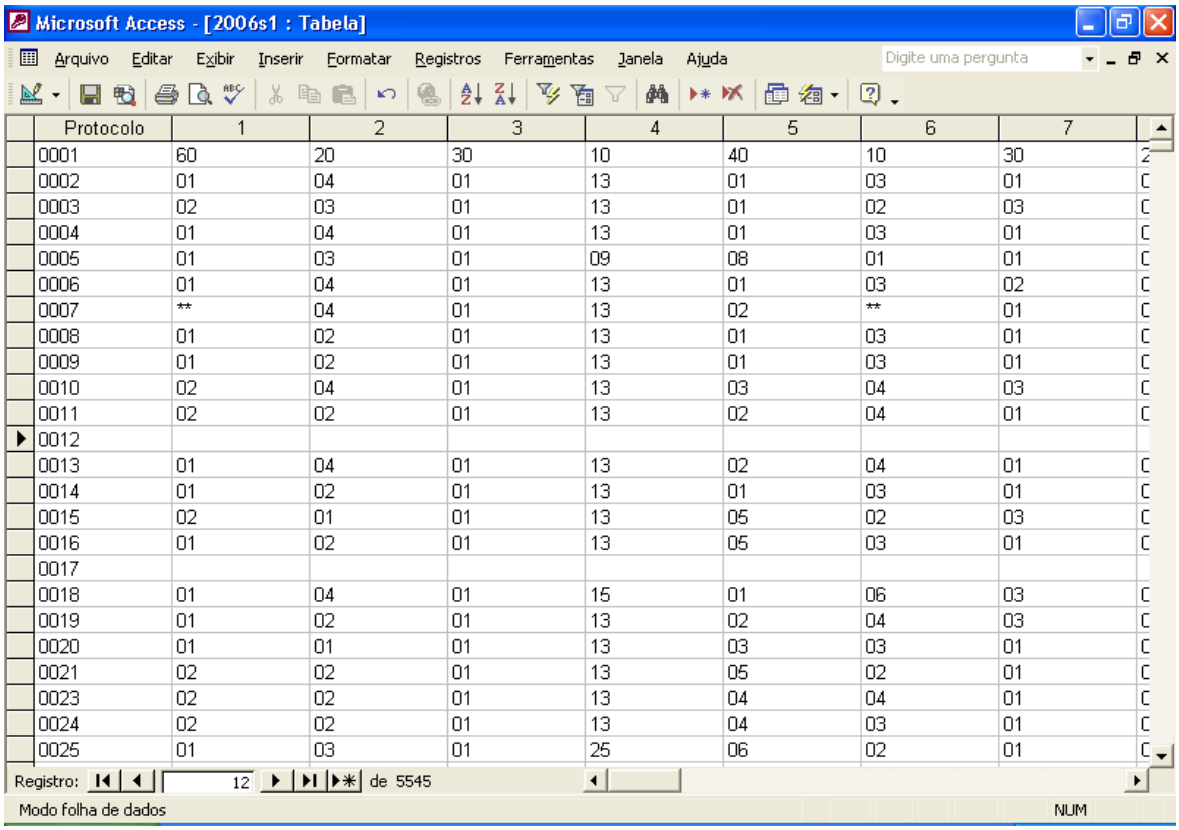

Figura 3-1: Banco de Dados Access Referente ao Questionário Sócio-Econômico.

Analisando a Figura 3-1, cada linha da base de dados corresponde às respostas fornecidas por cada vestibulando e cada coluna corresponde a uma pergunta do questionário sócio-econômico. Nota-se que alguns dados não foram preenchidos pelo candidato, como os dados cujo protocolo é 0012 e 0017. Outros dados foram inseridos com ruído na base de dados, como, por exemplo, o dado do vestibulando cujo protocolo é 007, representado como ** na base de dados. Para contornar este problema, os dados referentes a estes vestibulandos, foram eliminados, já que a base de dados é bastante extensa, com mais de 4.000 cadastros, e a eliminação destes elementos não iria prejudicar a etapa mineração de dados.

Logo após a limpeza dos dados, estes foram convertidos para o formato ARFF (*Attribute-Relation File Format*), que é um formato padrão de arquivo texto utilizado para representar os dados no software WEKA, conforme ilustra a Figura 3-2.

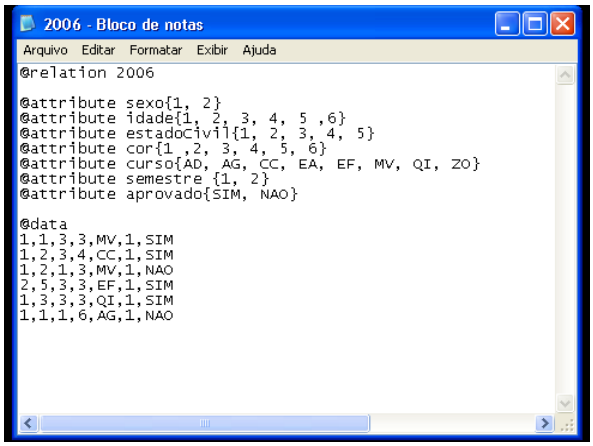

Figura 3-2: Dados no Formato ARFF.

Pela análise da Figura 3-2, um arquivo ARFF gerado, possui as seguinte estrutura:

Nome da Tabela, indicado pelo atributo @relation:

@relation 2006

Nome dos campos indicados pelo atributo @attribute e seus tipos

@attribute sexo {1, 2}
@attribute idade {1, 2, 3, 4, 5, 6}

Dados dos campos separados por virgula, logo após o atributo @data

@data
1,1,3,3,MV,1,SIM
1,2,3,4,CC,1,SIM
1,2,1,3,MV,1,NAO
2,5,3,3,EF,1,SIM
1,3,3,3,QI,1,SIM
1,1,1,6,AG,1,NAO

Após a realização desta etapa, os dados já estavam prontos para serem minerados.

3.1 Considerações Finais

Na aplicação de uma técnica de mineração de dados, a escolha da base de dados onde será efetuada a análise e da ferramenta a ser utilizada constitui atividades cruciais para o sucesso do trabalho, assim como a definição dos objetivos a serem alcançados é de suma importância para direcionar todo o processo.

Neste capítulo foi descritas a etapa de pré-processamento (definição do problema, seleção, limpeza e transformação dos dados). O objetivo geral foi detalhar os passos necessários para a aplicação de técnicas de mineração de dados utilizando a ferramenta WEKA.

4 TESTES E RESULTADOS

Neste capítulo são apresentados os testes e os resultados obtidos com a aplicação das técnicas de mineração de dados através da ferramenta WEKA, tendo como objetivo a obtenção do perfil dos candidatos que participaram do processo seletivo do vestibular da UFLA, realizados no ano de 2006.

Os resultados apresentados foram obtidos a partir da aplicação das técnicas de mineração de dados: Visualização Gráfica dos Dados, Árvore de Decisão, Regras de Associação e Redes Neurais.

O tipo de conhecimento esperado, com a realização deste trabalho, é a possibilidade de analisar o perfil dos candidatos ao processo seletivo da UFLA, bem como encontrar regras interessantes a esse respeito.

4.1 Visualização Gráfica dos Dados

Inicialmente, foi verificado se existe alguma relação entre as respostas fornecidas pelos vestibulandos, com o auxílio da técnica de visualização gráfica da ferramenta Weka, que classifica os dados de forma visual.

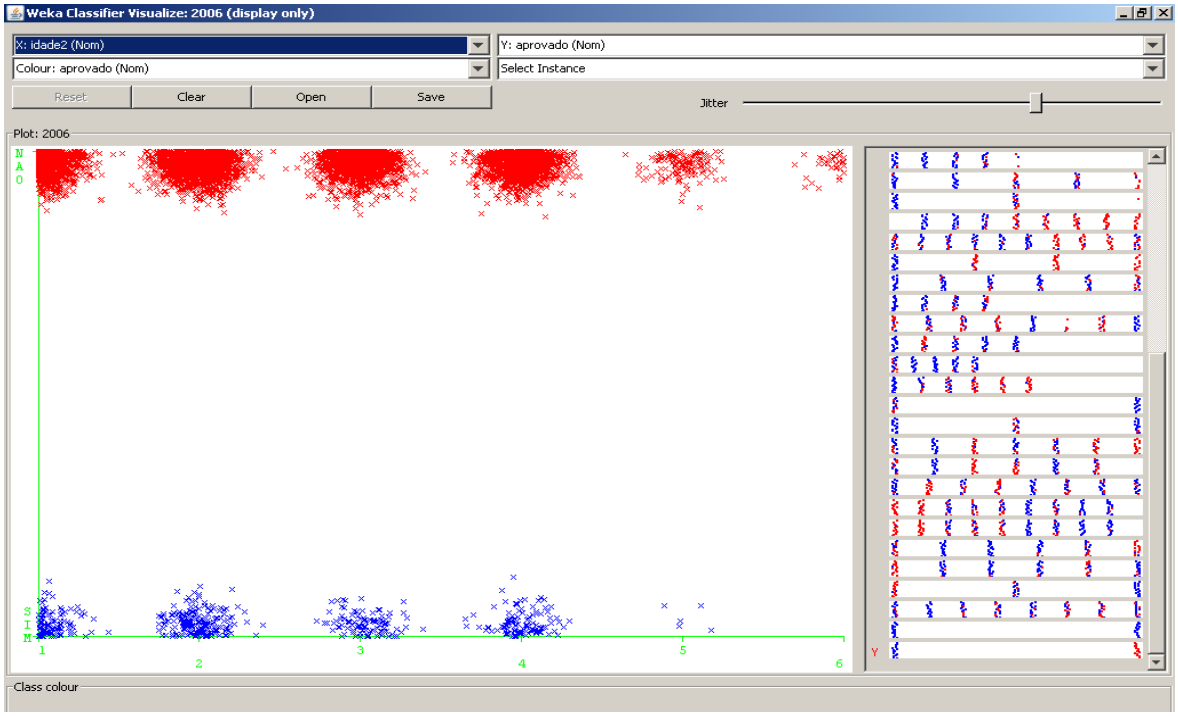

Figura 4-1: Visualização Gráfica dos Dados na ferramenta WEKA.

Analisando a Figura 4-1, observa-se que os pontos em azul correspondem aos candidatos que foram aprovados no vestibular e os pontos em vermelho, os reprovados. A partir desta Figura, pode-se concluir que 0.42% dos candidatos acima de 25 anos, correspondendo ao numero 5 do eixo x, passaram no vestibular, ou seja, os candidatos aprovados no processo seletivo de 2006 é jovem.

Pela análise gráfica de outros dados, utilizando este mesmo procedimento, outras informações interessantes foram descobertas:

1) 98% dos candidatos aprovados é solteiro, destes, 78% residem a uma distância superior a 200 Km de Lavras, cuja cidade de origem tem mais de 30.000 habitantes;

2) Dos candidatos que prestam vestibular e têm deficiência, têm paralisia permanente nas pernas, e, destes, apenas dois foram aprovados;

3) 88% dos candidatos aprovados é de cor branca ou parda, tem até quatro pessoas que compõe sua família, cuja renda é de até 10 salários mínimos;

4) 74% dos candidatos aprovados freqüentou cursinho por pelo menos um semestre, já foi classificado em um vestibular e reside a mais de 50 Km de Lavras;

5) 60% dos candidatos provém de escola pública e fizeram o ensino médio em turno diurno, sendo que a maioria das reprovações no ensino médio ocorreu nestes tipos de instituições de ensino;

6) A principal fonte de informações do candidato ainda é o telejornal e a revista;

7) 40% dos candidatos que trabalham, não fizeram cursinho e estão prestando vestibular na UFLA pela primeira vez. A principal fonte de informação destes é o telejornal e a escolaridade dos pais destes é o ensino fundamental incompleto (até a quarta série) e a principal ocupação exercida pela mãe destes candidato é do lar.

8) Dos candidatos que concluíram o ensino médio em até dois anos atrás, sua faixa etária varia de 19 a 24 anos;

9) 68% dos candidatos que concluíram o ensino médio em até três anos atrás não trabalha, estão tentando ingressar na UFLA há mais de um ano e já foram classificados em pelo menos um vestibular.

4.2 Árvore de Decisão

O experimento realizado com o algoritmo de classificação J48 tem como objetivo principal gerar árvores de decisão utilizando os dados da base de dados preparada. Selecionou-se, então, os atributos, de acordo o estabelecimento dos objetivos a serem atingidos, que foram definidos na etapa de compreensão e definição do domínio da aplicação, do processo de KDD.

Foram executados vários testes com o algoritmo J48 no ambiente WEKA. Inicialmente, utilizou-se alguns atributos da base de dados, tais como, o sexo, a idade e o curso que o candidato se inscreveu. Em outros testes realizados, foram utilizados estes mesmos atributos selecionados, juntamente com alguns outros atributos sócio-econômicos e culturais, tais como, o tipo de escola cursada (pública ou particular), o turno cursado e se o candidato fez cursinho ou não. Foram realizados, ainda, outros testes com outros atributos selecionados. Pode-se constatar que, em alguns casos, as árvores de decisão geradas, ficaram grandes, aumentando as dificuldades de sua interpretação.

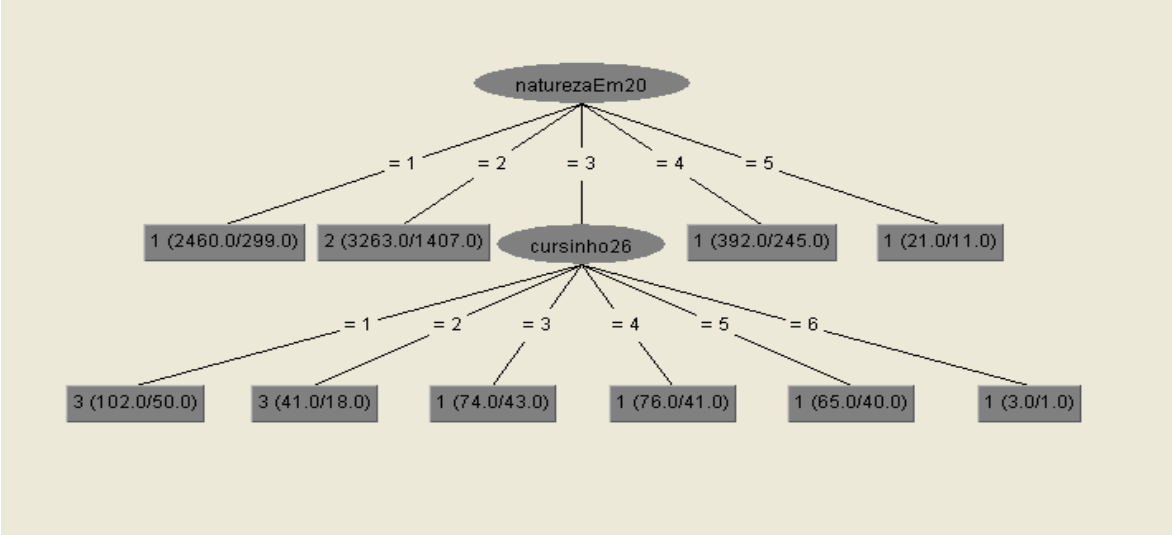

Figura 4-2: Árvore de Decisão gerada pela ferramenta WEKA.

Pela análise da árvore de decisão da Figura 4-2, pode-se concluir que:

1) Dos candidatos que fizeram o ensino médio em escola pública (naturezaEm= 3), e fizeram cursinho (cursinho= 2, 3, 4, 5, 6), 143 (18 + 43 + 41 + 40 + 1) fizeram o ensino fundamental em escola pública;

2) 392 candidatos fizeram o ensino médio em escola particular (naturezaEm= 4), destes, 245 fizeram o ensino fundamental em escola pública;

3) Dos candidatos que não fizeram cursinho (cursinho= 1), 50 fizeram a maior parte do ensino fundamental em escola pública.

Pela análise das diversas Árvore de Decisão geradas, outras informações foram descobertas:

1) Dos candidatos que fizeram o ensino médio em escola pública, 477 reside em Lavras, e desses 477, 311 têm idade entre 20 e 24 anos.

2) Dos candidatos aprovados que vieram da zona rural, 3 fizeram o ensino fundamental em escola particular, dos reprovados, 17 fizeram o ensino fundamental em escola pública.

3) Dos candidatos que prestaram mais de dois vestibulares e não foi classificado, o tempo que estão tentando ingressar no vestibular é de dois anos, e o ensino dos pais é o segundo grau completo.

4) 74% dos candidatos que prestam os cursos de Administração, Ciência da Computação e Química, a escolaridade dos pais destes é o ensino médio completo. Já nos cursos de Agronomia, Engenharia Florestal e Zootecnia, a escolaridade de 75% dos pais destes candidatos é o ensino superior completo. E, no curso de Engenharia Agrícola, a escolaridade de 71% dos pais deste candidatos é o ensino fundamental incompleto (até a quarta série).

5) Dos candidatos que fizeram o ensino médio em escola particular, 67%, reside a uma distancia a mais de 100 Km de Lavras, e, destes, 34% têm idade de 18 anos.

6) Dos alunos aprovados no vestibular, e que vieram de cidades entre 30.000 e 100.000 habitantes, 60% fizeram todo o ensino fundamental em escola pública.

7) A renda do candidato que trabalha, é de até um salário mínimo e do candidato que trabalha em atividades da família e mora sozinho, sua renda é de até 10 salários mínimos.

8) Dos candidatos que prestam vestibular, e são do sexo masculino, 40% prestam para o curso de Engenharia Florestal, seguidos de Ciência da Computação, 30% e Agronomia, 20%.

9) Dos candidatos que prestam vestibular, e são do sexo feminino, 40% prestam para o curso de Agronomia, seguidos de Engenharia Florestal, com 25% e Administração, com 20%.

10) Dos candidatos que estão tentando ingressar na UFLA pela primeira vez, 60% não

fizeram cursinho.

11) Dos candidatos que já foram classificados em um vestibular, estão tentando ingressar em um curso superior entre 1 e 2 anos e a escolaridade dos pais é o ensino superior completo.

4.3 Regras de Associação

Utilizando a técnica de Regras de Associação juntamente com o algoritmo Apriori, as seguintes relações foram geradas:

1) rendaFamiliar = 3 ensFundamental = 1 => naturezaEm = 1 (0.94)

Esta regra indica que 94% dos candidatos cuja renda familiar se enquadra na faixa de 1 a 2 salários mínimos (resposta 3 do questionário sócio-econômico), e cursou e ensino fundamental na escola pública (resposta 1 do questionário sócio-econômico), implica que estes candidatos fizeram o ensino médio também na escola pública.

2) suaCidade = 3 ensFundamental = 1 => aprovado = NAO (0.64)

Esta regra indica que 64% dos candidatos que reside em cidades de 30.000 a 100.000 habitantes e cursou o ensino fundamental em escola pública, não foram aprovados no vestibular.

3) nPFam = 5 rendaFamiliar = 5 => trabalha = 1 (0.92)

Esta regra indica que 92% dos candidatos que possui até quatro pessoas que compõe sua família e possui renda familiar entre 5 a 10 salários mínimos, não trabalham.

4) cursinho = 5 => vestibulares = 5 (0.71)

Esta regra indica que 71% dos candidatos que freqüentaram cursinho pré-vestibular por um ano ou mais, já foram classificados em um vestibular.

5) vestibulares = 5 => revistaLe = 2 (0.74)

Esta regra indica que 74% dos candidatos que já foram classificados em um vestibular, lêem revistas informativas (Veja, Exame, Isto é etc.).

6) possuiComputador = 1 curso = AG => usaComputador = 3 (0.99)

Esta regra indica que 99% dos candidatos que possui computador e são do curso de Agronomia, eles usam computador para trabalhos escolares.

7) reside = 1 => usaComputador = 3 (0.88)

Esta regra indica 88% dos candidatos que residem em Lavras, usam computador para lazer, trabalhos escolares e/ou profissionais.

8) rendaFamiliar = 3 => usaComputador = 3 (0.80)

Esta regra indica que 80% dos candidatos que possuem renda familiar entre 1 e 2 salários mínimos, usam computador.

9) curso = AG => pqUFLA = 1 (0.66)

10) curso = EF => pqUFLA = 1 (0.60)

Estas duas regras acima indicam que, dos candidatos que optaram por fazer os cursos de Agronomia e Engenharia Florestal, 66% e 60% destes, respectivamente, escolheram a UFLA pelo alto nível dos cursos. Isto evidencia que estes cursos são bastantes conhecidos pela sua qualidade.

Pela análise das regras acima extraídas, pode-se verificar que a renda familiar está diretamente relacionada com o local onde o candidato fez o ensino médio e com o ato de trabalhar. Não existe relação entre a renda familiar do candidato com o uso de computador, sendo que 80% dos candidatos que possuem renda familiar entre 1 e 2 salários mínimos, usam computador.

4.4 Redes Neurais

A aplicação desta técnica teve como objetivo descobrir se alguma pergunta do questionário sócio-econômico tem bastante ou pouca influência na aprovação do candidato. Para isto, foi utilizado o algoritmo MLP (*MultiLayerPerceptron*) disponível na ferramenta WEKA. A seguir, segue um exemplo da aplicação desta técnica, baseado nas Figuras 4-3 e 4-4.

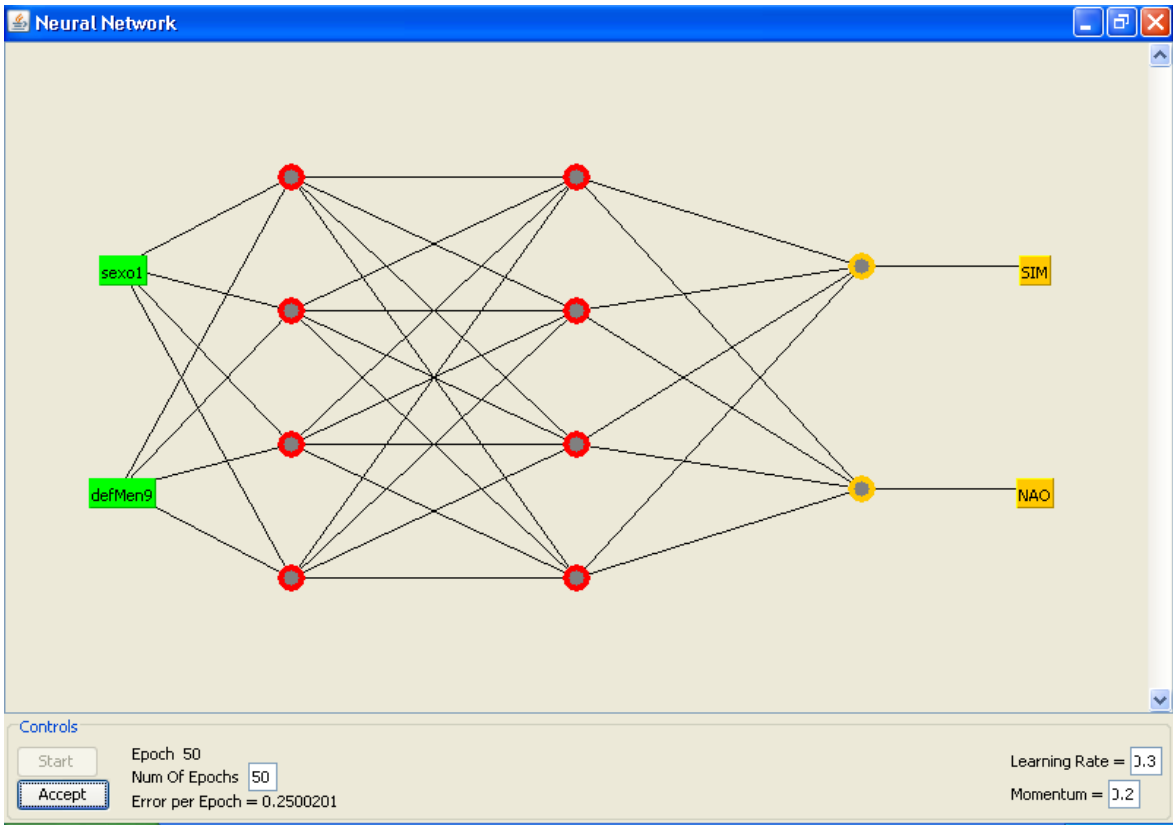

Figura 4-3: Rede Neural no ambiente WEKA.

A Figura 4-3 ilustra a rede neural no ambiente WEKA. Os retângulos em verde, são a entrada da rede neural, que são os atributos sexo e deficiência mental. Os retângulos em amarelo são a saída da rede, que no caso só podem assumir o valores SIM ou NAO, ou seja, se o candidato será aprovado ou não devido às suas características sócio-econômicas. Os círculos em vermelho são os neurônios das camadas escondidas. No exemplo, existem duas camadas escondidas com quatro neurônios para cada camada. Os círculos em amarelo são os neurônios da camada de saída, que no caso, são dois. Na parte inferior da Figura, encontra-se o numero de épocas com que a rede neural foi treinada, o erro por época, a taxa de aprendizado e o momento.

A Figura 4-4 mostra os pesos atribuídos para cada entrada do neurônio depois que a rede neural foi treinada, ou seja, pela analise desta Figura, o neurônio 2, que corresponde ao primeiro neurônio da camada escondida da rede neural ilustrada na Figura 4.3, recebeu como peso de entrada referente ao atributo sexo o valor de 0.22 e o valor de 3.28 referente ao atributo deficiência mental. O valor -0.25 corresponde a variação da taxa do erro durante a fase de treinamento. Pela análise desta Figura, pode-se concluir que, se o peso referen-

te a cada atributo de entrada for muito pequeno em relação aos demais atributos, isto indica que este atributo influência pouco na aprovação do candidato.

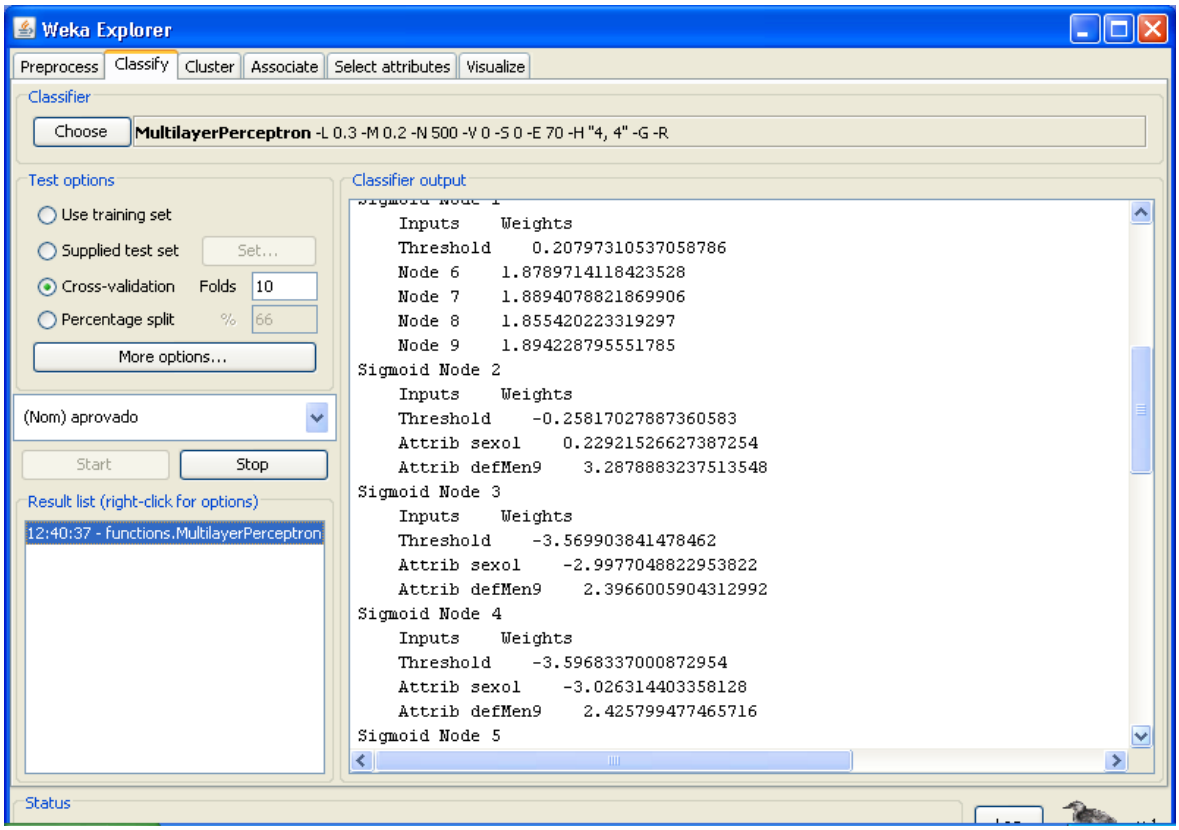

Figura 4-4: Resultado da Rede Neural treinada no ambiente WEKA.

Para a resolução deste trabalho, utilizou-se uma rede neural com as 40 perguntas do questionário sócio-econômico como parâmetros de entrada e duas camadas escondidas de 80 neurônios. A rede neural foi treinada com 70% dos dados, durante 15 horas, com 500 épocas, com taxa de aprendizado de 0,3 e com momento de 0,2.

Após o treinamento da rede neural, os pesos associados a cada neurônio da primeira camada escondida foram normalizados. Com isso, verificou-se que os atributos:

1) Capacidade de enxergar (pergunta 10 do questionário sócio-econômico);

2) Tempo tentando ingressar em um curso superior (pergunta 27);

3) Onde cursou o ensino médio (pergunta 24);

4) Onde cursou o ensino fundamental (pergunta 19); tiveram pouca influência na aprovação do candidato.

E os atributos:

1) Atividade de trabalho remunerado (pergunta 15);

2) Número de universidades que está prestando vestibular (pergunta 34);

3) Usa computador (pergunta 39);

4) Possui computador (pergunta 40); tiveram bastante influência na aprovação do candidato.

Se o objetivo da análise do questionário for prever se o candidato foi aprovado a partir das suas características sócio-econômicas, sugere-se que as perguntas que pouco influenciam na aprovação deste sejam excluídas do questionário.

4.5 Considerações Finais

Na aplicação do algoritmo J48 da ferramenta WEKA, pode-se constatar regras interessantes para futuras ações da UFLA, referente ao perfil dos candidatos que prestaram vestibular no ano de 2006.

Observou-se que nos cursos mais concorridos os dados sócio-econômicos e culturais do candidato são relevantes para o seu bom desempenho. A mesma característica não aparece como determinantes nos cursos menos concorridos.

A análise dos resultados obtidos através das regras geradas pela ferramenta em questão na base de dados do vestibular da UFLA do ano de 2006 facilitará os gestores de Instituições de Ensino Superior, de cursos pré-vestibulares, de escolas particulares e de escolas públicas na implementação de ações pedagógicas e administrativas para melhorar o desempenho dos seus alunos.

5 CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS

5.1 Conclusões

No processo de descoberta de conhecimento em banco de dados, todas as etapas, desde a preparação dos dados até a extração de conhecimento, são de extrema importância e exigem que a mesma atenção seja atribuída para cada uma delas. O sucesso de uma etapa depende exclusivamente do bom desenvolvimento das etapas anteriores.

A utilização da ferramenta WEKA para a mineração dos dados foi bastante útil, demonstrando a riqueza de informações ocultas em base de dados, reafirmando a necessidade do conhecimento.

O algoritmo Apriori desta ferramenta transformou os dados da base de dados preparada para este trabalho em regras com informações claras e relevantes, criando afinidades e dependências entre os dados. Descobrir e utilizar uma ferramenta que possa apontar soluções de forma clara e simples aos usuários na descoberta do conhecimento é um bom começo para compreender a importância do estudo da mineração de dados.

Dentre as varias técnicas existentes para a análise de dados, sem dúvida as técnicas estatísticas são as mais íntimas às técnicas de mineração de dados. Cabena et al. (1998) citam que grande parte das análises feita pelas técnicas de mineração de dados era feita pelas técnicas estatísticas. Entretanto, estes autores contemplavam que quase tudo do que é feito com a mineração de dados poderia ser feito, eventualmente, com analises estatísticas. Porém, Cabena et al. (1998) enfatizam que, o que está atraindo vários analistas para a mineração de dados é a facilidade relativa com que podem ser obtidos conhecimentos interessantes e mais elaborados em relação às aproximações estatísticas tradicionais.

Em relação às regras obtidas com a aplicação das técnicas de mineração de dados, é interessante destacar os seguintes resultados:

1) Algumas regras descobertas em uma técnica foram confirmadas ou complementadas com a utilização de outra técnica. Por exemplo, quando se utilizou a técnica de visualização gráfica dos dados, descobriu-se que a maioria dos candidatos que concluíram o ensino médio em até três anos atrás não trabalha, estão tentando ingressar na UFLA há mais de um ano e já foram classificados em pelo menos um vestibular. Esta regra foi confirmada

utilizando a técnica de redes neurais, pois a atividade de trabalho remunerado tem bastante influência na aprovação do candidato.

2) Verificou-se que o ato de fazer cursinho tem pouca influência na aprovação dos candidatos, pois 58% dos candidatos aprovados não fez cursinho, e a principal fonte de informação destes ainda é o telejornal.

3) A escolaridade dos pais influencia na escolha do curso, sendo que nos cursos da área de agrárias, a escolaridade de grande parte dos pais é o ensino superior completo.

4) Poucos candidatos acima de 25 anos passaram no vestibular, ou seja, os candidatos aprovados no processo seletivo de 2006 da UFLA são jovens.

Finalmente, com base nos estudos do processo de KDD, o objetivo desta pesquisa foi extrair conhecimentos interessantes, através das técnicas de mineração de dados, sobre os candidatos ao processo seletivo do vestibular da UFLA no ano de 2006, caracterizando as diferenças existentes entre estes, tendo como base os dados sócio-econômicos e culturais preenchidos pelos mesmos.

5.2 Sugestões para Trabalhos Futuros

O tipo de conhecimento esperado, com a realização deste trabalho, foi a possibilidade de analisar o perfil dos candidatos ao processo seletivo da UFLA, bem como encontrar regras interessantes a esse respeito. Os resultados obtidos poderão ser utilizados pela UFLA para definir novas regras para os próximos vestibulares, implementar ações visando melhorar a qualidade de ensino, diminuindo evasões e, ainda, direcionar melhor o candidato ao vestibular na escolha do curso baseado no seu perfil.

Sugere-se também, a aplicação deste processo em outras bases de dados, como a acadêmica, visando encontrar conhecimento útil sobre o histórico escolar dos alunos no decorrer dos períodos e descobrir prováveis razões de cunho positivo ou negativo que influenciaram neste histórico.

BIBLIOGRAFIA

ADDRIANS, P. ZANTINGE, D., Data Mining. Addison Wesley Longman, England, 1996.

AGRAWALL, R.; GEHRKE, J.; GUNOPULOS, D.; RAGAVAN, P., Automatic subspace clustering of right data for data mining applications. In: SIGMOD Conference, 1998, p.940-945.

AGRAWALL, R.; SRIKANT, R.; VU, Q., Mining association rules with item constrains. In: **Future generations computer system,** Elservir: Netherlands, v. 13, n. 2-3, nov., 1997, p. 161-180.

BRACHMAN, A., Data Warehousing, Data Mining, and OLAP. USA: McGraw-Hill, 1996.

BRASIL. Revista do Ensino Médio. Brasília, n° 4, ano II, 2004 (a).

BRASIL. Informativo do Ministério da Educação. Brasília, nov. 2004 (b).

BERRY, M. J. A.; LINOFF, G. Data Mining techniques- for marketing, sales and customer spport. United Satates: Wiley Computer Publishing, 1997.

CABENA, P. Discovering data mining: from concept to implementation. Upper Saddle River: Prentice-Hall PTR, 1998.

CARD S. K., Information visualization. In S. Card, Readings in Information Visualization – Using Visualization to Think, pp. 1-34 San Fransisco, CA, 1999.

CARVALHO, L. A. V. Data mining – a mineração de dados no marketing, medicina, economia, engenharia e Administração. São Paulo: Érica, 2002.

CARVALHO, D. R. Data mining através de introdução de regras e algoritmos genéricos, 2003. Dissertação Mestrado – PUCPR, Curitiba.

CORREA, Eduardo Gonçalves. Extração de Árvores de Decisão com a Ferramenta de Data Mining WEKA. Disponível em **http://www.devmedia.com.br/artibles/wiewcomp.asp?comp=3388**, acessado em 12/05/2007.

DIAS, M. M. Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados. 2001. Tese de Doutorado do Programa de Pós-Graduação em Engenharia de Produção UFSC. Florianópolis, Santa Catarina.

FAYYAD, U. From data mining to knowledge discovery: an overview. In: Advances in Knowledge discovery and data mining, AAA Press / The Mit Press, MIT, Cambridge, England, 1996, p.1-34.

FELDENS, M. A. Descoberta de conhecimento aplicada à detecção de anomalias em base de dados. Porto Alegre: PPGCC da UFRGS, 1996.

GOEBEL, M. A suvery of data mining and knowledge discovery software tools. In: SIGKDD Explorations, June, 1999.

HARJINDER, G. E., DWBrasil, Que Corporation, 2004.

HOUTSMA, M.; SWAMI A. Set-oriented mining of association rules; Research report RJ9567, IBM Almaden Research Center, San Jose, California, 1993.

IBGE. Instituto Brasileiro de Geografia e Estatística. Disponível em **http://www.ibge.gov.br**, acessado em 12/05/2007

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Disponível em **http://www.inep.gov.br/imprensa/noticias/censo/superior**, acessado em 12/05/2007.

OLIVEIRA, Aracele Garcia de. GARCIA, Denise Ferreira. Mineração da Base de Dados de um Processo seletivo Universitário. INFOCOMP – UFLA. Lavras-MG. Novembro de 2004.

PANIZZI, W. M. A Universidade Publica no Brasil de hoje. SOS universidade pública-Reforma ou demolição? Revista da Associação dos Docentes da Unicamp, ano 6, n° 2, setembro de 2004, p. 60-66.

PANSANTO, K. A.; SOARES, J. F. Desempenho dos alunos no ENEM e no vestibular da UFMG. I jornada Latino-Americana de Estatística Aplicada, 1999. São Carlos-SP p. 137-143.

PANIZZI, W. M. A Universidade Publica no Brasil de hoje. SOS universidade pública-Reforma ou demolição? Revista da Associação dos Docentes da Unicamp, ano 6, n° 2, setembro de 2004, p. 60-66.

RESENDE S. O. Sistemas inteligentes: fundamentos e aplicações. 1° ed. Barueri: Editora Manole, 2003.

SILVA, G. Estudo de técnicas de data mining e aplicação de uma das técnicas estudadas em uma base de dados da área de saúde. Trabalho de Conclusão de Curso – (Graduação em Superior em Tecnologia em Informática). Universidade Luterana do Brasil, Campus Canoas. 2003.

SOARES, J. F.; FONSECA, J. A. Fatores socioeconômicos e o desempenho no vestibular da UFMG-97, 1998. Disponível em **http://www.est.ufmg.br/proav**, acessado em 12/05/2007.

THOMÉ, A. C. G., Redes Neurais, uma ferramenta para KDD e Data Mining. Artigo publicado em jan/2007. Universidade Federal do Rio de Janeiro.

VIANNA, R. Mineração de dados: Introdução e Aplicações. Artigo SQL Magazine, Ed. 10, Ano 1.

WITTEN, I. H. ; FRANK, E. Data mining – pratical machine learning tools and Techniques with Java implementations. Morgan Kaufmann Publishers. San Fransisco, CA. 2000.

ANEXO

QUESTIONÁRIO SÓCIO-ECONÔMICO E CULTURAL

01. QUAL O SEU SEXO?
1. Masculino **2.** Feminino.

02. QUAL A SUA IDADE?
1. Até 17 anos **2.** 18 anos **3.** 19 anos **4.** 20 a 24 anos **5.** 25 a 29 anos **6.** 30 anos ou mais.

03. QUAL O SEU ESTADO CIVIL?
1. Solteiro **2.** Casado **3.** Viúvo **4.** Separado judicialmente ou divorciado **5.** Outro.

04. EM QUAL ESTADO VOCÊ RESIDE?
1. Acre **2.** Alagoas **3.** Amapá **4.** Amazonas **5.** Bahia **6.** Ceará **7.** Distrito Federal **8.** Espírito Santo **9.** F. de Noronha **10.** Goiás **11.** Maranhão **12.** Mato Grosso **13.** Mato Grosso do Sul **14.** Minas Gerais **15.** Pará **16.** Paraíba **17.** Paraná **18.** Pernambuco **19.** Piauí **20.** Rio Grande do Norte **21.** Rio Grande do Sul **22.** Rio de Janeiro **23.** Rondônia **24.** Roraima **25.** Santa Catarina **26.** São Paulo **27.** Sergipe **28.** Tocantins **29.** Resido fora do Brasil.

05. VOCÊ RESIDE:
1. Em Lavras **2.** Até 50 km de Lavras **3.** Entre 50 e 100 km de Lavras **4.** Entre 100 e 200 km de Lavras **5.** Entre 200 e 300 km de Lavras **6.** Entre 300 e 400 km de Lavras **7.** Entre 400 e 500 km de Lavras **8.** A mais de 500 km de Lavras.

06. CLASSIFIQUE SUA CIDADE
1. Capital de Estado **2.** Cidade do interior com mais de 100.000 habitantes **3.** Cidade entre 30.000 e 100.000 habitantes **4.** Cidade ou vila com menos de 30.000 habitantes. **5.** Zona rural. **6.** Outra situação.

07. VOCÊ SE CONSIDERA
1. Branco **2.** Negro **3.** Pardo **4.** Oriental **5.** Indígena **6.** Não deseja declarar.

08. VOCÊ TEM ALGUMA DAS SEGUINTES DEFICIÊNCIAS?
1. Paralisia permanente total **2.** Paralisia permanente das pernas **3.** Paralisia permanente de um dos lados do corpo **4.** Falta de perna, mão, braço, pé ou dedo polegar **5.** Nenhuma das enumeradas.

09. POSSUI ALGUMA DEFICIÊNCIA MENTAL PERMANENTE QUE LIMITE SUAS ATIVIDADES?
1. Sim **2.** Não.

10. COMO AVALIA SUA CAPACIDADE DE ENXERGAR?
1. Incapaz **2.** Grande dificuldade permanente **3.** Alguma dificuldade permanente **4.** Nenhuma dificuldade.

11. COMO AVALIA SUA CAPACIDADE DE OUVIR?
1. Incapaz **2.** Grande dificuldade permanente **3.** Alguma dificuldade permanente **4.** Nenhuma dificuldade.

12. COMO AVALIA SUA CAPACIDADE DE CAMINHAR/SUBIR ESCADAS?
1. Incapaz **2.** Grande dificuldade permanente **3.** Alguma dificuldade permanente **4.** Nenhuma dificuldade.

13. QUAL A SUA SITUAÇÃO FAMILIAR?
1. Pais vivos **2.** Pai falecido **3.** Mãe falecida **4.** Pais falecidos **5.** Situação materna desconhecida **6.** Situação paterna desconhecida.

14. INCLUINDO VOCÊ, QUAL O NÚMERO DE PESSOAS QUE MORAM NA SUA CASA E QUE COMPÕEM SUA FAMÍLIA?
1. Uma pessoa **2.** Duas pessoas **3.** Três pessoas **4.** Quatro pessoas **5.** Cinco pessoas **6.** Seis pessoas **7.** Mais de seis pessoas.

15. ATIVIDADE DE TRABALHO REMUNERADO
1. Não trabalha **2.** Trabalha para se sustentar **3.** Trabalha para sustentar a família **4.** Trabalha para ajudar no orçamento familiar **5.** Trabalha em atividades de família.

16. QUAL A SUA PARTICIPAÇÃO NA VIDA ECONÔMICA DO SEU GRUPO FAMILIAR?
1. Não trabalho e sou sustentado pela família ou por outras pessoas **2.** Trabalho, mas recebo ajuda financeira da família ou de outras pessoas **3.** Trabalho e sou responsável apenas pelo meu próprio sustento **4.** Trabalho é responsável pelo meu próprio sustento e contribuo, parcialmente, para o sustento da família **5.** Trabalho e sou o principal responsável pelo sustento da família.

17. EM QUE FAIXA MELHOR SE ENQUADRA A RENDA BRUTA MENSAL (SEM DESCONTOS) DO SEU GRUPO FAMILIAR? (EM SALÁRIOS MÍNIMOS – SM)
1. Nenhuma **2.** Até 1 SM **3.** Entre 1,1 e 2 SM **4.** Entre 2,1 e 5 SM **5.** Entre 5,1 e 10 SM **6.** Entre 10,1 e 15 SM **7.** Entre 15,1 e 20 SM **8.** Entre 20,1 e 30 SM **9.** Acima de 30 SM.

18. NÚMERO TOTAL DOS MEMBROS DA FAMÍLIA QUE DEPENDEM DA RENDA DO GRUPO FAMILIAR
1. Um **2.** Dois a três **3.** Quatro a cinco **4.** Seis a sete **5.** Oito a nove **6.** Acima de nove.

19. ONDE CURSOU O ENSINO FUNDAMENTAL?
1. Todo em escola pública **2.** Todo em escola particular **3.** Maior parte em escola pública **4.** Maior parte em escola particular **5.** Escolas comunitárias.

20. ONDE CURSOU OU CURSA O ENSINO MÉDIO?
1. Escola Federal **2.** Escola Estadual **3.** Escola Municipal **4.** Escola Particular **5.** Outra natureza.

21. EM QUE TURNO VOCÊ FEZ A MAIOR PARTE DO ENSINO MÉDIO?
1. Diurno **2.** Noturno **3.** Parte diurno e parte noturno.

22. QUANDO VOCÊ CONCLUIU OU CONCLUIRÁ O ENSINO MÉDIO?
1. Ainda não concluído **2.** Concluído este ano **3.** Concluído há um ano **4.** Concluído há dois anos **5.** Concluído há três anos **6.** Concluído há quatro anos **7.** Concluído há cinco anos **8.** Concluído de seis a dez anos **9.** Concluído há mais de 10 anos.

23. QUE CURSO DE ENSINO MÉDIO VOCÊ CONCLUIU OU CONCLUIRÁ?
1. Profissionalizante **2.** Não profissionalizante **3.** Supletivo **4.** Outro equivalente.

24. LOCAL DE CONCLUSÃO DO ENSINO MÉDIO
1. Ainda não concluído **2.** Lavras **3.** Belo Horizonte **4.** Sul de Minas Gerais **5.** Outras regiões de Minas Gerais **6.** São Paulo **7.** Rio de Janeiro **8.** Espírito Santo **9.** Distrito Federal **10.** Outros.

25. REPROVAÇÃO NO ENSINO MÉDIO
1. Nenhuma **2.** Uma **3.** Duas **4.** Mais de duas.

26. VOCÊ FREQÜENTA OU FREQÜENTOU CURSINHO PRÉ-VESTIBULAR?
1. Não **2.** Sim, menos de um semestre **3.** Sim, por um semestre **4.** Sim, por um ano **5.** Sim, por mais de um ano **6.** Sim, integrado ao ensino médio.

27. HÁ QUANTOS ANOS ESTÁ TENTANDO INGRESSAR EM UM CURSO SUPERIOR?
1. Este é o primeiro ano **2.** Há um ano **3.** Há dois anos **4.** Há três anos ou mais.

28. VESTIBULARES PRESTADOS ANTERIORMENTE
1. Nunca prestou **2.** Prestou um vestibular e não foi classificado **3.** Prestou dois vestibulares e não foi classificado **4.** Prestou mais de dois vestibulares e não foi classificado **5.** Já foi classificado em um vestibular.

29. O QUE O LEVOU A ESCOLHER O SEU CURSO?
1. Vocação ou tendência espontânea **2.** Indicação através de teste vocacional **3.** Interesse despertado pelo curso **4.** Maior oportunidade no mercado de trabalho **5.** Prestígio profissional **6.** Prestígio social **7.** Importância do curso no cenário nacional **8.** Pais, parentes ou amigos **9.** A competição pelas vagas é menor do que nos outros cursos **10.** Outras razões.

30. QUAL A RAZÃO PRINCIPAL QUE O LEVOU A ESCOLHER A UFLA?

1. Alto nível dos cursos **2.** Influencia de amigos ou parentes que trabalham ou estudam na UFLA **3.** Reside em Lavras **4.** Reside em cidade próxima de Lavras **5.** É uma universidade federal **6.** Oferece dois vestibulares por ano **7.** Falta de condições financeiras para se manter em outra cidade **8.** Uma opção a mais na tentativa de se ingressar em um curso superior **9.** Outras razões.

31. COMO VOCÊ FICOU SABENDO DOS CURSOS OFERECIDOS PELA UFLA?

1. Jornal **2.** Rádio **3.** Televisão **4.** Internet **5.** Correios **6.** Divulgação pela UFLA em sua escola **7.** Através de alunos e ex-alunos da UFLA **8.** Por intermédio de professores e funcionários da UFLA **9.** Eventos de divulgação/visita a UFLA **10.** Publicações e folhetos sobre a UFLA.

32. COMO FICOU SABENDO DO VESTIBULAR DA UFLA?

1. Através de parentes e amigos **2.** Através de alunos da UFLA **3.** Através de ex-alunos da UFLA **4.** Através de colégios **5.** Através de "cursinho" **6.** Através de cartazes e folhetos **7.** Através da Internet **8.** Através de rádio **9.** Através de televisão **10.** Através de jornal ou revista **11.** Através de contato direto com a UFLA **12.** Porque reside em Lavras **13.** Por outra forma.

33. QUAL FOI A SUA PRINCIPAL FONTE DE INFLUÊNCIA AO ESCOLHER A UFLA?

1. A família **2.** A escola ou cursinho **3.** Alguém conhecido **4.** Os meios de comunicação **5.** Outros.

34. INCLUINDO A UFLA, EM QUANTAS UNIVERSIDADES VOCÊ ESTÁ PRESTANDO VESTIBULAR?

1. Apenas na UFLA **2.** Em duas Instituições **3.** Em três Instituições **4.** Em quatro Instituições **5.** Em cinco ou mais Instituições.

35. QUANTOS VESTIBULARES JÁ PRESTOU NA UFLA?

1. Não prestou **2.** Um **3.** Dois **4.** Três **5.** Quatro **6.** Mais de quatro.

36. CASO A UFLA OFEREÇA NOVOS CURSOS, QUAL A SUA SUGESTÃO PARA O TURNO DO CURSO?

1. Diurno **2.** Noturno **3.** Tanto faz.

37. CASO A UFLA OFERECESSE OS CURSOS ABAIXO, VOCE TERIA INTERESSE EM FAZER ALUGM DELES?

1. Direito **2.** Nutrição **3.** Gastronomia **4.** Física **5.** Economia **6.** Jornalismo **7.** Engenharia Ambiental **8.** Engenharia de Controle e Automação **9.** Fisioterapia. **10.** Museologia.

38. QUAL A SUA SUGESTAO PARA O CURSO?

1. Não tenho sugestão de curso **2.** Administração **3.** Agrimensura **4.** Agronegócio **5.** Agropecuária **6.** Arquitetura e Urbanismo **7.** Artes Cênicas **8.** Biblioteconomia **9.** Biologia **10.** Biomedicina **11.** Bioquímica **12.** Biotecnologia **13.** Ciências Biológicas **14.** Ciências Contábeis **15.** Ciências Econômicas **16.** Ciências Humanas **17.** Ciências Sociais **18.** Cinema **19.** Comercio Exterior **20.** Comunicação Social **21.** Contabilidade **22.** Decoração **23.** Desenho Industrial **24.** Direito **25.** Ecologia **26.** Economia **27.** Educação Física **28.** Enfermagem **29.** Engenharia Agrícola **30.** Engenharia Ambiental **31.** Engenharia Astronáutica **32.** Engenharia Biomédica **33.** Engenharia Civil **34.** Engenharia da Computação **35.** Engenharia de Agrimensura **36.** Engenharia de Automação **37.** Engenharia de Materiais **38.** Engenharia de Minas **39.** Engenharia de Nanotecnologia **40.** Engenharia de Petróleo **41**. Engenharia de Produção **42.** Engenharia de Tecelagem **43.** Engenharia Elétrica **44.** Engenharia Física **45.** Engenharia Florestal **46.** Engenharia Genética **47.** Engenharia Geológica **48.** Engenharia Hídrica **49.** Engenharia Industrial **50.** Engenharia Mecânica **51.** Engenharia Metalúrgica **52.** Engenharia Naval **53.** Engenharia Química **54.** Eqüinocultura **55.** Farmácia **56.** Farmácia/Bioquímica **57.** Filosofia **58.** Física **59.** Fisioterapia **60.** Gastronomia **61.**Geografia **62.** Geologia **63.** Gestão Ambiental **64.** Historia **65.** Hotelaria **66.** Ilustração Cientifica **67.** Jornalismo **68.** Laticínios **69.**Letras **70.** Matemática **71.** Mecatronica **72.** Medicina **73.** Medicina Veterinária **74.** Microbiologia **75.** Mídia **76.** Moda **77.** Musica **78.** Nutrição **79.** Oceanografia **80.** Odontologia **81.** Pedagogia **82.** Psicologia **83.** Publicidade **84.** Publicidade e Propaganda **85.** Química (Bacharelado) **86.** Química Ambiental **87.** Química Forense **88.** Química Industrial **89.** Radiologia **90.** Relações Internacionais **91.** Resposta Incompatível **92.** Saúde Ambiental **93.** Secretariado **94.** Sistemas de Informação **95.** Sociologia **96.** Telecomunicações 97. Terapia Ocupacional **98.** Turismo **99.** Zoologia.

39. USA COMPUTADOR?

1. Não **2.** Sim, só para lazer **3.** Sim, para lazer, trabalhos escolares e/ou profissionais e acesso à internet.

40. POSSUI COMPUTADOR EM SUA RESIDÊNCIA?

1. Sim, com acesso à Internet **2.** Sim, sem acesso à Internet **3.** Não.

41. QUAL A SUA PRINCIPAL FONTE DE INFORMAÇÃO SOBRE OS ACONTECIMENTOS ATUAIS?

1. Jornal escrito **2.** Telejornal (TV) **3.** Jornal falado (rádio) **4.** Internet **5.** Revistas **6.** Outras fontes **7.** Não me mantenho informado.

42. DOS TIPOS DE REVISTAS ABAIXO, QUAL VOCÊ MAIS LÊ?

1. Humor e/ou quadrinhos **2.** Informativas (Veja, Exame, Isto é, etc.) **3.** Fotonovelas **4.** Esportivas **5.** Científicas **6.** Generalidades (Caras, Nova, etc).

43. COM QUAL ATIVIDADE ABAIXO VOCÊ OCUPA MAIS TEMPO?

1. TV **2.** Indo ao teatro/cinema **3.** Ouvido música **4.** Indo a bares, boates, etc**. 5.** Lendo **6.** Praticando esportes **7.** Navegando na internet **8.** Nenhuma destas.

44. NÍVEL DE ESCOLARIDADE DO SEU PAI:
1. Nenhum **2.** Ensino Fundamental incompleto (até a 4° série) **3.** Ensino Fundamental incompleto (após a 4° série) **4.** Ensino Fundamental completo **5.** Ensino Médio incompleto **6.** Ensino Médio completo **7.** Superior incompleto **8.** Superior completo **9.** Pós-graduação.

45. NÍVEL DE ESCOLARIDADE DA SUA MÃE: (Utilize os mesmos códigos da questão 44)

46. OCUPAÇÃO PRINCIPAL EXERCIDA PELO SEU PAI (Localize sua resposta nos agrupamentos de ocupações, usando o código correspondente, mesmo que seu pai seja aposentado ou falecido).
1. Agrupamento I **2.** Agrupamento II **3.** Agrupamento III **4.** Agrupamento IV **5.** Agrupamento V **6.** Agrupamento VI.

46. OCUPAÇÃO PRINCIPAL EXERCIDA PELO SUA MÃE (Localize sua resposta nos agrupamentos de ocupações, usando o código correspondente, mesmo que seu pai seja aposentado ou falecido).
1. Agrupamento I **2.** Agrupamento II **3.** Agrupamento III **4.** Agrupamento IV **5.** Agrupamento V **6.** Agrupamento VI.

AGRUPAMENTOS DE OCUPAÇÕES

AGRUPAMENTO I: Banqueiro, deputado, senador, diplomata, capitalista, alto posto militar (como general e marechal), alto cargo de chefia ou gerência em grandes organizações, alto posto administrativo no serviço público, pecuarista, grande industrial (empresas com mais de 100 empregados), grande proprietário rural (com mais de 2.000 hectares), e outras ocupações com características semelhantes.
AGRUPAMENTO II: Profissional liberal de nível universitário (como médico, engenheiro, arquiteto, advogado, dentista, etc.), cargo técnico-científico (como pesquisador, químico-industrial, professor de universidade, jornalista ou outra ocupação de nível superior), cargo de chefia ou gerência em empresa comercial ou industrial de porte médio (10 a 100 empregados), posto militar de tenente, capitão, major ou coronel, grande comerciante, dono de propriedade rural (de 200 a 2.000 hectares); e outras ocupações com características semelhantes.
AGRUPAMENTO III: Bancário, oficial de justiça, professor do ensino fundamental e médio, despachante, representante comercial, auxiliar administrativo, auxiliar de escritório ou outra ocupação que exija curso de 1º grau (ensino fundamental) completo, incluindo funcionário público com esse nível de instrução e exercendo atividades semelhantes, posto militar de sargento, subtenente e equivalentes pequeno industrial (até 10 empregados), comerciante médio, proprietário rural (de 20 a 200 hectares), e outras ocupações com características semelhantes.
AGRUPAMENTO IV: Datilógrafo, telefonista, mecanógrafo, contínuo, recepcionista, motorista (empregado), cozinheiro e garçom de restaurante, costureiro, operário qualificado (que tem um mínimo de aprendizado profissional, como mecânico, gráfico, metalúrgico, ferramenteiro), porteiro, chefe-de-turma, mestre de produção fabril, serralheiro, marceneiro,

comerciário (como balconista, empregado de loja de artigos finos ou de estabelecimento comercial de grande porte – casa de roupa, sapataria, joalheria, farmácia, drogaria, loja de aparelhos domésticos, imobiliária), funcionário público no exercício de atividades semelhantes, posto militar de soldado, cabo e equivalentes, pequeno comerciante, sitiante, pequeno proprietário rural (até 20 hectares), e outras ocupações com características semelhantes.

AGRUPAMENTO V: Operário (não qualificado), servente, carregador, empregada doméstica (como cozinheira, passadeira, copeira, lavadeira, arrumadeira), lixeiro, biscateiro, faxineiro, lavrador, garrafeiro, pedreiro, garçom de bar ou botequim, lavrador ou agricultor (assalariado), meeiro, caixeiro de armazém ou de outro pequeno estabelecimento comercial varejista (quitanda, mercearia, peixaria, lanchonete, lojas de ferragens), e outras ocupações com características semelhantes.

6. AGRUPAMENTO VI: Do lar.