SERGIO MARTINHAGO

# DESCOBERTA DE CONHECIMENTO SOBRE O PROCESSO SELETIVO DA UFPR

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciências, do Programa de Pós-Graduação em Métodos Numéricos em Engenharia, na Área de Concentração em Programação Matemática, Departamento de Matemática, Setor de Ciências Exatas e Departamento de Construção Civil, Setor de Tecnologia da Universidade Federal do Paraná.

Orientador: Profº Drº Celso Carnieri

Curitiba
2005

**TERMO DE APROVAÇÃO**

SERGIO MARTINHAGO

# DESCOBERTA DE CONHECIMENTO SOBRE O PROCESSO SELETIVO DA UFPR

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciências, do Programa de Pós-Graduação em Métodos Numéricos em Engenharia, na Área de Concentração em Programação Matemática, Departamento de Matemática, Setor de Ciências Exatas e Departamento de Construção Civil, Setor de Tecnologia da Universidade Federal do Paraná:

Profº Drº Celso Carnieri
Departamento de Matemática, UFPR
(Orientador)

Profª Drª Maria Madalena Dias
Departamento de Informática, UEM

Profº Drº Júlio César Nievola
Departamento de Ciência da Computação, PUC-PR

Profª Drª Maria Teresinha Arns Steiner
Departamento de Matemática, UFPR

Curitiba, 6 de julho de 2005

Ao meu amor, Margareth, luz da minha vida. Companheira em todos os momentos.

Aos meus filhos, Jorge Miguel e Mateus e, a Naiara, que souberam compreender com maturidade as vezes que não pude lhes dar atenção.

# AGRADECIMENTOS

Aos meus pais pela vida.

À professora Maria Madalena Dias, da Universidade Estadual de Maringá (UEM) pela dedicação na co-orientação deste trabalho, sem medir esforços, deixando seus compromissos para me auxiliar.

Ao professor Celso Carnieri, meu orientador, por acreditar na minha capacidade e pela paciência.

A todos os professores que participaram deste programa de mestrado.

Aos meus colegas de curso, com carinho, sem distinção.

Em especial ao companheiro de curso Douglas, pela dedicação e pelos valiosos conhecimentos transmitidos a todos os colegas que participaram deste mestrado.

A professora Débora Ribeiro de Carvalho, da Universidade Tuiuti do Paraná, que na hora do desespero me estendeu as mãos com seus ensinamentos e orientações.

A FECILCAM, em especial ao Departamento de Matemática, pela iniciativa de viabilizar esse convênio com a UFPR, possibilitando a oportunidade de cursar esse mestrado.

A UFPR, em especial ao Núcleo de Concursos (NC) que gentilmente disponibilizou a base de dados necessária para este trabalho e os resultados das freqüências de ocorrências.

Ao Marcus do NC que em todos os momentos que precisei sempre esteve à disposição para tirar as dúvidas sobre a base de dados.

Aos colegas professores do Colégio Estadual Marechal Rondon - EFM pela torcida e incentivo.

À Direção do Colégio Estadual Marechal Rondon Pela colaboração na elaboração de um horário adequado, para que eu pudesse fazer este mestrado.

Enfim, a todas as pessoas, que de uma maneira ou de outra, contribuíram para o sucesso deste trabalho.

# SUMÁRIO

## LISTA DE FIGURAS

## LISTA DE TABELAS

# RESUMO

A informação vem desempenhando um papel fundamental no desenvolvimento e sucesso das grandes organizações. Os sistemas de suporte à decisão, cada vez mais presentes na realidade dessas organizações, tornam mais confiáveis as tarefas de coletar, tratar, interpretar e utilizar informações, resultando em um processo eficaz. 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 serem de volume menor, geralmente elas não estão disponíveis e exigem que sejam extraídas a partir de grandes quantidades 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 grandes bases de dados. Mineração de dados (ou *Data Mining*), refere-se a uma determinada etapa deste processo e consiste na aplicação de algoritmos e técnicas estatísticas e de aprendizagem de máquina em grandes bases de dados para encontrar tendências ou padrões em dados que possam dar suporte à tomada de decisões. Este trabalho apresenta uma aplicação prática do processo de *KDD* na base de dados sobre os candidatos ao processo seletivo do vestibular ocorrido em dezembro de 2003 da UFPR. Neste trabalho, utilizando-se de uma ferramenta chamada *WEKA (Waikato Environment for Knowledge Analysis)*, foi aplicada a técnica de mineração de dados Árvore de Decisão, através dos algoritmos de classificação *J48.J48* e *J48.PART*. Os resultados obtidos poderão ser usados para traçar perfis dos candidatos ao processo seletivo do vestibular da UFPR, 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, Árvores de Decisão e Sistema de Suporte à Tomada de decisão.

# ABSTRACT

Information has been having a fundamental role on companies' growth, development and success. The making-decision supporting systems, available at these companies, make the work of collecting, treating, analyzing and using information more efficient, improving the whole process. 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 extraction of knowledge from a huge database amounts. Data Mining refers to a specific phase of this process and consists in the application of algorithms and statistical techniques and learning of machines in big databases to find tendencies or standardized patterns in data that can give support to making-decision process. This study demonstrates a practical application of KDD Process to the database of 2003 UFPR's Entrance Examination or Selective Process. Coherently to WEKA research tool – Waikato Environment for Knowledge Analysis, the Decision Tree Data research technique was applied, through the algorithms classification J48.J48 and J48.PART. The results can be used to build up the candidates profile, in order to extract and point out important information that offer support to the educational institutions on the making-decision process.

Keywords: Knowledge Discovery in Databases, Data Mining, Decision Tree and Making-decision Support.

# 1 INTRODUÇÃO

## 1.1 Tema e problema de pesquisa

O Brasil vive uma tendência, que hoje é mundial, de valorização da formação e da educação formal pela sociedade e pelo mundo do trabalho. Essa 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 a 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 por 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), revelam que apenas 9,7% dos cerca de 19,6 milhões de jovens entre 18 e 24 anos chegam à universidade e, somente 1,3% nessa faixa etária concluem uma faculdade.

A forma de ingresso no ensino superior brasileiro se efetiva através da seleção dos candidatos num exame classificatório. O exame para o ingresso neste nível de ensino tornou-se obrigatório em 1911. A lei exigia o exame de admissão, a difusão dos critérios das provas, a existência de bancas, do calendário e das taxas de inscrição.

Em 1915, de acordo com o Decreto 11.530, as provas passaram a se chamar “vestibular”. Na época, as escolas realizavam os testes em duas etapas. A primeira etapa constituía-se de prova escrita e dissertativa, a segunda, era oral. Essa forma de seleção foi utilizada até meados dos anos 60, quando surgiram as questões de múltipla escolha. Processados em computadores, os testes facilitavam a correção, cada vez mais complexa pelo volume crescente de candidatos. Porém, o critério de nota mínima aprovava candidatos acima do limite de vagas que eram destinadas aos primeiros colocados. O restante aguardava a expansão de ofertas.

Em 1968, através da Lei 5.540, o governo instituiu o sistema classificatório, com corte por notas máximas. A Lei de Diretrizes e Bases da Educação Nacional aprovada em 1996 permite que cada instituição de educação universitária opte por critérios próprios. A tendência atual é valorizar a prova dissertativa. No novo milênio, com quase 3 milhões de inscritos ao vestibular por ano, o mesmo recupera traços do modelo do início do século passado.

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 pós-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 e 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

### 1.2.1 Objetivo geral

Esta pesquisa tem por objetivo geral delinear o perfil do candidato ao processo seletivo de admissão para o ensino superior da Universidade Federal do Paraná, *campus* Curitiba, através da aplicação da técnica de mineração de dados Árvore de Decisão, utilizando os algoritmos de classificação *J48.J48* e *J48.PART* implementados na ferramenta *WEKA*. Foi utilizada a base de dados do vestibular realizado em dezembro de 2003 para o ano letivo de 2004, que contém os dados coletados no questionário sócio-educacional preenchido pelos candidatos no momento da inscrição, os dados do cadastro geral dos candidatos contendo o registro das notas obtidas pelo candidato nas provas e na redação, a opção pelo ENEM, a nota do ENEM, a média das notas do candidato e o status (resultado do vestibular).

### 1.2.2 Objetivos específicos

- Identificar e selecionar o banco de dados dos alunos inscritos ao concurso vestibular que serão utilizados pelo sistema de mineração de dados.

- Apresentar um modelo de mineração de dados para ser aplicado a uma instituição de ensino superior, mais especificamente, na base de dados dos inscritos no vestibular, através de suas fichas sócio-econômicas e culturais.

- Aplicar a estrutura de solução ao caso da Universidade Federal do Paraná, *campus* Curitiba.

- Elaborar um instrumento de pesquisa capaz de fornecer os dados necessários ao delineamento do perfil dos inscritos no vestibular.

- Caracterizar as diferenças sócio-econômicas e culturais existentes entre os candidatos.

- Estudar e aplicar técnicas/algoritmos de mineração de dados para descobrir padrões de comportamento do vestibulando.

- Analisar e identificar características específicas dos alunos da instituição, decorrentes da aplicação das técnicas e ferramentas de mineração de dados, procurando aumentar o conhecimento sobre os mesmos.

## 1.3 Justificativa

Os dados do Instituto Brasileiro de Geografia e Estatística (IBGE), demonstram que as instituições de educação superior do Brasil têm 3,9 milhões de estudantes em curso de graduação. Os referidos dados são do Censo da Educação Superior[1] realizado no ano de 2003. O levantamento coletou informações de 1859 instituições públicas e privadas, que tinham, pelo menos, um curso com data de início de funcionamento até 30 de outubro de 2003[2].

De acordo com o Inep/MEC, houve um aumento de 11,7% no número de matrículas em relação ao ano de 2002, sendo que no setor privado, que conta com 2.750.652 estudantes, o crescimento foi de 13,3%, e no setor público, de 8,1%.

O levantamento revela que o número de instituições da educação superior, registradas até 2003 é de 1859. Destas instituições registradas, 207 são públicas, representando 11,1% e 1652 são privadas, perfazendo 88,9%. O crescimento registrado em relação a 2002 é de 13,6%. No setor privado, o aumento foi de 14,6% e, no público, de 6,2%.

O Inep/MEC publicou que o número de cursos de graduação registrado é de 16.453, com aumento de 14,3% em relação a 2002. Nas instituições privadas, o crescimento

[1] O Censo faz parte do Sistema Nacional de Avaliação da Educação Superior (Sinaes).

[2] Os dados do Censo da Educação Superior foram divulgados pelo Ministério da Educação em 13 de outubro de 2004 e estão disponíveis na Internet no endereço HYPERLINK “http://www.inep.gov.br/

foi de 18% e nas públicas, de 7,8%. Do total de cursos existentes no Brasil, 10.791 (65,6%) estão no setor privado e 5.662 (34,4%) em instituições públicas.

Segundo o Inep/MEC, os dados revelam que, pela primeira vez, o número de vagas oferecidos na educação superior foi maior que o número de alunos concluintes do ensino médio. Apesar disso, a ociosidade do sistema alcançou 42,2% das vagas oferecidas pelas instituições privadas, e nas públicas 5,1%.

Diante dos dados sobre a 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)[3]. O Programa foi lançado no dia 13 de abril de 2004 e permitirá que, em cinco anos, 300 mil estudantes de baixa renda[4] e professores públicos sem formação superior ingressem na universidade.

A UFPR instituiu em 2003, o Processo de Ocupação de Vagas Remanescentes (PROVAR), com o objetivo de ocupar as vagas ociosas da instituição. O processo de ocupação das vagas ociosas promoveu uma discussão acadêmica acerca do projeto pedagógico e o perfil dos alunos de cada curso.

O Ministério da Educação, através do informativo de novembro de 2004, admite que, apesar do aumento do número de matrículas registradas no ensino superior, apenas 9% dos jovens brasileiros de 18 a 24 anos estão na universidade.

---

[3] O ingresso das instituições privadas de ensino superior no ProUni será formalizado mediante termo de adesão com o MEC. As instituições que aderirem ao programa ficarão isentas do pagamento do Imposto de Renda de Pessoa Jurídica, da Contribuição Social sobre Lucro Líquido, do Programa de Integração Social (PIS) e da Contribuição para o Financiamento da Seguridade Social (Cofins). Em contrapartida, deverão oferecer 10% de suas vagas em bolsas de estudo. No caso das filantrópicas, os 20% de gratuidade que já são exigidos por lei deverão ser concedidos exclusivamente por meio de bolsas de estudo, e não mais com outros tipos de atendimento, de difícil controle contábil. (BRASIL, 2004a, p. 10).

[4] Os critérios de seleção dos candidatos serão os resultados do Exame Nacional do Ensino Médio (ENEM) e o perfil socioeconômico dos candidatos. As bolsas integrais são para estudantes com renda familiar *per capita* de até um e meio salário mínimo, e as parciais, de 50%, para aqueles com renda familiar de até três salários mínimos por pessoa. Os professores da rede pública de ensino básico, sem curso superior, poderão participar do programa nos cursos de Licenciatura e Pedagogia, independente da renda familiar. (BRASIL, 2004b, p. 5).

De acordo com o professor Otaviano Helene[5] (2004), o Brasil engaja no ensino superior 15% dos seus jovens[6]. Esse autor considera baixo o percentual de alunos matriculados nesse nível de ensino e compara com dados de outros países. Segundo Helene (2004), essa taxa nos países vizinhos supera a brasileira. A Argentina possui 48%; o Chile 38%; o Uruguai 36%; Bolívia 36% e o Paraguai 17% de jovens matriculados no ensino superior. O referido autor afirma que se a comparação for com países mais desenvolvidos, o Brasil fica numa situação mais inferior ainda, como por exemplo, se for comparado com os Estados Unidos (75%), Canadá (59%) e Coréia (78%).

Considerando esses números, o Brasil tem um percentual muito reduzido de alunos matriculados no ensino superior e, na concepção de Helene (2004), o Plano Nacional de Educação, elaborado pelas entidades e associações ligadas a educação brasileira, apresentado à Assembléia Legislativa Nacional em 1998, tinha por objetivo triplicar em dez anos o número de alunos no ensino superior brasileiro. Porém, o Plano sofreu vetos no governo Fernando Henrique Cardoso e, de acordo com esse autor, "o Plano Nacional de Educação, na prática, não entrará em vigor, pois sobraram, após os vetos, apenas algumas declarações de princípios gerais e genéricos. Assim, se nada for mudado, a chance de se superar essa situação atual está praticamente descartada" (HELENE, 2004, p. 111).

Para a professora Wrana Panizzi (2004), não é apenas o ingresso numa instituição de ensino superior que merece maior atenção dos órgãos governamentais. A permanência do jovem que ingressou numa instituição de ensino superior também precisa ser acompanhada.

---

[5] Otaviano Helene é Professor Titular do Instituto de Física da USP. Ex-presidente e ex-vice presidente da Adusp. Ex-presidente do INEP (jan. a jul. de 2003). Colaborou na preparação do Plano Nacional de Educação e do Plano estadual de Educação de São Paulo.

[6] Otaviano Helene esclarece que, a taxa de engajamento que a UNESCO e outros órgãos usam para calcular qual o percentual de jovens que já se matricularam no ensino superior "é dividir o número de matrículas total de nível de ensino independente da idade, pela população num corte etário correspondente de 5 anos" (HELENE, 2004, p. 110)

O debate acerca dos problemas enfrentados pela educação superior no Brasil é acirrado e vem de longa data. Embora o tema seja pertinente e de extrema importância, este estudo não tem a intenção de abordar todas as discussões referentes aos problemas estruturais do sistema educacional brasileiro.

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 o processo de seleção do vestibular da UFPR e seus candidatos e apresentar a relação existente entre as variáveis sócio-econômicas com o desempenho dos candidatos nas provas do vestibular de dezembro de 2003, e também, servir de base para outros trabalhos da área.

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

### 1.4 Limitações do problema

As limitações para o desenvolvimento dessa pesquisa estão situadas na base de dados. A base de dados fornecida a respeito do vestibular não contemplava informações sobre a efetivação da matrícula do aluno aprovado no vestibular. Este fato se constituiu numa limitação porque após a aplicação das técnicas, 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.

Outra limitação para esta pesquisa foi o fato de serem corrigidas apenas as redações dos candidatos classificados de acordo com os critérios do guia do candidato ao processo seletivo 2004[7]. Dos 46.531 candidatos inscritos, apenas 12.228 redações foram corrigidas. Este número reduzido de redações corrigidas diminui a precisão das regras com um bom fator de confiança.

[7] O guia do candidato ao Processo Seletivo de 2004 está disponível no site: www.nc.ufpr.br

## 1.5 Estrutura do trabalho

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

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

No capítulo II está a revisão da literatura de descoberta de conhecimento e mineração de dados, a origem dos dados utilizados na pesquisa, as etapas do processo de *KDD*, o conceito de mineração de dados, as tarefas de mineração de dados, as técnicas de mineração de dados, as ferramentas de mineração de dados, a apresentação do software *WEKA*, e algumas aplicações da mineração de dados.

No Capítulo III é apresentado um breve histórico da UFPR, além disso, são descritas as etapas seguidas no processo de descoberta de conhecimento no banco de dados do processo seletivo do vestibular de 2004.

No Capítulo IV são apresentados os testes realizados e os resultados da aplicação das técnicas de mineração de dados.

Finalmente, no Capítulo V, é apresentada as conclusões e as sugestões de trabalhos futuros.

# 2 REVISÃO DA LITERATURA DE KDD E DATA MINING

As duas últimas décadas têm demonstrado um crescente aumento no número de dados armazenados em meio eletrônico e, em especial, os que as organizações, em suas operações diárias, geram e coletam. Porém, este grande volume de dados não é aproveitado plenamente porque as informações úteis estão, geralmente, implícitas e são de difícil acesso e compreensão pelos tomadores de decisão.

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 ferramentas de análise de dados automatizadas. Neste contexto, está o processo de descoberta de conhecimento em banco de dados (*Knowledge Discovery in Databases* – *KDD*), no qual Mineração de Dados (*Data Mining*) é a principal etapa.

A descoberta de conhecimento em banco de dados 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 de sistemas *KDD* e mineração de dados (MD); são relacionadas as etapas do processo *KDD*, as principais tarefas e técnicas de MD, ferramentas de MD e exemplos de aplicações de técnicas 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 procurar 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 MD refere-se a uma das etapas deste processo. A relação existente entre KDD e MD pode ser visualizada graficamente através da Figura 2.1.

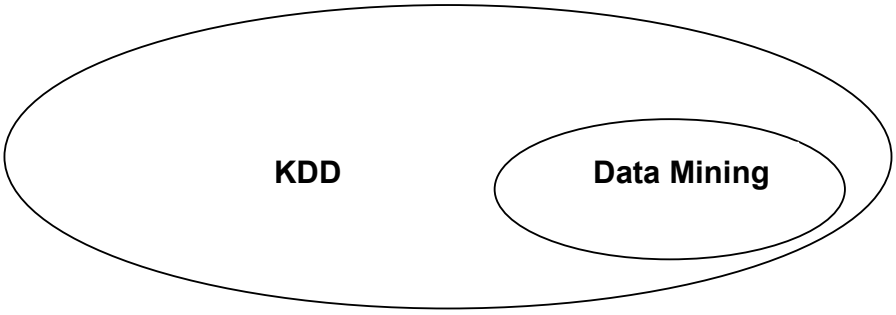


Figura 2.1: interligação entre *KDD* e *Data Mining*
Fonte: Carvalho (2002).

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 em 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 KDD é um processo iterativo porque o conhecimento descoberto apresentado ao usuário pode ser usado como base para a medida de avaliação a ser aprimorada, a mineração ser refinada, novos dados serem selecionados ou transformados, ou ainda, novas fontes de dados serem integradas para adquirir resultados diferentes e mais apropriados.

De acordo com Carvalho (1999) *apud* Adriaans e Zantinge (1996), a Descoberta de Conhecimento é interdisciplinar e envolve diversas áreas, entre elas, estatística e matemática, banco de dados, aprendizado de máquina, sistemas especialistas e reconhecimento de padrões. O processo *KDD* combina técnicas, algoritmos e definições de todas estas áreas com o objetivo principal de extrair conhecimento a partir de grandes bases de dados.

A Figura 2.2 demonstra a interdisciplinaridade das áreas envolvidas na descoberta do conhecimento.

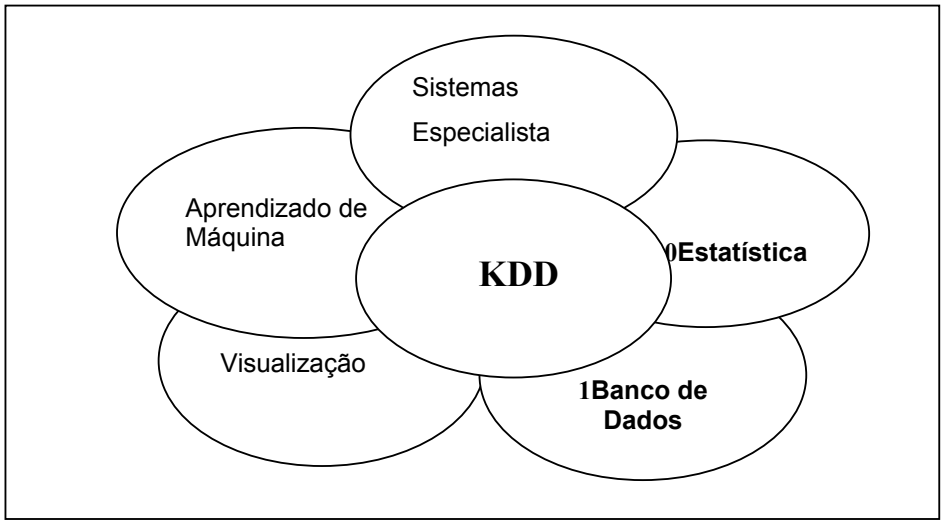


Figura 2.2: Áreas interdisciplinares do processo *KDD*.
Fonte: Adriaans e Zantinge (1996).

Na interpretação da Figura 2.2, Carvalho (1999) argumenta que na área do aprendizado de máquina são utilizados modelos cognitivos ou estratégicos de aprendizado de máquina, bem como os paradigmas para a aquisição automática de conhecimento. De acordo com a autora, na área de banco de dados existem

tecnologias específicas, bem como uma série de pesquisas que objetivam melhor explorar as características dos dados a serem trabalhados.

Os modelos matemáticos ou estatísticos são construídos para a geração de regras, padrões e regularidades. No caso específico da estatística, é disponibilizado um grande número de procedimentos técnicos e resultados de testes para as tarefas de MD*,* como, por exemplo, para verificar se estimativas e procedimentos de pesquisa estão consistentes sob determinados critérios de avaliação e identificar o grau de incerteza (CARVALHO, 1999).

Para Carvalho (1999), os sistemas especialistas são programas de Inteligência Artificial[8] criados para resolver problemas do mundo real. Inicialmente, estes sistemas ofereciam apenas mecanismos para a representação do conhecimento, raciocínio e explicações. Posteriormente foram incorporadas ferramentas para a aquisição do conhecimento.

Completando a interpretação da Figura 2.2, Carvalho (1999), afirma que a visualização de dados assume um papel importante já que em vários momentos existe a necessidade de interação entre o processo de descoberta e o ser humano. Pode-se citar como exemplo, a análise prévia dos dados que vão ou não fazer parte do processo, onde são realizadas algumas consultas usando ferramentas de análise ou mesmo de visualização de dados. Para a visualização, pode-se recorrer a distintas formas, tais como: gráficos ícones e figuras.

## 2.2 Origem dos Dados

As técnicas de mineração de dados, de acordo com Dias (2001), podem ser aplicadas “sobre bancos de dados operacionais ou sobre *Data Warehouse* (*DW*) ou *Data Mart*, nos quais geralmente resulta uma informação melhor, pois os dados

[8] Segundo Levine *et al*. (1992), a *Inteligência Artificial* (IA) é simplesmente uma maneira de fazer o computador pensar inteligentemente. Isto é conseguido estudando como as pessoas pensam quando estão tentando tomar decisões e resolver problemas, dividindo esses processos de pensamento em etapas básicas e desenhando um programa de computador que solucione problemas usando essas mesmas etapas. IA, então, fornece método simples e estruturado de se projetar programas complexos de tomada de decisão.

normalmente são preparados antes de serem armazenados no *DW* ou *data mart"* (DIAS, 2001, p. 8). Para a autora, as técnicas de MD podem ser aplicadas, também, "sobre um *Data Set*, que pode ser definido como um 'banco de dados' (em um sentido fraco do termo) contendo apenas o conjunto de dados específicos para um tipo de investigação a ser realizada" (DIAS, 2001, p. 8).

Uma definição sobre *DW* é dada por Inmon (1997, p.33), considerado um "guru" no assunto, "... *DW* é um conjunto de dados baseado em assuntos, integrado, não-volátil e variante em relação ao tempo, de apoio às decisões gerenciais".

Conforme Dias (2001), no princípio, a expressão representava simplesmente um armazém de dados, como é a tradução de *DW*; porém, ao longo do tempo, vem recebendo diversos incrementos em sua estrutura.

Argumenta ainda Dias (2001, p. 8) que um *DW* tem por objetivo oferecer organização, gerenciamento e integração de bancos de dados, assim como ferramentas de exploração dos mesmos, para se obter vantagens competitivas no mercado. De acordo com a autora, o *DW* é construído "tendo como base outros bancos de dados operacionais que podem estar implementados em diferentes plataformas na organização. É usado, geralmente, em aplicações de suporte à tomada de decisão" (DIAS, 2001, p. 8). *Data mart* é um *DW* departamental, ou seja, um *DW* construído para uma área específica da organização (DIAS, 2001 *apud* INMON, 1997), facilitando a tomada de decisões em nível departamental e permitindo dados relacionais ou multidimensionais não voláteis.

### 2.3 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, conforme ilustra a Figura 2.3.

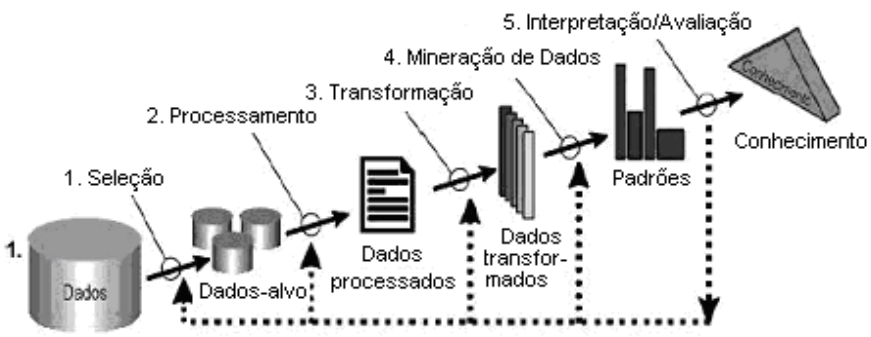


Figura 2.3: Processo *KDD*
Fonte: Traduzido de Fayyad *et al*.(1996)

O primeiro passo no processo de *KDD* é entender o domínio da aplicação, identificar o problema e definir os objetivos a serem atingidos. O processo inicia com os dados brutos e finaliza com a extração de conhecimento.

De acordo com Dilly (1995), a seleção dos dados é a extração dos dados visando a aplicação. Nesta etapa pode ser necessário integrar e compatibilizar as bases de dados.

Na atividade limpeza de dados, da etapa de pré-processamento, as informações consideradas desnecessárias são removidas. Adotam-se estratégias para manusear dados faltantes ou inconsistentes (DILLY, 1995; GONÇALVES, 2000). Se os erros não forem descobertos neste estágio, poderão contribuir para a obtenção de resultados de baixa qualidade (LUBEL, 1998).

A transformação dos dados consiste em desenvolver um modelo sólido de dados de maneira que possam ser utilizados por um algoritmo de extração de conhecimento. As transformações são ditadas pela operação e técnica a ser adotada. São conversões de um tipo de dados para outro, definição de novos atributos, etc. (GONÇALVES, 2000; IBM, 1997).

A mineração de dados é o núcleo do processo. Aplicam-se algoritmos para extrair padrões dos dados ou gerar regras que descrevam o comportamento da base de dados (BERRY e LINOFF, 1997; DILLY, 1995). Para isto, utiliza-se uma ou mais técnicas para se extrair o tipo de informação desejada. Durante esse procedimento,

pode ser necessário acessar dados adicionais e/ou executar outras transformações nos dados originalmente selecionados (IBM, 1997).

A interpretação e a avaliação dos resultados consiste em validar o conhecimento extraído da base de dados, identificar padrões e interpretá-los, transformando-os em conhecimentos que possam apoiar as decisões (DILLY, 1995). O objetivo de interpretar os resultados é filtrar as informações que serão apresentadas aos tomadores de decisão.

Se os resultados não forem satisfatórios, faz-se necessário repetir a etapa de MD ou retomar a qualquer um dos estágios anteriores. Somente após a avaliação e validação dos resultados é que se encontra conhecimento.

## 2.4 Conceito de Mineração de dados

Como já foi visto, *KDD* refere-se ao processo completo de descoberta de conhecimento, enquanto que a Mineração de dados é uma de suas etapas voltada a aplicar algoritmos específicos e a produzir padrões sobre uma base de dados (FAYYAD *et.al*.,1996).

Mineração de dados, segundo Han e Kamber (2001), é um campo multidisciplinar que inclui as seguintes áreas: banco de dados, inteligência artificial, aprendizado de máquina, redes neurais, estatísticas, reconhecimento de padrões, sistemas baseados em conhecimento, aquisição de conhecimento, recuperação de informação, computação de alto desempenho e visualização de dados.

De acordo com a DWBrasil (2004), a MD descende fundamentalmente de três linhagens. A mais antiga delas é a estatística clássica. A estatística clássica envolve conceitos como distribuição normal, variância, análise de regressão, desvio simples, análise de conjuntos, análise de discriminantes e intervalos de confiança, todos usados para estudar os dados e os relacionamentos entre eles.

A segunda linhagem da MD é a Inteligência Artificial. Essa área, que é construída a partir dos fundamentos de procedimentos heurísticos, em oposição à estatística, tenta imitar a maneira como o homem pensa na resolução dos problemas estatísticos.

E a terceira e última linhagem de MD é a chamada aprendizagem de máquina (*machine learning*), que pode ser descrita como o "casamento" entre a estatística e a Inteligência Artificial. A aprendizagem de máquina tenta fazer com que os programas de computador "aprendam" com os dados que eles estudam, de tal modo que esses programas tomem decisões diferentes, baseadas nas características dos dados estudados, usando a estatística para os conceitos fundamentais e adicionando heurística avançada da Inteligência Artificial e algoritmos para alcançar os seus objetivos.

De muitas formas, salienta DWBrasil (2004), a MD é fundamentalmente a adaptação das técnicas da aprendizagem de máquina para as aplicações de negócios. Desse modo, pode-se descrevê-lo como a união dos históricos e dos recentes desenvolvimentos em estatística, em Inteligência Artificial *e* em *machine learning*. Essas técnicas são usadas juntas para estudar os dados e achar tendências e padrões nos mesmos. Hoje, a MD tem experimentado uma crescente aceitação nas ciências e nos negócios que precisam analisar grandes volumes de dados e achar tendências que eles não poderiam achar de outra forma.

A Statistical Package for the Social Sciences - SPSS[9] (2004), aponta ainda que, usando técnicas estatísticas em MD, é possível influenciar significativamente todas as áreas de uma organização. Na arena empresarial de hoje, é um desafio constante manter o ritmo das tendências de mercado e predizer resultados futuros. Técnicas estatísticas ajudam a reagir rapidamente às mudanças de mercado, tornando a organização mais produtiva, mais competitiva, além de auxiliar a tomar decisões baseadas em fatos.

[9] SPSS é uma empresa de software. No contexto deste trabalho é uma ferramenta de técnicas estatísticas para MD disponível no site: http://www.spss.com/datamine/index.htm.

De todas as várias técnicas em análise de dados tradicional, Cabena *et al.* (1998) citam que as técnicas estatísticas são mais íntimas às técnicas de MD. Segundo estes autores, as técnicas estatísticas foram tradicionalmente usadas para muitas das análises que são agora feitas com MD, como construir modelos preditivos ou descobrir associações em bancos de dados. Afirmam também que para cada uma das áreas principais do esforço de MD há, de um modo geral, uma abordagem estatística equivalente e é provavelmente verdade que muito, se não tudo, do que é feito em MD poderia ser feito, eventualmente, com análises estatísticas. Entretanto, Cabena *et al.* (1998), complementam que o que está atraindo muitos analistas para as técnicas de MD é a facilidade relativa com que podem ser ganhas visões novas (embora não necessariamente condizentes) em comparação às abordagens estatísticas tradicionais.

A acessibilidade da mineração de dados é mostrada em vários caminhos. Por exemplo, mineração de dados é, geralmente, uma abordagem livre de hipóteses, enquanto que técnicas estatísticas mais populares requerem o desenvolvimento de uma hipótese com antecedência, os estatísticos têm que desenvolver manualmente as equações que se adaptam às hipóteses. Em contraste, algoritmos de mineração de dados podem desenvolver estas equações automaticamente (CABENA *et al.*, 1998). Para esses autores, as análises estatísticas representam uma função importante na maioria dos ambientes de mineração de dados e, embora os fatores distintivos pareçam favorecer a área de MD ao invés de técnicas estatísticas tradicionais, a melhor estratégia é sempre usar análises estatísticas e mineração de dados como abordagens complementares.

A tecnologia de MD tem grande potencial para auxiliar as organizações na extração de informações importantes provenientes dos seus bancos de dados, predizendo padrões e comportamentos futuros, respondendo a questões que tomariam muito tempo para serem resolvidas, o que possibilita as melhores decisões de negócio apoiadas em conhecimento. Para Lubel (1998), MD é um recurso em ascensão que se tornará obrigatório aos mercados competitivos.

Na prática, os objetivos de MD são a predição ou a descrição. A predição envolve a utilização de algumas variáveis (atributos) da base de dados para predizer valores

desconhecidos ou futuros de outras variáveis de interesse. A descrição procura por padrões que descrevem os dados interpretáveis pelos seres humanos (FAYYAD *et al.*, 1996).

Segundo Dias (2001), a MD é a exploração e análise, por meios automáticos ou semiautomáticos, de grandes quantidades de dados para descobrir padrões interessantes e ou regras. Envolvem a transformação dos dados em informação, a informação em ação e a ação em valor. Sendo assim, MD é o processo que permite "descobrir correlações significantes, padrões e tendências, através de filtragem de grandes quantidades de dados, pelo uso de tecnologias de reconhecimento desses padrões, bem como de técnicas estatísticas e matemáticas" (DIAS, 2001, p. 5).

### 2.4.1 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[10] de *KDD* que são, principalmente, dependentes do domínio da aplicação e do interesse do usuário. Cada tarefa de *KDD* extrai um tipo diferente de conhecimento do banco de dados e pode requerer um algoritmo diferente para cada tarefa.

Como já foi visto, MD dispõe de tarefas básicas classificadas nas categorias descritivas e preditivas. A seguir as seguintes tarefas de MD são descritas: classificação, associação, segmentação (ou *clustering*), estimativa (ou regressão) e sumarização.

#### 1) Classificação

A classificação é uma das tarefas mais estudadas pela comunidade científica. Nessa tarefa cada tupla (registro), pertence a uma classe entre um conjunto pré-definido de classes. A classe de uma tupla ou registro é indicado por um valor especificado pelo

[10] No contexto desta pesquisa conceituamos tarefa como sendo um problema de descoberta de conhecimento a ser solucionado.

usuário em um atributo[11] meta, ou atributo objetivo. As tuplas consistem de atributos preditivos e um atributo objetivo, esse último indicando a que classe essa tupla pertence. O atributo objetivo é do tipo categórico, ou discreto, isto é, pode apresentar apenas um valor dentro de um conjunto de valores discretos, determinando classes ou categorias. Esse atributo pode ter valores discretos como SIM ou NÃO, um código pertencente a um intervalo de números inteiros, tais como {1..10}, etc.

De acordo com Freitas (1998), o princípio da tarefa de classificação é descobrir algum tipo de relacionamento entre os atributos preditivos e o atributo objetivo, de modo a descobrir um conhecimento que possa ser utilizado para prever a classe de uma tupla desconhecida, ou seja, que ainda não possui uma classe definida. Por exemplo, suponha que uma editora de livros publicou um livro chamado “Um guia para restaurantes franceses na Inglaterra”. O livro é publicado em inglês, francês e alemão, de acordo com o país onde ele está sendo vendido. Suponha também que a editora tem um banco de dados contendo dados sobre seus clientes nos três países, Inglaterra, França e Alemanha. Seria interessante utilizar esses dados para prever que tipos de clientes estariam mais interessados em comprar esse novo livro. A editora pode então concentrar os esforços de vendas nesses clientes.

Para prever se o cliente irá ou não comprar o livro quando eles receberem um material de propaganda, a editora necessita de alguns dados sobre o efeito dessa técnica de propaganda em alguns de seus clientes na sua base de dados. A partir desses dados, um algoritmo de classificação pode descobrir regras que prevêem se um novo cliente irá ou não comprar esse novo livro. Para coletar esses dados a editora pode enviar o material de propaganda para alguns de seus clientes e monitorá-los para saber se eles compram ou não o livro. Essa informação é então armazenada em um novo atributo, nesse caso o atributo objetivo. Seu valor pode assumir dois possíveis valores: SIM, significando a compra do livro, ou NÃO, caso contrário. Uma vez esse atributo determinado, o próximo passo é selecionar um subconjunto de atributos preditivos entre todos os atributos dos clientes no banco de dados. Claramente alguns atributos, tal como: nome do cliente, é de modo geral

[11] Atributos podem ser classificados em discretos ou contínuos. Atributos discretos são aqueles em que os valores assumidos podem ser previstos e representados em um conjunto de possíveis valores, como *sexo*, que pode assumir somente dois valores possíveis. Atributos contínuos são aqueles em que o conjunto de valores é infinito como, por exemplo, *salário_anual.*

irrelevante para a previsão da compra ou não do livro. No exemplo abaixo, contido na tabela 2.1, serão considerados apenas os atributos SEXO, PAÍS e IDADE dos clientes como relevantes para a previsão.

Nesta Tabela 2.1, Freitas (1998), mostra os valores dos atributos preditivos selecionados, junto com valor do atributo objetivo, COMPRAR. Esses são dados de dez clientes, aos quais algum material de propaganda foi enviado sobre o novo livro. Um algoritmo de classificação pode analisar os dados da Tabela 2.1 para determinar que valores dos atributos preditivos tendem a ser relacionados, com cada um dos atributos objetivos. Esta descoberta de conhecimento pode então ser aplicada para prever se um cliente da base de dados da editora comprará ou não o novo livro. Note que esse conhecimento será aplicado nos clientes para o qual o valor do atributo objetivo ainda é desconhecido.

| **SEXO** | **PAÍS** | **IDADE** | **COMPRAR** |
|---|---|---|---|
| Masculino | França | 25 | **Sim** |
| Masculino | Inglaterra | 21 | Sim |
| Feminino | França | 23 | Sim |
| Feminino | Inglaterra | 34 | Sim |
| Feminino | França | 30 | Não |
| Masculino | Alemanha | 21 | Não |
| Masculino | Alemanha | 20 | Não |
| Feminino | Alemanha | 18 | Não |
| Feminino | França | 34 | Não |
| Masculino | França | 55 | Não |

Tabela 2.1: Entrada de dados para a tarefa de classificação
Fonte: Freitas (1998)

O conhecimento descoberto é freqüentemente representado na forma de regras **SE-ENTÃO**. Essas regras são interpretadas da seguinte maneira: "**SE** os atributos preditivos de uma tupla satisfazem as condições no antecedente da regra, **ENTÃO** a tupla tem a classe indicada no conseqüente da regra". A Figura 2.4 mostra as regras extraídas através de um algoritmo de classificação utilizando os dados da Tabela 2.1.

Se (PAÍS=Alemanha) então COMPRAR = Não
Se (PAÍS=Inglaterra) então COMPRAR = Sim
Se (PAÍS=França e IDADE $\leq$ 25) então COMPRAR = Sim
Se (PAÍS= França e IDADE>25) então COMPRAR = Não

Figura 2.4: Regras de classificação
Fonte: Freitas (1998)

Mais informações sobre a tarefa de classificação podem ser encontradas em (MEHTA *et al*., 1996), (SHAFER *et al*., 1996), (WEISS *et al*., 1991) e (MICHALEWICZ, 1994).

**2) Associação**

A tarefa de associação ou afinidade de grupos visa combinar itens (diversos artigos) importantes, de tal forma que, a presença de um item em uma determinada transação (compra ou venda de itens) pressupõe a de outro na mesma transação. Isto foi inicialmente proposto por Agrawal *et al*., em 1993.

As aplicações de técnicas de associação têm seu uso mais difundido na área de *marketing*, em que se pretende descobrir as associações existentes entre os produtos vendidos. A tecnologia possibilitou às organizações coletar e armazenar grandes quantidades de dados, como é o caso da tecnologia de código de barras sobre os dados de vendas (AGRAWAL *et al*., 1993). As grandes redes varejistas estudam as compras dos clientes para descobrir quais as vendas são normalmente realizadas ao mesmo tempo, chamando isso de *market basket analysis*. Essa análise pode determinar, por exemplo, os produtos que devem estar expostos juntos, objetivando incrementar as vendas (BUSINESS OBJECTS, 1997).

**3) Segmentação**

A tarefa de MD denominada segmentação é um exemplo de aprendizado não supervisionado ou indireto, cujo objetivo é agrupar tipos similares de dados ou identificar exceções (GROTH, 1998). O sistema tem que descobrir suas próprias classes, isto é, agrupar os dados e descobrir subconjuntos de objetos relacionados

ao conjunto de treinamento, encontrando descrições de cada um destes subconjuntos (DILLY, 1995).

Um *cluster* pode ser definido como um conjunto de objetos agrupados pela similaridade ou proximidade e, a segmentação pode ser definida como a tarefa de segmentar uma população heterogênea em um número de subgrupos (ou *clusters*) mais homogêneos possíveis, de acordo com alguma medida (BERRY e LINOFF, 1997; DILLY, 1995). Quando o processo é bem sucedido, os objetos do *cluster* têm alta homogeneidade interna e alta heterogeneidade externa. Um exemplo disso é a geração de *clusters* de sintomas de pacientes, que podem indicar diferentes doenças baseadas nas suas características.

A Figura 2.5 mostra um exemplo de segmentação em que foram encontrados 4 *clusters.*

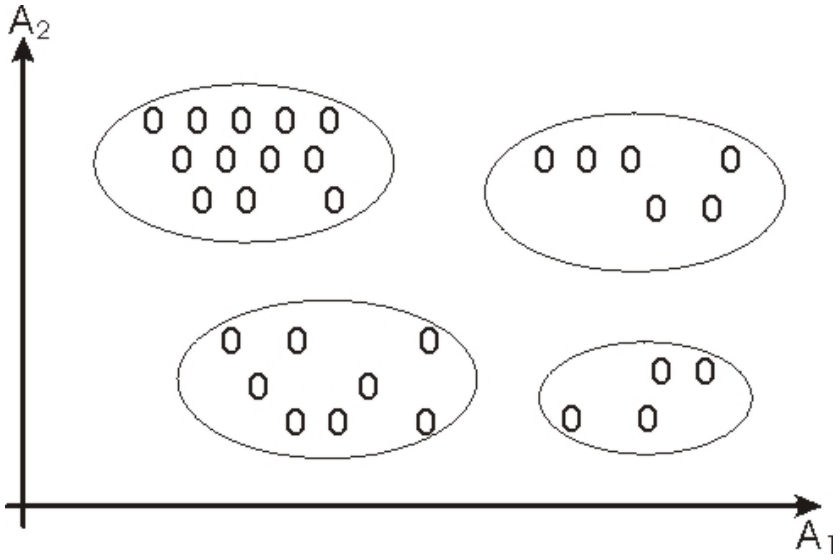


FIGURA 2.5: Exemplo de *Clustering*
Fonte: Carvalho (2002)

Na segmentação, diferentemente da classificação, não há classes pré-definidas. Na classificação, a população é subdividida e associa cada registro a uma classe pré-definida, com base no modelo desenvolvido através de treinamento e exemplos pré-classificados. A segmentação é mais geral e freqüentemente realizada como primeira etapa de outros métodos de MD ou de modelagem. Assim, aplica-se o modo direto para reconhecer relações nos dados e o indireto para explicar estas relações (BERRY e LINOFF, 1997).

A segmentação pode ser, por exemplo, aplicada em atividades de marketing para identificar os segmentos de mercado, para encontrar estruturas significativas nos dados e na descoberta de fraudes ou dados incorretos (GROTH, 1998).

**4) Estimativa (ou Regressão)**

A estimativa é usada para definir um valor para alguma variável contínua desconhecida como, por exemplo, receita, altura ou saldo de cartão de crédito (DIAS, 2001 *apud* HARRISON, 1998). Ela lida com resultados contínuos, enquanto que a classificação lida com resultados discretos. Ela pode ser usada para executar uma tarefa de classificação, convencionando-se que diferentes faixas (intervalos) de valores contínuos correspondem a diferentes classes. Para Fayyad (1996), a "regressão é aprender uma função que mapeia um item de dado para uma variável de predição real estimada" (FAYYAD *et al*., 1996, p. 13).

**5) Sumarização**

Segundo Fayyad *et al*. (1996), a tarefa de sumarização envolve métodos para encontrar uma descrição compacta para um subconjunto de dados. Um simples exemplo desta tarefa poderia ser tabular o significado e desvios padrão para todos os itens de dados. Métodos mais sofisticados envolvem a derivação de regras de sumarização.

As tarefas de mineração de dados, descritas acima, são apresentadas de forma resumida na Tabela 2.2.

| TAREFA | DESCRIÇÃO | EXEMPLOS |
|---|---|---|
| Classificação | Constrói um modelo de algum tipo que possa ser aplicado a dados não classificados a fim de categoriza-los em classes | • Classificar pedidos de crédito<br>• Esclarecer pedidos de seguros fraudulentos<br>• Identificar a melhor forma de tratamento de um paciente |
| Estimativa (ou Regressão) | Usada para definir um valor para alguma variável contínua desconhecida | • Estimar o número de filhos ou a renda total de uma família<br>• Estimar o valor em tempo de vida de um cliente<br>• Estimar a probabilidade de que um paciente morrerá baseando-se nos resultados de diagnósticos médicos<br>• Prever a demanda de um consumidor para um novo produto |
| Associação | Usada para determinar quais itens tendem a co-ocorrerem (serem adquiridos juntos) em uma mesma transação | • Determinar quais os produtos costumam ser colocados juntos em um carrinho de supermercado |
| Segmentação (ou *Clustering*) | Processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos | • Agrupar clientes por região do país<br>• Agrupar clientes com comportamento de compra similar<br>• Agrupar seções de usuários *Web* para prever comportamento futuro de usuário |
| Sumarização | Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados | • Tabular o significado e desvios padrão para todos os itens de dados<br>• Derivar regras de síntese |

Tabela 2.2: Tarefas Realizadas por Técnicas de Mineração de Dados
Fonte: Dias (2001)

### 2.4.2 Técnicas de Mineração de Dados

As técnicas de MD, de acordo com Rezende (2003), descrevem um paradigma de extração de conhecimento e vários algoritmos podem seguir esse 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 e a escolha é uma arte. 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.

MD é um campo que compreende, atualmente, muitas ramificações importantes. Cada tipo de tecnologia tem suas próprias vantagens e desvantagens, do mesmo

modo que uma mesma tecnologia não consegue atender todas as necessidades em todas as aplicações (LEMOS, 2003).

A seguir são descritas as técnicas de mineração de dados normalmente usadas.

**1) Redes Neurais Artificiais**

Segundo Dias (2001), as redes neurais utilizam um conjunto de elementos de processamento (ou nós) análogos aos neurônios no cérebro. Esses elementos de processamento são interconectados em uma rede que pode identificar padrões nos dados uma vez exposta aos mesmos, ou seja, a rede aprende através da experiência, tais como as pessoas.

Redes neurais 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 complicados ou imprecisos e podem ser utilizadas para extrair padrões e detectar tendências que são muito complexas para serem percebidas tanto por humanos quanto por outras técnicas computacionais (DWBRASIL, 2004).

De acordo com Silva (2003), as redes neurais podem ser do tipo supervisionada (as classes são conhecidas) e não supervisionada (as classes não são conhecidas). No primeiro tipo são algoritmos usados para construir modelos preditivos que podem capturar interações não lineares entre os atributos. As não supervisionadas são usadas para dividir em agrupamentos de acordo com certas regras pré-definidas.

A vantagem principal da utilização de Redes Neurais, conforme Adriaans e Zantinge (1996), é a versatilidade e o resultado satisfatório em áreas complexas com entradas incompletas ou imprecisas. Tem 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 e Zantinge (1996), afirmam 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 informação que justifique as conclusões obtidas. As redes neurais não podem provar uma teoria a partir do que aprenderam. Elas são simples “caixas pretas” que produzem respostas, mas não demonstram claramente o desenvolvimento de como chegaram aos resultados.

Exemplos de redes neurais: Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB (AZEVEDO, 2000), (BRAGA, 2000), (HAYKIN, 2001).

**2) Descoberta de 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 conjuntos de itens da base de dados e $X \cap Y = \varnothing$; *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 (DILLY,1995). O significado desta regra é que as transações da base que contêm *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; AGRAWAL *et al*., 1993).

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 os conjuntos 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 à significância estatística da regra (AGRAWAL *et al*., 1993).

A confiança é expressa pelo percentual de registros que satisfaz o antecedente (*X*) e o conseqüente (*Y*) em relação ao número de registros que satisfaz o antecedente, medindo a força da regra ou sua precisão (AGRAWAL *et al*., 1993). No exemplo anteriormente citado, 30% é o fator de confiança e 2% é o suporte da regra.

Berry e Linoff (1997) definem 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 a combinação ocorrer raramente.

Para Agrawal *et al.* (1997), o problema das regras de associação é encontrar todas as regras que possuem 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 algoritmos a serem utilizados na descoberta de regras de associação foram o *AIS* (AGRAWAL *et al*., 1993) e *SETM* (HOUTSMA e SWAMI, 1993). Porém, depois desta data, vários algoritmos foram criados. Um dos algoritmos, atualmente, mais referenciados para este método é o *Apriori* (AGRAWAL e SRIKANT, 1994), nas diversas variações, tais como, o *AprioriTid*, *DHP* e *Partition*.

**3) Árvore de Decisão**

Possui este nome porque a sua estrutura se assemelha a uma árvore. A sua estrutura é fácil de entender e de 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 de dados ou a predição de saídas. É conveniente usar árvore de decisão quando o objetivo for categorizar dados.

Na árvore, cada nó especifica um teste de algum 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 para um nível abaixo, que corresponde ao valor do atributo no exemplo dado. Este processo é repetido para a sub-árvore, enraizada pelo novo nó.

A Figura 2.6 ilustra uma árvore para tomar a decisão de jogar *golf*, ou seja, com o predicado meta "Jogar *Golf*", classificando as manhãs de sábado em agradável ou não para se jogar *golf*, dependendo de alguns atributos. Por exemplo, a instância: (visual=ensolarado e umidade=alta então jogar g*olf*=não), pode ser classificado como exemplo negativo (o predicado meta Jogar *golf*=não) ou (visual=nublado então jogar *golf*=sim), pode ser classificado como um exemplo positivo (o predicado meta jogar *golf*=sim).

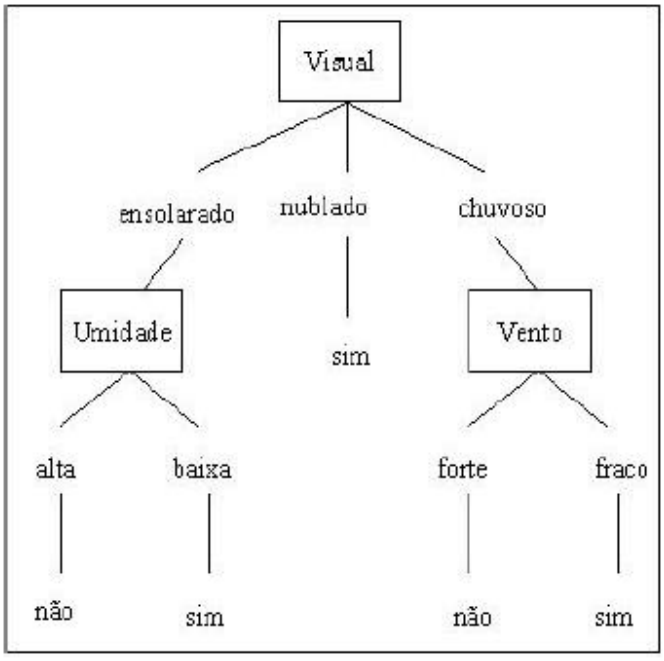


Figura 2.6: Árvore de Decisão para o Problema "Jogar Golf".
Fonte: Adaptado de Mitchell, 1997.

Houve uma onda de interesse em produtos baseados em árvores de decisão, principalmente, pelo fato de serem mais fáceis de compreender o seu funcionamento e a maneira como são obtidos os resultados. 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 e as desvantagens estão na possibilidade de erros na classificação quando existem muitas classes e o tratamento de dados contínuos.

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 do espaço definido pelos atributos em subespaços. Para Witten e Frank (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.

A técnica de árvore de decisão tem aplicação em geral nas tarefas de classificação.

Dias (2001) destaca alguns exemplos de algoritmos para a construção de uma árvore de decisão, que são: CART (BERRY e LINOFF, 1997), CHAID (BERRY e LINOFF, 1997), ID3 (QUINLAN, 1983), C4.5 (QUINLAN, 1993), SLIQ (METHA *et al*, 1996) E SPRINT (SHAFER *et al*, 1996).

**4) Raciocínio Baseado em Casos**

Também conhecido como MBR (*Memory-Based Reasoning* – raciocínio baseado em memória), o raciocínio baseado em casos tem por base o método do vizinho mais próximo. “O *MBR* procura os vizinhos mais próximos nos exemplos conhecidos e combina seus valores para atribuir valores de classificação ou de previsão” (HARRISON, 1998, p. 195). Tenta solucionar um dado problema fazendo uso direto de experiências e soluções passadas. A distância dos vizinhos dá uma medida da exatidão dos resultados.

Na aplicação do MBR, segundo Berry e Linoff (1997), existem quatro passos importantes: 1) escolher o conjunto de dados de treinamento; 2) determinar a função de distância; 3) escolher o número de vizinhos mais próximos; e 4) determinar a função de combinação.

A técnica de raciocínio baseado em casos é apropriada às seguintes tarefas: classificação e segmentação. Os seguintes algoritmos implementam a técnica de raciocínio baseado em casos (DIAS, 2001): BIRCH (ZHANG *et al*., 1996), CLARANS (CHEN et al., 1996) e CLIQUE (AGRAWAL *et al*., 1998).

**5) Algoritmos Genéticos**

Algoritmos genéticos são algoritmos de busca baseados na seleção natural dos seres vivos. Segundo Goldberg (1989), a cada geração, novos indivíduos (*strings*) são gerados a partir dos indivíduos velhos. Cada indivíduo representa os parâmetros para solução do problema e possui também um valor de *fitness*, o qual indica o quão satisfatório ele é como solução do problema. Os principais operadores genéticos são *crossover* e mutação. A seleção é um processo que ocorre antes da aplicação dos operadores genéticos e consiste na escolha de indivíduos dentre a população. A seleção dos indivíduos é feita baseada no seu valor de *fitness*. Indivíduos com melhor *fitness* possuem maior probabilidade de serem selecionados para gerar descendentes. *Crossover* é o processo que combina dois indivíduos selecionados. O *crossover* efetua escolha de uma posição aleatória na *string* e troca partes correspondentes das duas *strings* selecionadas, criando dois novos indivíduos.

O operador de mutação apenas efetua a troca de bits de uma *string* (indivíduo) substituindo 0 por 1 e vice-versa ou outra simbologia de acordo com a situação. Ele é importante para tentar recuperar material genético útil que pode ter sido perdido com as operações de seleção ao longo das gerações. Para um estudo mais detalhado recomenda-se a leitura de Goldberg (1989).

De um modo geral a técnica de algoritmos genéticos é usada nas tarefas de classificação e segmentação. Alguns exemplos de algoritmos genéticos são encontrados na literatura tais como: Algoritmo Genético Simples (GOLDBERG, 1989), Genitor e CHC (WHITLEY, 1993), Algoritmos de Hillis (HILLIS, 1997), GA-Nuggets (FREITAS, 1999) GS-PVMINER (ARAÚJO *et al.*, 1999).

Para melhor compreensão das técnicas de mineração de dados citadas na presente pesquisa, é apresentada a Tabela 2.3 contendo algumas delas.

| TÉCNICA | DESCRIÇÃO | TAREFAS | EXEMPLOS |
|---|---|---|---|
| Descoberta de Regras de Associação | Estabelece uma correlação estatística entre atributos de dados e conjuntos de dados | • Associação | Apriori, AprioriTid, AprioriHybrid, AIS, SETM (Agrawal e Srikant, 1994) e DHP (Chen et al., 1996). |
| Árvores de Decisão | Hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos | • Classificação | CART, CHAID, C5.0/See5, Quest (Two Crows, 1999); ID-3 (Chen *et al.*, 1996); C4.5 (Quinlan, 1993) SLIQ (Mehta *et al.*, 1996); SPRINT (Shafer *et al.*, 1996). |
| Raciocínio Baseado em Casos ou MBR | Baseado no método do vizinho mais próximo, combina e compara atributos para estabelecer hierarquia de semelhança | • Classificação<br>• Segmentação | BIRCH (Zhang *et al.*, 1996); CLARANS (Chen et al., 1996); CLIQUE (Agrawal *et al.*, 1998). |
| Algoritmos Genéticos | Métodos gerais de busca e otimização, inspirados na Teoria da Evolução, onde a cada nova geração, soluções melhores têm mais chance de ter "descendentes" | • Classificação<br>• Segmentação | Algoritmo Genético Simples (Goldberg, 1989); Genitor, CHC (Whitley, 1993); Algoritmo de Hillis (Hillis, 1997); GA-Nuggets (Freitas, 1999); GA-PVMINER (Araújo *et al.*, 1999). |
| Redes Neurais Artificiais | Modelos inspirados na fisiologia do cérebro, onde o conhecimento é fruto do mapa das conexões neuronais e dos pesos dessas conexões | • Classificação<br>• Segmentação | Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB (Azevedo *et al.*, 2000), (Braga et al., 2000), (Haykin, 2001) |

Tabela 2.3: Técnicas de Mineração de Dados.
Fonte: Dias (2001).

## 2.5 Ferramentas de Mineração de Dados

De acordo com Dias (2001) *apud* Goebel e Gruenwald (1999), muitas ferramentas atualmente disponíveis são ferramentas genéricas da Inteligência Artificial ou da comunidade de 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-processamento 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 e Gruenwald (1999), as características a serem consideradas na escolha de uma ferramenta de descoberta de conhecimento devem ser as seguintes (DIAS, 2001 *apud* GOEBEL E GRUENWALD, 1999):

- A habilidade de acesso a uma variedade de fontes de dados, de forma *on-line* e *off-line*;
- A capacidade de incluir modelos de dados orientados a objetos ou modelos não padronizados (tal como multimídia, espacial ou temporal);
- A capacidade de processamento com relação ao número máximo de tabelas/tuplas/atributos;
- A capacidade de processamento com relação ao tamanho do banco de dados;
- Variedade de tipos de atributos que a ferramenta pode manipular; e
- Tipo de linguagem de consulta.

Existem ferramentas que implementam uma ou mais técnicas de mineração de dados. A Tabela 2.3 relaciona algumas dessas ferramentas, fornecendo informações tais como: a empresa fornecedora, as técnicas implementadas de mineração de dados e exemplos de aplicações.

| **FERRAMENTA/ EMPRESA FORNECEDORA** | *TÉCNICAS DE MINERAÇÃO DE DADOS* | **APLICAÇÕES** |
|---|---|---|
| AIRA/ Hycones IT (1998) | Regras de associação | Gerenciamento de relacionamento de cliente, marketing, detecção de fraude, controle de processo e controle de qualidade. |
| Alice 5.1/ Isoft AS. (1998) | Árvore de decisão<br>Raciocínio baseado em casos | Política de crédito, marketing, saúde, controle de qualidade, recursos humanos. |
| Clementine/ Integral Solutions Limited (ISL, 1996) | Indução de regras<br>Árvores de decisão<br>Redes neurais | Marketing direto, identificação de oportunidades de venda cruzada, retenção de cliente, previsão de lucro do cliente, detecção de fraude, segmentação e lucro do cliente. |
| DataMind / DataMind Technology Center (1998), (Groth, 1998) | (abordagem própria) | Não identificadas. |
| Decision Series/ Neovista Solutions Inc. (1998) | Árvore de decisão<br>Métodos estatísticos<br>Indução de regras<br>Redes neurais | Marketing direcionado, detecção de fraude, retenção de cliente, análise de risco, segmentação de cliente, análise de promoção. |
| Intelligent Miner/ IBM (1997) | Árvores de decisão<br>Redes neurais | Segmentação de cliente, análise de conjunto de itens, detecção de fraude. |
| KnowledgeSEEKER/ Angoss IL (Groth, 1998) | Árvores de decisão<br>Indução de regras | Lucro e segmentação de cliente para detecção de fraude e análise de risco, controle de processo, marketing direto. |
| MineSet/ Silicon Graphics Computer Systems (2000) | Métodos estatísticos<br>Árvores de decisão<br>Indução de regras | Áreas da saúde, farmacêutica, biotecnologia e química. |
| NeuralWorks Predict/ NeuralWare (Groth, 1998) | Rede neural | Indústria. |
| PolyAnalyst/ Megaputer Intelligence Ltd. (1998) | Algoritmo genético<br>Métodos estatísticos<br>Indução de regras | Marketing direto, pesquisa médica, análise de conjunto de itens. |

Tabela 2.4: Ferramentas de Mineração de Dados
Fonte: Dias (2001)

### 2.5.1 WEKA

A ferramenta *WEKA* (*Waikato Environment for Knowledge Analysis*), tem sido bastante utilizada na realização da etapa de mineração de dados. Esta ferramenta foi implementada na linguagem Java e desenvolvida no meio acadêmico da Universidade de Waikato, na Nova Zelândia, em 1999. Tem como vantagem o fato de ser de Domínio Público estando disponível para *download* 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 MD (Witten *e* Frank, 2000). A Figura 2.7 mostra as interfaces principais da ferramenta *WEKA*.

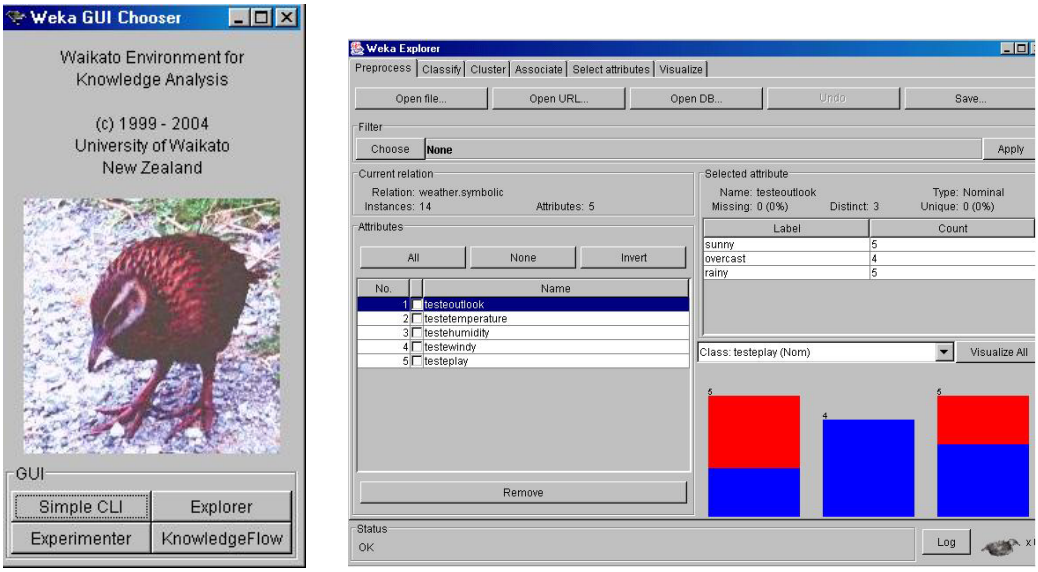


FIGURA 2.7: INTERFACE DA FERRAMENTA *WEKA*

É composto 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 linha de comando ou de uma interface gráfica.

## 2.6 Aplicações da Mineração de Dados

A aplicação de MD em *marketing* direcionado é utilizada para descobrir quais clientes têm maior probabilidade de comprar determinado produto. A base é organizada a partir do histórico de vendas do produto e os dados de clientes compradores. A MD identifica padrões de comportamento dos consumidores, encontra suas características de acordo com a região demográfica e prevê consumidores atingidos nas campanhas de *marketing*.

Essa típica tarefa de MD é usada por grandes lojas de departamentos e administradoras de cartões de crédito que utilizam os dados das compras dos clientes no passado recente para traçar seus perfis de consumo. Informações como idade, sexo, estado civil, salário, moradia (própria ou alugada), bairro e cidade também são importantes, pois permitem a setorização ainda mais fina dos clientes. Conhecer o perfil de seus clientes é fundamental para que uma empresa possa se manter no mercado. Muito investimento deve ser feito para que o cliente continue fiel à empresa e outros sejam conquistados. Para tanto, as empresas precisam realizar os desejos e necessidades do cliente cuidando do estoque, da distribuição dos produtos nas prateleiras e das promoções criativas a fim de propiciar compras "casadas" de produtos.

Um exemplo clássico de utilização de MD por uma empresa é o da cadeia americana de supermercados *Wal-Mart*. As técnicas de MD identificaram um hábito curioso dos consumidores. Ao procurar eventuais relações entre o volume de vendas e os dias da semana, o processo de MD identificou que, nas sextas-feiras, as vendas de cervejas cresciam na mesma proporção que as fraldas. Uma investigação detalhada revelou que, ao comprar fraldas para seus bebês, os pais aproveitavam para abastecer o estoque de cerveja para o final de semana.

De acordo com Carvalho (2001), ao analisar os bancos de dados da empresa, é necessário relacionar os diferentes produtos comprados ao mesmo tempo pelos clientes para que as ofertas cruzadas possam ser estabelecidas. Conhecendo os dados pessoais dos clientes, a empresa ainda poderá supor suas necessidades e oferecer, por exemplo, em sistema de crediário, produtos como mobília e decoração a clientes que possuam casa própria.

A Figura 2.8 ilustra as três áreas gerais de negócios onde MD é aplicado hoje e lista algumas aplicações comuns em cada área.

| Administração e Mercado | Administração de Risco | Administração de Fraudes |
|---|---|---|
| -Mercado alvo<br>-Gerenciamento do Relacionamento com Clientes<br>-Análise de cesta de mercado<br>-Venda cruzada<br>-Segmentação de mercado | -Previsão<br>-Retenção de clientes<br>-Controle de qualidade<br>-Análise competitiva | -Detecção de fraudes |

Figura 2.8: Áreas de Aplicação de Mineração de Dados.
Fonte: Adaptado de CABENA *et al.* (1998).

Na medicina é usado para prever qual paciente tem maior probabilidade de contrair certa doença, em função de dados históricos de pacientes e doenças. Na área de seguros e planos de saúde, a MD pode determinar procedimentos médicos requisitados ao mesmo tempo, pode prever consumidores que comprarão novas apólices e até identificar comportamentos fraudulentos. Na telecomunicação, serve para identificar fraudes em ligações telefônicas (particularmente em celulares), dentre um enorme número de ligações efetuadas pelos clientes. No mercado financeiro, pode prever as ações subirão ou descerão na bolsa de valores, em função de dados históricos com preços de ações e valores de índices financeiros.

De acordo com Santos *et al.* (1999), o *Security Pacific/Bank of America* está usando MD no suporte à decisões na área dos empréstimos bancários e para prevenir fraudes. Em Portugal, por exemplo, o Banco Privado Português – BPP, sentiu a necessidade de um suporte à decisão de avaliação de perfis de riscos para os investimentos financeiros dos seus clientes.

Cabe ressaltar que, embora os exemplos mais comuns se refiram a clientes, compras e vendas, as áreas de aplicação de MD são praticamente ilimitadas. Em vez de tentar prever qual cliente comprará um determinado tipo de produto, pode-se tentar prever se um paciente desenvolverá ou não uma doença, se uma ligação telefônica que está sendo feita de um celular é uma fraude (está usando um número roubado de celular) ou não.

## 2.7 Trabalhos relacionados

Na Universidade Federal de Minas Gerais (UFMG), as técnicas de MD vêm sendo utilizadas na determinação do perfil dos alunos, com base nos dados de pesquisa econômico-social preenchidos quando da admissão. Essas informações vêm sendo correlacionadas com o desempenho dos alunos no vestibular e mesmo e durante o curso de graduação (CESAR, 2000).

O artigo apresentado por Freitas Junior *et al.* (2001), mostra um estudo de caso no qual foram aplicadas técnicas de mineração, mais especificamente, as regras de associação, sobre uma base de dados do quadro de docentes da Universidade Estadual de Maringá (UEM). O objetivo da MD foi descobrir o perfil do corpo docente da UEM, extraindo conhecimentos tais como, a relação entre titulação e produção acadêmica, a relação entre regime de trabalho e produção acadêmica.

O artigo de Garcia (2000) apresenta um estudo do método de classificação utilizando a técnica de árvore de decisão e seus algoritmos. As bases de dados utilizadas contêm as Autorizações de Internação Hospitalar (AIHs), que registram internações, procedimentos e diagnósticos realizados em instituições de saúde ligadas ao Sistema Único de Saúde.

A pesquisa de Pansanato e Soares (1999) avaliou a capacidade preditiva do resultado do ENEM em relação ao desempenho dos candidatos ao vestibular da UFMG de 1999. Os resultados obtidos demonstram que a utilização dos resultados de ENEM, em substituição à primeira etapa do vestibular, não traria prejuízo para a universidade ou para os alunos envolvidos.

Em 1998, a pesquisa de Soares e Fonseca teve como principal objetivo analisar o desempenho dos candidatos ao vestibular da UFMG em 1997. Para isto, foram utilizadas várias características sócio-econômicas coletadas através do questionário respondido pelos candidatos que se inscreveram ao vestibular. Os resultados verificaram que estes fatores estão, como previsto, fortemente associados com o desempenho dos alunos, embora tenham no conjunto, pequena capacidade preditiva.

A pesquisa desenvolvida por Cabral Junior *et al.* (2002), teve por objetivo estudar os algoritmos de aprendizagem de máquinas baseados na construção de árvores de decisão e regras de classificação “SE-ENTÃO”**,** aplicados a alguns bancos de dados visando a descoberta de padrões e conhecimento.

O artigo publicado por Carvalho (2004), apresenta e discute alguns experimentos realizados através da aplicação de algoritmos de MD, usando as informações e o conhecimento extraído da base de dados referente ao aproveitamento dos acadêmicos da Universidade Tuiuti do Paraná, tais como, notas bimestrais, faltas e condição de aproveitamento por disciplina cursada. A pesquisa teve como objetivo auxiliar o processo decisório de coordenadores e colegiados de cursos de graduação.

A pesquisa de Dias (2001), apresenta um modelo de formalização do processo de desenvolvimento de Sistemas de Descobertas de Conhecimento em Banco de Dados, cujo objetivo principal é gerar informações relevantes à tomada de decisão, através da aplicação de técnicas de MD. O modelo proposto foi aplicado na plataforma de informações da pós-graduação brasileira, a partir dos dados da CAPES de 1998.

## 2.8 Considerações finais

Atualmente, os sistemas de descoberta de conhecimento são empregados 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 para o 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 aplicação ou implementação da mineração de dados, dificilmente haverá uma técnica que resolva todos os problemas de uma empresa ou uma instituição de ensino. Ter conhecimento das técnicas 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.

Neste capítulo foi feita uma abordagem sobre o processo de descoberta de conhecimento em banco de dados e descritas suas principais etapas. Em especial foram abordados os conceitos básicos sobre a mineração de dados, sua contribuição e as vantagens competitivas ganhas quando inserida em outras tecnologias.

Também foram descritas as tarefas, as técnicas e as ferramentas de mineração de dados e, em particular, a ferramenta *WEKA*, aplicada na base de dados deste trabalho. Foram, também, relacionados alguns exemplos de aplicações de mineração de dados em diversas áreas, bem como, alguns trabalhos correlatos.

# 3 MINERAÇÃO DE DADOS APLICADOS AO PROCESSO SELETIVO DO VESTIBULAR DA UFPR

Este capítulo tem por finalidade a descrição da aplicação de técnicas de mineração de dados a base de dados do processo seletivo do vestibular da UFPR. Para o propósito desta pesquisa foi focada uma maior atenção na mineração de dados preditiva (classificatória). Inicialmente, explicitam-se as motivações que levaram a este trabalho, assim como os objetivos pretendidos. Em seguida, evidencia-se a instituição de ensino que disponibilizou a base de dados utilizada neste trabalho. Após a descrição global da proposta, apresenta-se, detalhadamente, cada uma das etapas do processo de descoberta de conhecimento, a forma como os dados foram ajustados para o objetivo pretendido, salientando-se as principais contribuições da pesquisa. Finalmente, faz-se uma descrição das técnicas e ferramentas de mineração de dados utilizadas para a obtenção dos resultados esperados nesta pesquisa.

## 3.1 Histórico da Instituição

A Universidade Federal do Paraná[12](UFPR) é a mais antiga universidade do Brasil e símbolo da cidade de Curitiba. Desde 1912, a UFPR é referência no ensino Superior para o estado e para o Brasil. A universidade demonstra sua importância e excelência através dos cursos de graduação, especialização, mestrado e doutorado, além de suas áreas de extensão e pesquisa. A responsabilidade social da universidade, enquanto instituição pública, também é valorizada em suas ações perante a comunidade paranaense.

Além dos *campi* em Curitiba, a UFPR dispõe de outras instalações no interior e no litoral do Estado, facilitando o acesso à educação e integrando culturalmente o Paraná.

A História da UFPR é marcada por grandes feitos e está muito ligada à história de desenvolvimento do Estado do Paraná. Sua história começou em 1812, quando o

[12] Texto extraído do site www.ufpr.br

político Rocha Pombo lançou na Praça Ouvidor Pardinho a pedra fundamental da Universidade do Paraná. Porém, devido ao Movimento Federalista, o projeto não foi adiante. Em 1912, as lideranças políticas se mobilizaram em prol da criação da Universidade do Paraná.

No dia 19 de dezembro de 1913, Victor Ferreira do Amaral e Silva liderou a criação efetiva da Universidade do Paraná. Era uma época próspera da economia paranaense, devido à abundante produção e ao bom comércio da erva-mate. A Universidade iniciou seus trabalhos como instituição particular. Os primeiros cursos ofertados foram Ciências Jurídicas e Sociais, Engenharia, Medicina e Cirurgia, Comércio, Odontologia, Farmácia e Obstetrícia. Após ter fundado a Universidade do Paraná, Victor Ferreira do Amaral – que foi seu primeiro reitor – fez empréstimos e iniciou a construção do Prédio Central, na Praça Santos Andrade, em terreno doado pela Prefeitura de Curitiba.

Na década seguinte veio a Primeira Guerra Mundial e com ela a recessão econômica e as primeiras dificuldades. Entre elas uma lei que determinava o fechamento das universidades, pois o Governo Federal não recebia bem as iniciativas de forma independente nos estados. A alternativa encontrada para evitar o fechamento da Universidade do Paraná foi desmembrar a instituição em faculdades. Durante mais de trinta anos buscou-se novamente a restauração da Universidade que aconteceu no inicio da década de 50, quando as faculdades foram reunidas e novamente formada a Universidade do Paraná.

Restaurada a Universidade, a próxima batalha visou sua federalização. Na época o reitor Flávio Suplicy de Lacerda mobilizou as lideranças do Estado e, em 1950, a Universidade do Paraná tornou-se uma instituição pública e gratuita. Após a federalização, deu-se a fase da expansão da Universidade. A construção do Hospital de Clinicas em 1953, do Complexo da Reitoria em 1958 e do Centro Politécnico em 1961 comprovaram a consolidação da instituição que conta com 92 anos de história.

## 3.2 O Processo de *KDD*

A seguir são descritas as atividades realizadas em cada etapa do processo de descoberta de conhecimento em banco de dados.

### 3.2.1 Problema a ser tratado

O primeiro passo no processo de descoberta de conhecimento em banco de dados é a definição do problema a ser tratado. De acordo com Dias (2001), pode não existir um problema real a ser solucionado, considerando que a mineração de dados pode ser aplicada como um processo de descoberta, no qual nem sempre é feito algum tipo de suposição antecipada.

No caso desta pesquisa, o objetivo é analisar o perfil dos candidatos ao processo seletivo do vestibular da UFPR. As bases de dados utilizadas contêm informações sócio-econômico-culturais, dados cadastrais e o desempenho dos candidatos inscritos no processo seletivo do vestibular para o ano letivo de 2004 da UFPR.

Os dados foram disponibilizados em duas planilhas *Excel*, onde uma das planilhas contém as respostas dos itens propostos no questionário sócio-educacional aplicado aos candidatos no ato da inscrição e a outra denotada como cadastro geral dos candidatos, contém os registros das notas nas provas, notas do ENEM e outros dados cadastrais tais como, o protocolo, a inscrição, o sexo, o estado civil, a data de nascimento, o curso, a língua, o bairro, a cidade, a UF, o CEP. O cadastro geral contém, também, a média das notas obtidas pelo candidato e o status (resultado do vestibular). O status é a classe alvo (atributo meta) da presente pesquisa.

### 3.2.2 Seleção dos dados

O questionário sócio-educacional aplicado aos candidatos ao processo seletivo do vestibular para o ano letivo de 2004 é composto de 31 itens, e o cadastro geral de 24 itens, que formam duas bases distintas. Como nas duas planilhas eletrônicas citadas constam os protocolos dos candidatos, decidiu-se pela junção das duas planilhas, formando uma só base de dados com 55 itens (atributos) que se constitui na base referencial para a aplicação das técnicas de mineração de dados que serão detalhadas a seguir.

### 3.2.3 Limpeza dos Dados

Em uma análise inicial na base de dados, detectou-se vários itens de dados em branco e alguns com erros de digitação ou valores absurdos. Os itens de dados em branco foram preenchidos pela letra N (significando nulo). Isto porque a ferramenta *WEKA* utilizada nesta pesquisa exige que todos os registros estejam preenchidos.

Os demais registros foram adequados através de uma avaliação criteriosa, manualmente, com exceção de alguns que foram excluídos por falta de condições de serem corrigidos. Os itens inconsistentes eliminados, em análise manual, não causaram prejuízos para a base de dados por se mostrarem de maneira aleatória na base.

Foi necessária, também, a eliminação de alguns atributos por não serem considerados relevantes nesta pesquisa, tais como: protocolo, inscrição, bairro, cidade, UF, CEP, a opção pelo ENEM e a média das notas.

Na junção das duas planilhas citadas anteriormente, notou-se que o atributo estado civil estava em duplicidade e, portanto eliminou-se um deles.

### 3.2.4 Transformação dos Dados

A primeira transformação realizada nesta etapa foi a transformação da data de nascimento para idade, usando uma função do *Microsoft Excel 2000*.

Após a limpeza dos dados e esta transformação inicial realizada, a base de dados ficou então com 46 atributos e foi denominada base_final. Esses atributos estão relacionados no Anexo A.

Para a obtenção de melhores resultados com a aplicação de técnicas de mineração de dados é recomendável que os dados sejam discretizados. A discretização foi feita utilizando-se do *Microsoft Excel 2000* pelos recursos oferecidos e pelo fato da base de dados já estar em planilha *excel*. Foram criadas faixas para os atributos de acordo com as suas freqüências e, em seguida, os atributos originais foram substituídos pelas faixas (novos atributos) através de uma função criada no próprio *excel*.

Foram discretizados o atributo idade e aqueles atributos que envolvem as notas de todas as disciplinas, da redação e do ENEM. Portanto, foram definidas, para cada

atributo, faixas de valores e, aplicados rótulos a essas faixas, como por exemplo, a idade na faixa de 18 a 20 foi rotulada como “idade_18_a_20”.

As faixas de valores para cada atributo discretizado foram definidas da seguinte forma:

1. Foram transferidos os valores do atributo para uma outra planilha;
2. Foi realizada a classificação desses valores em ordem crescente para que os valores iguais fossem agrupados;
3. Foi calculada a freqüência de ocorrência de cada valor do atributo;
4. Foi realizada novamente a classificação, em ordem crescente, dos valores dos atributos juntamente com os valores das freqüências de ocorrência;
5. Foi aplicada uma heurística baseando-se nas freqüências de ocorrência e na delimitação dos valores mínimo e máximo do atributo, de acordo com as convenções descritas abaixo:
    - o valor máximo de cada faixa não deveria ser maior que o dobro do mínimo, salvo algumas exceções;
    - qualquer valor de freqüência de ocorrência maior que o triplo da média de freqüência (contabilizando de cima para baixo) seria uma espécie de "salto" que oferece um ponto singular para delimitar uma divisão de escopo de faixas distintas.

Após a discretização dos atributos descrita acima, optou-se pela categorização desses atributos como, por exemplo, a faixa de idade “idade_15_a_15” foi convertida para o caractere 1, a faixa “idade_16_a_16”, para o caractere 2 e assim sucessivamente até a faixa “idade_60_a_71” que foi convertida para o caractere 12. O mesmo procedimento foi adotado para os demais atributos discretizados. Os atributos não discretizados ficaram no formato original, conforme pode ser verificado no anexo A.

A base de dados preparada foi dividida em várias outras bases de acordo com os critérios estabelecidos por esta pesquisa, que são: os onze cursos mais concorridos

(Medicina, Comunicação Social-Publicidade e Propaganda, Direito-D, Direito-N, Comunicação Social-Jornalismo, Arquitetura e Urbanismo, Administração-N, Medicina Veterinária-Curitiba, Turismo-N, Desenho Industrial-Programação Visual e Psicologia); os onze cursos menos concorridos (Filosofia-Bacharelado com Licenciatura Plena, Estatística, Letras-Francês-N, Física-Bacharelado, Matemática Industrial, Educação Artística-Desenho, Engenharia Cartográfica, Ciências Econômicas, Gestão da Informação, Física Licenciatura e Ciências Sociais); alguns cursos escolhidos por serem mais concorridos (Medicina) ou menos concorridos (Matemática Industrial) e algumas áreas específicas (exatas, tecnológicas, humanística e biológica).

Após a realização das atividades descritas acima na etapa de pré-processamento, foi necessário formatar a base de dados resultante de acordo com o formato de entrada da ferramenta *WEKA*, obtendo assim um arquivo no formato ARFF. Um arquivo *ARFF* consiste de uma lista de todas as instâncias, com os valores de atributos para cada instância separados por vírgula. Esta formatação foi realizada através de recursos do *Microsoft Excel*. O Anexo B mostra a seqüência passo a passo para transformar a base de dados em um arquivo *ARFF*, usando como exemplo a base de dados considerada como referência para esta pesquisa.

Portanto, todas as bases de dados geradas foram transformadas para o formato "arquivo.*arff*". As novas bases ficaram com a seguinte denominação, base_+concorridos.*arff*, base_-concorridos.*arff*, base_mat_licenc.*arff* e assim sucessivamente com todas as bases das quais este trabalho fez uso.

### 3.2.5 Mineração de Dados (*Data Mining*)

A etapa mineração de dados, considerada como o núcleo do processo de descoberta de conhecimento em banco de dados, consiste na efetiva aplicação da técnica de MD através do algoritmo escolhido sobre os dados a serem analisados com o objetivo de localizar os padrões. É nesta etapa que, após a base de dados já estar preparada, acontece a busca de padrões, classificações, regularidades, ou seja, descobrir novas relações, não identificáveis a "olho nu".

Os algoritmos e as técnicas utilizadas na criação de modelos a partir de dados, normalmente, provêm de áreas como Aprendizado de Máquinas (AM),

Reconhecimento de Padrões e Estatística. Estas técnicas, muitas vezes, podem ser combinadas para se obter resultados melhores.

A presente pesquisa está focada nas técnicas de AM, as quais envolvem métodos para aquisição de conhecimento implícito a partir de um conjunto de fatos. Esses métodos são vistos como sistemas de aprendizado que possuem o objetivo de descobrir conhecimento de dados através da construção de um modelo de predição ou de descrição. A descrição e a predição são obtidas selecionando-se as tarefas, os algoritmos e as técnicas de mineração de dados.

Como auxílio nas análises preliminares desta pesquisa foi utilizada a tabela de distribuição de freqüências sobre os dados originais dos candidatos que se encontram nas planilhas fornecidas pela UFPR, ou seja, o questionário sócio-educacional e, o cadastro geral, descritos anteriormente.

Na tabela da distribuição de freqüências, pode-se observar a proporção dos indivíduos em cada atributo e o grau de preenchimento de cada variável. No anexo C encontra-se como exemplo a tabela de distribuição de freqüências dos dados do questionário sócio-educacional referente aos 46.532 candidatos que participaram do processo seletivo do vestibular para o ano letivo de 2004.

Como o objetivo do trabalho é a meta preditiva, pelo fato de constar na base de dados utilizada um atributo-meta forte, ou seja, o status, que é o resultado do desempenho do candidato no vestibular e, ainda por existir poucos trabalhos relacionados com a base do vestibular usando o atributo-meta, optou-se pela tarefa de classificação, descrita no Capítulo II, a técnica de mineração árvore de decisão e as regras de classificação geradas a partir da árvore de decisão.

Entre as ferramentas existentes no mercado de software já mencionadas no capítulo II, no item 2.5, a ferramenta *WEKA* foi escolhida por ser uma ferramenta de domínio público, pela sua praticidade de utilização e por possuir algoritmos de classificação que implementam a técnica árvore de decisão e as regras de classificação. Nas seções seguintes são descritas a técnica árvore de decisão e as regras de classificação.

### 3.2.5.1 A técnica Árvores de Decisão

Dentre as técnicas de classificação mais utilizadas está a árvore de decisão, que é uma representação simples do conhecimento e um meio eficiente de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados.

As árvores de decisão são uma evolução das técnicas que apareceram durante o desenvolvimento das disciplinas de *machine learning* (aprendizado de máquina-AM). Elas surgiram a partir do uso de uma análise chamada Detecção de Interação Automática (*DIA*), desenvolvida na Universidade de Michigan. Essa análise trabalha testando automaticamente todos os valores dos dados para identificar aqueles que são fortemente associados com os itens de saída selecionados para exame. Os valores que são encontrados com forte associação são os prognósticos chaves ou fatores explicativos, usualmente chamados de regras sobre os dados (DW BRASIL, 2004).

Um antigo algoritmo chamado *CHAID* foi desenvolvido estendendo as capacidades do *DIA*, sendo um pouco através da adição da fórmula estatística chi-quadrado (*Chi squared*). Mas foi o professor Ross Quinlan na Austrália que desenvolveu a tecnologia que permitiu o aparecimento das árvores de decisão.

Muitas pessoas na indústria de DM consideram Ross Quinlan, da Universidade de Sydney, Austrália, como o “pai das árvores de decisão”. A contribuição de Quinlan foi um novo algoritmo chamado ID3, desenvolvido em 1983. O ID3 e suas evoluções (ID4, ID6, C 4.5, C5.0/See 5) são muito bem adaptadas para usar em conjunto com as árvores de decisão, na medida em que eles produzem regras ordenadas pela importância. Essas regras são, então, usadas para produzir um modelo de árvore de decisão dos fatos que afetam os itens de saída (DW BRASIL, 2004).

A aprendizagem por árvores de decisão é um dos métodos mais usados e práticos para inferência indutiva. A indução mediante árvores de decisão é uma das formas mais simples de algoritmos de aprendizagem e também a de maior sucesso. Recebe como entrada um objeto ou uma situação descrita por um conjunto de propriedades ou atributos, e dá como saída uma decisão. Em termos de árvore de decisão, um exemplo é descrito pelos valores dos atributos e o valor do predicado meta. O valor

do predicado meta é chamado classificação do exemplo. Se o predicado meta é verdadeiro para algum exemplo, o chamamos de exemplo positivo, caso contrário, exemplo negativo. O conjunto completo de exemplos é chamado conjunto de treinamento (MITCHEL,1997).

O algoritmo C4.5 deriva do algoritmo simples que usa o esquema "dividir-pra-conquistar[13]", porém, estendido para considerar problemas no mundo real como, por exemplo, valores ausentes e valores numéricos. Este algoritmo possui habilidade para trabalhar com valores contínuos, para selecionar uma medida apropriada para a seleção de atributos, para manusear dados com diferentes valores, melhorar a eficiência computacional e resolver os casos em que a quantidade de dados é limitada e o algoritmo simples ID3 produz um overfitting[14]. A seção seguinte apresenta a construção de uma árvore de decisão utilizando o algoritmo C4.5, uma extensão do algoritmo ID3 para tratar de instâncias com alguns atributos contínuos.

#### 3.2.5.1.1 Algoritmo para Indução de Árvores de Decisão

Para Carvalho (2002), um algoritmo para indução de árvores de decisão é um exemplo de algoritmo de estrutura TDIDT - *Top-Down Induction of Decision Trees.* Este algoritmo utiliza a estratégia "dividir-para-conquistar", ou seja, um problema complexo é decomposto em subproblemas mais simples.

Segundo Mitchel (1997), esse algoritmo consiste dos procedimentos descritos na Figura 3.1, os quais criam uma árvore que classifica todos os exemplos do conjunto de treinamento corretamente. A idéia básica do algoritmo (Figura 3.1) deve seguir os seguintes passos:

a) escolha de um atributo;

b) estender a árvore adicionando um ramo para cada valor do atributo;

c) passar os exemplos para as folhas (tendo em conta o valor do atributo escolhido);

[13] Método de geração de árvore de decisão (Mitchel, 1997)

[14] *Overffiting:* ocorre quando a taxa de acertos no conjunto de treinamento é muito alta, mas no conjunto de testes é baixa.

d) para cada nó folha (se todos os exemplos são da mesma classe), associar esta classe ao nó folha, caso contrário, repetir os passos (a), (b) e (c).

Porém, a árvore assim construída pode estar ajustada demais (*overfitted*) aos dados de treinamento. Uma árvore de decisão ***a*** está ajustada demais aos dados se existir uma árvore ***a'*** tal que ***a*** tem menor erro que ***a'*** no conjunto de treinamento, porém ***a'*** tem menor erro no conjunto de teste.

Para corrigir o fato de uma árvore estar ajustada demais, deve-se executar um procedimento de poda da árvore, como será explicado posteriormente. Antes disso, porém, serão apresentados os principais conceitos usados na construção da árvore.

O passo principal de um algoritmo que constrói uma árvore de decisão é a escolha de um atributo para rotular o nó atual da árvore. Deve-se escolher o atributo que tenha o maior poder de discriminação entre as classes para os exemplos no nó atual. Para isso, deve-se utilizar uma medida de poder de discriminação de classes. A seguir são discutidas medidas baseadas na Teoria da Informação (QUINLAN, 1993), (COVER, 1991), as quais são usadas pelo algoritmo C4.5.

```
/*  Conj_Exemplos  representa o conjunto de treinamento */
/*  Atributo_Meta  é o atributo a ser predito pela árvore */
/*  Lista_Atributos representa a lista dos outros atributos a serem testados*/

INICIO1 (Conj_Exemplos , Atributo_Meta , Lista_Atributos)
Selecionar o melhor atributo para o nó raiz da árvore, de acordo com função de avaliação
    SE todos os exemplos em Conj_Exemplos são de uma única classe
       ENTÃ0
          Retornar um único nó com valor da classe
       CASOCONTRÁRIO
            SE Lista_Atributos = ϕ
                 ENTÃO
                      Retornar um único nó com o valor de Atributo_Meta mais freqüente em Conj_Exemplos
                 CASOCONTRÁRIO
                     INICIO2
                         A ← o atributo de Lista_Atributos que melhor classifica Conj_Exemplos
                         PARA cada valor (v_i) possível de A
                            Adicionar uma nova ramificação, A = v_i
                            Criar o subconjunto Conj_Exemplos_vi contendo os exemplos de Conj_Exemplos que
                                  satisfazem o teste A = v_i
                           SE Conj_Exemplos_vi = ϕ
                               ENTÃO
                                    Criar uma ramificação subordinada ao novo nó com o valor de Atributo_Meta
                                        mais  freqüente
                               CASOCONTRARIO
                                     INICIO1(Conj_Exemplos_vi, Atributo_Meta, Lista_Atributos – {A})
                            FIMSE
                         FIMPARA
                    FIMINICIO2
            FIMSE
    FIMSE
    Retornar raiz
FIMINICIO1
```

Figura 3.1: Procedimentos para a Construção da Árvore de Decisão
Fonte: Mitchel (1997)

**1) Ganho de Informação**

O ganho de informação mede a redução da entropia causada pela partição dos exemplos de acordo com os valores do atributo, ou seja, o ganho de informação representa a diferença entre a quantidade de informação necessária para uma predição correta e as correspondentes quantidades acumuladas dos segmentos resultantes após a introdução de um novo teste para o valor de determinado atributo (EIJKEL, 1999). Para a avaliação do quanto é oportuno à introdução de um novo teste, são considerados dois momentos: primeiro, antes da inserção deste novo teste, que constitui uma nova ramificação (partições dos dados com base nos valores dos atributos (KUBAT, 1998)) e, o outro, depois da sua inserção. Se a quantidade de informação requerida é menor depois que a ramificação é introduzida, isso indica que a inclusão deste teste reduz a entropia (desordem) do segmento original.

A entropia é uma medida bem-definida da "desordem" ou da informação encontrada nos dados. A construção de uma árvore de decisão é guiada pelo objetivo de diminuir a entropia. A introdução da entropia no processo de construção de árvores de decisão visa a criação de árvores menores e mais eficazes na classificação (CARVALHO, 2002 *apud* WU, 95).

A forma de obtenção da entropia é dada por Carvalho (2002):

- T = *PE* $\cup$ *NE* onde *PE* é o conjunto de exemplos positivos e *NE* é o conjunto de exemplos negativos;
- *p* = |*PE*| e *n* = |*NE*| onde |*PE*| e |*NE*| representam a cardinalidade de *PE* e *NE* respectivamente;
- para cada nó da árvore serão determinadas as probabilidades de um exemplo pertencente àquele nó ser um exemplo positivo ou negativo, calculadas como *p*/(*p*+*n*) e *n/(p+n)*, respectivamente.

Assim, a entropia é definida pela quantidade de informação necessária para decidir se um exemplo pertence a *PE* ou a *NE*, segundo a expressão 3.1:

$$entropia\ (p,n) = -\ p\ /\ (p+n)\ \log_2 (p\ /\ (p+n)) - n\ /\ (p+n)\ \log_2 (n/(p+n)) \text{ para } p \neq 0 \text{ e } n \neq 0$$

$$entropia\ (p,n) = 0 \qquad \text{caso contrário} \qquad (3.1)$$

Note-se que entropia (p, n) depende apenas de *p* e de *n*. A expressão 3.1 assume que há apenas duas classes, mas ela pode ser facilmente generalizada para o caso de *K* classes, com $K > 2$.

A entropia dos segmentos descendentes de um nó pai da árvore é acumulada de acordo com o peso de suas contribuições na entropia total da ramificação, ou seja, de acordo com o número de exemplos cobertos pela ramificação.

A métrica que é usada para escolher o melhor teste deve avaliar o quanto de "desordem" será reduzido com o novo segmento e como será a ponderação da "desordem" em cada segmento (CARVALHO, 2002 *apud* BERSON, 1997).

Para avaliar o quanto de "desordem" é reduzido através de um novo teste, basta calcular a entropia em cada novo segmento (nó filho) criado por cada ramo, onde cada ramo é associado com um valor do atributo sendo testado.

Se o atributo *X* com um domínio $\{v_1, .... , v_N\}$ é usado como raiz da árvore de decisão, a árvore terá então *N* partições de *T*, $\{T_1, .... , T_N\}$, onde $T_i$ conterá aqueles exemplos em *T* que possuam o valor $v_i$ de *X*. Dado que $T_i$ contém $p_i$ exemplos de *PE* (positivos) e $n_i$ exemplos de *NE* (negativos), a expectativa de informação requerida para a subárvore $T_i$ é dada pela entropia($p_i$, $n_i$) (CARVALHO, 2002 *apud* WU, 1995).

A medida de ganho de informação, ganho*(X),* obtida pela partição associada com o atributo em *X,* é dada pela expressão 3.2:

$$\text{ganho}(X) = \text{entropia}(p,n) - \text{entropia\_ ponderada}(X) \quad (3.2)$$

onde a entropia ponderada de *X* é dada pela expressão 3.3:

$$\text{entropia\_ponderada}(X) = \sum_{i=1}^{N} ((p_i+n_i) / (p+n))\text{entropia}(p_i,n_i) \quad (3.3)$$

onde *N* é o número de partições (segmentos) criadas pelo teste.

**2) Taxa de Ganho**

Outra medida do poder de discriminação (entre classes) de um atributo é a taxa de ganho de informação. Esta medida é definida pela expressão 3.4 (QUINLAN, 1993):

$$\text{taxa\_ganho}(X) = \text{ganho}(X) / \text{informação\_corte}(X) \quad (3.4)$$

onde ganho(*X*) é definido pela expressão 3.2 e

$$\text{informação-corte}(X) = -\sum_{i=1}^{N} (|T_i| / |T|) * \log_2 (|T_i| / |T|) \quad (3.5)$$

onde: informação_corte *(X)* é a quantidade de informação em potencial associada com o fato de um teste do atributo *X* particionar *T* em *N* subconjuntos.

**3) Observações sobre ganho de informação e taxa de ganho**

Segundo Carvalho (CARVALHO, 2002 apud QUINLAN, 1993), o critério ganho de informação, apesar de apresentar bons resultados, tem um *bias* que beneficia os testes com muitas saídas (isto é, atributos com muitos valores). Este problema pode ser corrigido através da normalização deste ganho aparente atribuído ao teste com várias saídas.

O critério taxa de ganho expressa a proporção de informação gerada pela ramificação que parece ser útil para o processo de classificação. Se a ramificação for trivial (no sentido que cada ramo é associado com apenas um exemplo), o valor informação-corte será muito pequeno e a taxa de ganho será instável. Para evitar esta situação, o critério taxa de ganho seleciona um teste que maximize o seu próprio valor, sujeito à restrição que o teste escolhido tenha um ganho de informação pelo menos maior que a média de ganho de informação sobre todos os testes avaliados (QUINLAN, 1993).

O C4.5 examina todos os atributos previsores candidatos, escolhe o atributo *X* que maximize a taxa de ganho*(X)* para rotular o nó atual da árvore, e repete o processo de forma recursiva para dar continuação à construção da árvore de decisão nos subconjuntos residuais $T_1$, ... , $T_N$ .

**4) Poda em Árvores de Decisão**

Conforme mencionado anteriormente, geralmente uma árvore construída pelo algoritmo C4.5 deve ser podada, a fim de reduzir o excesso de ajuste (*overfitting*) aos dados de treinamento.

Carvalho (2002), afirma que existem duas possibilidades de realização da poda em árvores de decisão: parar com o crescimento da árvore mais cedo (pré-poda) ou crescer uma árvore completa e, em seguida, podar a árvore (pós-poda). Segundo Quinlan (1987), “a pós-poda é mais lenta, porém mais confiável que a pré-poda”.

No C4.5 foram desenvolvidos mecanismos de poda sofisticados para tratar desta questão. Um dos mecanismos de poda em árvores de decisão adotado pelo C4.5 é baseado na comparação das taxas de estimativa de erro[15] de cada subárvore e do nó folha. São processados sucessivos testes a partir do nó raiz da árvore, de forma que, se a estimativa de erro indicar que a árvore será mais precisa se os nós descendentes (filhos) de um determinado nó *n* forem eliminados, então estes nós descendentes serão eliminados e o nó *n* passará a ser o novo nó folha (CARVALHO, 2002).

### 3.2.6 Regras de Classificação

Apesar de serem, a princípio, uma forma de representação de conhecimento intuitiva pelo usuário, em alguns casos as árvores de decisão "crescem muito", o que aumenta a dificuldade de sua interpretação (CARVALHO, 2002 *apud* QUINLAN, 1993). Para combater esse problema, alguns algoritmos transformam as árvores de decisão em outras formas de representação, tais como as regras de classificação.

De acordo com Carvalho (2002), essa transformação é simples. Basicamente, cada percurso da árvore de decisão, desde o nó raiz até um nó folha, é convertido em uma regra, onde a classe do nó folha corresponde à classe prevista pelo conseqüente (parte "então" da regra) e as condições ao longo do caminho correspondem às condições do antecedente (parte "se" da regra).

As regras de classificação que resultam da transformação de árvores de decisão podem ter as seguintes vantagens:

- são uma forma de representação do conhecimento amplamente utilizada em sistemas especialistas;
- em geral são de fácil interpretação pelo ser humano;
- em geral melhoram a precisão preditiva pela eliminação das ramificações que expressam peculiaridades do conjunto de treinamento que são pouco generalizáveis para dados de teste.

[15] Pode-se definir a taxa de estimativa de erro da seguinte forma: se *N* exemplos são cobertos por determinado nó folha e *E* dentre estes *N* são classificados de forma incorreta, então a taxa de estimativa de erro desta folha é *E / N* (BERSON, 1997).

Na classificação, as regras identificam as definições ou as descrições dos conceitos de cada classe (CARVALHO, 2002 *apud* HOLSHEIMER, 1994).

O conjunto de regras pode ser usado para descrever a relação entre os conceitos (ou classes) e as propriedades (ou atributos previsores). Um conjunto de regras consiste de uma coleção de declarações do tipo se... então ..., que são chamadas de regras de classificação ou simplesmente regras. O antecedente da regra corresponde a uma descrição de conceito, e o conseqüente da regra especifica a classe prevista pela regra para os exemplos que satisfazem a respectiva descrição de conceitos.

É de relevância que as regras sejam acompanhadas de medidas relativas à sua precisão (ou confiança) e a sua cobertura. A precisão informa o quanto a regra é correta, ou seja, qual a porcentagem de casos que, se o antecedente é verdadeiro, então o conseqüente é verdadeiro. Uma alta precisão indica uma regra com uma forte dependência entre o antecedente e o conseqüente da regra.

### 3.2.7 Implementação Computacional

Para este trabalho, como já foi mencionado anteriormente, foram selecionados dois algoritmos de aprendizagem para a tarefa de classificação disponíveis na ferramenta *WEKA*, o *Weka.classifiers.J48.J48*, que gera a árvore de decisão e o *Weka.classifiers.J48.PART*, que transforma a árvore de decisão em regras de classificação. A seguir será apresentada uma breve descrição desses algoritmos.

De acordo com Queiroz *et al*. (2002), o algoritmo *J48.J48* (a última publicação da família de algoritmos de código aberto que geram árvores de decisão) é uma implementação em Java do algoritmo C4.5, da versão mais adiantada e leve do C4.5, chamada C4.5 revisão 8. O C4.5 release 8, o mais popular dos algoritmos da WEKA, foi a última publicação da família de algoritmos antes do C5.0/See5, a versão mais recente e disponível apenas comercialmente.

Na utilização do algoritmo *J48.J48* é necessário conhecer alguns parâmetros que podem ser modificados para proporcionar melhores resultados, tais como, *U* (usa a árvore sem poda), *C* (*confidence*: escolhe o fator de confiança inicial para a *poda-Default*: 0.25), *M* (escolhe o número mínimo de instâncias por *folha-Default*: 2), *R* (usa a poda com redução de erro), *N* (escolhe o número de partições para a poda com

redução de erro, onde uma partição é utilizada como conjunto de *poda-Default*: 3), *B* (usa árvore binária), *S* (não utiliza subárvore de poda) e *L* (não apaga a árvore depois de construída).

#### 3.2.7.1 Geração de regras a partir da Árvore de Decisão

O processo de geração de regras para classificação de sistemas normalmente atua em dois estágios: as regras são induzidas inicialmente e posteriormente refinadas. Isto é feito através de dois métodos, através da geração das árvores de decisão e o posterior mapeamento da árvore em regras e, então, aplicando processos de refinamento, ou pela utilização do paradigma "separar - pra – conquistar[16]". Assim como na árvore de decisão esse processo também possui um estágio de otimização das regras geradas.

Witten e Frank (2000), combinam estas duas aproximações em um algoritmo chamado *J48.PART* (onde *PART* significa *partial decision trees*). O algoritmo *J48.PART* é uma variação do *J48.J48*, que constrói regras de classificação a partir da árvore de decisão gerada pelo algoritmo *J48.J48 da WEKA*.

O algoritmo *J48.PART* trabalha construindo a regra e estimando sua cobertura como no processo de "separar – pra – conquistar" repetidamente até que todas as instâncias estejam cobertas. A diferença entre esse processo e o de "dividir-para-conquistar" é que neste caso uma árvore de decisão parcialmente podada deve ter sido construída anteriormente. Os ramos com a mais alta cobertura são transformados em regras e a árvore é descartada (WITTEN e FRANK, 2000).

## 3.3 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 foram descritas as atividades realizadas nesta pesquisa: etapas de pré-processamento (definição do problema, seleção, limpeza e transformação de

[16] Método utilizado por algoritmos de cobertura (WITTEN e FRANK, 2000).

dados) e mineração de dados do Processo de Descoberta de Conhecimento em Banco de Dados.

O objetivo geral foi detalhar os passos necessários para a preparação dos dados e a formatação do arquivo para a aplicação de técnicas de mineração de dados utilizando a ferramenta *WEKA*. Também foi descrito o algoritmo que constrói uma árvore de decisão e apresentado o conceito de regras de classificação.

Por fim, foram apresentadas as justificativas da escolha do algoritmo *J.48.PART* disponibilizado pela *WEKA*, onde regras de classificação são geradas a partir da árvore de decisão construída pelo algoritmo *J48.J48*, também da *WEKA*.

# 4 TESTES E RESULTADOS

Neste capítulo são apresentados os testes e os resultados obtidos com a aplicação de 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 UFPR, realizado em dezembro de 2003 para o ano letivo de 2004. Para tanto, utilizou-se as bases de dados preparadas para o *software WEKA*, que contém os dados pertinentes ao processo seletivo do vestibular da Universidade em questão, conforme exposto no Capítulo III.

## 4.1 Aplicação das técnicas de Mineração de Dados

Os resultados apresentados foram obtidos a partir da aplicação das técnicas de mineração de dados Árvore de Decisão e Regras de Classificação que realizam a tarefa de classificação no *software WEKA*, através dos algoritmos *J48.J48 (C4.5 release 8)* e *J48.PART*, respectivamente.

O tipo de conhecimento esperado, com a realização deste trabalho, é a possibilidade de analisar o perfil dos candidatos ao processo seletivo do vestibular da UFPR, bem como encontrar regras interessantes a esse respeito. Assim sendo, o conhecimento obtido através dessas regras tornar-se-á importante para que a UFPR possa implementar ações visando melhorar a qualidade do ensino, diminuir possíveis evasões e definir novas regras para os próximos vestibulares.

### 4.1.1 Classificador J48. J48

Os experimentos realizados com o algoritmo de classificação *J48.J48* tem como objetivo principal gerar árvores de decisão utilizando os dados da base de dados preparada.

Inicialmente, selecionou-se os atributos necessários para gerar regras que pudessem ser analisadas e, a partir delas o perfil do candidato ao vestibular da UFPR fosse caracterizado.

Foram executados vários testes com o algoritmo J48.J48 no ambiente *WEKA*. Num primeiro momento, utilizou-se todos os atributos constantes da base_final.arff, número total de 46 atributos. Num segundo momento, foram realizados testes, selecionando alguns atributos da base de dados, tais como, o sexo, a idade, o curso e as notas. Em outro teste realizado, foram utilizados esses mesmos atributos selecionados, juntamente com alguns 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 as árvores de decisão geradas, na maioria dos testes realizados, ficaram grandes e apareceram várias regras com valor nulo nas instâncias, como por exemplo, NOTA_GEO = 1: 00 (0.0), aumentando as dificuldades de sua interpretação.

A ferramenta *WEKA* tem vários algoritmos que transformam as árvores de decisão em regras de classificação. Conforme citado no Capítulo III, optou-se pelo algoritmo de classificação *J48.PART*. As regras geradas são mais fáceis de serem interpretadas em relação às árvores de decisão.

De acordo com Mitchel (1997), o conjunto de regras de classificação é um dos modelos mais compreensíveis e legíveis para o ser humano. Na seção seguinte são descritos os testes realizados com o algoritmo *J48.PART*, utilizando as mesmas bases de dados preparadas para o *J48.J48*.

### 4.1.2 Classificador *J48.PART*

O algoritmo de classificação *J48.PART* foi executado várias vezes utilizando como entrada as bases de dados preparadas, conforme descrito no Capítulo III. Várias regras[17] foram geradas em cada execução deste algoritmo. A seguir é apresentada a análise dos resultados obtidos usando cada uma das bases de dados consideradas.

**1) Base de dados contendo dados dos candidatos aos onze cursos mais concorridos (nomeada base_+concorrridos.arff)**

A nota na redação é importante porque está diretamente relacionada ao resultado do vestibular, ou seja, o candidato a um curso bastante concorrido que obtém uma boa nota de redação tem grande probabilidade de ser classificado.

O conjunto de atributos com notas categorizadas com valores na faixa 5 também colaboram para que o candidato seja classificado no vestibular. Esses atributos e sua pontuação são: redação (2,40 a 5,73), língua portuguesa (2,31 a 8,10), geografia (1,91 a 5,70) e química (4,91 a 6,90).

Juntamente com as notas da língua portuguesa e redação, as notas de matemática e química, em geral, influenciam na classificação ou aprovação no vestibular.

Observou-se, também, como regra relevante nos cursos mais concorridos, que os candidatos que obtiveram no ENEM nota na faixa 6 (4,208 a 8,969 pontos) ou acima alcançaram sucesso no vestibular.

A nota na disciplina de Física abaixo da faixa 3 (0,51 a 1,10 pontos), influenciou na eliminação dos candidatos na maioria dos cursos escolhidos entre os mais concorridos.

Nota-se, por exemplo, pela regra *NOTA_LEM = 6 AND NOTA_POR = 5 AND NOTA_QUI = 4 AND CURSO = 41: 96 (216.0/6.0)*, obtida na execução do algoritmo

[17] Alguns exemplos de regras encontram-se no Anexo D.

*J48.PART*, para o curso de Direito a nota de Química é fundamental, ou seja, 97% dos candidatos que obtiveram nota de Química na faixa 4 (1,31 a 4,90 pontos) juntamente com as notas de Língua Estrangeira na faixa 6 (5,31 a 9,80 pontos) e Língua Portuguesa na faixa 5 (2,31 a 8,10 pontos) foram eliminados por falta de vagas (status 96 = elim_vagas).

Pode-se concluir que o candidato que presta vestibular para o curso de Direito e obtém nota em Matemática, Química e Biologia acima da faixa 5 (4,91 a 6,30 pontos, 4,91 a 6,90 pontos, 2,31 a 6,90 pontos, respectivamente) é, geralmente, classificado e tem chances de ser aprovado após a análise da redação.

O candidato que obteve nota no ENEM na faixa 6 (4,208 a 8,969 pontos) e prestou vestibular para o curso de Comunicação Social, foi aprovado.

No curso de Comunicação Social, juntamente com as notas de Língua Estrangeira, Língua Portuguesa, Redação e História, a nota de Biologia influencia na aprovação.

Em geral, para o curso de Medicina por ser o mais concorrido, as notas obtidas nas provas são os principais fatores para a aprovação do candidato. Os fatores sócio-econômicos e culturais, ao contrário do que se propaga nas pesquisas do Inep, não aparecem como relevantes nos testes realizados. Nas pesquisas do Inep, os seguintes fatores são considerados determinantes no resultado do vestibular para o curso de Medicina: renda familiar, escolaridade dos pais, ter estudado em escola pública ou particular no período diurno ou noturno, ter estudado em curso pré-vestibular.

**2) Base de dados contendo somente dados dos candidatos ao curso de medicina**

Executando o algoritmo *J48.PART* na *WEKA* com a base de dados com dados apenas dos candidatos ao curso de Medicina, e da mesma forma utilizando todos os atributos anteriormente mencionados ou parte deles, foi constatado que, como já foi observado no conjunto de cursos mais concorridos, as notas são fatores

preponderantes para aprovação. Porém, tendo esta base de dados como entrada, apareceram algumas regras que envolvem os fatores sócio-econômicos e culturais do candidato. Grande parte dos candidatos aprovados moram com os pais, optaram pelo curso de medicina pela possibilidade de realização pessoal, prestaram vestibular por mais de uma vez, concluíram o Ensino Médio em 2003, estudaram em escola particular no período diurno, fizeram curso pré-vestibular e os pais têm como escolaridade o curso superior completo. Portanto, pode-se concluir que as condições sócio-econômicas interferem na aprovação dos candidatos ao curso de medicina.

**3) Base de dados contendo dados dos candidatos aos onze cursos menos concorridos**

As regras geradas através dos testes realizados com os onze cursos menos concorridos demonstram que a nota de Geografia, Redação, Língua Estrangeira e História, influencia na aprovação dos candidatos. No âmbito geral das regras, percebeu-se que as notas de matemática, química e física, não são determinantes para a aprovação.

O nível de escolaridade dos pais dos candidatos à vaga nos cursos menos concorridos aparece em algumas regras como importante, porém não é uma regra geral como ocorre nos cursos mais concorridos.

Um fator interessante nos cursos menos concorridos é que existe uma quantidade razoável de candidatos aprovados que concluíram o Ensino Médio há mais de 5 anos.

Nestes cursos é freqüente a aprovação ou classificação de candidatos provenientes da Escola Pública, ao contrário dos cursos mais concorridos, cuja relação não é freqüente.

O candidato ter feito ou não curso pré-vestibular parece não influenciar no resultado dos candidatos aprovados em cursos menos concorridos. Esta afirmação baseia-se no fato de terem sido obtidas várias regras onde tanto os candidatos que fizeram

curso pré-vestibular, quanto os que não fizeram, aparecem com as mesmas notas de Geografia e História e foram aprovados.

**4) Base de dados contendo dados de todos os candidatos ao vestibular**

Dentre as regras geradas utilizando a base de dados com dados de todos os candidatos ao vestibular, notou-se que a maioria dos candidatos reside com os pais, não trabalha, e está na faixa etária entre 17 e 20 anos.

Um conhecimento interessante extraído mostra que nos Cursos de Estatística, Matemática Industrial e Física, da Área de Exatas, as notas de Geografia, Língua Estrangeira e História influenciam na aprovação do candidato. Já no Curso de Ciências Sociais, como por exemplo, da Área Humanística, acontece ao contrário, ou seja, as notas de Matemática, Física e Química contribuem para a aprovação do candidato. Pode-se concluir que a pontuação obtida pelo candidato nas disciplinas da área do curso prestado no vestibular não tem tanta influência no resultado como a pontuação obtida nas disciplinas das outras áreas.

Outro conhecimento interessante extraído mostra que a pontuação obtida na prova de Língua Estrangeira, sempre em conjunto com as provas de Geografia e História, é relevante para a aprovação do candidato, independente do conjunto de atributos selecionado a cada execução do algoritmo *J48.PART*.

## 4.2 Considerações Finais

Na aplicação do algoritmo J48.PART da ferramenta WEKA pode-se constatar regras relevantes para futuras ações da UFPR, referente ao perfil dos candidatos que prestaram o processo seletivo do vestibular de janeiro de 2004.

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 determinante nos cursos menos concorridos.

A respeito do desempenho nas notas dos candidatos foi extraído conhecimento relevante, como por exemplo, os candidatos que prestaram vestibular para cursos da área de Exatas e foram aprovados, obtiveram as melhores pontuações nas disciplinas das outras áreas. As disciplinas que mais se destacaram foram Geografia, História e Biologia. Enquanto que na área Humanística acontece o contrário, ou seja, as notas de Matemática, Física e Química contribuem para a aprovação do candidato.

Percebeu-se, também, através das regras geradas pelo algoritmo J48.PART da ferramenta WEKA que a nota de Língua Estrangeira é um atributo importante no sucesso do candidato em praticamente todos cursos, independente da área.

A análise dos resultados obtidos através das regras geradas pela ferramenta em questão na base de dados do vestibular da UFPR para o ano letivo de 2004 facilitará aos gestores de Instituições do Ensino Superior, de cursos pré-vestibulares, de escolas particulares e das escolas públicas na implementação de ações pedagógicas e administrativas para melhorar o desempenho do alunado.

# 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 conhecimentos, são de extrema importância e exigem que a mesma atenção seja dispensada para cada uma delas. O sucesso de uma etapa depende exclusivamente do bom desenvolvimento das etapas anteriores.

O principal objetivo deste trabalho foi extrair conhecimentos interessantes com a finalidade de traçar o perfil dos candidatos ao processo seletivo do vestibular da 0UFPR, através da aplicação de técnicas de mineração de dados. Os resultados obtidos poderão ser utilizados pela UFPR 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.

É importante salientar que o presente trabalho poderá ser utilizado como um referencial para outros trabalhos relacionados à extração de conhecimento em banco de dados por apresentar os passos necessários para a aplicação de técnicas de mineração de dados, através de um exemplo prático.

Além disso, se os padrões dos dados mudarem significativamente ao longo do tempo, surgindo outros atributos na base de dados, o modelo proposto poderá ser revisto levando em conta a influência desses novos atributos.

Todas as etapas foram realizadas com seriedade e empenho no desejo de alcançar os principais objetivos e intensificar o apoio ao processo de tomada de decisão. A utilização de ferramentas de mineração de dados foi de extrema importância e

absolutamente interessante, demonstrando a riqueza de informações ocultas em bases de dados, reafirmando a necessidade do conhecimento.

O algoritmo *J48.PART* da ferramenta *WEKA*, 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. A ferramenta *WEKA* possibilitou uma boa observação e compreensão dos resultados por parte de todos os envolvidos na realização deste trabalho. Todos os resultados e conhecimentos obtidos foram considerados satisfatórios.

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 várias técnicas existentes para análise de dados tradicional, 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 complementam que quase tudo do que é feito com a mineração de dados poderia ser feito, eventualmente, com análises 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.

A mineração de dados é considerada uma tecnologia eficaz, capaz de lidar com grande volume de dados de uma forma mais eficiente que a estatística, conseguindo reconhecer padrões para fenômenos ditos complexos que necessitam de muitos parâmetros.

Para Cabena *et al*. (1998) e Pyle (1999) a estatística é orientada para verificar e validar hipóteses. A medida das técnicas estatísticas requer o desenvolvimento de uma hipótese prévia, ou seja, os estatísticos têm que desenvolver equações que

“casem” com as hipóteses. De um modo geral, a estatística tenta provar hipóteses sobre os dados e a mineração de dados tenta criar hipóteses.

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

- nos cursos mais concorridos, os dados sócio-econômicos e culturais do candidato são relevantes para o seu bom desempenho; o mesmo não ocorre nos cursos menos concorridos;
- as pontuações obtidas nas provas de outras áreas diferentes daquelas a qual o curso pertence influenciaram na aprovação dos candidatos a cursos das áreas de Exatas e Humanística;
- Percebeu-se, também, através das regras geradas, que a nota de Língua Estrangeira é um atributo importante no sucesso do candidato em praticamente todos cursos, independente da área.
- Notou-se que a nota de Redação, pelo fato de estar presente apenas nos registros dos candidatos classificados ou aprovados, prejudicou a análise das regras geradas.

Com base nos estudos do processo de *KDD*, o objetivo desta pesquisa foi extrair conhecimentos interessantes sobre os candidatos ao processo seletivo do vestibular da UFPR, tendo como base os dados cadastrais e sócio-econômicos e culturais dos candidatos. Nesta extração foram utilizadas a técnica de árvore de decisão e as regras de classificação aplicadas através da ferramenta *WEKA*.

Finalmente, o trabalho apresentado tem por objetivo contribuir para a análise do perfil e o desempenho do candidato ao vestibular da UFPR. Acredita-se que ele pode complementar no repertório das teorias utilizadas pelos profissionais responsáveis pela análise e definições de ações para melhorar o vestibular da UFPR.

## 5.2 SUGESTÕES PARA TRABALHOS FUTUROS

Como trabalhos futuros podem ser sugeridos:

- Aplicação de outras técnicas de mineração de dados sobre a base de dados utilizada neste trabalho;
- Estudo mais aprofundado dos recursos oferecidos pela ferramenta *WEKA*;
- Uso de outras bases de dados, como por exemplo, o IRA (Índice de Rendimento Acadêmico), na aplicação de técnicas de mineração de dados visando avaliar o desempenho dos acadêmicos pós-ingresso na universidade em relação aos resultados obtidos no vestibular e à situação sócio-econômicos e culturais do candidato ao vestibular.

# REFERÊNCIAS

ADRIAANS, P.; ZANTINGE, D. **Data mining**. Addison Wesley Longman, England, 1996.

AGRAWAL, R.; IMICLINSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. In: **Proccedings of the ACM SIGMOD conference**, Washington, D.C., May, 1993. Disponível na Internet: http://www.cs.bham.ac.uk/~anp/bibtex/kdd.bib.html Acessado em 20 de abril de 2004.

AGRAWAL, R.; SRIKANT, R. **Fast algorithms for mining association rules.** Proceedings of the 20th VLDB Conference. Chile: Santiago, 1994.

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

AGRAWAL, R.; GEHRKE, J.; GUNOPULOS, D.; RAGHAVAN, P. **Automatic subspace clustering of high dimensional data for data mining applications**. In: *SIGMOD Conference*, 1998, p. 94-105.

ARAÚJO, D. L. A.; LOPES, H. S.; FREITAS, A. A. A parallel genetic algorithm for rule Discovery in large databases. In: **IEEE systems, man and cybernetics conf**., v. III.Tokyo,1999, p. 940-945.

AZEVEDO, F.M.; BRASIL, L.M.; OLIVEIRA, R. C. L. **Redes neurais com aplicações em controle e em sistemas especialistas**. Visual Books, 2000.

BERRY, M. J. A.; LINOFF, G. **Data Mining techniques** – for marketing, sales, and customer support. United States: Wiley Computer Publishing, 1997.

BERSON, A.; SMITH, S.J. **Data Warehousing, Data Mining, and OLAP**. USA: McGraw-Hill, 1997.

BRAGA, A.P.; LUDERMIR, T.B.; CARVALHO, A.C.P.L.F. **Redes neurais artificiais**: teoria e aplicações. Livros Técnicos e Científicos Editora S.A., 2000.

BRASIL. Revista do Ensino Médio. Brasília, nº 4, ano II, 2004(a).

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

BUSINESS OBJECTS. **Introducing Business Miner:** business mining for decision support insights. White Paper, 1997.

CABENA, P.; HANDJINIAN, P.; STADLER, R.; VERHEES, J.; ZANASI, A. **Discovering data mining**: from concept to implementation. Upper Saddle River: Prentice-Hall PTR, 1998.

CABRAL JUNIOR, J. E.; COLMAN, R.; JORGE, R. P. **Paradigma simbólico de aprendizagem aplicado ao banco de dados do vestibular da UFMS**. 2002. Disponível em: www.dct.ufms.br/~mzanusso/producao/JedRodrRog.pdf. Acesso em 10 Mar. 2005.

CARVALHO, D. R. **Data mining através de introdução de regras e algoritmos genéticos**, 1999. Dissertação Mestrado – PUCPR, Curitiba.

_____. **Um método híbrido árvore de decisão / algoritmo genético para data mining**, 2002. Tese Doutorado – PUCPR, Curitiba.

_____. **Gestão pedagógica de cursos de graduação a partir de data mining**, 2004. Disponível em deborah@ipardes.gov.br acesso em: 20 de Ago. de 2004.

CARVALHO, L. A. V. de. **Data mining** – a mineração de dados no marketing, medicina, economia, engenharia e Administração. São Paulo: Érica, 2001.

CESAR - Centro de Estudos e Sistemas Avançados do Recife - Ano II - Número 32 - Maio/Junho de 2000. Disponível em: http://www.cesar.org.br/analise/n_32/fra_areas.html. Acesso em: 20 fev. de 2005.

CHEN, M. S.; HAN, J.; YU, P. S**. Data mining**: an overview from database perspective.TKDE 8(6). 1996.

COVER, T.M.; THOMAS, J.A. **Elements of information theory**. New York: John Wiley & Sons, Inc. 1991.

DataMind Technology Center. Agent network technology.

URL:http://datamindcorp.com/paper_agetnetwork.html, 1998.

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.

DILLY, R. **Data mining -** an introduction. Parallel Computer Centre - Queen's University of Belfast, 1995.
http://www.pcc.qub.ac.uk/tec/coursers/datamining/stu_notes/dm_book_2.html.
Acesso em 10 de Julho de 2004.

DWBrasil. Disponível em: http://www.dwbrasil.com.br/html/dmining.html. Acesso em: 27 Jun. 2004.

EIJKEL, G.C. Rule Induction, In Berthold, M., Hand, D.J., (Eds.), **Intelligent data analysis**, Berkeley, CA: Springer- Verlag. 1999, p.196-216.

EXCEL 2000. Planilha Eletrônica. Microsoft Corporation. Plataforma Windows.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery: an overview. In: **Advances in knowledge discovery and data mining**, AAAI Press / The MIT Press, MIT, Cambridge, Massachusets, and London, 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.

FREITAS, .A. On objective measures of rule surprisingness. **Principles of data mining & knowledge discovery** (Proc. 2nd European Symp., PKDD'98. Nantes, France, Sep. 1998). Lecture Notes in Artificial Intelligence 1510, 1-9. Springer-Verlag. 1998.

FREITAS, A.A. A genetic algorithm for generalized rule induction. In: R. Roy et al Advances in Soft Computing – Engineering Design and Manufaturing, 340-353.
(Proc. WSC3, 3rd On-Line World Conference on Soft Computing, hosted on the Internet, July 1998.) Springer-Verlag, 1999.

FREITAS JUNIOR, O. de G.; MARTINS, J. G.; RODRIGUES, A. M.; BARCIA, R. M. **Sistema de apoio à decisão usando a tecnologia data mining com estudo de caso da Universidade Estadual de Maringá**. I Congresso Brasileiro de Computação. Maringá, 2001.

GARCIA, S. C.. **O uso de árvore de decisão na descoberta de conhecimento na área da saúde**. 2000. Disponível em: www.inf.ufrgs.br/pos/SemanaAcadêmica/Semana 2000/Simone Garcia. Acesso em 10 de Out. de 2004.

GOEBEL, M.; GRUENWALD, L. **A survey of data mining and knowledge discovery software tools. In:** SIGKDD Explorations**, June 1999.**

GOLDBERG, D. E. **Genetic algorithms in search, optimization and machine learning**. Reading, MA: Addison Wesley, 1989.

GONÇALVES, A. L. **Utilização de técnicas de mineração de dados na análise dos grupos de pesquisa no Brasil**. 2000. Dissertação Mestrado em Engenharia de Produção - Engenharia de Produção e Sistemas UFSC. Florianópolis, Santa Catarina.

GROTH, R. **Data mining:** a hands-on approach for business professionals. Prentice Hall, New Jersey, 1998.

HAN, J.; KAMBER, M. **Data mining** – concepts and techniques. United States: Morgan Kaufmann Publishers, 2001.

HARRISON, T.H. **Intranet data warehouse**. Editora Berkeley, 1998.

HAYKIN, S. **Redes neurais**: princípios e prática. Bookman, 2001.

HELENE, O. Financiamento do Ensino Superior no Brasil. **SOS universidade pública** - reforma ou demolição? Revista da Associação dos Docentes da Unicamp, ano 6, nº 2, setembro de 2004, p. 108 – 114.

HILLIS, D.B. Using a genetic algorithm for multi-hypothesis tracking. In: **9th International Conference on Tools with Artificial Intelligence** (ICTAI'97), 1997.

HOLSHEIMER, M.; SIEBES, A. **Data mining the search for knowledge in databases, technical report CS-R9406.** CWI. Amsterdã, 1994.

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: www.ibge.gov.br. Acesso em 08 de dez de 2004.

IBM. Intelligent Miner.URL:http://scanner-group.mit.edu/DATAMINING/Datamining/ibm.html,1997.

Hycones Information Technology. AIRA – Data mining tool. URL:http://www.hycones.com.br/portuguese.aira.htm, 1998.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Disponível em: http://www.inep.gov.br/imprensa/noticias/censo/superior/ news 04-05-imp.htm. Acesso em 07 de Jan. de 2005.

INMON, W.H. **Como construir o data warehouse**. Editora Campus. 1997.

ISL Decision Systems Inc. Clementine – visual tools. URL:http://www.isl.co.uk/toolkit.html, 1996.

ISOFT SA. Alice 5.1. URL: http://www.alice-soft.com, 1998.

KUBAT, M.; BRATKO, I.; MICHALSKI, R.S. A Review of Machine Learning Methods, in Michalski, R.S., Bratko, I. and Kubat, M. (Eds.), **Machine learning and data mining**: methods and applications. London: John Wiley & Sons. 1998, p. 3-69.

LEMOS, E. P. **Análise de crédito bancário com o uso de data mining**: redes neurais e árvore de decisão. 2003. Dissertação Mestrado. Curitiba: UFPR.

LEVINE, R.; DRANG, D.E.; EDELSON, B. **Inteligência artificial e sistemas especialistas**. McGraw Hill, 1992.

LUBEL, K. S. **Data mining:** a new way to find answers. University of Maryland European Division, 1998. http://faculty.ed.umuc.ed/~jmeinke/inss690/lubel.htm. Acesso em 07 de Agosto de 2004.

Megaputer Intelligence Ltd. PolyAnalyst 3.5. http://www.megaputer.com, 1998.

MEHTA, M.; AGRAWAL, R.; RISSANEN, J.; SLICK. A fast scalable classifier for data mining. In: **Fifth In'tl conference on extending database technology**. Avignon, France, Mar, 1996.

MICHALEWICZ, Z. **Genetic algorithms + data structures = evolution programs**. Springer Verlag, 1994.

MITCHELL, T. M. **Machine Learning**. New York, United States of America: McGraw-Hill, 1997.

NeoVista Solutions Inc. Decision Series 3.0. http://www.neovista.com, 1998.

PANSANATO, 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. Disponível em: www.fac.ufmg.br/gama/artigos.html. Acesso em 20 de Jan. de 2005.

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

PYLE, Dorian. **Data preparation for data mining**. San Francisco: Morgan Kaufmann Publishers, 1999.

QUEIROZ, A. E. de M.; GOMES, A. S.; CARVALHO, F. de A. T. de. **Mineração de Dados de IHC para Interface Educativas,** 2002. Recife-PE. Disponível em: www.sbc.org.br/reic/edicoes/2002e4/cientificos/MineracaoDeDadosDeIHCParaInterfacesEducativas.pdf.

QUINLAN, J. R. **Learning efficient classification procedures and their application to chess end games**. In J. G. Carbonell, R. S. Michalski, and T. M. Mitchell, editors: Machine Learning, v. 1, Tioga, Palo Alto, USA, 1983.

_____. Simplifying decision trees. **International journal of man-machine studies**, 12. 1987, p. 221-234.

_____. *C4.5* **Programs for machine learning**. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

REZENDE, S. O. **Sistemas inteligentes**: fundamentos e aplicações. 1ª ed. Barueri: Editora Manole, 2003.

SANTOS, J.; HENRIQUES, N.; REIS, V. Universidade Nova de Lisboa - Faculdade de Ciências e Tecnologia 1999. Disponível em: http://students.fct.unl.pt/users/jms/dmdw/datamining/ Acesso em: 21 fev. 2005.

SHAFER, J.C.; AGRAWAL, R.; MEHTA, M.; SPRINT, A. **Scalable parallel classifier for data mining**, VLDB-96, Bombay, India,1996.

Silicon Graphics Computer Systems. MineSet. http://www.sgi.com/chembio/apps/datamine.html, 2000.

SILVA, G. **Estudo de técnicas de data mining e aplicação de uma das técnicas estudadas em uma base de dados da área da 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: www.est.ufmg.br/proav/proav.html. Acesso em 18 de Jan. de 2005.

SPSS. **Data mining and statistics** - gain a competitive advantage. Disponível em: http://www.spss.com/datamine/index.htm. Acesso em: 12 Jun. 2004.

Two Crows Corporation. **Introduction to data mining and knowledge discovery**. Third Edition, 1999.

UFPR. Disponível no site: www.ufpr.br

UFPR/PROGRAD/NC – Guia do Candidato / Processo Seletivo 2004. Disponível em: http://www.nc.ufpr.br. Acesso em 30 de Mar. 2004.

WEISS, S.M.; KULIKOWSKY, C.A., **Computer systems that learn**: Classificationn and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufman, 1991.

WEKA Data Mining System. Weka Experiment Environment. Disponível em: http://www.cs.waikato.ac.nz/weka. 1999. Acesso em: 10 Março de 2004.

WHITLEY, D. **A genetic algorithm tutorial**. Technical Report, 1993.

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

WORD 2000. Editor de Textos. Microsoft Corporation. Plataforma Windows.

WU, X. **Knowledge acquisition from databases**. USA: Ablex Publishing Corporation. 1995.

ZHANG, T.; RAMAKRISHNAN, R.; LIVNY, M. **BIRCH**: an efficient data clustering method for very large database. SIGMOD, 1996, p. 103-114.

# ANEXO A

## A1 Algumas considerações a respeito dos arquivos do processo seletivo do vestibular-2004 da UFPR:

Segue algumas observações a respeito das planilhas eletrônicas em excel tomadas como base de dados para esta pesquisa.

1. O total de candidatos é de 46531.
2. Os itens do sócio-educacional que constam do guia iniciam no item 03 do arquivo, sendo que os itens 01 e 02 são obtidos a partir das informações do sexo e data de nascimento contidas no cadastro geral dos candidatos.
3. As notas das provas estão numa escala de 0 a 1000, porém no desempenho do candidato elas aparecem na escala de 0 a 10. Basta dividir por 100 para obter o mesmo resultado. O mesmo serve para a média geral do candidato, que está na escala de 0 a 1000.
5. A nota do ENEM foi considerada somente como critério de desempate, não sendo incluída no computo da média do candidato. Esta nota está na escala de 0 a 100, conforme consta no cadastro do INEP (órgão que repassa esta informação), mas esta pesquisa considera a nota do ENEM na escala de 0 a 10, ou seja, basta dividir por 10 para obter o mesmo resultado. Os candidatos que optaram pela utilização da nota do ENEM como critério de desempate estão marcados no campo ENEM com 1, no arquivo com o cadastro geral.
6. No arquivo de cadastro segue a situação atual do candidato, considerando até a 7ª Chamada Complementar de 2004. Alunos desistentes, cancelados ou cancelados por resolução foram candidatos aprovados, seja na chamada geral ou em chamadas posteriores, conforme tabela A1.
7. Candidatos faltantes possuem notas em branco em uma ou mais provas, dependendo do dia em que faltaram. Existem casos em que os candidatos faltamno primeiro dia, vem no segundo e faltam no terceiro dia.

8**.** Será considerado desclassificado o candidato nas seguintes condições[18]:

-Se tiver média das oito provas objetivas inferior ao valor mínimo (Vmin) assim calculado: Vmin = $\overline{X}$ x 0,8, onde:

- Vmin = valor mínimo
- $\overline{X}$ = média aritmética das médias das notas dos candidatos presentes em cada uma das oito provas objetivas, na escala numérica de zero a dez, com precisão de milésimos.

-Se, mesmo tendo média igual ou superior ao valor mínimo, não se encontrar entre os primeiros colocados, até o limite de três vezes o número de vagas de seu curso.
9. Não foram corrigidas as provas de Redação dos candidatos desclassificados.

**A2 Tabela com o status da base de dados**:

A tabela A1 contém a descrição do atributo status (atributo-meta) com os itens considerados para este trabalho.

| STATUS | DESCRIÇÃO |
|---|---|
| 00 | CLASSIFICADO |
| 10 | APROVADO NA CHAMADA GERAL |
| 11 | APROVADO EM CHAMADA COMPLEMENTAR |
| 20 | DESISTENTE |
| 21 | CANCELADO |
| 22 | CANCELADO POR RESOLUÇÃO |
| 91 | FALTANTE |
| 92 | ZERADO |
| 93 | ELIMINADO |
| 96 | ELIMINADO VAGAS |

Tabela A1: Status da base de dados do vestibular da UFPR de 2004.
Fonte**:** UFPR/PROGRAD/NC – Guia do Candidato / Processo Seletivo 2004.

[18]Informações extraídas: UFPR/PROGRAD/NC – Guia do Candidato / Processo Seletivo 2004.

## A3 QUESTIONÁRIO SOCIOEDUCACIONAL

Os itens abaixo relacionados são provenientes do questionário sócio-educacional exigido pela UFPR no ato da inscrição para o vestibular de 2004. Os itens foram aproveitados na integra como atributos para base de dados deste trabalho.

**Item 1 - Qual o seu estado civil?**
|1| Solteiro(a)
|2| Casado(a)
|3| Outro

**Item 2 - Qual o Estado em que nasceu?**
|1| Paraná
|2| Santa Catarina Sul
|3| Rio Grande do Sul
|4| São Paulo
|5| Mato Grosso do Sul
|6| Outro

**Item 3 - Qual o local de sua residência?** (Trata-se de residência permanente e não temporária, para fins de estudo)
|1| Curitiba
|2| Demais municípios da Região Metropolitana
|3| Interior do Paraná
|4| Santa Catarina
|5| Rio Grande do Sul
|6| São Paulo
|7| Mato Grosso do Sul
|8| Outro

**Item 4 - Qual a sua situação quanto à moradia?**
|1| Mora em casa dos pais, quitada ou financiada
|2| Mora em casa dos pais, alugada
|3| Mora em casa própria, quitada ou financiada
|4| Mora em casa alugada, paga por você
|5| Mora em república, casa de estudante, pensão ou pensionato
|6| Mora em casa de parentes ou amigos
|7| Mora em casa alugada para você, paga por seus pais

**Item 5 - Qual o nível de instrução do seu pai?**
|1| Sem escolaridade
|2| Ensino fundamental incompleto
|3| Ensino fundamental completo
|4| Ensino médio incompleto
|5| Ensino médio completo
|6| Superior incompleto
|7| Superior completo
|8| Não sei informar

**Item 6 - Qual o nível de instrução da sua mãe**?
Responda conforme os quesitos do item anterior.

**Item 7 - Qual a principal ocupação de seu pai?**
|01| Funcionário público da administração direta ou indireta do governo Federal, Estadual ou Municipal
|02| Empregado de empresa comercial, industrial, bancária, agrícola ou prestadora de serviços
|03| Sócio ou proprietário de empresa comercial, industrial, bancária, agrícola ou prestadora de serviço
|04| Trabalho remunerado por conta própria, com auxílio de parentes e/ou de familiares
|05| Trabalho remunerado por conta própria, com empregados
|06| Artista (pintor, escultor, músico, cantor, ator, etc.)
|07| Trabalha em entidade, organização ou instituição não governamental de cunho filantrópico,

assistencial, religioso, de lazer ou outro
|08| Parlamentar ou cargo eleitoral, diplomata, militar
|09| Atleta profissional
|10| Trabalha em casa e/ou não tem atividade remunerada
|11| Não Trabalha
|12| Outros

Atenção: Se seu pai ou responsável for aposentado ou falecido, indicar a ocupação que exerceu a maior parte de sua vida.

**Item 8 - Qual a principal ocupação de sua mãe?**
Responda conforme os quesitos do item anterior.

**Item 9 - A renda total mensal de sua família se situa na faixa:**
|1| Até R$ 240,00
|2| De R$ 241,00 a R$ 500,00
|3| De R$ 501,00 a R$ 1.000,00
|4| De R$ 1.001,00 a R$ 1.500,00
|5| De R$ 1.501,00 a R$ 2.000,00
|6| De R$ 2.001,00 a R$ 3.000,00
|7| De R$ 3.001,00 a R$ 4.000,00
|8| De R$ 4.001,00 a R$ 5.000,00
|9| Acima de R$ 5.001,00

**Item 10 - Quantas pessoas contribuem para a obtenção da renda familiar?**
|1| Uma |4| Quatro
|2| Duas |5| Cinco
|3| Três |6| Seis ou mais

**Item 11 - Quantas pessoas são sustentadas com a renda familiar?**
|1| Uma |4| Quatro
|2| Duas |5| Cinco
|3| Três |6| Seis ou mais

**Item12 - Com que idade você começou a exercer atividade remunerada?**
|1| Antes dos 14 anos
|2| Entre 14 e 16 anos
|3| Entre 16 e 18 anos
|4| Após 18 anos
|5| Nunca trabalhei

**Item 13- Durante o curso, você terá obrigatoriamente que trabalhar?**
|1| Sim, mas apenas nos últimos anos
|2| Sim, desde o primeiro ano, em tempo parcial
|3| Sim, desde o primeiro ano, em tempo integral
|4| Não sei
|5| Não

**Item 14 - Como fez seus estudos do ensino fundamental?**
|1| Todos em escola pública
|2| Todos em escola particular
|3| Maior parte em escola pública
|4| Maior parte em escola particular
|5| Em escolas comunitárias/CNEC ou outro

**Item 15 - Em que ano você concluiu (ou concluirá) o curso do ensino médio?**
*Marque os dois últimos algarismos do ano de conclusão.

**Item 16 - Como fez seus estudos de ensino médio?**
|1| Integralmente em escola pública
|2| Integralmente em escola particular
|3| Maior parte em escola pública
|4| Maior parte em escola particular
|5| Em escolas comunitárias/CNEC ou outro

**Item 17 - Com relação à sua formação de ensino médio e sua atividade atual, você:**
|1| Concluiu o curso de magistério
|2| Concluiu outro curso técnico (agrícola, contábil, mecânico, etc.)
|3| Não se enquadra nas alternativas anteriores

**Item 18 - Em que turno você fez o curso de ensino médio?**
|1| Todo diurno
|2| Todo noturno
|3| Maior parte diurno
|4| Maior parte noturno
|5| Outro

**Item 19 - Você fez "terceirão" ou cursinho preparatório?**
|1| Fiz apenas o "terceirão"
|2| Fiz apenas cursinho
|3| Fiz "terceirão" e cursinho
|4| Não fiz nem "terceirão" nem cursinho

**Item 20 - Por quanto tempo você fez cursinho?**
|1| Por menos de um semestre
|2| Por um semestre
|3| Por um ano
|4| Por mais de um ano
|5| Não fiz cursinho

**Item 21 - Por que você fez cursinho?**
|1| Para atualizar meus conhecimentos, porque parei de estudar há muito tempo
|2| Para aprender "macetes"
|3| Para complementar os conhecimentos adquiridos no colégio
|4| Por outro motivo
|5| Não fiz cursinho

**Item 22 - Você já fez o vestibular em outros anos?**
(Não leve em conta a possível situação de "treineiro")
|1| Sim, este é o segundo ano que faço vestibular
|2| Sim, este é o terceiro ano que faço vestibular
|3| Sim, este é o quarto ano que faço vestibular
|4| Sim, faço vestibular há mais de quatro anos.
|5| Não, este é o primeiro ano em que faço vestibular.

**Item 23 - Você já iniciou algum curso superior?**
|1| Sim, mas não concluí
|2| Sim, estou cursando
|3| Sim, mas já concluí
|4| Não

**Item 24 - Qual o principal motivo que o levou a inscrever-se no Processo Seletivo da UFPR?**
|1| Por se tratar de universidade pública e gratuita
|2| Pela qualidade do ensino
|3| Pelo horário do curso que pretendo fazer
|4| Pela localização
|5| Por ser a única na cidade que oferece o curso que desejo
|6| Outro motivo

**Item 25 - Qual o motivo que o levou a escolher o curso para o qual está se candidatando?**
|1| Mercado de trabalho e possibilidades salariais
|2| Possibilidade de contribuir para a sociedade
|3| Possibilidade de realização pessoal
|4| Gosto pela profissão a que o curso me habilita
|5| Gosto pelas matérias do curso
|6| Baixa concorrência pelas vagas
|7| Permite conciliar aula e trabalho
|8| Outro motivo

**Item 26 - Quem ou o que mais o influenciou na escolha do curso?**
|1| A família
|2| Colegas e amigos
|3| Professor ou escola
|4| Teste vocacional
|5| Imprensa e televisão
|6| Outros

**Item 27 - Quando você se decidiu pelo curso a que está se candidatando?**
|1| Às vésperas da inscrição no processo seletivo
|2| Há alguns meses
|3| Há um ano ou pouco mais
|4| No início do ensino médio
|5| No ensino fundamental

**Item 28 - Quanto à sua escolha pelo curso, você se considera:**
|1| Absolutamente decidido |4| Indeciso
|2| Muito decidido |5| Muito indeciso
|3| Decidido

**Item 29 - O que você espera, EM PRIMEIRO LUGAR, de um curso universitário?**
|1| Aquisição de cultura geral ampla
|2| Formação profissional, voltada para o trabalho
|3| Formação teórica, voltada para a pesquisa
|4| Formação acadêmica para melhorar a atividade prática que já estou desempenhando
|5| Aquisição de conhecimentos que me permitam compreender melhor o mundo em que vivemos
|6| Aquisição de conhecimentos que permitam melhorar meu nível de instrução
|7| Diploma de nível superior

**Item 30 - Como você se informou sobre o Processo Seletivo 2003 da UFPR?**
|1| TV |6| Cartaz em ônibus
|2| Rádio |7| Outro cartaz
|3| Jornal |8| No colégio/cursinho
|4| Internet |9| Outros
|5| Folder

**Item 31 - A sua cor ou raça é?**
|1| Branca |4| Parda
|2| Preta |5| Indígena
|3| Amarela

## A4 CADASTRO GERAL DOS CANDIDATOS AO VESTIBULAR UFPR-2004

Os itens relacionados abaixo são referentes ao cadastro geral dos candidatos ao vestibular da UFPR-2004, dispostos numa planilha de excel. Foram considerados apenas os atributos usados neste trabalho, onde os atributos sobre as notas e a idade estão na forma discretizada e os demais mantidos na forma original, conforme mencionado no Capítulo III. Os atributos discretizados também foram categorizados para ficarem no padrão da base.

**1.Qual o seu sexo?**
1- Masculino
**2-** Feminino

**2.Qual a sua IDADE?**
1- Idade_15_a_15
2- Idade_16_a_16
3- Idade_17_a_17
4- Idade_18_a_20
5- Idade_21_a_23
6- Idade_24_a_27
7- Idade_28_a_32
8- Idade_33_a_40
9- Idade_41_a_48
10-Idade_49_a_53
11-Idade_54_a_59
12-Idade_60_a_71

**3.Qual o seu Curso?**
Relação em anexo, abaixo.

**4.Qual a língua estrangeira?**
1- Inglês
2- Francês
3- Alemão
4- Italiano
5- Espanhol

**5.Qual a NOTA_BIO**
1- Nota_BIO_0_a_0,90
2- Nota_BIO_0,91_a_1,30
3- Nota_BIO_1,31_a_1,70
4- Nota_BIO_1,71_a_2,30
5- Nota_BIO_2,31_a_6,90
6- Nota_BIO_6,91_a_8,70
7- Nota_BIO_8,71_a_9,40
8- Nota_BIO_9,41_a_10,00

**6.Qual a NOTA_QUI**
1- Nota_QUI_0_a_0,30
2- Nota_QUI_0,31_a_0,70
3- Nota_QUI_0,71_a_1,30
4- Nota_QUI_1,31_a_4,90
5- Nota_QUI_4,91_a_6,90
6- Nota_QUI_6,91_a_8,90
7- Nota_QUI_8,91_a_9,40
8- Nota_QUI_9,41_a_10,00

**8.Qual a NOTA_HIS**
1- Nota_HIS_0_a_0,30
2- Nota_HIS_0,31_a_0,70
3- Nota_HIS_0,71_a_1,50
4- Nota_HIS_1,51_a_5,10
5- Nota_HIS_5,11_a_6,10
6- Nota_HIS_6,11_a_7,30
7- Nota_HIS_7,31_a_8,10
8- Nota_HIS_8,11_a_8,60
9- Nota_HIS_8,61_a_10,00

**9.Qual a NOTA_MAT**
1- Nota_MAT_0_a_0,50
2- Nota_MAT_0,51_a_0,90
3- Nota_MAT_0,91_a_1,50
4- Nota_MAT_1,51_a_4,90
5- Nota_MAT_4,91_a_6,30
6- Nota_MAT_6,31_a_7,90
7- Nota_MAT_7,91_a_8,70
8- Nota_MAT_8,71_a_9,80
9- Nota_MAT_9,80_a_10,00

**10.Qual a NOTA_FIS**
1- Nota_FIS_0_a_0,10
2- Nota_FIS_0,11_a_0,50
3- Nota_FIS_0,51_a_1,10
4- Nota_FIS_1,11_a_3,70
5- Nota_FIS_3,71_a_4,90
6- Nota_FIS_4,91_a_6,70
7- Nota_FIS_6,71_a_8,90
**8-** Nota_FIS_8,91_a_10,00

**11.Qual a NOTA_GEO**
1- Nota_GEO_0_a_0,50
2- Nota_GEO_0,51_a_0,90
3- Nota_GEO_0,91_a_1,30
4- Nota_GEO_1,31_a_1,90
5- Nota_GEO_1,91_a_5,70
6- Nota_GEO_5,71_a_7,50
7- Nota_GEO_7,51_a_8,30
8- Nota_GEO_8,31_a_8,90
9- Nota_GEO_8,91_a_9,40
10-Nota_GEO_9,41_a_10,00

**12.Qual a NOTA_POR**

1- Nota_POR_0_a_0,10
2- Nota_POR_0,11_a_0,90
3- Nota_POR_0,91_a_1,50
4- Nota_POR_1,51_a_2,30
5- Nota_POR_2,31_a_8,10
6- Nota_POR_8,11_a_9,10
7- Nota_POR_9,11_a_9,80
8- Nota_POR_9,81_a_10,00

**13.Qual a NOTA_LEM**

1- Nota_LEM_0_a_0,10
2- Nota_LEM_0,11_a_0,50
3- Nota_LEM_0,51_a_0,90
4- Nota_LEM_0,91_a_150
5- Nota_LEM_1,51_a_5,30
6- Nota_LEM_5,31_a_9,80
7- Nota_LEM_9,81_a_10,00

**14.Qual a NOTA_RED**

1- Nota_RED_0_a_0,31
2- Nota_RED_0,32_a_0,73
3- Nota_RED_0,74_a_1,56
4- Nota_RED_1,57_a_2,39
5- Nota_RED_2,40_a_5,73
6- Nota_RED_5,70_a_6,56
7- Nota_RED_6,57_a_7,18
8- Nota_RED_7,19_a_7,81
9- Nota_RED_7,82_a_10,00

**15.Qual a NOTA_ENEM**

1- Nota_ENEM_0_a_0,08
2- Nota_ENEM_0,081_a_1,667
3- Nota_ENEM_1,668_a_2,302
4- Nota_ENEM_2,303_a_3,098
5- Nota_ENEM_3,099_a_4,207
6- Nota_ENEM_4,208_a_8,969
7- Nota_ENEM_8,97_a_9,445
8- Nota_ENEM_9,446_a_9,762
9- Nota_ENEM_9,763_a_9,921
10-Nota_ENEM_9,922_a_10,0

**16.Qual o STATUS?**

1- 00-Clasificado
2- 10-Aprov_Cham_Geral
3- 11-Aprov_Cham_Complem
4- 20-Desistente
5- 21-Cancelado
6- 22-Cancel_Resolução
7- 91-Faltante
8- 92-Zerado
9- 93-Eliminado
10- 96-Elim_Vagas

## A5 TABELA DOS CURSOS E VAGAS

A tabela A2 descreve os cursos oferecidos pela UFPR e suas respectivas vagas.

| Área | Curso | Turno | Nome do Curso | Vagas 1º Sem | Vagas 2º Sem | Total de Vagas |
|---|---|---|---|---|---|---|
| TECNOLÓGICA | 01 | MT | AGRONOMIA | 66 | 66 | 132 |
| | 02 | MT | ENGENHARIA FLORESTAL | 66 | 0 | 66 |
| | 03 | MT | ENGENHARIA ELÉTRICA (ELETRÔN., ELETROTÉC., TELECOM.) | 44 | 44 | 88 |
| | 04 | MT | ENGENHARIA MECÂNICA | 44 | 44 | 88 |
| | 07 | MT | ARQUITETURA E URBANISMO | 44 | 0 | 44 |
| | 10 | MT | ENGENHARIA CIVIL | 176 | 0 | 176 |
| | 11 | MT | ENGENHARIA QUÍMICA | 88 | 0 | 88 |
| | 59 | MT | ENGENHARIA INDUSTRIAL MADEIREIRA (3) | 60 | 0 | 60 |
| | 60 | MT | ENGENHARIA AMBIENTAL (3) | 45 | 0 | 45 |
| | 61 | MT | ENGENHARIA DE BIOPROCESSOS E BIOTECNOLOGIA (3) | 30 | 0 | 30 |
| EXATAS E DA TERRA | 05 | M | FÍSICA (Bacharelado) | 35 | 35 | 70 |
| | 06 | N | FÍSICA (Licenciatura) | 35 | 35 | 70 |
| | 08 | TN | BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO | 55 | 55 | 110 |
| | 12 | N | ESTATÍSTICA | 66 | 0 | 66 |
| | 13 | T | MATEMÁTICA (Bacharelado e Licenciatura) | 44 | 0 | 44 |
| | 14 | N | MATEMÁTICA (Licenciatura) | 44 | 0 | 44 |
| | 18 | MT | QUÍMICA | 66 | 0 | 66 |
| | 62 | T | MATEMÁTICA INDUSTRIAL (3) | 40 | 0 | 40 |
| | 09 | MT | ENGENHARIA CARTOGRÁFICA | 44 | 0 | 44 |
| | 15 | M | GEOGRAFIA | 30 | 0 | 30 |
| | 16 | N | GEOGRAFIA | 36 | 0 | 36 |
| | 17 | MT | GEOLOGIA | 33 | 0 | 33 |
| | 63 | MT | CIÊNCIAS DO MAR – Pontal do Paraná (3) | 30 | 0 | 30 |
| BIOLÓGICA | 19 | MTN | CIÊNCIAS BIOLÓGICAS | 50 | 50 | 100 |
| | 20 | MT | EDUCAÇÃO FÍSICA | 120 | 0 | 120 |
| | 21 | MT | ENFERMAGEM | 28 | 27 | 55 |
| | 22 | MT | FARMÁCIA | 54 | 54 | 108 |
| | 23 | MT | MEDICINA | 88 | 88 | 176 |
| | 24 | MT | MEDICINA VETERINÁRIA – CURITIBA | 48 | 0 | 48 |
| | 25 | MT | MEDICINA VETERINÁRIA – PALOTINA | 60 | 0 | 60 |
| | 26 | MT | NUTRIÇÃO | 33 | 33 | 66 |
| | 27 | MT | ODONTOLOGIA | 46 | 46 | 92 |
| | 64 | MT | ZOOTECNIA (3) | 45 | 0 | 45 |
| | 65 | MT | TERAPIA OCUPACIONAL (3) | 30 | 0 | 30 |
| HUMANÍSTICA | 28 | M | ADMINISTRAÇÃO | 55 | 0 | 55 |
| | 29 | N | ADMINISTRAÇÃO | 55 | 0 | 55 |
| | 30 | N | ADMINISTRAÇÃO INTERNACIONAL DE NEGÓCIOS | 55 | 0 | 55 |
| | 31 | M | GESTÃO DA INFORMAÇÃO (3) | 50 | 0 | 50 |
| | 32 | N | CIÊNCIAS CONTÁBEIS | 110 | 0 | 110 |
| | 33 | M | CIÊNCIAS ECONÔMICAS | 110 | 0 | 110 |
| | 34 | N | CIÊNCIAS ECONÔMICAS | 110 | 0 | 110 |
| | 35 | M | CIÊNCIAS SOCIAIS | 80 | 0 | 80 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | 36 | MN | COMUNICAÇÃO SOCIAL – JORNALISMO | 30 | 0 | 30 |
| | 37 | MN | COMUNICAÇÃO SOCIAL – PUBLICIDADE E PROPAGANDA | 30 | 0 | 30 |
| | 38 | MN | COMUNICAÇÃO SOCIAL – RELAÇÕES PÚBLICAS | 30 | 0 | 30 |
| | 39 | M | DESENHO INDUSTRIAL – PROGRAMAÇÃO VISUAL | 33 | 0 | 33 |
| | 40 | M | DESENHO INDUSTRIAL – PROJETO DO PRODUTO | 33 | 0 | 33 |
| | 41 | M | DIREITO | 84 | 0 | 84 |
| | 42 | N | DIREITO | 88 | 0 | 88 |
| | 43 | T | EDUCAÇÃO ARTÍSTICA – ARTES PLÁSTICAS | 16 | 0 | 16 |
| | 44 | T | EDUCAÇÃO ARTÍSTICA – DESENHO | 16 | 0 | 16 |
| | 67 | T | MÚSICA – PRODUÇÃO SONORA (Bacharelado) (3) | 20 | 0 | 20 |
| | 68 | T | MÚSICA – EDUCAÇÃO MUSICAL (Licenciatura) (3) | 20 | 0 | 20 |
| | 46 | M | FILOSOFIA (Bacharelado com Licenciatura Plena) | 75 | 0 | 75 |
| | 66 | N | FILOSOFIA (Bacharelado com Licenciatura Plena) | 50 | 0 | 50 |
| | 47 | T | HISTÓRIA | 60 | 0 | 60 |
| | 48 | M | LETRAS – INGLÊS OU PORTUGUÊS COM INGLÊS | 20 | 0 | 20 |
| | 49 | M | LETRAS – PORT. OU/COM ALEM OU/COM ITAL OU/COM GREGO OU/COM LATIM | 25 | 0 | 25 |
| | 71 | M | LETRAS – ESPANHOL OU PORTUGUÊS COM ESPANHOL | 25 | 0 | 25 |
| | 50 | N | LETRAS – PORTUGUÊS | 40 | 0 | 40 |
| | 51 | N | LETRAS – INGLÊS | 20 | 0 | 20 |
| | 72 | N | LETRAS – FRANCÊS | 10 | 0 | 10 |
| | 52 | M | PEDAGOGIA | 70 | 0 | 70 |
| | 53 | N | PEDAGOGIA | 100 | 0 | 100 |
| | 54 | MT | PSICOLOGIA | 80 | 0 | 80 |
| | 55 | N | TURISMO | 44 | 0 | 44 |
| ET | 69 | N | Tecnologia em Informática (1) | 50 | 0 | 50 |
| | 70 | T | Tecnologia em Informática (1) | 50 | 0 | 50 |
| CFO | 91 | MTN | Oficial Policial Militar (2) | 15 | 0 | 15 |
| | 92 | MTN | Oficial Policial Militar (feminino) (2) | 1 | 0 | 1 |
| | | | TOTAL: | 3583 | 577 | 4160 |

Tabela A2: Cursos oferecidos pela UFPR e números de vagas por cursos
Fonte: UFPR-Guia do Candidato. Site www.nc.ufpr.br

# ANEXO B

Para se finalizar a transformação do arquivo *DBASE* para o arquivo *ARFF*, basta carregar o arquivo em um editor de texto, adicionar o nome do banco de dados usando a expressão @*relation*, a informação dos atributos usando @*attribute* e uma linha @*data*. Em seguida, salvar o arquivo como somente texto. Para tal, foi utilizado o *software Word* (*Microsoft Corporation*, 2000). E, por último, renomear o arquivo .txt para .*arff*.

@*relation* base_final

@attribute SEXO {1, 2, N}
@attribute IDADE {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, N}
@attribute CURSO REAL
@attribute LINGUA {1, 2, 3, 4, 5, N}
@attribute NOTA_BIO {1, 2, 3, 4, 5, 6, 7, 8, N}
@attribute NOTA_QUI {1, 2, 3, 4, 5, 6, 7, 8, N}
@attribute NOTA_HIS {1, 2, 3, 4, 5, 6, 7, 8, 9, N}
@attribute NOTA_MAT {1, 2, 3, 4, 5, 6, 7, 8, 9, N}
@attribute NOTA_FIS {1, 2, 3, 4, 5, 6, 7, 8, N}
@attribute NOTA_GEO {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, N}
@attribute NOTA_POR {1, 2, 3, 4, 5, 6, 7, 8, N}
@attribute NOTA_LEM {1, 2, 3, 4, 5, 6, 7, N}
@attribute NOTA_RED {1, 2, 3, 4, 5, 6, 7, 8, 9, N}
@attribute NOTA_ENEM {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, N}
@attribute ITEM01 {1, 2, 3, N}
@attribute ITEM02 {1, 2, 3, 4, 5, 6, N}
@attribute ITEM03 {1, 2, 3, 4, 5, 6, 7, 8, N}
@attribute ITEM04 {1, 2, 3, 4, 5, 6, 7, N}
@attribute ITEM05 {1, 2, 3, 4, 5, 6, 7, 8, N}
@attribute ITEM06 {1, 2, 3, 4, 5, 6, 7, 8, N}
@attribute ITEM07 {01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, N}
@attribute ITEM08 {01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, N}
@attribute ITEM09 {1, 2, 3, 4, 5, 6, 7, 8, 9, N}
@attribute ITEM10 {1, 2, 3, 4, 5, 6, N}
@attribute ITEM11 {1, 2, 3, 4, 5, 6, N}
@attribute ITEM12 {1, 2, 3, 4, 5, N}
@attribute ITEM13 {1, 2, 3, 4, 5, N}
@attribute ITEM14 {1, 2, 3, 4, 5, N}
@attribute ITEM15 REAL
@attribute ITEM16 {1, 2, 3, 4, 5, N}

```
@attribute ITEM17 {1, 2, 3, N}
@attribute ITEM18 {1, 2, 3, 4, 5, N}
@attribute ITEM19 {1, 2, 3, 4, N}
@attribute ITEM20 {1, 2, 3, 4, 5, N}
@attribute ITEM21 {1, 2, 3, 4, 5, N}
@attribute ITEM22 {1, 2, 3, 4, 5, N}
@attribute ITEM23 {1, 2, 3, 4, N}
@attribute ITEM24 {1, 2, 3, 4, 5, 6, N}
@attribute ITEM25 {1, 2, 3, 4, 5, 6, 7, 8, N}
@attribute ITEM26 {1, 2, 3, 4, 5, 6, N}
@attribute ITEM27 {1, 2, 3, 4, 5, N}
@attribute ITEM28 {1, 2, 3, 4, 5, N}
@attribute ITEM29 {1, 2, 3, 4, 5, 6, 7, N}
@attribute ITEM30 {1, 2, 3, 4, 5, 6, 7, 8, 9, N}
@attribute ITEM31 {1, 2, 3, 4, 5, N}
@attribute STATUS {00, 10, 11, 20, 21, 22, 91, 92, 93, 96, N}

@data

1,4,29,1,5,4,4,4,4,6,5,6,5,N,1,1,1,1,5,5,03,02,4,3,3,3,3,4,02,4,3,1,3,4,3,1,4,1,1,2,2,3,2,8,1,00
2,5,39,5,5,5,4,4,5,5,5,6,5,N,1,1,1,1,3,2,02,10,5,3,5,3,2,1,99,1,3,1,2,4,3,4,4,5,4,2,4,1,4,4,1,00
2,4,23,1,N,N,N,N,N,N,N,N,N,N,1,6,8,5,7,5,05,12,6,2,5,5,5,4,02,2,3,1,2,3,3,1,4,1,3,4,5,1,2,4,3,91
2,4,10,5,4,3,4,3,2,5,4,5,N,N,1,1,2,1,5,5,02,11,3,3,5,3,2,1,02,1,3,3,4,5,5,1,4,1,4,6,4,1,2,4,1,93
1,7,06,5,5,4,4,4,3,5,5,6,3,N,1,2,4,4,4,3,02,04,4,3,3,1,3,3,94,3,3,2,4,5,5,1,4,2,4,6,5,1,5,4,1,00
2,5,08,5,5,5,4,4,5,3,5,6,N,N,1,1,1,1,4,2,02,11,4,2,5,4,2,1,98,1,2,1,2,2,3,4,4,2,3,6,4,1,4,8,1,96
1,4,23,5,5,4,4,4,5,5,5,6,N,6,1,1,1,1,7,7,10,01,3,2,3,1,2,3,02,3,3,1,3,4,3,2,4,1,2,6,2,3,3,8,1,96
2,4,41,1,5,4,4,4,4,5,5,5,N,N,1,1,1,2,7,6,02,02,4,2,5,2,2,2,02,2,3,1,3,3,3,2,4,1,3,4,2,1,1,3,3,96
2,5,67,5,5,4,4,4,4,5,5,5,N,N,1,1,2,1,2,2,12,10,4,1,4,5,4,1,00,1,3,1,2,3,1,1,4,5,4,6,3,1,4,4,1,93
2,6,23,1,5,5,4,4,4,5,5,5,N,N,1,1,1,1,5,2,03,10,5,1,4,5,5,2,99,2,3,1,3,4,3,4,4,2,4,6,5,1,1,8,1,96
. . . . . .
. . . . . .
. . . . . .
```

Figura B1: Arquivo *ARFF* para a base_final.

No exemplo seguinte será mostrada a seqüência passo a passo para transformar os arquivos *DBASE* em arquivo *ARFF*.

Primeiramente, o arquivo base_final foi carregado pelo *Microsoft Excel* 2000. Então, selecionou-se o item “Salvar Como” do menu Arquivo, na caixa de diálogo selecionou-se CSV (separado por vírgulas), digitou-se um nome para o arquivo e clicou-se no botão Salvar (uma mensagem informativa apareceu, bastando clicar no botão Sim para encerrar).

Em seguida, esse arquivo foi carregado pelo *Word* (*Microsoft Corporation*, 2000). As filas da planilha eletrônica original foram convertidas em linhas de texto e os elementos deveriam estar separados uns dos outros através de vírgulas. Entretanto, ao invés de vírgulas, apareceu ponto e vírgula entre os campos da tabela. Porém, sanar esse problema foi uma tarefa simples, bastou utilizar o item Localizar e Substituir do menu Editar no *Word* (*Microsoft Corporation*, 2000) para converter os ponto e vírgula em apenas vírgulas.

Devemos acrescentar o cabeçalho do arquivo *ARFF*. Temos então na primeira linha o nome do conjunto de dados atribuído pelo comando @ *relation nome_do_conjuto_de_dados*, em seguida temos a relação dos atributos, onde colocamos o nome do atributo e tipo ou seus possíveis valores, definido por *@attribute nome_do_atributo tipo ou {valores}*, na seção dos dados colocamos o comando *@data* e nas próximas linhas colocamos os registros onde cada linha representa um registro. Para visualizar melhor veja a Figura A1.

O passo seguinte é gravar o arquivo, para isso selecionamos no menu *Arquivo* a opção *Salvar como...* no menu *Salvar como tipo* selecione a opção *Somente texto com quebra de linha*, em *Nome do arquivo* digite o nome do arquivo com a extensão *arff* . Sugere-se renomear o arquivo para, por exemplo, base_final.*arff* para indicar que está no formato *ARFF*.

Após esta fase pode-se inicializar a analise dos dados usando o algoritmo apropriado, assumindo que já esteja instalado o *WEKA* e o *JAVA*.

# ANEXO C

Segue a tabela das freqüências de ocorrências dos 31 itens do sócio-educacional (dados sócio-econômicos e culturais) e mais 02 itens (os atributos sexo e idade) do cadastro geral dos candidatos da UFPR, bem como, a freqüência de ocorrência do desempenho do candidato no processo seletivo de 2004.

| **Item: 01** | **Qual o seu sexo?** | Inscritos | | **Aprovados** | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | **C** | **%** | **A** | **%** | **%+** |
| 1 | Masculino | 21455 | 46.11 | 02289 | 55.02 | 10.67 |
| 2 | Feminino | 25076 | 53.89 | 01871 | 44.98 | 07.46 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| **Item: 02** | **Quantos anos você completará em 2004?** | **Inscritos** | | **Aprovados** | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | **C** | **%** | **A** | **%** | **%+** |
| 15 | Menos de 16 anos | 00078 | 00.17 | 00007 | 00.17 | 08.97 |
| 16 | 16 anos | 01978 | 04.25 | 00134 | 03.22 | 06.77 |
| 17 | 17 anos | 13753 | 29.56 | 01362 | 32.74 | 09.90 |
| 18 | 18 anos | 09709 | 20.87 | 00943 | 22.67 | 09.71 |
| 19 | 19 anos | 05098 | 10.96 | 00455 | 10.94 | 08.93 |
| 20 | 20 anos | 03432 | 07.38 | 00267 | 06.42 | 07.78 |
| 21 | 21 anos | 02498 | 05.37 | 00193 | 04.64 | 07.73 |
| 22 | 22 anos | 01903 | 04.09 | 00154 | 03.70 | 08.09 |
| 23 | 23 anos | 01330 | 02.86 | 00083 | 02.00 | 06.24 |
| 24 | Mais de 23 anos | 06752 | 14.51 | 00562 | 13.51 | 08.32 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| **Item: 03** | **Qual o seu estado civil?** | **Inscritos** | | **Aprovados** | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | **C** | **%** | **A** | **%** | **%+** |
| 1 | Solteiro(a) | 42836 | 92.06 | 03874 | 93.13 | 09.04 |
| 2 | Casado(a) | 02745 | 05.90 | 00217 | 05.22 | 07.91 |
| 3 | Outro | 00950 | 02.04 | 00069 | 01.66 | 07.26 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| **Item: 04** | **Qual o Estado em que nasceu?** | **Inscritos** | | **Aprovados** | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | **C** | **%** | **A** | **%** | **%+** |
| 1 | Paraná | 35499 | 76.29 | 03084 | 74.13 | 08.69 |
| 2 | Santa Catarina | 02389 | 05.13 | 00227 | 05.46 | 09.50 |
| 3 | Rio Grande do Sul | 01006 | 02.16 | 00103 | 02.48 | 10.24 |
| 4 | São Paulo | 04257 | 09.15 | 00413 | 09.93 | 09.70 |
| 5 | Mato Grosso do Sul | 00418 | 00.90 | 00021 | 00.50 | 05.02 |
| 6 | Outro | 02962 | 06.37 | 00312 | 07.50 | 10.53 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 05 | Qual o local de sua residência? (Trata-se de residência permanente e não temporária, para fins de estudo) | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Curitiba | 28994 | 62.31 | 02985 | 71.75 | 10.30 |
| 2 | Demais municípios da Região Metropolitana | 06253 | 13.44 | 00420 | 10.10 | 06.72 |
| 3 | Interior do Paraná | 05820 | 12.51 | 00332 | 07.98 | 05.70 |
| 4 | Santa Catarina | 01608 | 03.46 | 00142 | 03.41 | 08.83 |
| 5 | Rio Grande do Sul | 00177 | 00.38 | 00023 | 00.55 | 12.99 |
| 6 | São Paulo | 01890 | 04.06 | 00140 | 03.37 | 07.41 |
| 7 | Mato Grosso do Sul | 00202 | 00.43 | 00010 | 00.24 | 04.95 |
| 8 | Outro | 01587 | 03.41 | 00108 | 02.60 | 06.81 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 06 | Qual a sua situação quanto à moradia? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Mora em casa dos pais, quitada ou financiada | 31291 | 67.25 | 02957 | 71.08 | 09.45 |
| 2 | Mora em casa dos pais, alugada | 04653 | 10.00 | 00373 | 08.97 | 08.02 |
| 3 | Mora em casa própria, quitada ou financiada | 04014 | 08.63 | 00278 | 06.68 | 06.93 |
| 4 | Mora em casa alugada, paga por você | 02074 | 04.46 | 00175 | 04.21 | 08.44 |
| 5 | Mora em república, casa de estudante, pensão ou pensionato | 00879 | 01.89 | 00064 | 01.54 | 07.28 |
| 6 | Mora em casa de parentes ou amigos | 02427 | 05.22 | 00195 | 04.69 | 08.03 |
| 7 | Mora em casa alugada para você, paga por seus pais | 01193 | 02.56 | 00118 | 02.84 | 09.89 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 07 | Qual o nível de instrução do seu pai? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Sem escolaridade | 00957 | 02.06 | 00063 | 01.51 | 06.58 |
| 2 | Ensino Fundamental incompleto | 08632 | 18.55 | 00506 | 12.16 | 05.86 |
| 3 | Ensino Fundamental completo | 03885 | 08.35 | 00252 | 06.06 | 06.49 |
| 4 | Ensino Médio incompleto | 03273 | 07.03 | 00206 | 04.95 | 06.29 |
| 5 | Ensino Médio completo | 10445 | 22.45 | 00815 | 19.59 | 07.80 |
| 6 | Superior incompleto | 03752 | 08.06 | 00404 | 09.71 | 10.77 |
| 7 | Superior completo | 13791 | 29.64 | 01790 | 43.03 | 12.98 |
| 8 | Não sei informar | 01796 | 03.86 | 00124 | 02.98 | 06.90 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 08 | Qual o nível de instrução da sua mãe? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Sem escolaridade | 01018 | 02.19 | 00060 | 01.44 | 05.89 |
| 2 | Ensino Fundamental incompleto | 08702 | 18.70 | 00531 | 12.76 | 06.10 |
| 3 | Ensino Fundamental completo | 04499 | 09.67 | 00290 | 06.97 | 06.45 |
| 4 | Ensino Médio incompleto | 03718 | 07.99 | 00259 | 06.23 | 06.97 |
| 5 | Ensino Médio completo | 11623 | 24.98 | 00935 | 22.48 | 08.04 |
| 6 | Superior incompleto | 03703 | 07.96 | 00405 | 09.74 | 10.94 |
| 7 | Superior completo | 12417 | 26.69 | 01620 | 38.94 | 13.05 |
| 8 | Não sei informar | 00851 | 01.83 | 00060 | 01.44 | 07.05 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 09 | Qual a principal ocupação do seu pai? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Funcionário público da administração direta ou indireta do governo Federal, Estadual | 06907 | 14.84 | 00728 | 17.50 | 10.54 |
| 2 | Empregado de empresa comercial, industrial, bancária, agrícola ou prestadora de serviços | 13903 | 29.88 | 01332 | 32.02 | 09.58 |
| 3 | Sócio ou proprietário de empresa comercial, industrial, bancária, agrícola ou prestadora de serviços | 07750 | 16.66 | 00782 | 18.80 | 10.09 |
| 4 | Trabalho remunerado por conta própria, com auxílio de parentes e/ou de familiares | 03901 | 08.38 | 00308 | 07.40 | 07.90 |
| 5 | Trabalho remunerado por conta própria, com empregados | 03977 | 08.55 | 00337 | 08.10 | 08.47 |
| 6 | Artista (pintor, escritor, músico, cantor, ator, etc.) | 00172 | 00.37 | 00024 | 00.58 | 13.95 |
| 7 | Trabalha em entidade, organização ou instituição não governamental de cunho filantrópico, assistencial, religioso, de lazer ou outro | 00320 | 00.69 | 00028 | 00.67 | 08.75 |
| 8 | Parlamentar ou cargo eleitoral, diplomata, militar | 00513 | 01.10 | 00036 | 00.87 | 07.02 |
| 9 | Atleta profissional | 00027 | 00.06 | 00006 | 00.14 | 22.22 |
| 10 | Trabalha em casa e/ou não tem atividade remunerada | 01009 | 02.17 | 00076 | 01.83 | 07.53 |
| 11 | Não trabalha | 01199 | 02.58 | 00085 | 02.04 | 07.09 |
| 12 | Outros | 06853 | 14.73 | 00418 | 10.05 | 06.10 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 10 | Qual a principal ocupação da sua mãe? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Funcionária pública da administração direta ou indireta do governo Federal, Est./Mun. | 08575 | 18.43 | 00931 | 22.38 | 10.86 |
| 2 | Empregada de empresa comercial, industrial, bancária, agrícola ou prest. de serviços | 07408 | 15.92 | 00693 | 16.66 | 09.35 |
| 3 | Sócia ou proprietária de empresa comercial, industrial, bancária, agrícola ou prestadora de serviços | 04006 | 08.61 | 00410 | 09.86 | 10.23 |
| 4 | Trabalho remunerado por conta própria, com auxílio de parentes e/ou familiares | 02364 | 05.08 | 00203 | 04.88 | 08.59 |
| 5 | Trabalho remunerado por conta própria, com empregados | 01612 | 03.46 | 00154 | 03.70 | 09.55 |
| 6 | Artista (pintora, escritora, música, cantora, atriz, etc.) | 00425 | 00.91 | 00053 | 01.27 | 12.47 |
| 7 | Trabalha em entidade, organização ou instituição não governamental de cunho filantrópico, assistencial, religioso, de lazer ou outro | 00425 | 00.91 | 00029 | 00.70 | 06.82 |
| 8 | Parlamentar ou cargo eleitoral, diplomata, militar | 00033 | 00.07 | 00002 | 00.05 | 06.06 |
| 9 | Atleta profissional | 00028 | 00.06 | 00001 | 00.02 | 03.57 |
| 10 | Trabalha em casa e/ou não tem atividade remunerada | 10507 | 22.58 | 00913 | 21.95 | 08.69 |
| 11 | Não trabalha | 06398 | 13.75 | 00448 | 10.77 | 07.00 |
| 12 | Outros | 04750 | 10.21 | 00323 | 07.76 | 06.80 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |
| | | | | | | |

| Item: 11 | A renda total mensal de sua família se situa na faixa: | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Até R$ 240,00 | 00923 | 01.98 | 00053 | 01.27 | 05.74 |
| 2 | De R$ 241,00 a R$ 500,00 | 03849 | 08.27 | 00161 | 03.87 | 04.18 |
| 3 | De R$ 501,00 a R$ 1.000,00 | 09475 | 20.36 | 00572 | 13.75 | 06.04 |
| 4 | De R$ 1.001,00 a R$ 1.500,00 | 08152 | 17.52 | 00629 | 15.12 | 07.72 |
| 5 | De R$ 1.501,00 a R$ 2.000,00 | 06494 | 13.96 | 00572 | 13.75 | 08.81 |
| 6 | De R$ 2.001,00 a R$ 3.000,00 | 06353 | 13.65 | 00705 | 16.95 | 11.10 |
| 7 | De R$ 3.001,00 a R$ 4.000,00 | 03834 | 08.24 | 00463 | 11.13 | 12.08 |
| 8 | De R$ 4.001,00 a R$ 5.000,00 | 02855 | 06.14 | 00364 | 08.75 | 12.75 |
| 9 | Acima de R$ 5.001,00 | 04596 | 09.88 | 00641 | 15.41 | 13.95 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 12 | Quantas pessoas contribuem para a obtenção da renda familiar? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Uma | 14911 | 32.05 | 01332 | 32.02 | 08.93 |
| 2 | Duas | 23404 | 50.30 | 02197 | 52.81 | 09.39 |
| 3 | Três | 05698 | 12.25 | 00459 | 11.03 | 08.06 |
| 4 | Quatro | 01834 | 03.94 | 00130 | 03.13 | 07.09 |
| 5 | Cinco | 00484 | 01.04 | 00032 | 00.77 | 06.61 |
| 6 | Seis ou mais | 00200 | 00.43 | 00010 | 00.24 | 05.00 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 13 | Quantas pessoas são sustentadas com a renda familiar? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Uma | 01252 | 02.69 | 00108 | 02.60 | 08.63 |
| 2 | Duas | 04155 | 08.93 | 00352 | 08.46 | 08.47 |
| 3 | Três | 08638 | 18.56 | 00722 | 17.36 | 08.36 |
| 4 | Quatro | 17288 | 37.15 | 01624 | 39.04 | 09.39 |
| 5 | Cinco | 10989 | 23.62 | 01036 | 24.90 | 09.43 |
| 6 | Seis ou mais | 04209 | 09.05 | 00318 | 07.64 | 07.56 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 14 | Com que idade você começou a exercer atividade remunerada? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Antes dos 14 anos | 02630 | 05.65 | 00175 | 04.21 | 06.65 |
| 2 | Entre os 14 e 16 anos | 07677 | 16.50 | 00514 | 12.36 | 06.70 |
| 3 | Entre os 16 e 18 anos | 08786 | 18.88 | 00638 | 15.34 | 07.26 |
| 4 | Após os 18 anos | 04889 | 10.51 | 00493 | 11.85 | 10.08 |
| 5 | Nunca trabalhei | 22549 | 48.46 | 02340 | 56.25 | 10.38 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |
| **Item: 15** | **Durante o curso, você terá obrigatoriamente que trabalhar?** | **Inscritos** | | **Aprovados** | | |
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Sim, mas apenas nos últimos anos | 02487 | 05.34 | 00282 | 06.78 | 11.34 |
| 2 | Sim, desde o primeiro ano, em tempo parcial | 11187 | 24.04 | 00912 | 21.92 | 08.15 |
| 3 | Sim, desde o primeiro ano, em tempo integral | 08167 | 17.55 | 00571 | 13.73 | 06.99 |
| 4 | Não sei | 15823 | 34.01 | 01552 | 37.31 | 09.81 |
| 5 | Não | 08867 | 19.06 | 00843 | 20.26 | 09.51 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 16 | Como fez seus estudos do Ensino Fundamental ? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Todos em escola pública | 21507 | 46.22 | 01299 | 31.23 | 06.04 |
| 2 | Todos em escola particular | 14619 | 31.42 | 02004 | 48.17 | 13.71 |
| 3 | Maior parte em escola pública | 05702 | 12.25 | 00422 | 10.14 | 07.40 |
| 4 | Maior parte em escola particular | 04593 | 09.87 | 00425 | 10.22 | 09.25 |
| 5 | Em escolas comunitárias/CNEC ou outro | 00110 | 00.24 | 00010 | 00.24 | 09.09 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 17 | Em que ano você concluiu (ou concluirá) o curso do Ensino Médio ? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 0 | 00 | 03231 | 06.94 | 00263 | 06.32 | 08.14 |
| 1 | 01 | 04479 | 09.63 | 00412 | 09.90 | 09.20 |
| 2 | 02 | 08631 | 18.55 | 00895 | 21.51 | 10.37 |
| 3 | 03 | 19844 | 42.65 | 01647 | 39.59 | 08.30 |
| 4 | 04 | 01871 | 04.02 | 00119 | 02.86 | 06.36 |
| 93 | Antes de 94 | 02266 | 04.87 | 00256 | 06.15 | 11.30 |
| 94 | 94 | 00391 | 00.84 | 00044 | 01.06 | 11.25 |
| 95 | 95 | 00486 | 01.04 | 00057 | 01.37 | 11.73 |
| 96 | 96 | 00646 | 01.39 | 00076 | 01.83 | 11.76 |
| 97 | 97 | 00922 | 01.98 | 00089 | 02.14 | 09.65 |
| 98 | 98 | 01417 | 03.05 | 00117 | 02.81 | 08.26 |
| 99 | 99 | 02347 | 05.04 | 00185 | 04.45 | 07.88 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 18 | Como fez seus estudos de Ensino Médio ? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Integralmente em escola pública | 19693 | 42.32 | 01301 | 31.27 | 06.61 |
| 2 | Integralmente em escola particular | 17832 | 38.32 | 02157 | 51.85 | 12.10 |
| 3 | Maior parte em escola pública | 05696 | 12.24 | 00473 | 11.37 | 08.30 |
| 4 | Maior parte em escola particular | 03073 | 06.60 | 00218 | 05.24 | 07.09 |
| 5 | Em escolas comunitárias/CNEC ou outro | 00237 | 00.51 | 00011 | 00.26 | 04.64 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 19 | Com relação à sua formação de Ensino Médio e sua atividade atual, você: | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Concluiu curso de magistério | 01840 | 03.95 | 00159 | 03.82 | 08.64 |
| 2 | Concluiu outro curso técnico (agrícola, contábil, mecânico, etc.) | 04457 | 09.58 | 00408 | 09.81 | 09.15 |
| 3 | Não se enquadra nas alternativas anteriores | 40234 | 86.47 | 03593 | 86.37 | 08.93 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 20 | Em que turno você fez o curso de Ensino Médio ? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Todo diurno | 31126 | 66.89 | 03226 | 77.55 | 10.36 |
| 2 | Todo noturno | 06710 | 14.42 | 00310 | 07.45 | 04.62 |
| 3 | Maior parte diurno | 05464 | 11.74 | 00421 | 10.12 | 07.71 |
| 4 | Maior parte noturno | 02715 | 05.83 | 00158 | 03.80 | 05.82 |
| 5 | Outro | 00516 | 01.11 | 00045 | 01.08 | 08.72 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 21 | Você fez "terceirão" ou cursinho preparatório? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Fiz apenas o terceirão | 10697 | 22.99 | 01239 | 29.78 | 11.58 |
| 2 | Fiz apenas cursinho | 14892 | 32.00 | 01438 | 34.57 | 09.66 |
| 3 | Fiz "terceirão" e cursinho | 07965 | 17.12 | 00878 | 21.11 | 11.02 |
| 4 | Não fiz nem "terceirão" nem cursinho | 12977 | 27.89 | 00605 | 14.54 | 04.66 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 22 | Por quanto tempo você fez cursinho ? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Por menos de um semestre | 03594 | 07.72 | 00299 | 07.19 | 08.32 |
| 2 | Por um semestre | 08841 | 19.00 | 00784 | 18.85 | 08.87 |
| 3 | Por um ano | 09007 | 19.36 | 01050 | 25.24 | 11.66 |
| 4 | Por mais de um ano | 03367 | 07.24 | 00410 | 09.86 | 12.18 |
| 5 | Não fiz cursinho | 21722 | 46.68 | 01617 | 38.87 | 07.44 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 23 | Por que você fez cursinho ? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Para atualizar meus conhecimentos, porque parei de estudar há muito tempo | 04691 | 10.08 | 00436 | 10.48 | 09.29 |
| 2 | Para aprender "macetes" | 00899 | 01.93 | 00080 | 01.92 | 08.90 |
| 3 | Para complementar os conhecimentos adquiridos no colégio | 16836 | 36.18 | 01711 | 41.13 | 10.16 |
| 4 | Por outro motivo | 02255 | 04.85 | 00305 | 07.33 | 13.53 |
| 5 | Não fiz cursinho | 21850 | 46.96 | 01628 | 39.13 | 07.45 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 24 | Você já fez vestibular em outros anos | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Sim, este é o segundo ano que faço vestibular | 14839 | 31.89 | 01776 | 42.69 | 11.97 |
| 2 | Sim, este é o terceiro ano que faço vestibular | 05478 | 11.77 | 00659 | 15.84 | 12.03 |
| 3 | Sim, este é o quarto ano que faço vestibular | 01747 | 03.75 | 00209 | 05.02 | 11.96 |
| 4 | Sim, faço vestibular há mais de quatro anos | 01135 | 02.44 | 00128 | 03.08 | 11.28 |
| 5 | Não, este é o primeiro ano em que faço vestibular | 23332 | 50.14 | 01388 | 33.37 | 05.95 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 25 | Você já iniciou algum curso superior? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Sim, mas não concluí | 03842 | 08.26 | 00520 | 12.50 | 13.53 |
| 2 | Sim, estou cursando | 02143 | 04.61 | 00243 | 05.84 | 11.34 |
| 3 | Sim, mas já concluí | 01092 | 02.35 | 00170 | 04.09 | 15.57 |
| 4 | Não | 39454 | 84.79 | 03227 | 77.57 | 08.18 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 26 | Qual o principal motivo que o levou a inscrever-se no Processo Seletivo da UFPR? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Por se tratar de universidade pública e gratuita | 26638 | 57.25 | 02250 | 54.09 | 08.45 |
| 2 | Pela qualidade do ensino | 16137 | 34.68 | 01509 | 36.27 | 09.35 |
| 3 | Pelo horário do curso que pretendo fazer | 00187 | 00.40 | 00021 | 00.50 | 11.23 |
| 4 | Pela localização | 00333 | 00.72 | 00032 | 00.77 | 09.61 |
| 5 | Por ser a única na cidade que oferece o curso que desejo | 00917 | 01.97 | 00117 | 02.81 | 12.76 |
| 6 | Outro motivo | 02319 | 04.98 | 00231 | 05.55 | 09.96 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 27 | Qual o motivo que o levou a escolher o curso para o qual esta se candidatando? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Mercado de trabalho e possibilidades salariais | 07231 | 15.54 | 00480 | 11.54 | 06.64 |
| 2 | Possibilidade de contribuir para a sociedade | 03510 | 07.54 | 00274 | 06.59 | 07.81 |
| 3 | Possibilidade de realização pessoal | 11578 | 24.88 | 01116 | 26.83 | 09.64 |
| 4 | Gosto pela profissão a que o curso me habilita | 18079 | 38.85 | 01495 | 35.94 | 08.27 |
| 5 | Gosto pelas matérias do curso | 02168 | 04.66 | 00338 | 08.13 | 15.59 |
| 6 | Baixa concorrência pelas vagas | 00419 | 00.90 | 00073 | 01.75 | 17.42 |
| 7 | Permite conciliar aula a trabalho | 00946 | 02.03 | 00102 | 02.45 | 10.78 |
| 8 | Outro motivo | 02600 | 05.59 | 00282 | 06.78 | 10.85 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 28 | Quem ou o que mais o influenciou na escolha do curso? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | A família | 11876 | 25.52 | 00845 | 20.31 | 07.12 |
| 2 | Colegas e amigos | 02969 | 06.38 | 00214 | 05.14 | 07.21 |
| 3 | Professor ou escola | 02373 | 05.10 | 00319 | 07.67 | 13.44 |
| 4 | Teste vocacional | 03242 | 06.97 | 00218 | 05.24 | 06.72 |
| 5 | Imprensa e televisão | 02012 | 04.32 | 00113 | 02.72 | 05.62 |
| 6 | Outros | 24059 | 51.71 | 02451 | 58.92 | 10.19 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 29 | Quando você se decidiu pelo curso a que está se candidatando? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Às vesperas da inscrição no processo seletivo | 07071 | 15.20 | 00877 | 21.08 | 12.40 |
| 2 | Há alguns meses | 13652 | 29.34 | 01397 | 33.58 | 10.23 |
| 3 | Há um ano ou pouco mais | 11168 | 24.00 | 01033 | 24.83 | 09.25 |
| 4 | No início do ensino médio | 09234 | 19.84 | 00558 | 13.41 | 06.04 |
| 5 | No ensino fundamental | 05406 | 11.62 | 00295 | 07.09 | 05.46 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 30 | Quanto à sua escolha pelo curso, você se considera: | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Absolutamente decidido | 21170 | 45.50 | 01645 | 39.54 | 07.77 |
| 2 | Muito decidido | 07722 | 16.60 | 00844 | 20.29 | 10.93 |
| 3 | Decidido | 14748 | 31.70 | 01332 | 32.02 | 09.03 |
| 4 | Indeciso | 02217 | 04.76 | 00231 | 05.55 | 10.42 |
| 5 | Muito Indeciso | 00674 | 01.45 | 00108 | 02.60 | 16.02 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 31 | O que você espera, EM PRIMEIRO LUGAR, de um curso universitário? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Aquisição de cultura geral ampla | 05241 | 11.26 | 00556 | 13.37 | 10.61 |
| 2 | Formação profissional, voltada para o trabalho | 29753 | 63.94 | 02323 | 55.84 | 07.81 |
| 3 | Formação teórica, voltada para a pesquisa | 01543 | 03.32 | 00261 | 06.27 | 16.92 |
| 4 | Formação acadêmica para melhorar a atividade prática que já estou desempenhando | 02093 | 04.50 | 00194 | 04.66 | 09.27 |
| 5 | Aquisição de conhecimentos que me permitam compreender melhor o mundo em que vivemos | 02864 | 06.16 | 00361 | 08.68 | 12.60 |
| 6 | Aquisição de conhecimentos que permitam melhorar meu nível de instrução | 03782 | 08.13 | 00353 | 08.49 | 09.33 |
| 7 | Diploma de nível superior | 01255 | 02.70 | 00112 | 02.69 | 08.92 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 32 | Como você se informou sobre o Processo Seletivo 2004 da UFPR | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | TV | 03623 | 07.79 | 00216 | 05.19 | 05.96 |
| 2 | Rádio | 00361 | 00.78 | 00028 | 00.67 | 07.76 |
| 3 | Jornal | 01526 | 03.28 | 00096 | 02.31 | 06.29 |
| 4 | Internet | 09311 | 20.01 | 00792 | 19.04 | 08.51 |
| 5 | Folder | 02142 | 04.60 | 00160 | 03.85 | 07.47 |
| 6 | Cartaz em ônibus | 02478 | 05.33 | 00155 | 03.73 | 06.26 |
| 7 | Outro cartaz | 00394 | 00.85 | 00046 | 01.11 | 11.68 |
| 8 | No colégio/cursinho | 20417 | 43.88 | 02086 | 50.14 | 10.22 |
| 9 | Outros | 06279 | 13.49 | 00581 | 13.97 | 09.25 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

| Item: 33 | A sua cor ou raça é? | Inscritos | | Aprovados | | |
|---|---|---|---|---|---|---|
| Resposta | Descrição | C | % | A | % | %+ |
| 1 | Branca | 38828 | 83.45 | 03558 | 85.53 | 09.16 |
| 2 | Preta | 01258 | 02.70 | 00071 | 01.71 | 05.64 |
| 3 | Amarela | 01802 | 03.87 | 00199 | 04.78 | 11.04 |
| 4 | Parda | 04396 | 09.45 | 00316 | 07.60 | 07.19 |
| 5 | Indígena | 00247 | 00.53 | 00016 | 00.38 | 06.48 |
| | | **46531** | **100.00** | **04160** | **100.00** | **08.94** |

Tabela C1: Freqüência de ocorrência aplicada nos atributos sócio-educacionais da base de dados do vestibular de 2004 da UFPR.
Fonte: Núcleo de Concursos da UFPR.

# ANEXO D

Este anexo apresenta algumas regras geradas através da aplicação do classificador *J48_PART* executado pela ferramenta *WEKA.*

- **Regras referentes ao curso de Bacharelado em Matemática:**

weka.classifiers.rules.PART -M 2 -C 0.25 -N 3 -Q 1
Relation:base_mat.bach_final
Instances: 220
Attributes: 46

| | | | |
|---|---|---|---|
| SEXO | NOTA_RED | ITEM11 | ITEM23 |
| IDADE | NOTA_ENEM | ITEM12 | ITEM24 |
| CURSO | ITEM01 | ITEM13 | ITEM25 |
| LINGUA | ITEM02 | ITEM14 | ITEM26 |
| NOTA_BIO | ITEM03 | ITEM15 | ITEM27 |
| NOTA_QUI | ITEM04 | ITEM16 | ITEM28 |
| NOTA_HIS | ITEM05 | ITEM17 | ITEM29 |
| NOTA_MAT | ITEM06 | ITEM18 | ITEM30 |
| NOTA_FIS | ITEM07 | ITEM19 | ITEM31 |
| NOTA_GEO | ITEM08 | ITEM20 | STATUS |
| NOTA_POR | ITEM09 | ITEM21 | |
| NOTA_LEM | ITEM10 | ITEM22 | |

NOTA_RED = N AND
NOTA_MAT = 4: 93 (72.0)

NOTA_HIS = 6: 10 (4.0/1.0)

NOTA_RED = 3: 00 (5.0)

NOTA_RED = N AND
NOTA_GEO = 5: 93 (5.0)

NOTA_GEO = 5 AND
NOTA_LEM = 5: 00 (56.0/12.0)

NOTA_RED = 5 AND
NOTA_QUI = 5: 10 (15.0/3.0)

NOTA_HIS = 5 AND
ITEM08 = 02: 20 (2.0)

NOTA_RED = 5 AND
NOTA_FIS = 4 AND

ITEM15 = 03 AND
ITEM04 = 1: 00 (7.0/1.0)

NOTA_RED = 4: 00 (4.0)

NOTA_RED = 5 AND
NOTA_FIS = 3: 00 (4.0)

NOTA_RED = 5 AND
ITEM07 = 02 AND
ITEM25 = 4: 10 (5.0/1.0)

NOTA_RED = 5 AND
ITEM09 = 4: 10 (7.0/2.0)

NOTA_RED = 5 AND
ITEM29 = 2 AND
ITEM25 = 5: 00 (4.0)

NOTA_RED = 5 AND
ITEM03 = 2: 00 (9.0/3.0)

NOTA_GEO = 6 AND
ITEM10 = 2: 20 (3.0)

NOTA_MAT = 4 AND
ITEM22 = 5: 10 (3.0)

NOTA_MAT = 4: 11 (3.0/1.0)

NOTA_HIS = 5: 10 (2.0/1.0)

- **Regras referentes aos cursos mais concorridos:**

weka.classifiers.rules.PART -M 2 -C 0.25 -N 3 -Q 1
Relation: base_+concorridos-weka.filters.unsupervised.attribute.Remove-R14-16,18,24-27,31,34-35,37-45
Instances: 17234
Attributes: 26

| | | | |
|---|---|---|---|
| SEXO | NOTA_MAT | ITEM05 | ITEM16 |
| IDADE | NOTA_FIS | ITEM06 | ITEM18 |
| CURSO | NOTA_GEO | ITEM07 | ITEM19 |
| LINGUA | NOTA_POR | ITEM08 | ITEM22 |
| NOTA_BIO | NOTA_LEM | ITEM09 | STATUS |
| NOTA_QUI | NOTA_RED | ITEM14 | |
| NOTA_HIS | ITEM03 | ITEM15 | |

NOTA_LEM = 5 AND
NOTA_QUI = 4 AND
ITEM15 = 03 AND
IDADE = 3 AND
ITEM09 = 6 AND
CURSO = 24: 96 (17.0/3.0)

NOTA_RED = 5 AND
NOTA_LEM = 6 AND
NOTA_POR = 5 AND
NOTA_GEO = 5 AND
NOTA_QUI = 4: 00 (129.0/4.0)

NOTA_LEM = 5 AND
NOTA_QUI = 4 AND
ITEM15 = 04 AND
ITEM18 = 1 AND
ITEM19 = 4 AND
LINGUA = 1 AND
ITEM03 = 1 AND
ITEM07 = 02 AND
NOTA_HIS = 4: 96 (18.0/5.0)

NOTA_RED = 5 AND
NOTA_GEO = 5 AND
NOTA_MAT = 4: 00 (113.0/13.0)

ITEM03 = 1 AND
NOTA_POR = 5 AND
NOTA_BIO = 5 AND
CURSO = 42 AND
NOTA_MAT = 5: 00 (31.0)

CURSO = 41 AND
NOTA_QUI = 5: 00 (50.0/7.0)

weka.classifiers.rules.PART -M 2 -C 0.25 -N 3 -Q 1
Relation:base_+concorridos-weka.filters.unsupervised.attribute.Remove-R13-18,24-26,31,34-35,38-45
Instances: 17234
Attributes: 26

| | | | |
|---|---|---|---|
| SEXO | NOTA_MAT | ITEM07 | ITEM18 |
| IDADE | NOTA_FIS | ITEM08 | ITEM19 |
| CURSO | NOTA_GEO | ITEM09 | ITEM22 |
| LINGUA | NOTA_POR | ITEM13 | ITEM23 |
| NOTA_BIO | NOTA_LEM | ITEM14 | STATUS |
| NOTA_QUI | ITEM05 | ITEM15 | |
| NOTA_HIS | ITEM06 | ITEM16 | |

NOTA_LEM = 4 AND
NOTA_HIS = 4 AND
NOTA_FIS = 4 AND
NOTA_MAT = 4 AND
NOTA_GEO = 5 AND
NOTA_BIO = 5 AND
NOTA_QUI = 4 AND
NOTA_POR = 5 AND
ITEM19 = 4: 93 (35.0/7.0)

NOTA_LEM = 5 AND
NOTA_POR = 3: 93 (101.0/5.0)

NOTA_LEM = 5 AND
NOTA_QUI = 5 AND
NOTA_GEO = 5 AND
NOTA_BIO = 5 AND
CURSO = 23: 96 (161.0)

NOTA_LEM = 5 AND
NOTA_GEO = 5 AND
NOTA_POR = 5 AND
NOTA_QUI = 5 AND
NOTA_BIO = 5 AND
NOTA_MAT = 4: 96 (175.0/10.0)

NOTA_LEM = 5 AND
NOTA_GEO = 5 AND
NOTA_POR = 4: 93 (109.0/8.0)

NOTA_LEM = 6 AND
NOTA_POR = 5 AND
NOTA_BIO = 5 AND
NOTA_GEO = 5 AND
NOTA_QUI = 4 AND
NOTA_HIS = 4 AND
NOTA_MAT = 4 AND

NOTA_FIS = 4 AND
CURSO = 42: 96 (220.0/9.0)

NOTA_LEM = 6 AND
NOTA_POR = 5 AND
NOTA_GEO = 5 AND
NOTA_BIO = 5 AND
NOTA_HIS = 4 AND
CURSO = 23: 96 (428.0/4.0)

NOTA_LEM = 6 AND
NOTA_POR = 5 AND
NOTA_MAT = 4 AND
CURSO = 23 AND
NOTA_BIO = 6: 96 (304.0/1.0)

NOTA_LEM = 6 AND
NOTA_POR = 5 AND
NOTA_GEO = 5 AND
CURSO = 23 AND
NOTA_QUI = 5: 96 (199.0)

NOTA_LEM = 6 AND
NOTA_POR = 5 AND
NOTA_QUI = 4 AND
CURSO = 41: 96 (216.0/6.0)

NOTA_LEM = 6 AND
NOTA_POR = 5 AND
NOTA_HIS = 4 AND
CURSO = 23 AND
NOTA_BIO = 5: 96 (189.0)

- **Regras referentes aos cursos menos concorridos:**

weka.classifiers.rules.PART -M 2 -C 0.25 -N 3 -Q 1
Relation: base_-concorridos-weka.filters.unsupervised.attribute.Remove-R14-15,18,24-27,31,34-45
Instances: 2536
Attributes: 26

| | | | |
|---|---|---|---|
| SEXO | NOTA_MAT | ITEM03 | ITEM15 |
| IDADE | NOTA_FIS | ITEM05 | ITEM16 |
| CURSO | NOTA_GEO | ITEM06 | ITEM18 |
| LINGUA | NOTA_POR | ITEM07 | ITEM19 |
| NOTA_BIO | NOTA_LEM | ITEM08 | STATUS |
| NOTA_QUI | NOTA_RED | ITEM09 | |
| NOTA_HIS | ITEM02 | ITEM14 | |

NOTA_RED = 4 AND
NOTA_GEO = 5: 00 (173.0/19.0)

NOTA_RED = N AND
NOTA_BIO = N: 91 (165.0)

NOTA_RED = N: 93 (775.0/45.0)

NOTA_LEM = 5 AND
NOTA_QUI = 4 AND
NOTA_HIS = 4 AND
CURSO = 33: 00 (85.0/3.0)

NOTA_GEO = 4: 00 (10.0/1.0)

NOTA_GEO = 7 AND
IDADE = 4: 10 (9.0/1.0)

NOTA_GEO = 7 AND
ITEM14 = 1: 10 (7.0)

NOTA_GEO = 7 AND
NOTA_QUI = 4: 10 (6.0/1.0)

NOTA_GEO = 8: 10 (11.0/3.0)

NOTA_GEO = 6 AND
NOTA_LEM = 7: 10 (5.0/1.0)

NOTA_GEO = 7: 20 (4.0)

NOTA_GEO = 6 AND
NOTA_LEM = 5 AND
ITEM02 = 2: 10 (3.0)

NOTA_GEO = 6 AND

NOTA_LEM = 5 AND

ITEM02 = 4 AND
ITEM07 = 02: 10 (3.0)

NOTA_MAT = 3: 00 (21.0/2.0)

NOTA_MAT = 2: 00 (5.0/1.0)

NOTA_MAT = 7: 10 (5.0)

NOTA_MAT = 8: 10 (4.0)

NOTA_MAT = 6 AND
NOTA_RED = 7 AND
NOTA_HIS = 6: 10 (3.0)

NOTA_MAT = 6 AND
NOTA_RED = 5: 10 (31.0/10.0)

NOTA_HIS = 6: 10 (42.0/11.0)

NOTA_MAT = 5 AND
NOTA_QUI = 5: 10 (46.0/5.0)

NOTA_HIS = 5 AND
NOTA_RED = 5 AND
NOTA_LEM = 6 AND
IDADE = 4: 10 (24.0/3.0)

NOTA_GEO = 5 AND
NOTA_RED = 3: 00 (49.0/6.0)

NOTA_RED = 2: 00 (10.0)

NOTA_HIS = 4 AND
NOTA_LEM = 5 AND
CURSO = 06 AND
NOTA_GEO = 5: 00 (40.0/2.0)
NOTA_HIS = 4 AND
NOTA_QUI = 5 AND
ITEM03 = 1 AND

NOTA_LEM = 6 AND
IDADE = 4: 10 (26.0/1.0)

NOTA_HIS = 4 AND
NOTA_LEM = 5 AND
NOTA_QUI = 4 AND
CURSO = 05: 00 (46.0/4.0)

NOTA_GEO = 6 AND
NOTA_RED = 6 AND
NOTA_MAT = 4: 10 (9.0/1.0)

NOTA_HIS = 5 AND
NOTA_LEM = 6: 10 (59.0/16.0)

NOTA_FIS = 6: 10 (42.0/11.0)

IDADE = 2: 20 (33.0/8.0)

NOTA_RED = 6 AND
ITEM02 = 1 AND
NOTA_FIS = 4 AND
NOTA_LEM = 6: 10 (7.0/1.0)

NOTA_RED = 5 AND
NOTA_GEO = 6 AND
ITEM02 = 1 AND
ITEM06 = 7: 10 (16.0/2.0)

NOTA_RED = 5 AND
CURSO = 72: 00 (18.0/3.0)

NOTA_RED = 5 AND
CURSO = 31 AND
ITEM03 = 2 AND
NOTA_LEM = 5 AND
NOTA_QUI = 4: 00 (5.0/1.0)

NOTA_RED = 5 AND
CURSO = 31 AND
ITEM03 = 1 AND
NOTA_HIS = 4 AND
ITEM02 = 1 AND
NOTA_LEM = 5 AND
NOTA_GEO = 5: 00 (25.0/3.0)

NOTA_RED = 5 AND
CURSO = 31 AND
ITEM03 = 1 AND
NOTA_FIS = 4 AND
LINGUA = 1: 10 (15.0/3.0)

NOTA_RED = 5 AND
CURSO = 31 AND
ITEM03 = 1 AND
NOTA_FIS = 4 AND
ITEM02 = 1 AND
SEXO = 2 AND
ITEM19 = 2: 10 (10.0/4.0)

NOTA_RED = 5 AND
NOTA_FIS = 3 AND
ITEM14 = 2: 00 (6.0/1.0)

NOTA_RED = 5 AND
NOTA_FIS = 3 AND
ITEM14 = 3: 00 (6.0)

NOTA_RED = 5 AND
NOTA_FIS = 3: 00 (22.0/8.0)

NOTA_RED = 5 AND
NOTA_FIS = 4 AND
CURSO = 05 AND
LINGUA = 5: 00 (19.0/1.0)

NOTA_RED = 6 AND
ITEM02 = 1: 00 (5.0)

NOTA_RED = 5 AND
CURSO = 31 AND
NOTA_FIS = 4 AND
ITEM03 = 1 AND
ITEM02 = 1: 00 (11.0)

NOTA_RED = 5 AND
CURSO = 31 AND
ITEM03 = 2: 10 (6.0)

NOTA_RED = 6: 10 (5.0/2.0)

CURSO = 31 AND
ITEM02 = 1: 10 (10.0/1.0)

NOTA_QUI = 3: 00 (15.0/4.0)

CURSO = 31 AND
ITEM02 = 4: 10 (4.0)

CURSO = 35 AND
NOTA_HIS = 4 AND
NOTA_LEM = 5: 00 (32.0)

ITEM15 = 92: 10 (7.0/2.0)

ITEM15 = 94: 10 (5.0/1.0)

ITEM15 = 90: 10 (4.0/1.0)

NOTA_POR = 4: 00 (7.0)

ITEM15 = 89 AND
IDADE = 7: 00 (3.0)

ITEM15 = 99 AND
NOTA_HIS = 4: 00 (46.0/12.0)

ITEM15 = 93 AND
NOTA_FIS = 4: 00 (9.0/2.0)

ITEM15 = 01 AND
NOTA_MAT = 5: 10 (8.0/1.0)

ITEM15 = 01 AND
ITEM03 = 2 AND
NOTA_HIS = 4: 00 (7.0)

ITEM15 = 95 AND
ITEM08 = 10 AND
NOTA_HIS = 4: 10 (3.0/1.0)

ITEM15 = 01 AND
ITEM03 = 1 AND
CURSO = 33: 00 (6.0/1.0)

ITEM15 = 01 AND
ITEM03 = 1 AND
NOTA_LEM = 5: 00 (14.0/3.0)

ITEM15 = 01 AND
ITEM03 = 1 AND
CURSO = 35: 00 (5.0)

ITEM15 = 01: 10 (22.0/7.0)

ITEM15 = 98 AND
ITEM16 = 1: 00 (17.0/6.0)

ITEM15 = 98: 10 (10.0/4.0)

ITEM15 = 88: 10 (9.0/4.0)

ITEM15 = 00 AND
ITEM02 = 1 AND
ITEM08 = 02: 00 (8.0)

ITEM15 = 95: 00 (5.0/1.0)

ITEM15 = 99 AND
CURSO = 06: 00 (3.0)

ITEM15 = 91 AND
ITEM16 = 1: 10 (3.0/1.0)

ITEM15 = 99: 11 (4.0/2.0)

ITEM15 = 85 AND
SEXO = 1: 10 (3.0/1.0)
ITEM15 = 97 AND
LINGUA = 5: 00 (13.0/5.0)

ITEM15 = 00 AND
ITEM02 = 1 AND
ITEM06 = 5: 00 (11.0/3.0)

ITEM15 = 96 AND
NOTA_GEO = 5: 00 (17.0/6.0)

ITEM15 = 03 AND
CURSO = 06 AND
NOTA_FIS = 4: 00 (6.0/1.0)

ITEM15 = 03 AND
ITEM02 = 6 AND
LINGUA = 1: 00 (10.0/1.0)

ITEM15 = 97: 10 (6.0/1.0)

ITEM15 = 96: 10 (4.0)

ITEM15 = 91: 00 (3.0)

ITEM15 = 03 AND
ITEM02 = 2 AND
ITEM19 = 1: 00 (2.0)

ITEM15 = 00 AND
ITEM03 = 1 AND
ITEM14 = 1: 00 (9.0/2.0)

ITEM15 = 03 AND
ITEM02 = 2: 10 (4.0/1.0)

ITEM15 = 03 AND
ITEM02 = 5: 00 (2.0/1.0)

ITEM15 = 03 AND
ITEM06 = 2: 00 (23.0/3.0)

ITEM15 = 02 AND
NOTA_HIS = 4 AND
NOTA_LEM = 5 AND
ITEM06 = 5: 00 (6.0/2.0)

ITEM15 = 02 AND
NOTA_HIS = 4 AND
NOTA_LEM = 6 AND
ITEM03 = 2 AND
ITEM18 = 1: 10 (7.0/1.0)

NOTA_HIS = 5: 10 (13.0/4.0)

ITEM15 = 02 AND
NOTA_LEM = 6 AND
ITEM03 = 1 AND
ITEM08 = 02: 11 (10.0/5.0)

ITEM15 = 02 AND
NOTA_LEM = 6 AND
LINGUA = 1: 10 (19.0/9.0)

ITEM15 = 02 AND
NOTA_LEM = 6: 00 (20.0/5.0)

ITEM15 = 03 AND
ITEM02 = 1 AND
ITEM18 = 1 AND
CURSO = 62 AND
LINGUA = 1: 10 (6.0)

ITEM15 = 03 AND
ITEM02 = 1 AND
ITEM18 = 1 AND
CURSO = 35: 00 (11.0/2.0)

ITEM15 = 02 AND
NOTA_LEM = 5 AND
ITEM14 = 1: 10 (7.0/1.0)

ITEM15 = 02 AND
ITEM18 = 1 AND
LINGUA = 1: 00 (8.0/1.0)

ITEM15 = 02 AND
ITEM16 = 1: 11 (3.0/1.0)

ITEM15 = 03 AND
ITEM03 = 2 AND
ITEM02 = 1: 10 (10.0/3.0)

ITEM15 = 86: 00 (6.0/3.0)

ITEM15 = 02: 10 (3.0/1.0)

ITEM15 = 03 AND
ITEM03 = 2: 11 (2.0)

ITEM15 = 03 AND
ITEM03 = 3 AND
ITEM06 = 4: 10 (2.0)

ITEM15 = 03 AND
ITEM03 = 3 AND
ITEM06 = 5: 11 (2.0)

ITEM15 = 03 AND
ITEM03 = 6 AND
SEXO = 1: 11 (3.0/1.0)
ITEM15 = 03 AND
ITEM03 = 1 AND
ITEM09 = 2: 00 (4.0/1.0)

ITEM15 = 03 AND
ITEM03 = 1 AND
CURSO = 33 AND
ITEM05 = 7: 00 (7.0)

ITEM15 = 03 AND
ITEM03 = 1 AND
NOTA_LEM = 6: 10 (24.0/11.0)

ITEM15 = 03: 00 (28.0/12.0)

IDADE = 5: 00 (5.0/1.0)

ITEM15 = 00 AND
IDADE = 4 AND
ITEM05 = 2: 11 (2.0)

ITEM06 = 2: 10 (7.0/2.0)

ITEM06 = 7: 10 (7.0/2.0)

ITEM06 = 4: 11 (4.0/2.0)

: 00 (4.0/1.0)

- **Regras referentes ao curso de Medicina:**

weka.classifiers.rules.PART -M 2 -C 0.25 -N 3 -Q 1
Relation: base_medicina_final
Instances: 5611
Attributes: 46

| | | | |
|---|---|---|---|
| SEXO | NOTA_RED | ITEM11 | ITEM23 |
| IDADE | NOTA_ENEM | ITEM12 | ITEM24 |
| CURSO | ITEM01 | ITEM13 | ITEM25 |
| LINGUA | ITEM02 | ITEM14 | ITEM26 |
| NOTA_BIO | ITEM03 | ITEM15 | ITEM27 |
| NOTA_QUI | ITEM04 | ITEM16 | ITEM28 |
| NOTA_HIS | ITEM05 | ITEM17 | ITEM29 |
| NOTA_MAT | ITEM06 | ITEM18 | ITEM30 |
| NOTA_FIS | ITEM07 | ITEM19 | ITEM31 |
| NOTA_GEO | ITEM08 | ITEM20 | STATUS |
| NOTA_POR | ITEM09 | ITEM21 | |
| NOTA_LEM | ITEM10 | ITEM22 | |

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_LEM = 6 AND
NOTA_BIO = 5 AND
NOTA_GEO = 5 AND
ITEM01 = 1 AND
NOTA_HIS = 4: 96 (998.0/13.0)

NOTA_RED = 5 AND
NOTA_QUI = 6 AND
NOTA_MAT = 6 AND
NOTA_LEM = 6 AND
NOTA_FIS = 7 AND
ITEM12 = 5 AND
ITEM23 = 4: 00 (54.0/5.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_BIO = 6: 96 (1038.0/2.0)

NOTA_RED = 5 AND
NOTA_MAT = 5: 00 (43.0)

NOTA_RED = 6 AND
NOTA_FIS = 6: 00 (31.0/6.0)

NOTA_RED = 6 AND
NOTA_FIS = 7 AND
NOTA_ENEM = 7: 10 (16.0/2.0)

NOTA_RED = 6 AND
NOTA_FIS = 8: 10 (9.0)

NOTA_RED = 6 AND
NOTA_MAT = 5: 00 (8.0)

NOTA_RED = 6 AND
ITEM03 = 1 AND
ITEM17 = 3 AND
NOTA_HIS = 5: 00 (11.0)

NOTA_RED = 8 AND
ITEM13 = 4: 10 (13.0/2.0)

NOTA_RED = 8 AND
ITEM22 = 1: 10 (4.0/1.0)

NOTA_RED = 8: 00 (8.0)

NOTA_RED = 7 AND
ITEM03 = 3 AND
ITEM26 = 6: 00 (6.0)

NOTA_RED = 7 AND
ITEM04 = 7 AND
ITEM27 = 3: 10 (3.0)

NOTA_RED = 7 AND
ITEM04 = 1 AND
ITEM03 = 1 AND
ITEM02 = 1 AND
NOTA_HIS = 7: 10 (8.0)

NOTA_RED = 5 AND
NOTA_QUI = 6 AND
NOTA_FIS = 6: 00 (35.0/4.0)

NOTA_RED = 6 AND
ITEM17 = 3 AND
ITEM03 = 1 AND
ITEM27 = 4 AND
ITEM04 = 1: 10 (9.0/1.0)

NOTA_RED = 6 AND
ITEM05 = 7 AND
ITEM30 = 8 AND
ITEM24 = 2: 10 (9.0/3.0)

NOTA_RED = 6 AND
ITEM04 = 1 AND
NOTA_BIO = 6: 00 (10.0)

NOTA_RED = 5 AND
NOTA_POR = 5 AND
NOTA_ENEM = 6: 00 (32.0/2.0)

NOTA_RED = 7 AND
ITEM04 = 1 AND
ITEM12 = 5 AND
ITEM15 = 02: 10 (8.0/1.0)

NOTA_RED = N AND
NOTA_BIO = N: 91 (452.0)

NOTA_RED = N AND
NOTA_LEM = 6 AND
NOTA_GEO = 6: 96 (535.0)

NOTA_RED = N AND
NOTA_POR = 6: 96 (153.0)

NOTA_RED = N AND
NOTA_POR = 7: 96 (31.0)

NOTA_RED = N AND
NOTA_POR = 3 AND
ITEM29 = 2: 93 (22.0)

NOTA_RED = N AND
NOTA_POR = 4 AND
NOTA_LEM = 5 AND
ITEM29 = 2 AND
ITEM30 = 8: 93 (26.0)

NOTA_RED = N AND
NOTA_POR = 4 AND
NOTA_LEM = 5 AND
NOTA_ENEM = N: 93 (40.0/1.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_BIO = 7: 96 (68.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_BIO = 4 AND
ITEM17 = 3 AND
NOTA_QUI = 4 AND
ITEM21 = 5: 93 (28.0/1.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_BIO = 4 AND
ITEM17 = 3 AND
ITEM31 = 1: 93 (15.0/2.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_GEO = 6: 96 (121.0/1.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_GEO = 7: 96 (30.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_GEO = 3 AND
NOTA_ENEM = N: 93 (12.0/1.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_GEO = 5 AND
NOTA_QUI = 5: 96 (249.0/1.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_GEO = 4 AND
NOTA_FIS = 4 AND
NOTA_HIS = 3: 93 (12.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_GEO = 5 AND
NOTA_QUI = 4 AND
NOTA_LEM = 6 AND
NOTA_MAT = 4 AND
ITEM31 = 1: 96 (87.0/4.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_GEO = 5 AND
NOTA_QUI = 6: 96 (39.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_GEO = 5 AND
NOTA_QUI = 3: 93 (34.0/7.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_GEO = 5 AND
NOTA_LEM = 6: 96 (32.0/5.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_GEO = 5 AND
NOTA_LEM = 3 AND
LINGUA = 1: 93 (9.0/1.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_GEO = 5 AND
NOTA_LEM = 5 AND
NOTA_MAT = 5: 96 (41.0/1.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_GEO = 4 AND
NOTA_HIS = 4 AND
NOTA_QUI = 4 AND
ITEM18 = 1 AND
ITEM22 = 5 AND
ITEM12 = 5 AND
ITEM15 = 03: 93 (15.0/3.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_GEO = 5 AND
NOTA_LEM = 5 AND
NOTA_HIS = 5: 96 (38.0/1.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_MAT = 3 AND
NOTA_LEM = 5 AND
NOTA_ENEM = N: 93 (28.0/2.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_HIS = 4 AND
NOTA_LEM = 5 AND
NOTA_QUI = 4 AND
NOTA_FIS = 5 AND
NOTA_ENEM = 6: 96 (25.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_MAT = 6: 96 (9.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_MAT = 2 AND
ITEM14 = 1: 93 (8.0)

NOTA_RED = N AND
NOTA_POR = 2: 93 (6.0)

NOTA_RED = N AND
NOTA_POR = 5 AND
NOTA_MAT = 5: 96 (4.0)

NOTA_RED = N AND
NOTA_POR = 1: 92 (3.0)

NOTA_RED = N AND
NOTA_LEM = 4: 93 (51.0/13.0)

NOTA_RED = N AND
NOTA_LEM = 5 AND
NOTA_GEO = 5 AND
NOTA_POR = 4 AND
ITEM14 = 1: 93 (8.0/1.0)

NOTA_RED = N AND
NOTA_LEM = 5 AND
NOTA_GEO = 5 AND
NOTA_FIS = 5 AND
NOTA_ENEM = N AND
ITEM16 = 2: 96 (18.0)

NOTA_RED = N AND
NOTA_LEM = 5 AND
NOTA_GEO = 5 AND
NOTA_POR = 5 AND
NOTA_BIO = 5 AND
NOTA_HIS = 4 AND
NOTA_FIS = 6: 96 (12.0)

NOTA_RED = N AND
NOTA_LEM = 5 AND
NOTA_GEO = 5 AND
NOTA_POR = 5 AND
NOTA_BIO = 5 AND
NOTA_HIS = 4 AND
NOTA_FIS = 5 AND
ITEM08 = 11: 96 (7.0)

NOTA_RED = N AND
NOTA_LEM = 5 AND
NOTA_GEO = 5 AND
NOTA_POR = 5 AND
NOTA_BIO = 5 AND
NOTA_HIS = 4 AND
NOTA_FIS = 5 AND
SEXO = 1: 96 (8.0)

NOTA_RED = N AND
NOTA_LEM = 2: 93 (6.0)

NOTA_RED = N AND
NOTA_LEM = 1: 92 (5.0)

NOTA_RED = N AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_QUI = 4 AND
NOTA_FIS = 4 AND
NOTA_ENEM = 6 AND
ITEM15 = 03 AND
ITEM18 = 1 AND
ITEM12 = 5 AND
ITEM26 = 1 AND
ITEM14 = 2: 96 (20.0)

NOTA_RED = N AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_QUI = 4 AND
NOTA_FIS = 4 AND
NOTA_ENEM = 6 AND
ITEM15 = 03 AND
ITEM22 = 5 AND
ITEM18 = 1 AND
ITEM12 = 5 AND
ITEM25 = 4 AND
LINGUA = 1: 96 (27.0)

NOTA_RED = N AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_QUI = 4 AND
NOTA_FIS = 4 AND
ITEM15 = 01 AND
ITEM04 = 1 AND
ITEM30 = 8: 96 (15.0)

NOTA_RED = N AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_QUI = 4 AND
NOTA_FIS = 4 AND
ITEM24 = 6: 96 (32.0/3.0)

NOTA_RED = N AND
NOTA_POR = 4 AND
ITEM19 = 2: 96 (5.0)

NOTA_RED = N AND
NOTA_POR = 4 AND
ITEM11 = 5: 93 (7.0)
NOTA_LEM = 5 AND
NOTA_MAT = 3 AND
LINGUA = 1: 93 (5.0)

NOTA_LEM = 3: 96 (3.0/1.0)
NOTA_LEM = 5 AND
NOTA_MAT = 1: 96 (3.0/1.0)

NOTA_LEM = 5 AND
NOTA_MAT = 2 AND
ITEM27 = 2: 96 (2.0)

NOTA_LEM = 5 AND
NOTA_MAT = 3: 96 (3.0/1.0)

NOTA_LEM = 5 AND
NOTA_MAT = 2 AND
ITEM27 = 4: 96 (2.0)

NOTA_LEM = 5 AND
NOTA_MAT = 4 AND
NOTA_QUI = 2: 93 (7.0/1.0)

NOTA_LEM = 5 AND
NOTA_MAT = 4 AND
NOTA_POR = 5 AND
NOTA_QUI = 4 AND
NOTA_GEO = 5 AND
NOTA_HIS = 3 AND
NOTA_FIS = 4 AND
ITEM28 = 1 AND
ITEM14 = 1 AND
ITEM16 = 1: 93 (7.0/1.0)

NOTA_LEM = 5 AND
NOTA_MAT = 4 AND
NOTA_POR = 5 AND
NOTA_QUI = 3: 93 (6.0)

NOTA_LEM = 5 AND
NOTA_MAT = 4 AND
NOTA_POR = 5 AND
NOTA_GEO = 5 AND
NOTA_HIS = 3 AND
ITEM07 = 01: 96 (4.0)

NOTA_LEM = 5 AND
NOTA_MAT = 4 AND
NOTA_POR = 5 AND
NOTA_HIS = 2: 93 (8.0/2.0)

NOTA_LEM = 5 AND
NOTA_MAT = 4 AND
NOTA_POR = 5 AND
NOTA_HIS = 4 AND
NOTA_FIS = 4 AND

ITEM15 = 97 AND
ITEM29 = 2: 96 (6.0)

NOTA_LEM = 5 AND
NOTA_MAT = 4 AND
NOTA_POR = 5 AND
NOTA_HIS = 4 AND
NOTA_FIS = 4 AND
NOTA_GEO = 5 AND
ITEM15 = 99 AND
ITEM20 = 4: 96 (7.0)

NOTA_LEM = 5 AND
NOTA_MAT = 4 AND
NOTA_POR = 5 AND
NOTA_HIS = 4 AND
NOTA_FIS = 4 AND
NOTA_GEO = 4 AND
ITEM16 = 1: 93 (8.0)

NOTA_LEM = 5 AND
NOTA_MAT = 4 AND
NOTA_POR = 5 AND
NOTA_GEO = 5 AND
NOTA_HIS = 3 AND
ITEM28 = 1 AND
ITEM10 = 1: 93 (8.0/1.0)

NOTA_LEM = 5 AND
NOTA_MAT = 4 AND
NOTA_POR = 5 AND
NOTA_GEO = 5 AND
NOTA_HIS = 3 AND
ITEM28 = 1: 96 (9.0/2.0)

NOTA_LEM = 5 AND
NOTA_MAT = 4 AND
NOTA_POR = 5 AND
NOTA_HIS = 4 AND
NOTA_FIS = 4 AND
ITEM31 = 4 AND
ITEM01 = 1: 96 (42.0/6.0)

NOTA_LEM = 5 AND
NOTA_MAT = 4 AND
NOTA_POR = 5 AND
NOTA_HIS = 4 AND
NOTA_FIS = 4 AND
ITEM15 = 04 AND
ITEM16 = 2 AND
NOTA_ENEM = N AND
LINGUA = 1: 96 (28.0/5.0)

NOTA_LEM = 5 AND
NOTA_MAT = 4 AND
NOTA_POR = 5 AND
NOTA_HIS = 3: 93 (16.0/2.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 5 AND
ITEM27 = 2: 96 (5.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 5 AND
ITEM27 = 5: 93 (5.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 4 AND
ITEM15 = 01: 96 (25.0/8.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 5: 96 (7.0/2.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 4 AND
NOTA_ENEM = 6 AND
ITEM31 = 1 AND
ITEM15 = 03 AND
ITEM26 = 6 AND
ITEM12 = 5 AND
ITEM04 = 1 AND
ITEM22 = 5 AND
ITEM06 = 7: 93 (6.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 4 AND
NOTA_ENEM = 6 AND
ITEM26 = 6 AND
ITEM15 = 03 AND
ITEM18 = 1 AND
ITEM04 = 1 AND
ITEM12 = 5 AND
ITEM06 = 7: 96 (7.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 4 AND
NOTA_ENEM = 6 AND

ITEM31 = 1 AND
ITEM15 = 03 AND
ITEM26 = 6 AND
ITEM24 = 2: 96 (7.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 4 AND
NOTA_ENEM = 6 AND
ITEM31 = 1 AND
ITEM15 = 03 AND
ITEM26 = 6: 96 (24.0/9.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 4 AND
NOTA_ENEM = 6 AND
ITEM31 = 1 AND
ITEM30 = 4: 96 (12.0/3.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 4 AND
NOTA_ENEM = 5 AND
ITEM15 = 02: 96 (9.0/2.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_ENEM = 5: 93 (10.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 4 AND
NOTA_ENEM = 6 AND
ITEM31 = 3: 93 (5.0/1.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 4 AND
NOTA_ENEM = 4: 93 (4.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 4 AND
ITEM15 = 00 AND
ITEM04 = 1 AND
IDADE = 4: 96 (9.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM15 = 04 AND
NOTA_GEO = 5: 93 (9.0/2.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM15 = 03 AND
NOTA_FIS = 4 AND
ITEM26 = 6: 96 (36.0/9.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM15 = 96: 93 (4.0/1.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM15 = 98 AND
ITEM30 = 4: 93 (4.0/1.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM15 = 01: 93 (3.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 3 AND
ITEM09 = 7: 96 (5.0/1.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 4 AND
ITEM15 = 05: 96 (3.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
NOTA_FIS = 4 AND
ITEM24 = 2 AND
ITEM29 = 1 AND
ITEM27 = 5: 96 (6.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 2 AND
IDADE = 8: 96 (6.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND

NOTA_POR = 5 AND
ITEM01 = 2 AND
ITEM08 = 11: 96 (5.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 2 AND
ITEM20 = 2: 93 (3.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 2 AND
ITEM19 = 4: 93 (3.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
ITEM23 = 4 AND
NOTA_GEO = 5 AND
NOTA_FIS = 4 AND
ITEM21 = 3 AND
LINGUA = 1 AND
ITEM31 = 1 AND
ITEM04 = 1 AND
ITEM07 = 02: 96 (10.0/2.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
ITEM23 = 3: 93 (6.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
NOTA_GEO = 4: 96 (4.0)

NOTA_LEM = 5 AND
NOTA_GEO = 5 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
NOTA_FIS = 4 AND
ITEM24 = 2 AND
ITEM29 = 2 AND
ITEM05 = 7: 96 (13.0/3.0)

NOTA_LEM = 5 AND
NOTA_GEO = 5 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
NOTA_FIS = 3 AND
NOTA_ENEM = 6: 96 (9.0/1.0)

NOTA_LEM = 5 AND
NOTA_GEO = 5 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
NOTA_FIS = 3 AND
ITEM22 = 5: 93 (7.0)

NOTA_LEM = 5 AND
NOTA_GEO = 5 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
NOTA_FIS = 4 AND
ITEM08 = 11 AND
ITEM02 = 1 AND
ITEM28 = 1: 93 (11.0)

NOTA_LEM = 5 AND
NOTA_GEO = 5 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
NOTA_FIS = 4 AND
ITEM24 = 2 AND
ITEM29 = 5: 96 (11.0/3.0)

NOTA_LEM = 5 AND
NOTA_GEO = 5 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
NOTA_FIS = 4 AND
ITEM08 = 05: 96 (9.0/2.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
NOTA_FIS = 4 AND
ITEM08 = 01 AND
ITEM04 = 1 AND
ITEM15 = 02: 93 (6.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
NOTA_FIS = 4 AND
ITEM25 = 3: 96 (29.0/11.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
NOTA_FIS = 4 AND
ITEM29 = 1 AND
SEXO = 1: 96 (8.0/1.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
ITEM29 = 2 AND
NOTA_FIS = 4 AND
ITEM14 = 4: 96 (12.0/2.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
ITEM29 = 1: 93 (7.0/2.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
ITEM29 = 2 AND
ITEM25 = 2 AND
ITEM13 = 4: 93 (6.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
ITEM23 = 4 AND
ITEM29 = 2 AND
NOTA_FIS = 4 AND
ITEM26 = 6 AND
ITEM14 = 1: 96 (9.0/2.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
ITEM23 = 4 AND
ITEM29 = 2 AND
NOTA_FIS = 4 AND
ITEM21 = 3 AND
ITEM07 = 03 AND
ITEM14 = 2: 96 (5.0)

NOTA_LEM = 5 AND
NOTA_HIS = 4 AND
NOTA_POR = 5 AND
ITEM01 = 1 AND
ITEM29 = 2 AND
ITEM21 = 3: 93 (22.0/5.0)

NOTA_FIS = 8 AND
ITEM18 = 1: 10 (22.0/4.0)

NOTA_FIS = 6: 00 (16.0/4.0)

NOTA_FIS = 4 AND
NOTA_HIS = 4 AND
NOTA_LEM = 5 AND
ITEM01 = 1 AND
ITEM27 = 3: 96 (10.0)

NOTA_FIS = 4 AND
NOTA_HIS = 4 AND
NOTA_LEM = 5 AND
ITEM01 = 1 AND
ITEM29 = 2 AND
ITEM18 = 3: 96 (5.0)

NOTA_FIS = 4 AND
NOTA_HIS = 4 AND
NOTA_LEM = 6: 96 (5.0/1.0)

NOTA_FIS = 5: 96 (4.0/1.0)

NOTA_FIS = 3: 93 (3.0/1.0)

NOTA_FIS = 4 AND
NOTA_MAT = 4 AND
NOTA_LEM = 5 AND
ITEM01 = 1 AND
ITEM29 = 2 AND
ITEM06 = 3: 93 (3.0)

NOTA_FIS = 4 AND
NOTA_MAT = 4 AND
NOTA_LEM = 5 AND
ITEM01 = 1 AND
ITEM12 = 5 AND
LINGUA = 5: 96 (5.0)

NOTA_FIS = 4 AND
NOTA_MAT = 4 AND
NOTA_POR = 5: 93 (28.0/8.0)

NOTA_FIS = 4 AND
NOTA_POR = 5: 91 (3.0)

NOTA_FIS = 7 AND
NOTA_QUI = 8 AND
LINGUA = 1: 10 (19.0/3.0)

NOTA_FIS = 7 AND
NOTA_GEO = 6 AND
ITEM03 = 3: 00 (14.0/2.0)

NOTA_FIS = 7 AND
NOTA_QUI = 5: 00 (12.0/1.0)

NOTA_FIS = 7 AND
IDADE = 5 AND
NOTA_GEO = 6: 00 (8.0)

NOTA_FIS = 7 AND
IDADE = 5 AND
NOTA_POR = 6: 10 (6.0/1.0)

NOTA_FIS = 7 AND
IDADE = 3 AND
ITEM04 = 1 AND
NOTA_GEO = 7: 10 (7.0)

NOTA_FIS = 7 AND
IDADE = 3 AND
ITEM28 = 1: 10 (12.0/3.0)

NOTA_FIS = 7 AND
ITEM03 = 6 AND
ITEM26 = 6: 00 (5.0)

NOTA_FIS = 7 AND
ITEM03 = 1 AND
ITEM12 = 5 AND
ITEM02 = 1 AND
NOTA_HIS = 5: 10 (10.0/1.0)

NOTA_FIS = 7 AND
NOTA_ENEM = 6 AND
ITEM31 = 1 AND
ITEM13 = 4: 00 (10.0)

NOTA_FIS = 7 AND
ITEM03 = 1 AND
NOTA_HIS = 4: 00 (8.0)
NOTA_FIS = 4 AND
SEXO = 2: 91 (3.0/1.0)
NOTA_FIS = 7 AND
NOTA_ENEM = 6 AND
NOTA_POR = 7: 10 (5.0)

NOTA_FIS = 7 AND
NOTA_RED = 6 AND
ITEM13 = 5: 10 (5.0/1.0)

NOTA_FIS = 7 AND
NOTA_RED = 7: 00 (5.0/1.0)

NOTA_RED = 5 AND
NOTA_QUI = 6: 00 (31.0/12.0)

NOTA_FIS = 7 AND
ITEM14 = 2 AND
ITEM28 = 1: 21 (5.0)

NOTA_LEM = 6 AND
NOTA_FIS = 7 AND
ITEM28 = 1 AND
ITEM10 = 2: 00 (3.0/1.0)

NOTA_LEM = 6: 00 (8.0/3.0)

SEXO = 2: 93 (3.0/1.0)