

CENTRO UNIVERSITÁRIO FEEVALE

CLÁUDIO AURÉLIO DA SILVA

APLICAÇÃO DE MINERAÇÃO DE DADOS EM UMA  
ACADEMIA DE MUSCULAÇÃO

Novo Hamburgo, junho de 2009.

CLÁUDIO AURÉLIO DA SILVA

APLICAÇÃO DE MINERAÇÃO DE DADOS EM UMA  
ACADEMIA DE MUSCULAÇÃO

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

Professor Orientador: Juliano Varella de Carvalho

Novo Hamburgo, junho de 2009.

## AGRADECIMENTOS

Primeiramente, quero agradecer aos meus pais, Márcia e Wanderlei, pela educação oferecida e por sempre me incentivarem a estudar. Ao meu irmão Bruno, pelo companheirismo e ajuda ao longo do curso.

À minha avó Ernestina, por ter sido a grande incentivadora e por ter me dado a oportunidade de descobrir novos horizontes, ainda na adolescência, me propiciando o acesso à tecnologia e fazendo com que isso se tornasse um marco em minha vida.

Aos familiares que sempre confiaram em mim e me deram total apoio nas decisões tomadas, meu muito obrigado.

Aos meus amigos e professores, que de uma forma ou de outra me auxiliaram nesta longa caminhada, com lições diárias de companheirismo e sabedoria.

Especialmente ao mestre e amigo Juliano Varella por todo o apoio e atenção dedicados ao longo de todo o curso e mais especificamente neste trabalho.

## RESUMO

As técnicas de mineração de dados, de maneira concisa, são utilizadas para auxiliar a extração de conhecimento. Algumas bases de dados não são aproveitadas da melhor maneira possível, pois informações desconhecidas, que podem se tornar importantes para as organizações, deixam de ser exploradas pelos *decision makers*. O software Weka, através de algoritmos de *data mining*, possibilita um aproveitamento dos dados com maior eficácia. A boa compreensão desse software, das técnicas e dos algoritmos, permite que o conhecimento sobre o ambiente em questão seja realizado de maneira mais completa e eficiente. Este trabalho visa aplicar algoritmos de mineração de dados em uma base de dados de uma academia de musculação, a fim de obter um melhor conhecimento sobre os dados dos clientes e das atividades em geral dentro desta organização.

Palavras-chave: Mineração de Dados. Descoberta de Conhecimento. Análise de Dados. Classificação. Weka.

## ABSTRACT

The techniques of data-mining, in a concise way, are used to assist the extraction of knowledge. Some databases are not exploited in the best possible way, because unknown information that may become important for organizations are no longer explored by decision makers. The Weka software through data mining algorithms, allows a more efficient use of data. The good understanding of this software, the techniques and algorithms, allows that the knowledge about the environment in question is held in a more complete and efficient way. This paper aims to apply data-mining algorithms on a database of an academy of weight training, to obtain a better understanding of the data of customers and activities in general within this organization.

Key words: Data mining, Knowledge Discovery, Data analysis, Classification Method, Weka.

## LISTA DE FIGURAS

Figura 1.1 – Etapas do Processo de KDD _____	16
Figura 1.2 – Interdisciplinaridade da Mineração de Dados _____	19
Figura 2.1 – Tipos de Nodos de uma Árvore de Decisão _____	25
Figura 2.2 – Estrutura de uma Rede Neural Simples _____	29
Figura 3.1 – Interface Principal do <i>Sipina</i> _____	34
Figura 3.2 – Interface do <i>QwikNet</i> _____	35
Figura 3.3 – Regras Geradas pelo <i>Sipina</i> _____	36
Figura 3.4 – Localização do município de Capixaba, Acre _____	38
Figura 3.5 – Prevalência da Esquistossomose em 197 Municípios de Minas Gerais _____	42
Figura 3.6 – Árvore de Decisão Gerada pelo Algoritmo J4.8 _____	44
Figura 4.1 – Metodologia Utilizada para Realização do Trabalho _____	47
Figura 4.2 – Relacionamento das Tabelas e Atributos _____	49
Figura 4.3 - Primeiro Arquivo de Texto Gerado _____	50
Figura 4.4 – Exemplo de Cabeçalho e Registros de um Arquivo arff _____	56
Figura 4.5 – Gráfico Gerado pelo <i>Software Weka</i> _____	61
Figura 4.6 – Árvore de Decisão Gerada pelo Algoritmo J48 _____	62
Figura 5.1 – Árvore de Decisão da Variação da Cintura _____	64
Figura 5.2 – Árvore de Decisão da Variação da Frequência Cardíaca _____	66
Figura 5.3 – Árvore de Decisão da Variação do Peso _____	68
Figura 5.4 – Árvore de Decisão da Variação da Pressão Sistólica _____	70
Figura 5.5 – Árvore de Decisão da Raça _____	72

## LISTA DE TABELAS

Tabela 3.1 – Matriz de Erros Obtida com o Algoritmo de Árvore de Decisão.....	40
Tabela 3.2 – Matriz de Erros Obtida com o Algoritmo J4.8 .....	43
Tabela 4.1 – Faixa de Variação do Peso.....	54
Tabela 4.2 – Faixas Etárias.....	55
Tabela 4.3 – Primeiros Arquivos Gerados e seus Atributos.....	57
Tabela 4.4 – Arquivos de Treinamento e Testes Gerados e seus Atributos .....	58
Tabela 4.5 – Faixas de Variação da Cintura Reformuladas .....	59
Tabela 4.6 – Faixas Etárias Reformuladas .....	60
Tabela 5.1 – Matriz de Confusão Gerada pelo Teste da Frequência.....	67
Tabela 5.2 – Matriz de Confusão Gerada pelo Teste do Peso.....	69
Tabela 5.3 – Códigos e Raças presentes na Base de Dados .....	71

## LISTA DE GRÁFICOS

Gráfico 1.1 – Percentagem de esforço para cada etapa do processo de KDD.....	21
--	----



## LISTA DE ABREVIATURAS E SIGLAS

DDD	Discagem Direta a Distância
DDI	Discagem Direta Internacional
DM	Data Mining
EUA	Estados Unidos da América
FC	Frequência Cardíaca
IBGE	Instituto Brasileiro de Geografia e Estatística
ID3	Idemized Dichotomizer 3
KDD	Knowledge Discovery Database
LMT	Logistic Model Tree
MLP	Perceptron Multi-Camadas
PD	Pressão Diastólica
PHP	Hypertext Preprocessor
PS	Pressão Sistólica
RNA	Rede Neural Artificial
SGBD	Sistema Gerenciador de Banco de Dados
SIG	Sistemas de Informação Geográficos
SR	Sensoriamento Remoto
TM	Thematic Mapper
WEKA	Waikato Environment for Knowledge Analysis

## SUMÁRIO

<b>INTRODUÇÃO</b>	<b>12</b>
<b>1 KNOWLEDGE DISCOVERY DATABASE</b>	<b>15</b>
1.1 Pré-Processamento	16
1.1.1 Seleção dos Dados	17
1.1.2 Limpeza dos Dados	17
1.1.3 Transformação dos Dados	17
1.2 Mineração de Dados	18
1.3 Pós-Processamento	20
<b>2 CLASSIFICADORES</b>	<b>23</b>
2.1 Árvores de Decisão	23
2.1.1 Histórico	24
2.1.2 Conceitos	24
2.1.3 Vantagens e Desvantagens	26
2.1.4 Algoritmo C4.5	27
2.1.5 Algoritmo LMT	28
2.2 Redes Neurais	28
2.2.1 <i>Perceptron</i> Multi-Camadas	31
<b>3 APLICAÇÕES DE MINERAÇÃO DE DADOS</b>	<b>33</b>
3.1 Análise do perfil do usuário de serviços de telefonia utilizando técnicas de mineração de dados	33
3.2 Avaliação da exatidão do mapeamento da cobertura da terra em Capixaba, Acre, utilizando classificação por árvore de decisão	37
3.3 Uso de árvore de decisão para predição da prevalência de esquistossomose no Estado de Minas Gerais, Brasil	41
<b>4 ESTUDO DE CASO NA ACADEMIA DE MUSCULAÇÃO</b>	<b>46</b>
4.1 Pré-Processamento	47
4.1.1 Seleção dos Dados	48
4.1.2 Limpeza dos Dados	52
4.1.3 Transformação dos Dados	53
4.2 Mineração de Dados	60
<b>5 ANÁLISE E DISCUSSÃO DOS RESULTADOS</b>	<b>63</b>
5.1 Relação Entradas - Faixa Etária - Variação da Cintura	64
5.2 Relação Entradas - Variação Frequência - Faixa Etária	65

5.3 Relação Variação Peso - Entradas - Faixa Etária _____	67
5.4 Relação Entradas - Variação Pressão Sistólica - Faixa Etária _____	69
5.5 Relação Entradas - Raça - Variação Cintura _____	71
<b>CONCLUSÃO</b> _____	<b>73</b>
<b>REFERÊNCIAS BIBLIOGRÁFICAS</b> _____	<b>75</b>

## INTRODUÇÃO

Em 1989 foi formalizado um termo para denominar o abrangente conceito de buscar conhecimento a partir de bases de dados: o KDD – *Knowledge Discovery Database*. Historicamente a Descoberta de Conhecimento em Bases de Dados foi consolidada a partir de várias disciplinas, sendo que entre elas pode-se citar a Estatística, a Inteligência Artificial, o Reconhecimento de Padrões e Banco de Dados (GOLDSCHMIDT, 2005).

KDD é um processo, de várias etapas, para descoberta de informações que sejam úteis e estejam implícitas nas grandes bases de dados. O exponencial aumento no volume de dados que não podem ser restaurados de uma maneira adequada pelos limites e capacidades de consultas dos SGBD's atuais, faz com que a utilização do KDD tenha uma grande importância na atualidade (BORGES, 2006).

O uso de técnicas para exploração de grandes quantidades de dados, a fim de descobrir padrões e relações, que demandariam grande trabalho por parte do ser humano, é definido como *Data Mining* (CARVALHO, 2005). A Mineração de Dados é a principal etapa do processo de KDD, onde efetivamente é feita a busca por conhecimentos que possam se tornar úteis no conjunto da aplicação da Descoberta de Conhecimento em Bases de Dados (GOLDSCHMIDT, 2005).

Para a solução de diversos casos que relatam os problemas de mau aproveitamento nas bases de dados, a aplicação de técnicas de mineração auxilia na abstração de conhecimento. Carvalho (2005) cita um exemplo de caso de segmentação de mercados, onde a venda foi otimizada através da análise e mineração na base de dados que contém os registros de venda, a fim de indicar quais as tendências de aquisição de determinados clientes, considerando vários aspectos. Também exemplifica a situação de uma empresa que utilizou as técnicas de redes neuronais artificiais para previsão de mercados financeiros e se tornou uma

das mais importantes no ramo, nos EUA, durante sete anos consecutivos, tendo crescimento de sua carteira, neste período, de 25% a 100% ao ano.

Existem diversas técnicas para Descoberta de Conhecimento em Bases de Dados, sendo que neste trabalho será utilizada a de Classificação. A técnica tem como objetivo classificar casos em classes distintas, levando em conta os atributos comuns a um determinado conjunto de objetos, pertencentes a uma base de dados. Esse modelo gerado possibilita a descoberta das classes de novos objetos a serem adicionados na base (SOUSA, 1998).

O processo de classificação é definido por dois passos, onde no primeiro as características dos dados formam um modelo de classificação, e no segundo, esse modelo criado é empregado com o objetivo de classificar novos objetos. Segundo Passini e Toledo (2002), a construção do modelo pode ser dividida em três fases: o treinamento, onde são estabelecidos parâmetros para se treinar o modelo, a fase de teste, onde a precisão do mesmo é testada com a aplicação de dados diferentes, e a aplicação, que é responsável pela execução da técnica.

Dessa forma, este trabalho visa extrair conhecimento da base de dados de uma academia de musculação, através do uso de técnicas de classificação, a fim de obter dados sobre os clientes, além de encontrar alguma relação das atividades dos alunos com a sua frequência na academia, evolução do peso e medidas dos alunos de acordo com faixa etária, sexo entre outras informações. Com isso, é possível criar perfis de alunos e focar determinados tipos de serviços de acordo com as características desses perfis.

A base disponibilizada para execução deste trabalho está gravada no *software* Microsoft Office Access, contendo 42 tabelas no total. Possui desde dados sobre vendas de produtos até informações sobre os clientes. São no total 6.993 membros (alunos da academia que possuem suas informações armazenadas na base de dados) desde a data da implementação do sistema, no ano de 1997.

É realizada uma avaliação periódica nos alunos, que mede seus pesos, alturas, cinturas, e várias outras características. É possível obter um controle da assiduidade de cada aluno para levantamento de dados, uma vez que cada acesso dos membros é gravado na tabela Entradas.

Auxiliar a tomada de decisão na academia, promovendo ações que resultem em um conhecimento mais profundo de seus clientes e, por consequência, permitir uma análise de tendências é uma motivação adicional para a realização deste trabalho.

Sendo assim, no capítulo 1 será abordado todo o processo de KDD, desde a seleção dos dados até o pós-processamento, onde o conhecimento obtido é analisado e fica pronto para aplicação, se o resultado for considerado satisfatório. O capítulo 2 tratará de dois tipos de classificadores: árvores de decisão e redes neurais. Também será abordado neste capítulo o funcionamento de alguns algoritmos utilizados por estas técnicas. No terceiro capítulo serão apresentados três estudos de casos a fim de demonstrar que as técnicas de mineração de dados vem sendo cada vez mais utilizadas em diversas áreas e que seus resultados são altamente satisfatórios. O capítulo 4 descreverá a metodologia do trabalho e o conteúdo de alguns arquivos gerados. O quinto e último capítulo contém a análise e discussão dos resultados obtidos.

## 1 KNOWLEDGE DISCOVERY DATABASE

Atualmente, em vários segmentos do mercado, o volume de dados cresce exponencialmente, gerando grandes problemas para as organizações. Os responsáveis por armazenar essas informações perdem o controle sobre o conteúdo armazenado, precisando encontrar formas para resolver esse problema. Quanto maior a base de dados, mais difícil fica a extração de algo que possa ser útil à empresa, caso as informações não estejam armazenadas de maneira correta.

Tudo o que está armazenado nas bases de dados pode ser visto de maneira superficial pelos proprietários de empresas. Contudo, alguns dados não são aproveitados da melhor maneira possível, pois informações desconhecidas, que podem se tornar importantes para as organizações deixam de ser exploradas pelas pessoas responsáveis. Sendo assim, mostra-se necessária a criação e utilização de tecnologias para o processo de recuperação de informações.

O termo em inglês *Knowledge Discovery Database* (KDD), utilizado para referenciar a descoberta de conhecimento em bases de dados, segundo Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 6), “é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados”. De um modo geral, a tarefa mais difícil do processo de KDD é perceber e interpretar corretamente vários fatos observáveis durante o processo, e a dificuldade em conjugar dinamicamente essas interpretações de forma a tomar a decisão de quais ações devem ser realizadas em cada caso (GOLDSCHMIDT, 2005).

Uma característica muito relevante que deve ser levada em consideração é que a descoberta do conhecimento não se dá exclusivamente por métodos e algoritmos, é necessário que exista a interferência humana a fim de delimitar quais são os níveis e interpretar se as

respostas geradas serão úteis dentro do contexto (GONCHOROSKI, 2007). O KDD é caracterizado como um processo formado por várias etapas operacionais, conforme representado na figura 1.1 (GOLDSCHMIDT, 2005).

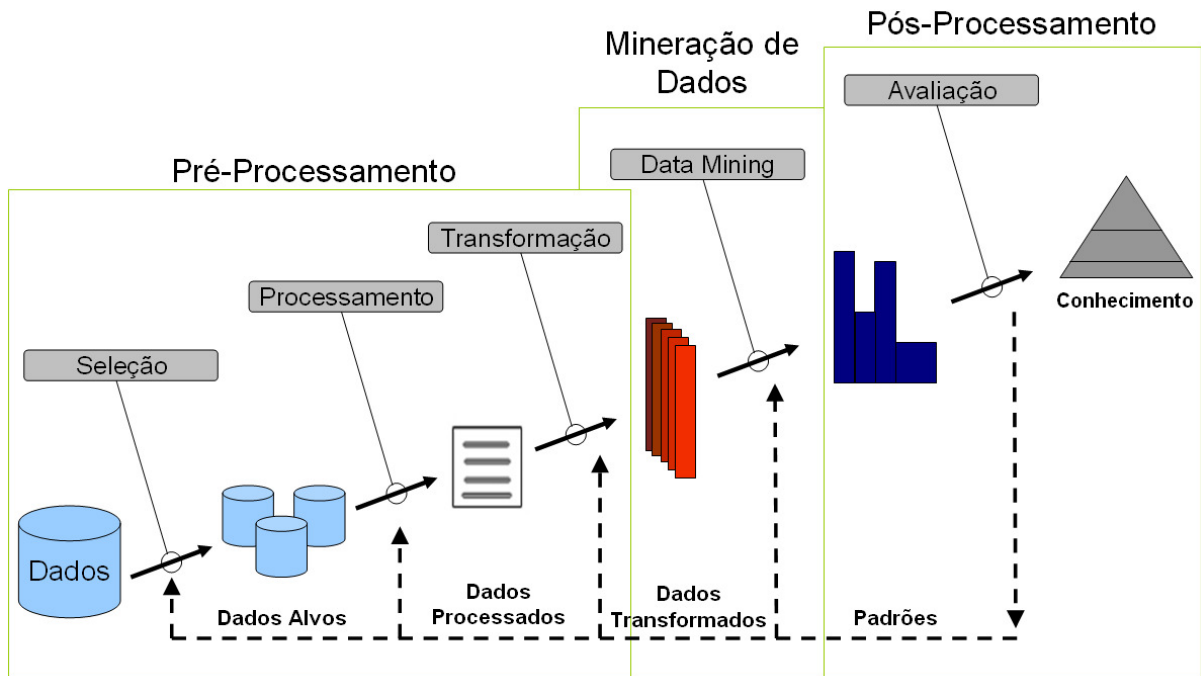


Figura 1.1 – Etapas do Processo de KDD  
Fonte: Adaptado de GONCHOROSKI, 2007, p. 30

De acordo com Borges (2006), o processo de KDD pode ser dividido em três grandes fases: o Pré-Processamento, que abrange todas as funções relacionadas à captação, organização e o tratamento dos dados; a Mineração de dados onde é feita a descoberta de padrões; e o Pós-Processamento que abrange os padrões e conhecimento descoberto. Em primeiro lugar, antes de executar as etapas do processo de KDD, é preciso que seja feita a análise e definição das metas a serem alcançadas com a extração do conhecimento no contexto da aplicação. É neste momento que são definidos itens importantes como a união entre o escopo de aplicação e a tecnologia KDD, visando ter a relação custo-benefício ao aplicar esta tecnologia (GONCHOROSKI, 2007).

### 1.1 Pré-Processamento

A etapa de pré-processamento possui um papel essencial no processo de descoberta de conhecimento. Nela é realizada desde a correção de dados obsoletos até a adequação da



formatação dos dados para os algoritmos de Mineração de Dados a serem empregados (GOLDSCHMIDT, 2005). Outro fator importante nesta etapa é a verificação de predominância de classes, sendo que uma vez constatada, é necessário excluir alguns dos registros da classe predominante ou adicionar registros de outras classes. Este processo objetiva balancear a base de dados de tal forma que, no processo do aprendizado, determinada classe não seja beneficiada, impedindo que o sistema fique tendencioso (ALMEIDA, 2003).

### **1.1.1 Seleção dos Dados**

Na etapa de seleção identificam-se as bases de dados e quais variáveis e tipos de dados serão extraídos na fase de Mineração de Dados. Por exemplo, alguns dados, como telefone, endereço e e-mail, poderiam ser descartados por não terem utilidade no contexto de associações entre compras de clientes. (CARVALHO, 2000). Esta etapa pode ter dois enfoques diferentes: a escolha de atributos ou a escolha de registros que devem ser levados em conta no processo de KDD (GOLDSCHMIDT, 2005.)

### **1.1.2 Limpeza dos Dados**

Presente na etapa de pré-processamento, a limpeza dos dados pode ser realizada utilizando o conhecimento do domínio. Pode-se localizar registros com valores nulos em algum atributo, granularidade incorreta ou exemplos errôneos, por exemplo. A limpeza pode também ser feita independente de domínio, como decisão da estratégia para tratamento de atributos incompletos, remoção de ruídos, entre outros (REZENDE, 2005). Os campos devem ser tratados por um analista que pode fazer interpolações, entrar códigos especiais nestes campos, ou simplesmente eliminar os registros com estas informações. Esta medida deve considerar o tipo de dados e seu impacto no processo de descoberta de conhecimento (BORGES, 2006).

### **1.1.3 Transformação dos Dados**

De acordo com Rezende (2005) algumas transformações comuns podem ser aplicadas aos dados, entre elas: resumo, onde dados sobre vendas, por exemplo, podem ser agrupados para formar relatórios diários; transformação de tipo: quando algum atributo é transformado em outro tipo de dados para ser aproveitado da melhor maneira possível por algum algoritmo de extração de padrões. Portanto, a principal finalidade desta etapa é

transformar os dados pré-processados, a fim de ajustá-los de acordo com a entrada de algum dos diversos algoritmos de Mineração de Dados (CARVALHO, 2000).

As próximas etapas do processo serão compostas pela própria Mineração dos Dados, onde será feita a escolha de quais algoritmos serão utilizados e, por fim, o pós-processamento, onde é realizada a análise dos resultados obtidos de modo a verificar se o conhecimento adquirido será útil ao contexto proposto. É no momento da análise que é feita a constatação se será necessário que o processo seja reiniciado, caso o resultado não esteja dentro dos objetivos definidos inicialmente (GONCHOROSKI, 2007).

A principal etapa do processo de KDD, chamada de Mineração de Dados, é composta por diversas técnicas conhecidas como seus algoritmos, com uma grande complexidade. Sendo assim o processo e algumas dessas técnicas serão abordadas a seguir.

## **1.2 Mineração de Dados**

Na atualidade, sistemas que usufruem do potencial de Mineração de Dados têm sido utilizados em vários ramos do mercado, com intuito de extrair conhecimento de grandes bases de dados. Qualquer algoritmo que gere um padrão a partir de dados é um algoritmo de Mineração de Dados. (CARVALHO, 2000)

O processo de Mineração de Dados baseia-se na interação entre várias classes de usuários, e grande parte do seu sucesso depende dessa interação. Existem três classes diferentes nas quais podem ser divididos os usuários deste processo: especialista do domínio, que deve oferecer apoio para a execução do processo e possuir grande conhecimento do domínio da aplicação; analista, que deve conhecer profundamente todas as etapas que fazem parte do processo e é o usuário especialista no processo de extração de conhecimento; e o usuário final, que utiliza o conhecimento obtido no processo para a tomada de decisão (REZENDE, 2005).

De acordo com Borges (2006) a Mineração de Dados ou *Data Mining* (DM) é a principal etapa do processo KDD, que tem como finalidade extrair padrões dos dados. Esta fase é considerada o centro de todo o processo onde a maior preocupação é ajustar modelos ou determinar padrões de acordo com os dados observados. Pode ser vista também como uma maneira de selecionar, explorar e modelar grandes quantidades de dados a fim de detectar padrões de comportamento. Já de acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), a

mineração de dados é o processo de extrair informações desconhecidas, válidas e acionáveis de grandes conjuntos de dados para então aplicar o conhecimento obtido em decisões cruciais no mundo dos negócios.

Na prática, os objetivos de DM são a predição ou a descrição. Compreende-se por predição a utilização de alguns atributos da base de dados para prever valores desconhecidos ou futuros de outras variáveis de interesse. Já a descrição procura por padrões que descrevem os dados interpretáveis pelos seres humanos (MARTINHAGO, 2005).

Mineração de Dados é uma área interdisciplinar que integra principalmente estatística, inteligência artificial e banco de dados. Pode-se afirmar isto, pois ao realizar várias medidas estatísticas, os algoritmos de *data mining* conseguem, por exemplo, classificar ou relacionar itens de uma base de dados. Os algoritmos podem também ser aplicados em um grande conjunto de dados armazenados aproveitando-se de métodos de indução com base na Inteligência Artificial. A Figura 1.2 ilustra a interdisciplinaridade da tecnologia de Mineração de Dados (CARVALHO, 2000).

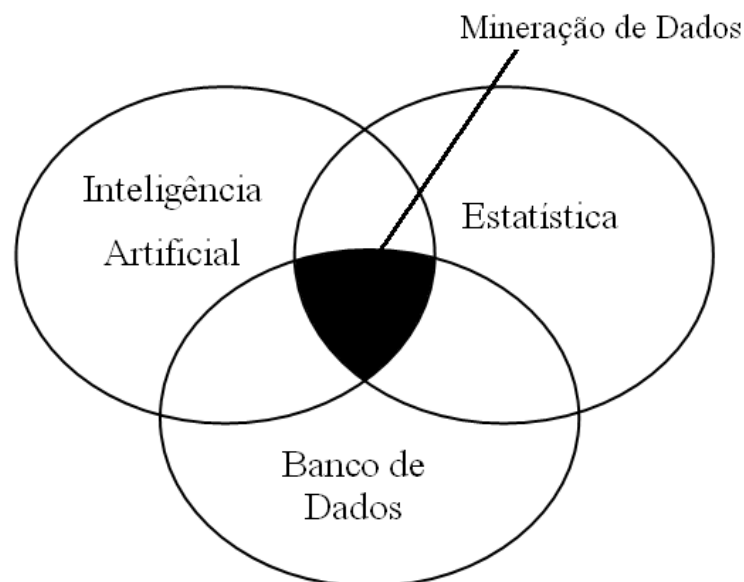


Figura 1.2 – Interdisciplinaridade da Mineração de Dados  
Fonte: Adaptado de CARVALHO, 2000, p. 25

Os algoritmos de Mineração de Dados são essenciais nesta etapa do processo de KDD, pois alguns deles têm capacidade de aprender a partir de exemplos. Tais algoritmos assimilam relacionamentos eventuais existentes entre os dados, utilizando o resultado deste aprendizado nos modelos de conhecimento gerados (GOLDSCHMIDT, 2005).

Abordando técnicas estatísticas em DM, é possível influenciar significativamente todas as áreas de uma organização. As técnicas estatísticas ajudam a assimilação e reação às mudanças de mercado, fazendo com que a organização se torne mais produtiva e competitiva, além de tomar decisões baseadas em fatos (BORGES, 2006).

A tecnologia de Mineração de Dados tem grande potencial para auxiliar as organizações a extrair importantes informações oriundas das suas bases de dados, formulando padrões e comportamentos futuros, ajudando a responder questões que demandariam muito tempo para serem resolvidas, possibilitando melhores decisões de negócio apoiadas no conhecimento extraído. Assim, é possível afirmar que a Mineração de Dados é um recurso em grande ascensão e se tornará obrigatória aos mercados mais competitivos (MARTINHAGO, 2005).

### **1.3 Pós-Processamento**

A última etapa do processo de KDD é de grande relevância, mesmo sendo a mais simples, uma vez que não exige grande esforço computacional e tempo do usuário em relação às outras etapas. É nesta etapa que o usuário que acompanhou todo o processo define se o conhecimento gerado será útil e aplicável. É muito comum que ao final do processo, os algoritmos utilizados apresentem ao usuário algumas informações não relevantes, sendo que cabe a este analisá-las e decidir aplicá-las ou não, dependendo do contexto da aplicação (GONCHOROSKI, 2007).

Segundo Santos (2008), esta etapa deve ser executada pelo analista de dados, responsável pelas etapas anteriores, e pelo analista de negócios, responsável por analisar se o conhecimento obtido será útil. Pode-se também contatar o executivo responsável para que o mesmo forneça esclarecimentos sobre o conhecimento descoberto, relacionando-os aos objetivos do negócio a fim de validá-los. Já de acordo com Martinhago (2005), a etapa de pós-processamento consiste na validação do conhecimento extraído da base de dados, identificação de padrões e interpretação dos mesmos, transformando-os em conhecimentos

apoiadores na tomada de decisão. O objetivo de interpretar os resultados é filtrar as informações que serão apresentadas aos tomadores de decisão.

Geralmente, a meta principal desta etapa é fazer com que o conhecimento descoberto seja compreendido da melhor maneira possível, validando-os através de medidas da qualidade da solução e da percepção de um analista de dados. O conhecimento gerado será consolidado em forma de relatórios, sendo feita a documentação e explicação das informações obtidas em cada etapa do processo de KDD (BORGES, 2006).

Visto que o interesse para um determinado padrão gerado no processo de extração de conhecimento varia de acordo com cada usuário e ramo de mercado, medidas subjetivas são necessárias. Quando um conjunto de regras interessantes é selecionado, estas medidas consideram que fatores específicos devem ser tratados, como o conhecimento do domínio e o interesse do usuário (REZENDE, 2005).

Durante as etapas do processo de KDD esforços diferentes são consumidos. A fase que mais consome esforço computacional e tempo do usuário é a de preparação dos dados, conforme mostra o gráfico 1.1 (CARVALHO, 2000).

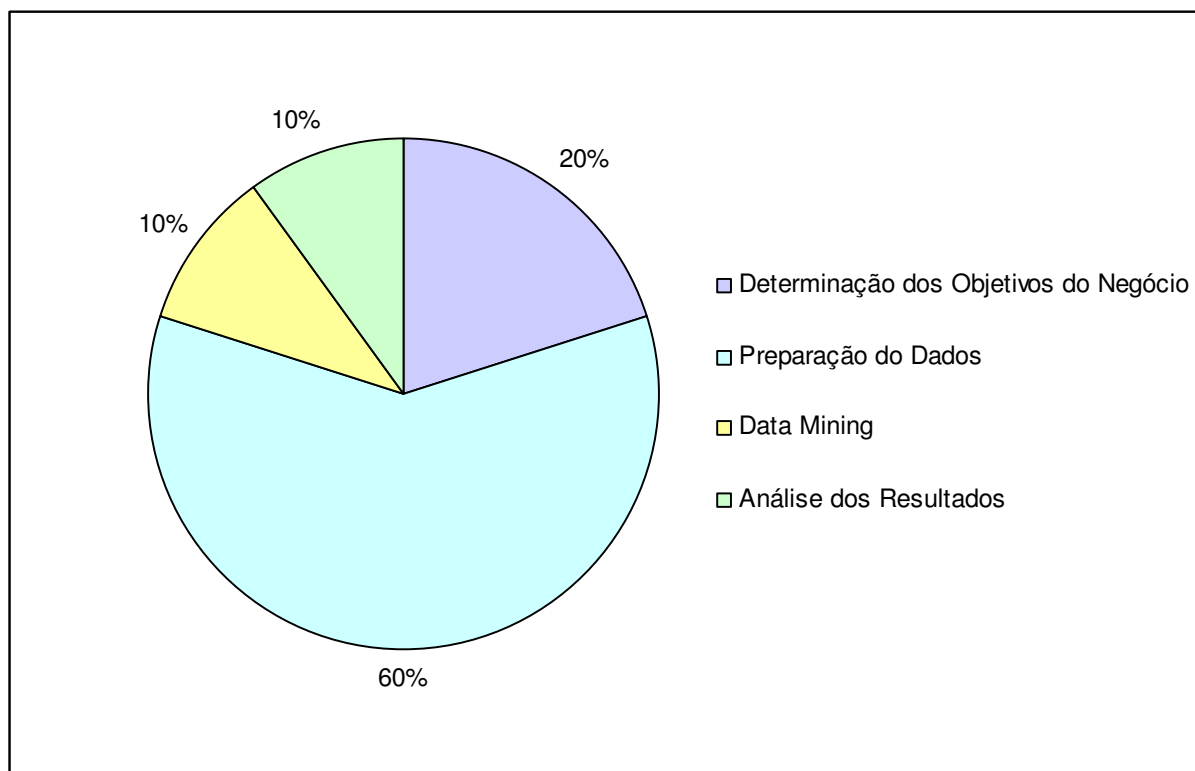


Gráfico 1.1 – Percentagem de esforço para cada etapa do processo de KDD

Fonte: Adaptado de CARVALHO, 2000, p. 24

A etapa de pós-processamento encerra o ciclo da descoberta do conhecimento e é nela que se coloca em ação todo o conhecimento adquirido durante as etapas anteriores. Após interpretar e avaliar o resultado obtido, o usuário vai identificar a necessidade ou não de reiniciar o processo e gerar outro tipo de regra ou informação. Se os resultados não forem satisfatórios, faz-se necessário repetir a etapa de Mineração de Dados ou retomar qualquer um dos estágios anteriores. O conhecimento é encontrado somente após a avaliação e validação dos resultados (MARTINHAGO, 2005).

Conforme já foi comentado, algumas técnicas e algoritmos são utilizados no processo de Descoberta de Conhecimento. Na etapa de Mineração de Dados um dos métodos mais conhecidos e utilizados é a Classificação, que dispõe de algumas técnicas para extrair conhecimento. Entre essas se podem citar as árvores de decisão e redes neurais como sendo as mais estudadas. Visando uma abordagem mais completa e detalhada sobre estas técnicas, torna-se necessário um estudo sobre o histórico, vantagens e desvantagens entre outras de suas características.

## 2 CLASSIFICADORES

Na tarefa de classificação, existem algumas técnicas que são utilizadas para a extração de conhecimento de bases de dados sendo que a seguir serão abordadas somente duas delas: árvores de decisão e redes neurais. Essas duas técnicas baseiam-se no aprendizado supervisionado, na qual os resultados obtidos necessitam de análise de um especialista que fará a avaliação de sua relevância, e geram modelos a partir de exemplos de uma base de dados, denominados conjunto de treinamento, representando uma amostra dos registros que serão analisados (GONCHOROSKI, 2007).

Objetiva-se com este estudo a comparação das duas técnicas, observando suas vantagens e desvantagens, de acordo com os resultados obtidos, permitindo a escolha da técnica que revela os resultados mais adequados dentro do contexto da aplicação.

### 2.1 Árvores de Decisão

As árvores de decisão possuem este nome devido a sua estrutura, muito compreensível e assimilativa, se assemelhar a uma árvore. Suas técnicas dividem os dados em subgrupos, baseadas nos valores das variáveis, sendo que o resultado disto é uma hierarquia de declarações do tipo “Se...então...” que são principalmente aplicadas quando o grande objetivo da mineração de dados é a classificação de dados ou a predição de saídas. (MARTINHAGO, 2005).

De acordo com Goldschmidt (2005, p. 109), uma árvore de decisão pode ser definida como “um modelo de conhecimento em que cada nó interno da árvore representa uma decisão sobre um atributo que determina como os dados estão particionados pelos seus nós filhos”. Já de acordo com Sousa (1998), métodos de árvore de decisão representam um tipo de algoritmo

de aprendizado de máquina, que fazem uso de uma abordagem dividir-para-conquistar para classificar casos, representando-os em forma de árvores.

### 2.1.1 Histórico

Muitas pessoas na área de Mineração de Dados consideram Ross Quinlan, da Universidade de Sydney, Austrália, o criador das árvores de decisão. Isto se deve, em grande parte, pela criação de um novo algoritmo chamado de ID3 (*Itemized Dichotomizer 3*), desenvolvido em 1983. O algoritmo ID3 e versões posteriores como o ID4 e o C 4.5, por exemplo, são estruturados de tal forma que se adaptam muito bem ao serem utilizados em conjuntos com árvores de decisão, visto que eles produzem regras ordenadas por importância. Essas regras são utilizadas na produção de um modelo de árvore de decisão dos fatos que afetam os itens de saída (JERONIMO, 2001).

Pode-se dizer que as árvores de decisão são uma evolução das técnicas que apareceram durante o desenvolvimento das disciplinas de *machine learning*. A partir da aproximação conhecida como Detecção de Interação Automática, desenvolvida na Universidade de Michigan, as árvores de decisão foram ganhando maior importância no meio científico (KRANZ, 2004).

### 2.1.2 Conceitos

Considerada uma ferramenta completa e bastante conhecida para classificação dos dados e apresentação dos resultados na forma de regras, as árvores de decisão são utilizadas frequentemente no processo de Descoberta de Conhecimento. A classificação é executada, na maioria das vezes, em duas fases no uso das árvores de decisão: construção da árvore e poda (OLIVEIRA, 2001). Nesta técnica, o usuário escolhe o atributo que quer avaliar para que o algoritmo procure as variáveis mais correlacionadas, gerando uma árvore de decisão com inúmeras ramificações. A árvore criada será utilizada na classificação de novas instâncias, de acordo com os valores dos atributos da nova instância (ARAÚJO, 2006).

Pode-se considerar as árvores de decisão como um algoritmo supervisionado, pois há a necessidade de ser informadas com antecedência as classes dos registros usadas no conjunto de treinamento. Uma árvore de decisão é formada por um conjunto de nós que são conectados através de ramificações, estes nós se dividem em três tipos conforme mostra a figura 2.1, onde



o nodo raiz é o início da árvore, os nodos comuns dividem um atributo e geram novas ramificações e o nodo folha possui as informações de classificação do registro (SANTOS, 2008).

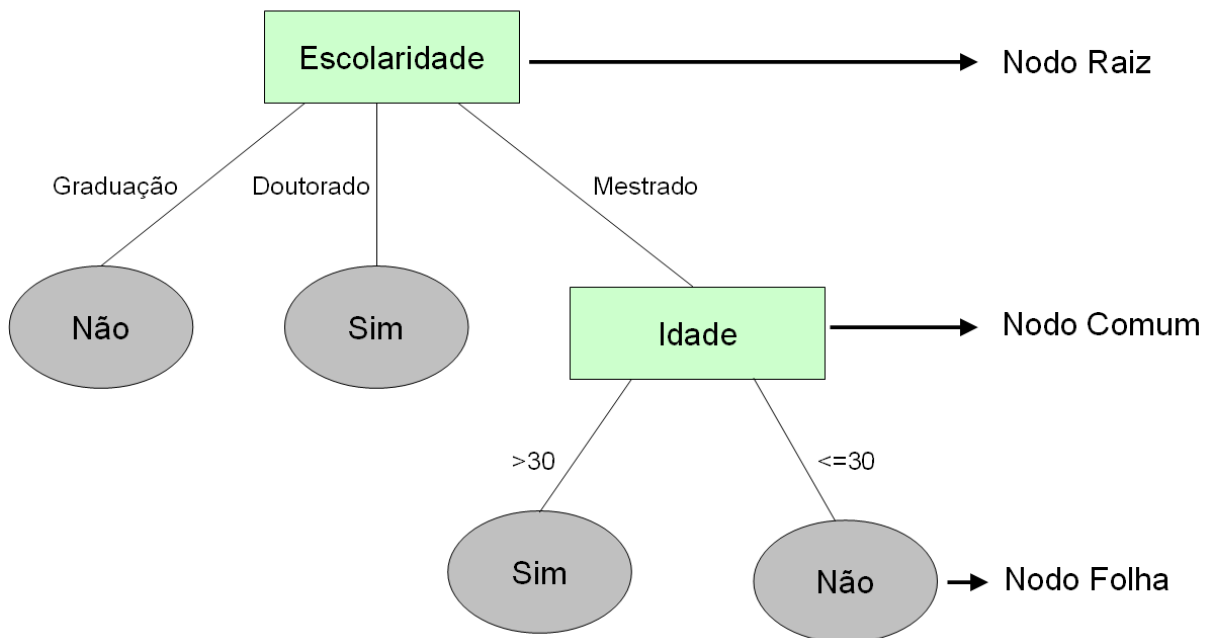


Figura 2.1 – Tipos de Nodos de uma Árvore de Decisão  
Fonte: SANTOS, 2008, p. 21

Na fase de construção da árvore, são realizadas ramificações na árvore através de sucessivas divisões dos dados com base nos valores dos atributos. Sendo assim, o processo é repetido recursivamente até que todos os registros pertençam a uma classe (OLIVEIRA, 2001).

Um registro entra na árvore pelo nó raiz. A partir deste nodo todos os outros nodos são percorridos até ser alcançado o nodo folha. Cada um dos nodos testa o valor de um único atributo e oferece arestas distintas a serem percorridas na árvore a partir deste nodo, para cada uma de suas valorações. Assim é determinado o próximo nodo no qual o registro irá se posicionar. Podem ser utilizados diferentes algoritmos na escolha do teste inicial, porém, todos têm o mesmo objetivo: escolher aquele que melhor descreve a classe alvo. Quando o algoritmo chega ao nodo folha, todos os registros que terminam na mesma folha são classificados da mesma forma. É importante salientar que existe somente um caminho da raiz

até cada folha, que significa a expressão utilizada para classificar os registros (BORGES, 2006).

Após a fase de crescimento da árvore, pode-se encontrar uma estrutura especializada que está super ajustada aos dados, sendo que desta maneira é oferecida mais estrutura que o necessário. Então, a poda passa a ter um papel crucial, fazendo com que sejam consideradas árvores menores e potencialmente de melhor precisão (SOUSA, 1998).

Na fase de poda, as ramificações que não tem valor significativo são removidas, a fim de criar um modelo de classificação, fazendo a seleção da sub-árvore que contém a menor taxa de erro estimada (OLIVEIRA, 2001).

Após a fase de poda, a árvore gerada pode representar uma estrutura complexa e de difícil compreensão. Nestes casos pode-se utilizar a extração de regras como uma fase final, visando extrair regras menores e menos complexas, porém, com precisão similar (SOUSA, 1998).

### **2.1.3 Vantagens e Desvantagens**

O uso de árvores de decisão possui algumas vantagens em relação às outras técnicas, dentre as quais se podem citar: facilidade de compreender o modelo obtido, uma vez que tem a forma de regras explícitas, possibilitando a avaliação dos resultados e a identificação dos seus atributos chaves no processo; facilidade de expressar as regras como instruções lógicas sendo aplicadas diretamente aos novos registros; árvores de decisão são relativamente mais rápidas em comparação às redes neurais, por exemplo, e na maioria das vezes se obtém mais precisão nos resultados quando comparadas a outras técnicas de classificação (OLIVEIRA, 2001).

Segundo Sousa (1998), as principais desvantagens no uso de árvores de decisão estão na necessidade de uma considerável quantidade de dados para desvendar estruturas complexas e na possibilidade de haver erros na classificação, no caso de existirem muitas classes, bem como o tratamento de dados contínuos.

Para melhor compreensão dos algoritmos de árvore de decisão, serão apresentados a seguir dois deles: o C4.5 e o *Logistic Model Tree* (LMT).

### 2.1.4 Algoritmo C4.5

Considerado um dos mais tradicionais algoritmos na tarefa de Classificação, o C4.5 foi inspirado no algoritmo ID3, sendo que seu método visa abstrair árvores de decisão seguindo uma abordagem recursiva de particionamento das bases de dados (GOLDSCHMIDT, 2005). Também desenvolvido pelo pesquisador australiano Ross Quinlan, em 1993, este algoritmo encontra-se disponível em diversos softwares de mineração. O C4.5 transforma a árvore de decisão em um conjunto de regras ordenadas pela sua importância, possibilitando ao usuário a identificação dos fatores mais relevantes em seus negócios (OLIVEIRA, 2001)

A principal vantagem do algoritmo C4.5 em relação ao ID3, é que ele tem o poder de lidar com a poda (*prunning*) da árvore, evitando o sobre-ajustamento, com a valoração numérica de atributos e com a presença de ruído nos dados (BORGES, 2006). Na maioria das vezes uma árvore originada do algoritmo C4.5 precisa ser podada pela necessidade de redução do excesso de ajuste (*overfitting*) aos dados de treinamento (MARTINHAGO, 2005).

Enquanto o algoritmo ID3 manipula apenas dados nominais, o C4.5 pode manipular também dados numéricos. Entretanto, trabalhar com dados numéricos não é tão simples, pois enquanto os atributos nominais são testados apenas uma vez em qualquer caminho da raiz às folhas, os atributos numéricos podem ser testados diversas vezes no mesmo percurso. Esta característica pode ser considerada uma possível desvantagem do C4.5, pois em alguns casos, a árvore gerada pode ser de difícil entendimento do usuário (BORGES, 2006).

Neste algoritmo é utilizada a abordagem “dividir para conquistar”, em que o problema original é dividido em partes semelhantes ao original, porém menores, fazendo com que os problemas sejam resolvidos e suas soluções formem uma combinação para o problema inicial. Ele ainda possui a capacidade de aprimorar a estimativa do erro utilizando uma técnica conhecida como *v-fold*, onde é realizada a validação cruzada com dois ou mais grupos (GONCHOROSKI, 2007).

De acordo com Oliveira (2001), o algoritmo cria uma árvore com uma quantidade aleatória de folhas por nodo e assume os valores das categorias como um divisor, diferentemente do que realiza algoritmos que produzem árvores binárias, por exemplo. Então o *prunning* é realizado de acordo com a taxa de erro de cada nodo e seus descendentes, sendo que a soma dessas taxas compõem a taxa de erro da árvore. Para a identificação do nodo raiz e

de seus descendentes são realizados os cálculos da entropia e do ganho de informação (OLIVEIRA, 2001).

Após a criação de um conjunto de regras, o algoritmo realiza o agrupamento das regras obtidas para cada classe e a eliminação das regras que não possuem relevância na precisão do conhecimento a ser extraído. Como resultado final, se obtém um pequeno conjunto de regras que podem ser facilmente entendidas, criadas pela combinação das regras que induzem à mesma classificação (MARTINHAGO, 2005).

### 2.1.5 Algoritmo LMT

O algoritmo *logistic model tree* (LMT) aplica os princípios das árvores em problemas de classificação, utilizando para a construção da árvore a regressão logística, que tem como objetivo saber quais variáveis independentes influenciam no resultado, utilizando uma equação para prever um resultado baseado nestas variáveis. Um processo de adaptação por etapas é empregado na construção dos modelos de regressão nos nodos folhas, realizando uma redefinição incremental àqueles construídos em níveis superiores da árvore (ARAÚJO, 2006).

Esse algoritmo normalmente é utilizado para a predição numérica, sendo que os nodos folhas gerados armazenam um modelo de regressão logística para geração do resultado. Após a construção da árvore, é aplicada uma regressão para cada nodo interior, utilizando os dados associados a este nodo e todos os atributos que participam nos testes na sub-árvore. Em seguida, os modelos de regressão logística são simplificados, utilizando a poda. Porém, a poda só acontecerá se o erro estimado para o modelo na raiz de uma sub-árvore for menor ou igual ao erro esperado para a sub-árvore. Após a poda é realizado um processo que forma o modelo final, colocando-o no nodo folha (LANDWEHR; HALL; FRANK, 2003).

## 2.2 Redes Neurais

Uma Rede Neural Artificial (RNA) é uma técnica computacional que cria um modelo matemático, emulado por computador, simulando um sistema neural biológico simplificado, que tem como principal característica a capacidade de aprendizado, generalização, associação e abstração (ARAÚJO, 2006). São consideradas as técnicas mais comuns utilizadas pelos processos de Mineração de Dados e possuem uma característica que

as diferenciam das outras técnicas: podem gerar saídas iguais às entradas, que não existiam durante a fase de treinamento (SANTOS, 2008).

Duas características fazem com que as redes neurais sejam semelhantes ao cérebro: o conhecimento é adquirido pela rede através de seu ambiente utilizando um processo de aprendizagem; e as forças de conexão entre os neurônios, mais conhecidas como pesos sinápticos, são usadas para o armazenamento do conhecimento obtido (HAYKIN, 2001). De acordo com Sousa (1998), os métodos baseados em RNA proporcionam métodos mais práticos para funções de aprendizado, que são representadas por atributos contínuos, discretos ou vetores.

A estrutura de uma rede neural consiste em uma quantidade de neurônios interconectados que são organizados em camadas. O conhecimento através destas camadas se dá através da modificação das conexões, que são responsáveis pela comunicação entre as camadas (ARAÚJO, 2006). A figura 2.2 apresenta a arquitetura de uma rede neural simples em que os círculos representam os neurônios, e as linhas representam os pesos das conexões.

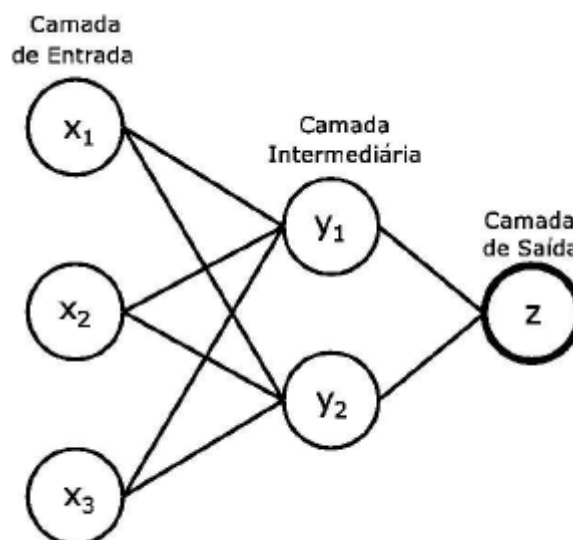


Figura 2.2 – Estrutura de uma Rede Neural Simples

Fonte: ARAÚJO, 2006, p. 30

Todas as camadas de uma rede neural possuem funções específicas. A camada de entrada é a que recebe os dados a serem analisados. A camada intermediária é responsável pelo processamento interno das informações e extraem características, permitindo que a rede

crie sua própria representação. É importante salientar que uma RNA pode conter várias camadas intermediárias, de acordo com a complexidade do problema. A camada de saída recebe os estímulos da camada intermediária, construindo o padrão que será a resposta para o problema em análise (ARAÚJO, 2006).

O processo de aprendizado de uma rede neural pode ser realizado de duas formas:

- Supervisionado: é utilizado um conjunto de pares de dados de entrada e saída desejada. A partir dos conjuntos de entrada, a rede neural cria um conjunto de valores de saída desejado. Na existência de grande diferença entre as saídas, os pesos sinápticos e os níveis de bias<sup>1</sup>, são acertados até que a diferença seja diminuída (SANTOS, 2008);

- Não supervisionado: o treinamento da rede se dá apenas através de valores de entrada. Assim, são realizados processos, chamados de competição e cooperação, entre os neurônios para a classificação dos dados, obtendo-se um reconhecimento de padrões (SANTOS, 2008).

Existem redes neurais com apenas um neurônio, chamadas de *perceptron*, que são a unidade mais simples de uma rede neural. Através de um vetor com números reais de entrada, é possível que o *perceptron* calcule uma combinação linear destes atributos para retornar um resultado. Essa rede de apenas uma camada é capaz de solucionar problemas que sejam linearmente separáveis (HAYKIN, 2001).

De acordo com Martinhago (2005), a grande vantagem da utilização de redes neurais é a grande versatilidade que elas possuem, sendo que o resultado é satisfatório até mesmo em áreas complexas, com entradas incompletas ou imprecisas. Além disso, as redes neurais possuem excelente desempenho em problemas de classificação e reconhecimento de padrões (SANTOS, 2008).

As desvantagens da utilização das redes neurais estão ligadas à solução final, que depende das condições finais estabelecidas na rede, uma vez que os resultados dependem dos valores aprendidos. Outra desvantagem das redes neurais é o fato de os resultados obtidos não terem uma comprovação, pois todo o conhecimento adquirido pelos neurônios não podem ser representados. Portanto, não é possível comprovar um resultado adquirido através da utilização de redes neurais (MARTINHAGO, 2005).

---

<sup>1</sup> Os níveis de bia possibilitam que a saída não seja nula mesmo que as entradas sejam.

Em comparação às árvores de decisão, os algoritmos de redes neurais normalmente necessitam de maior força computacional para serem utilizados. Os tempos de treinamento variam de acordo com o número de casos de treinamento, número de pesos na rede e das configurações dos vários parâmetros do algoritmo de aprendizado (SOUSA, 1998). A seguir será apresentado o *perceptron* multi-camadas e o algoritmo de retropropagação.

### **2.2.1 Perceptron Multi-Camadas**

A rede *perceptron* multi-camadas (MLP) é formada por múltiplas camadas de neurônios interconectadas, normalmente em forma de *feedward*, onde cada neurônio de uma camada tem conexões diretas aos neurônios da camada seguinte (HAYKIN, 2001). Essas redes possuem poder computacional elevado em relação as que não possuem camadas intermediárias e podem receber dados que não são linearmente separáveis (BRAGA, 2000).

O processamento realizado por cada neurônio neste tipo de rede é definido pela combinação dos processamentos efetuados pelos neurônios da camada anterior que estão conectados a ele (BRAGA, 2000). Essa rede representa uma generalização do *perceptron* de camada única, tendo o seu funcionamento descrito como uma seqüência de *perceptrons* (HAYKIN, 2001).

Os *perceptrons* de múltiplas camadas são aplicados com bastante sucesso na resolução de problemas difíceis, a partir de seu treinamento de forma supervisionada com um algoritmo bastante conhecido chamado retropropagação de erro. Baseado na regra de aprendizagem por correção de erro, este algoritmo é considerado uma generalização de outro algoritmo bastante conhecido chamado de algoritmo do mínimo quadrado médio (HAYKIN, 2001).

#### **2.2.1.1 Algoritmo de Retropropagação**

O surgimento da retropropagação se deu devido ao interesse por parte dos pesquisadores na resolução de alguns problemas existentes dentro do treinamento das redes neurais. Após seu surgimento, este algoritmo acabou tornando-se um dos mais populares para este tipo de treinamento, sendo considerado um dos responsáveis pelo ressurgimento do interesse nesta área (ARAÚJO, 2006).

O algoritmo de retropropagação utiliza pares para ajustar os pesos na rede, através de um mecanismo de correção de erros. Seu aprendizado baseia-se na propagação retrógrada do erro para níveis anteriores da rede, de acordo com o nível de participação que cada neurônio teve na camada superior (BRAGA, 2000).

O treinamento através deste algoritmo ocorre em duas fases, chamadas de *forward* e *backward*, sendo que em cada uma delas a rede é percorrida em um sentido diferente. Na fase *forward*, um padrão é apresentado à camada de entrada da rede. A atividade resultante percorre a rede, camada por camada, até que uma resposta seja produzida pela camada de saída. Na fase *backward*, é feita a comparação da saída obtida com a saída desejada para este padrão particular. Se o resultado não estiver correto, o erro é calculado, sendo o erro propagado a partir da camada de saída até a camada de entrada, modificando os pesos das conexões das unidades das camadas internas, de acordo com a retropropagação do erro (BRAGA, 2000).

A partir das técnicas estudadas, busca-se uma relação entre teoria e prática. Sendo assim, a seguir serão apresentados estudos de casos nos quais as soluções para os problemas existentes foram encontradas utilizando técnicas de mineração de dados.



### 3 APLICAÇÕES DE MINERAÇÃO DE DADOS

De acordo com o estudo feito através de artigos que utilizam Mineração de Dados e mais especificamente árvores de decisão, foram selecionados três com o objetivo de exemplificar alguns casos em que o uso de árvores de decisão auxiliou a extração de conhecimento de bases de dados em áreas distintas. Assim, é possível perceber que o uso de árvores de decisão se torna cada vez mais comum em diversas áreas, auxiliando a tomada de decisão por parte de homens de negócios e até mesmo por parte de organizações governamentais. O primeiro artigo utiliza a Mineração de Dados para criar perfis de usuários inadimplentes de empresas de telefonia, o segundo artigo utiliza árvores de decisão na tentativa de controlar o desmatamento da Amazônia na cidade de Capixaba, Acre. Por fim, o último artigo aborda a utilização de Mineração de Dados para predição da prevalência de esquistossomose no Estado de Minas Gerais, Brasil.

#### **3.1 Análise do perfil do usuário de serviços de telefonia utilizando técnicas de mineração de dados**

Segundo estudo de Junior e Perez (2006), o crescimento exponencial do prejuízo que as operadoras de telecomunicações absorvem devido à inadimplência e utilização ilícita dos seus recursos, fez com que as mesmas procurassem alternativas para diminuir esse problema. Com isso, percebeu-se que técnicas de mineração de dados, quando aplicadas neste setor, se tornam um poderoso recurso para identificar o perfil de usuários inadimplentes. Quanto mais rápido forem identificados esses usuários, menor será o prejuízo que a operadora de telefonia terá e, conseqüentemente, mais recursos poderão ser oferecidos aos usuários.

Por outro lado, conhecer o perfil dos bons pagadores através de suas preferências na utilização dos serviços, auxilia as empresas de telecomunicações a realizarem campanhas de

marketing, por exemplo, visando promoções e privilégios que as ajudam a manter e agregar clientes novos.

O objetivo principal deste artigo é apresentar o uso de duas técnicas de mineração de dados: redes neurais e árvores de decisão, a fim de avaliar qual delas se torna mais eficaz para identificar o perfil de um usuário inadimplente.

Como ferramenta de Mineração de Dados para geração de árvores de decisão os autores optaram por utilizar o algoritmo C4.5 do software Sipina, pois o mesmo possui licença para uso educacional, implementa o método de classificação e utiliza árvore de decisão para representar o conhecimento obtido. A figura 3.1 apresenta a interface principal do Sipina.

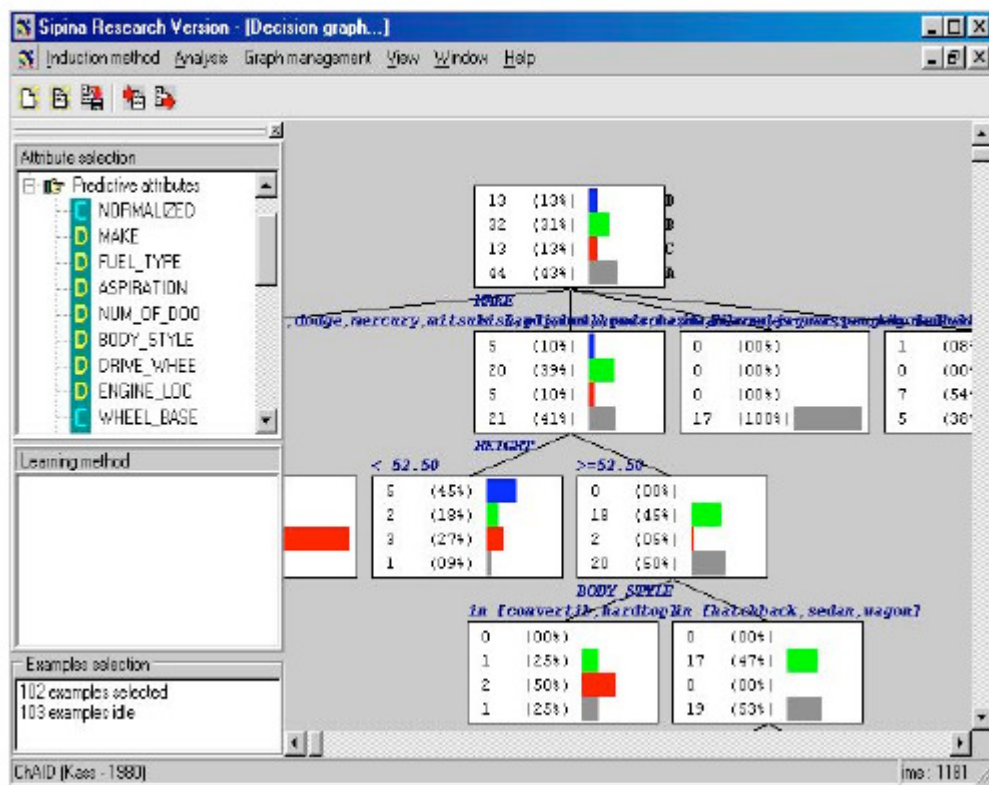


Figura 3.1 – Interface Principal do *Sipina*

Fonte: JUNIOR; PEREZ, 2006, p. 5

Já para a extração do conhecimento utilizando redes neurais foi escolhido o software QwikNet, que simula redes neurais executando vários métodos eficientes para treiná-las e testá-las e tem como característica oferecer uma relação flexível e intuitiva, permitindo

projetar, treinar e testar redes neurais em um ambiente gráfico. A interface do QwikNet é apresentada na figura 3.2.

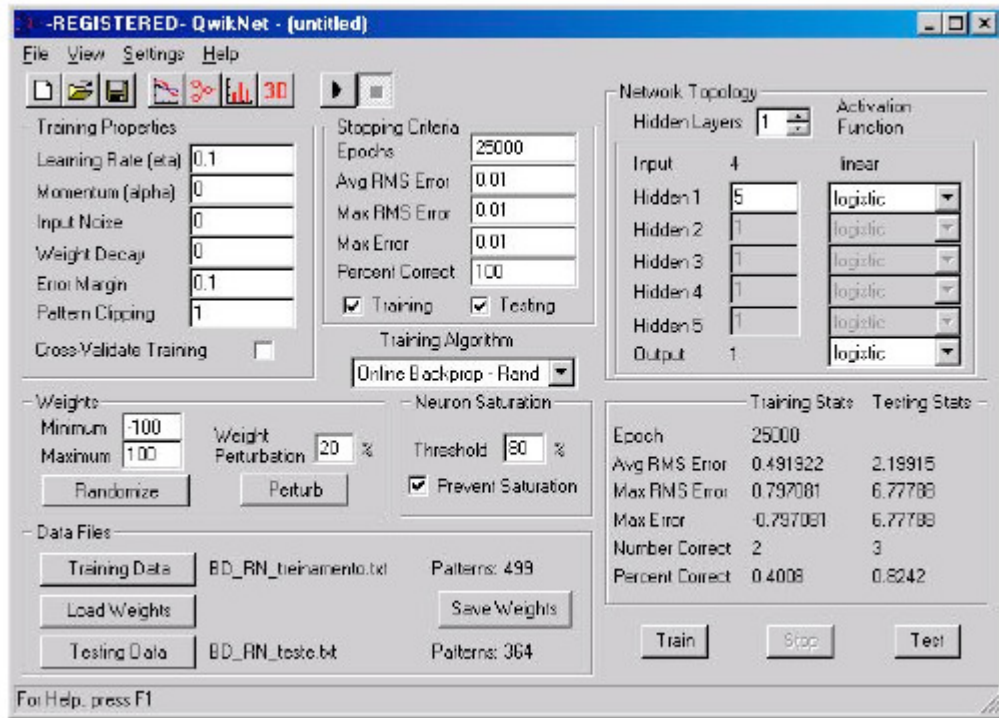


Figura 3.2 – Interface do *QwikNet*  
Fonte: JUNIOR; PEREZ, 2006, p. 6

Primeiramente, foi gerado um arquivo texto contendo os dados das chamadas telefônicas presentes no SGBD da companhia. Estes dados foram separados por tabulações e, após a sua limpeza, sofreram uma codificação, utilizando uma planilha que depois foi salva em formato “.txt”, que os enriqueceram e os prepararam para o processo de descoberta de conhecimento. Foram disponibilizados 63.534 registros para testes, com dados de chamadas telefônicas realizadas entre 01 de setembro e 31 de dezembro de 2005, sendo todas estas chamadas realizadas por assinantes inadimplentes de telefones fixos.

Não foi possível a utilização de um número maior de registros devido ao sigilo e também por esses dados serem pontos estratégicos das empresas no combate a inadimplência. Esses registros possuem os seguintes dados: dia da semana que a chamada foi executada, hora inicial da chamada, destino da chamada que identifica o tipo do destino (local, DDD, celular, DDI) e duração das chamadas.

Para a geração da árvore de decisão com o software *Sipina*, a classe principal foi criada com o atributo “dia da semana” como nodo principal, com o intuito de descobrir os dias da semana em que os usuários fazem mais ligações. Como nodos filhos foram especificados os atributos hora e destino, com a finalidade de descobrir o horário preferido das chamadas e o tempo utilizado nas conversações. Como resultado da aplicação deste algoritmo, pode-se considerar as regras apresentadas na figura 3.3.

- Regra 01: Os dias da semana com maior número de chamadas são quarta e quinta feira no horário entre 06:00h – 12:00h:  
 Quarta feira: 22%  
 Quinta feira: 31%  
 Na sexta feira o horário de maior tráfego é entre 18:00h – 24:00h: 21%
- Regra 02: Nas segundas feiras o horário entre 12:00h e 18:00h concentra chamadas para serviços especiais: 33%
- Regra 03: Nas quartas feiras o horário entre 12:00h e 18:00h concentra chamadas:  
 para telefone fixo (Local): 23%  
 para telefone celular (DDD): 22%
- Regra 04: Nas quintas feiras o horário entre 12:00h e 18:00h concentram-se chamadas para telefone fixo (DDI): 30%
- Regra 05: Nas sextas feiras o horário entre 12:00h e 18:00h concentram-se chamadas:  
 para telefone celular (Local): 22%  
 para telefone fixo (DDD): 24%

Figura 3.3 – Regras Geradas pelo *Sipina*

Fonte: JUNIOR; PEREZ, 2006, p. 5

Conforme as regras obtidas, pode-se definir o perfil geral dos usuários, obtendo-se o comportamento generalizado dos inadimplentes. Com essa definição, pode-se estabelecer um parâmetro comparativo, com o qual é realizada a verificação da semelhança do perfil dos usuários individuais com o perfil dos inadimplentes. Analisando os dados de um usuário específico, é feita a comparação dos resultados com as classes pré-determinadas, verificando se determinado usuário se encaixa em um dos perfis já encontrados.

Na realização dos testes utilizando redes neurais, todos os dados foram transformados, conforme realizado no teste com árvores de decisão. A rede neural foi treinada utilizando um arquivo com 499 linhas de dados com informações como dia da semana, horário, destino da chamada e duração, de usuários inadimplentes. Para o teste utilizou-se um

arquivo com 240 linhas com o arquivo de um único usuário para comparar sua semelhança com o perfil de usuário que a rede neural conseguiu aprender.

Foram utilizados quatro neurônios de entrada e um de saída na aplicação do teste, sendo a taxa de aprendizado de 0,1, momentum 0 e critério de parada com 25.000 épocas de treinamento. Dos 364 registros com os quais foram realizados os testes, representando uma pequena amostra do total de registros, três encaixam-se no perfil aprendido pela rede neural. Com isso, considera-se que esse indicador é bastante baixo quando aplicado para identificação de perfis de usuários inadimplentes ou fraudulentos.

As redes neurais são utilizadas para aprender com o histórico dos usuários, analisando o comportamento de cada um em diferentes períodos do dia. Realizando uma análise individual, dão condições às empresas de descobrir em tempo real alguma atividade suspeita, interrompendo-a rapidamente colaborando com o aumento da lucratividade da empresa.

Com a técnica de árvores de decisão os dados apresentados representam um padrão de comportamento, contudo foi gerado um número elevado de subdivisões, o que fez com que a leitura do resultado se tornasse um pouco demorada, mas de fácil compreensão. A técnica também permite a geração de regras que definem o padrão, o que facilita a procura de novos padrões quando aplicada em outro volume de dados.

### **3.2 Avaliação da exatidão do mapeamento da cobertura da terra em Capixaba, Acre, utilizando classificação por árvore de decisão**

Conforme artigo publicado por Carvalho e Figueiredo (2006), um problema real enfrentado nos dias de hoje é o desmatamento na Amazônia. As instituições brasileiras e internacionais, assim como pesquisadores, organizações não-governamentais e sociedade em geral passaram a ter uma preocupação maior, devido à gravidade deste problema. Diversas experiências e políticas públicas passaram a ser aplicadas na tentativa de reversão deste cenário, entre elas, as que têm como objetivo o monitoramento ambiental da cobertura florestal e que utilizam a aplicação de técnicas de classificação digital e sensoriamento remoto no mapeamento da cobertura da terra.

O objetivo principal deste estudo é realizar uma avaliação, com a maior exatidão possível, do mapeamento da cobertura da terra em Capixaba, utilizando a classificação digital

de imagens de sensoriamento remoto, através de um algoritmo de árvore de decisão. De acordo com dados do IBGE, a área de estudos está localizada no sudeste do Estado do Acre, e corresponde ao município de Capixaba, com a superfície territorial de 1.713 km<sup>2</sup>, equivalendo a 1,1% da área total do estado, conforme representado na figura 3.4.

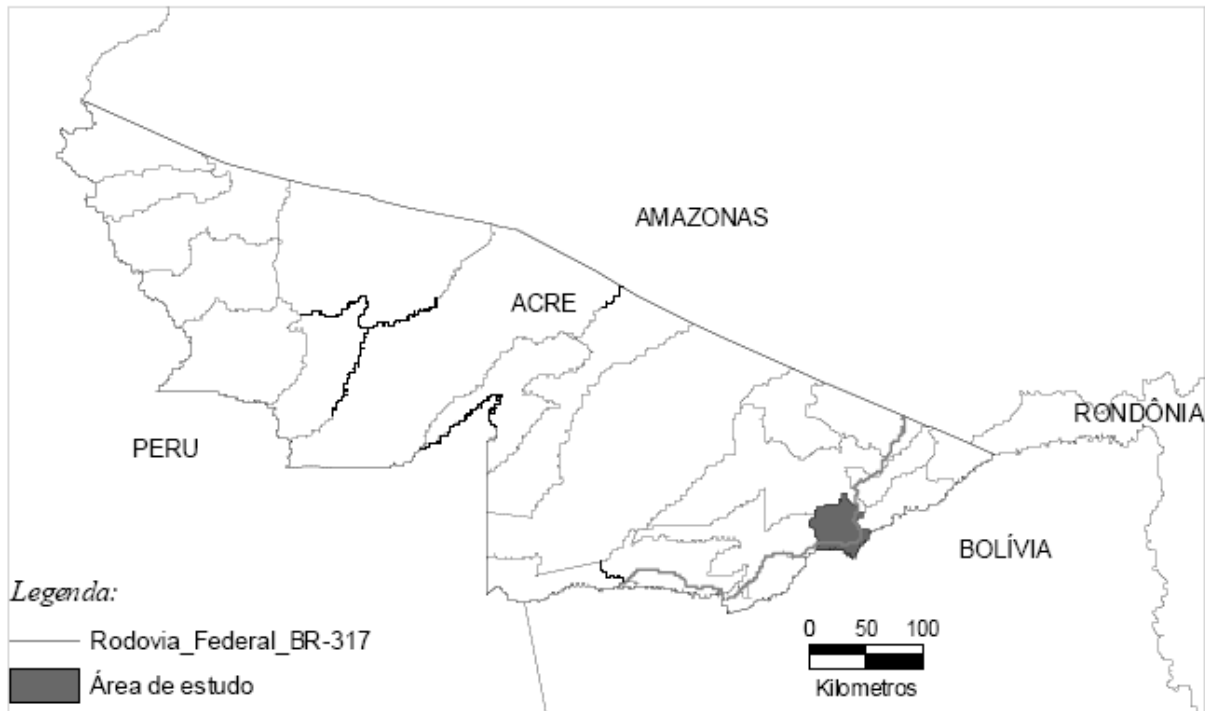


Figura 3.4 – Localização do município de Capixaba, Acre  
 Fonte: CARVALHO; FIGUEIREDO, 2006, p. 39

Foram utilizadas, para realização deste estudo, imagens multiespectrais, que são imagens de um mesmo objeto, tomadas com diferentes comprimentos de ondas eletromagnéticas, do sensor *Thematic Mapper* (TM) do satélite *Landsat 5*, referentes ao ano de 2003, sendo as imagens derivadas de técnicas de extração de informações. Foram utilizados aplicativos de mineração de dados e de processamento de imagens para a construção da árvore de decisão e geração do mapa temático. Para o mapeamento da cobertura da terra, foi necessária a definição de sete classes temáticas, de acordo com a vegetação e características do relevo das áreas a serem estudadas. Estas classes temáticas são as seguintes: Floresta, Capoeira, Pasto alto, Pasto baixo, Solo, Água e Queimada.

Para a classificação pelo algoritmo de árvore de decisão foi utilizado um aplicativo de mineração de dados, gerando assim um conjunto de regras e a árvore de decisão, sendo

posteriormente utilizado um aplicativo de processamento de imagens para classificação digital.

Nessa técnica de classificação, todos os atributos foram organizados em um único arquivo de imagem, sendo considerados os seguintes atributos: bandas 1, 2, 3, 4, 5 e 7 do *Landsat 5* com valores em número digital; imagens fração solo, sombra, vegetação e de erro geradas pelo modelo linear de mistura espectral e o índice de vegetação.

Para a construção da árvore de decisão foram utilizadas amostras de treinamento e amostras de teste. Cada *pixel* das amostras e seus respectivos valores nas onze imagens do arquivo de dados precisam ser analisados pelo algoritmo de aprendizado de máquina. Na fase de preparação dos dados utilizados para a implementação da mineração de dados, foram organizados três arquivos, denominados arquivo de nomes, arquivo de dados e arquivo de teste.

No arquivo de nomes está o nome das classes temáticas, dos atributos e valores dos atributos. A classe temática do caso e a descrição dos casos ou amostras de treinamento estão presentes no arquivo de dados. No mesmo formato do arquivo de dados, o arquivo de testes foi utilizado na avaliação do erro da árvore de decisão, criada pela mineração de dados. Foi aplicada também, a técnica convencional de classificação digital por meio do algoritmo de máxima verossimilhança e do algoritmo isodata, com o objetivo de avaliar seu desempenho em relação ao algoritmo de árvore de decisão.

Na construção da matriz de erro foram escolhidas as amostras de validação numa imagem de referência estratificada por classes temáticas. Para a realização da estratificação e produção do mapa temático de referência, foi utilizado um algoritmo de análise de agrupamento e migração.

Com a aplicação do algoritmo de árvores de decisão, foi possível gerar estimativas do percentual de área de cada uma das classes temáticas do município. Foi possível identificar também que a cobertura florestal representa 57,94% da área do município, além do percentual ocupado pela classe capoeira, que totaliza 10,56% da área. Pastagens de propriedades rurais e projetos de assentamento representam 29,05% da área de Capixaba, o que representa 50 mil hectares.

Os melhores desempenhos no emprego do algoritmo de árvores de decisão ficaram por conta das classes de floresta e água, com erros de inclusão ou omissão inferiores a 10%, o que demonstra a eficiência da técnica no mapeamento destas classes temáticas. Já nas áreas de capoeira, os erros de classificação ocorreram devido a semelhança das classes de pasto alto e floresta, o que ocasionou confusão na classificação das mesmas.

Porém, os erros mais significativos obtidos através do emprego do algoritmo de árvore de decisão foram verificados nas classes de pasto alto, pasto baixo e capoeira que obtiveram erros de 18,18%, 17,65% e 16,87%, respectivamente. As características da classe pasto baixo e de solo são semelhantes, gerando alguma confusão em sua classificação, porém a exatidão da classe solo foi altamente satisfatória, com índice de 84,62%.

A tabela 3.1 representa a exatidão global obtida pelo classificador de árvore de decisão com os erros de inclusão e omissão por classe de mapeamento do ano de 2003.

Tabela 3.1 – Matriz de Erros Obtida com o Algoritmo de Árvore de Decisão

Classificação	Amostras de validação (pixels)						Total	Inclusão	
	CA	FL	SO	AG	PA	PR			
Capoeira (CA)	69	1	0	0	5	0	75	8,0%	
Floresta (FL)	11	194	0	0	0	0	205	5,4%	
Solo (SO)	0	0	44	0	0	1	45	2,2%	
Água (AG)	0	0	0	77	0	0	77	0,0%	
Pasto alto (PA)	3	0	0	0	90	17	110	18,2%	
Pasto baixo (PB)	0	0	8	0	0	84	92	8,7%	
Total	83	195	52	77	95	102	604		
Omissão	16,9%	0,5%	15,4%	0,0%	5,3%	17,6%			
Exatidão global = 92,38%				Kappa = 0,9044					

Fonte: CARVALHO; FIGUEIREDO, 2006, p. 45

O resultado do mapeamento da cobertura da terra em Capixaba com algoritmo de árvore de decisão foi considerado excelente, sendo superior ao do algoritmo de máxima verossimilhança e ainda maior quando comparado com o método de classificação digital não



supervisionada isodata. Os melhores resultados foram obtidos com as classes de floresta e água, com exatidão superior a 94%. Apesar dos erros de classificação em algumas classes, os resultados demonstram que a técnica é altamente eficiente para aplicação de mapeamento de cobertura de terra.

### **3.3 Uso de árvore de decisão para predição da prevalência de esquistossomose no Estado de Minas Gerais, Brasil**

De acordo com Martins et al. (2007), no Brasil os hospedeiros da esquistossomose são moluscos límnicos do gênero *Biomphalaria*. Devido a problemas de saneamento domiciliar e ambiental e do baixo nível de educação em saúde da população que vive sob risco, a doença tem caráter social e comportamental. Uma vez que a doença é limitada no espaço e tempo por fatores ambientais, tornam-se extremamente importantes o uso de sistemas de informação geográficos (SIG) e o sensoriamento remoto (SR) na identificação dos fatores ambientais, permitindo que recursos sejam alocados nas áreas com mais risco de contaminação, auxiliando o combate a doença.

O objetivo principal deste trabalho é aplicar uma técnica de mineração de dados, utilizando árvores de decisão, para obter a estimativa da prevalência da esquistossomose no Estado de Minas Gerais. Através dos dados obtidos com sensoriamento remoto, derivadas climáticas e sócio-econômicas, pretende-se empregar ferramentas de SIG para dar auxílio no combate à doença e conscientização da população através dos órgãos competentes.

Como área de estudo, foi utilizado o Estado de Minas Gerais, com área de aproximadamente 590.000 km<sup>2</sup>, contendo 853 municípios, com aproximadamente 18 milhões de habitantes.

Para o estudo, foram utilizadas 197 amostras de um total de 853, que corresponde ao total de municípios do Estado, que possuíam dados de prevalência da doença. Com essa amostra, foi gerada a árvore de decisão, sendo possível, através da árvore selecionada, a extrapolação da estimativa da prevalência da esquistossomose para as demais amostras.

Todos os dados da prevalência foram fornecidos pela Secretaria de Vigilância em Saúde e Secretaria de Estado de Saúde de Minas Gerais. Das quarenta e quatro variáveis

utilizadas para amostragem da prevalência, vinte e duas são oriundas de dados de sensoriamento remoto, seis são climáticas e dezesseis são sócio-econômicas.

Para a aplicação da técnica de mineração de dados foi utilizado o software *Weka* (*Waikato Environment for Knowledge Analysis*), por ser um software de código aberto e sua licença gratuita. Para a criação da árvore de decisão foi empregado o algoritmo J4.8, que é a versão na linguagem Java do algoritmo C4.5, mencionado no capítulo anterior.

Como em todo processo de descoberta de conhecimento, foi necessária a realização do pré-processamento dos dados. Como o algoritmo de classificação necessita que a variável a ser explicada seja uma variável nominal, foi necessária a transformação dos dados, utilizando uma regra simples no *Excel*, que classificou os dados em quatro categorias: baixa (0 a 5%); média (>5 a 15%); alta (>15 a 25%); e muito alta (>25%). Essa classificação pode ser observada na figura 3.5.

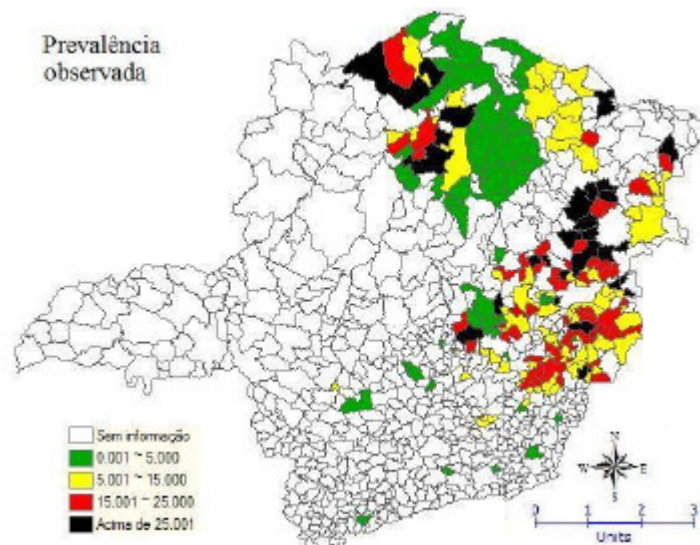


Figura 3.5 – Prevalência da Esquistossomose em 197 Municípios de Minas Gerais  
Fonte: MARTINS et al., 2007, p. 2844

Com a utilização de algoritmos de classificação, objetivou-se analisar as diferenças no padrão de comportamento das variáveis em relação à prevalência da doença. Visto que o algoritmo J4.8 gera as regras de decisão e uma matriz de erros, é possível detectar prováveis problemas na classificação e também a comparação entre as classes, conforme mostra a tabela 3.2

Tabela 3.2 – Matriz de Erros Obtida com o Algoritmo J4.8

Classificada \ Observada	Baixa	Média	Alta	Muito alta
Baixa (46)	42	3	1	0
Média (73)	17	47	9	0
Alta (51)	5	14	25	7
Muito alta (27)	2	4	8	13

Fonte: MARTINS et al. 2007, p. 2844

É possível observar também que das 197 amostras, 127 são classificadas corretamente, sendo das 70 amostras classificadas incorretamente, 58 classificadas com um erro de classe, 10 com dois erros e somente duas com três erros. Levando em consideração as 46 amostras da classe de prevalência baixa, 91,3% das amostras foram classificadas corretamente. Isto significa que o resultado pode ser considerado satisfatório, visto que os recursos para combate à doença são escassos e com este resultado dificilmente eles serão alocados em áreas com menor prevalência. Com as amostras de prevalência média, 64,4% são classificadas corretamente, já com as amostras de prevalência alta e muito alta, 49% e 48%, respectivamente, são classificadas corretamente.

A árvore de decisão para a prevalência da doença em relação a algumas variáveis preditivas, selecionadas pelo *software Weka* por conterem maior quantidade de informações, é apresentada na figura 3.6.

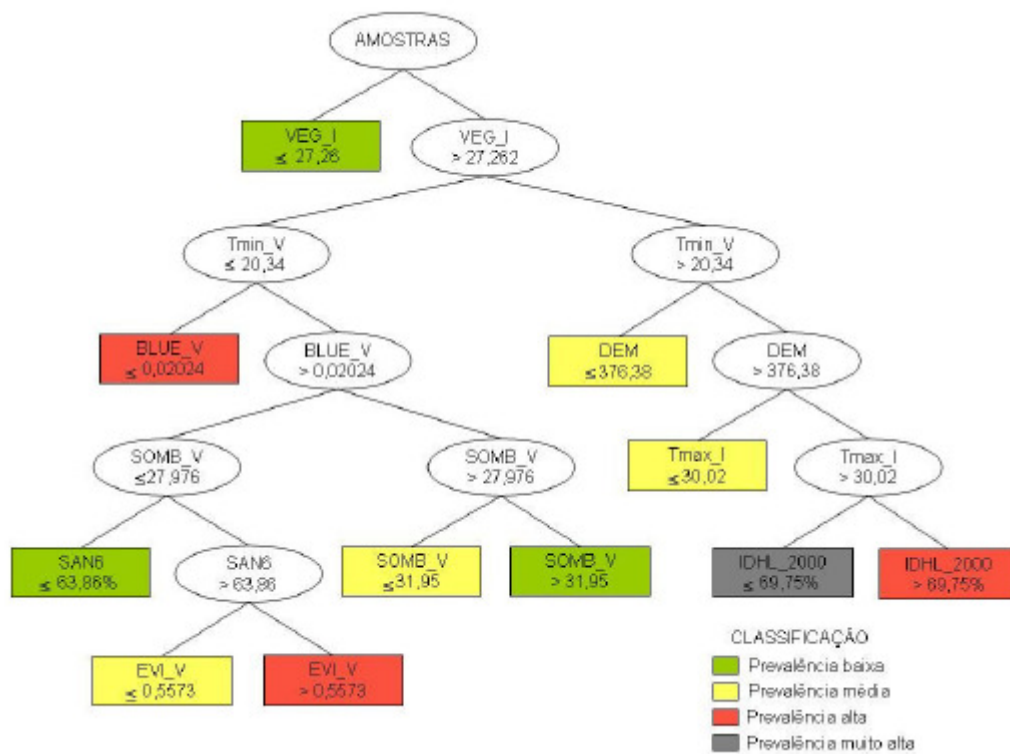


Figura 3.6 – Árvore de Decisão Gerada pelo Algoritmo J4.8  
 Fonte: MARTINS et al, 2007, p. 2845

Os resultados da utilização de árvore de decisão para verificação da prevalência de esquistossomose podem ser considerados satisfatórios, pois com isso foi possível enxergar os locais com maior índice de contaminação, onde os recursos e a conscientização da população devem ser tratados com maior prioridade. Também é possível verificar que o resultado é coerente, pois o habitat ideal para o caramujo e condições de vida das pessoas são fatores importantes para a existência da doença.

Assim, é possível ter uma melhor percepção da importância da mineração de dados em áreas diferentes, fazendo com que questões sociais e patrimoniais sejam resolvidas de uma forma inteligente e eficaz. O aproveitamento das informações desconhecidas presentes nas bases de dados faz com que conhecimento seja gerado e a utilização das técnicas seja adotada por um número cada vez maior de organizações.

Pretende-se apresentar uma proposta de utilização de técnicas de *data mining* na base de dados de uma academia de musculação, com o intuito de fornecer informações sobre os alunos aos responsáveis pela academia. Todas as etapas do processo de KDD serão realizadas para que isto seja possível e, através da utilização de algoritmos de classificação, se tenha um

maior conhecimento e domínio sobre os dados contidos nessa base, gerando árvores de decisão a fim de se analisar e interpretar o resultado delas.

## 4 ESTUDO DE CASO NA ACADEMIA DE MUSCULAÇÃO

Através de uma parceria do Centro Universitário Feevale com a Universidade de Moçambique foi obtida a base de dados da academia de musculação, da qual um professor dessa Universidade é proprietário. A base de dados está no formato Access e possui dados cadastrais dos alunos, informações financeiras, além do histórico das avaliações de medidas realizadas periodicamente nos membros.

De uma maneira geral, pretende-se extrair informações dessa base de dados e utilizá-las para criar arquivos e aplicá-los no *software* de mineração de dados Weka, maiores detalhes sobre este *software* podem ser encontrados em Gonchoroski (2007). Essas informações representam dados técnicos, como o peso, altura e pressão arterial, por exemplo, coletados através das avaliações periódicas realizadas com os alunos da academia.

Essa base de dados é chamada de GESPHY e é formada por três bases separadas que são unificadas no sistema da academia. A primeira delas é a PhyData, que possui informações cadastrais dos alunos, além dos planos e serviços prestados pela academia. A segunda é a PhyDesp que contém informações referentes a fornecedores e contas pagas pela organização. E por último, a PhyTecD que possui dados referentes às avaliações dos alunos.

Visto que essa última base possui somente os dados técnicos referentes aos alunos e também visando otimizar e acelerar a consulta dos dados, a tabela DadosResumo da base PhyTecD (1) foi importada para a base PhyData (2) conforme demonstra a figura 4.1. Essa importação foi realizada utilizando o *software* Microsoft Access e manteve os dados preservados, ou seja, sem alteração alguma.

Após isso, foi criada uma consulta em linguagem PHP (3), que extrai todas as informações pretendidas pelo programador e as grava em um arquivo formato texto (4). Esse

tipo de arquivo pode ser facilmente importado para outro *software* que possua mais recursos para edição e análise, sendo que nesse trabalho foi utilizado o Microsoft Excel (5), já que neste *software* os dados ficam divididos em células que podem ser manipuladas utilizando macros. Esse procedimento será explicado mais adiante.

Esse arquivo Excel é transformado em vários arquivos diferentes com extensão arff, (6) sendo que estes arquivos são carregados no *software* Weka (7), que realiza análise dos mesmos, gerando uma árvore de decisão. Essa árvore é analisada pelo *decision maker* (8), e caso o resultado dela não seja satisfatório, os parâmetros da consulta PHP devem ser alterados (9). Caso o resultado seja satisfatório, o conhecimento obtido pode ser aplicado na academia (10). Todos estes procedimentos serão explicados detalhadamente nas seções seguintes.

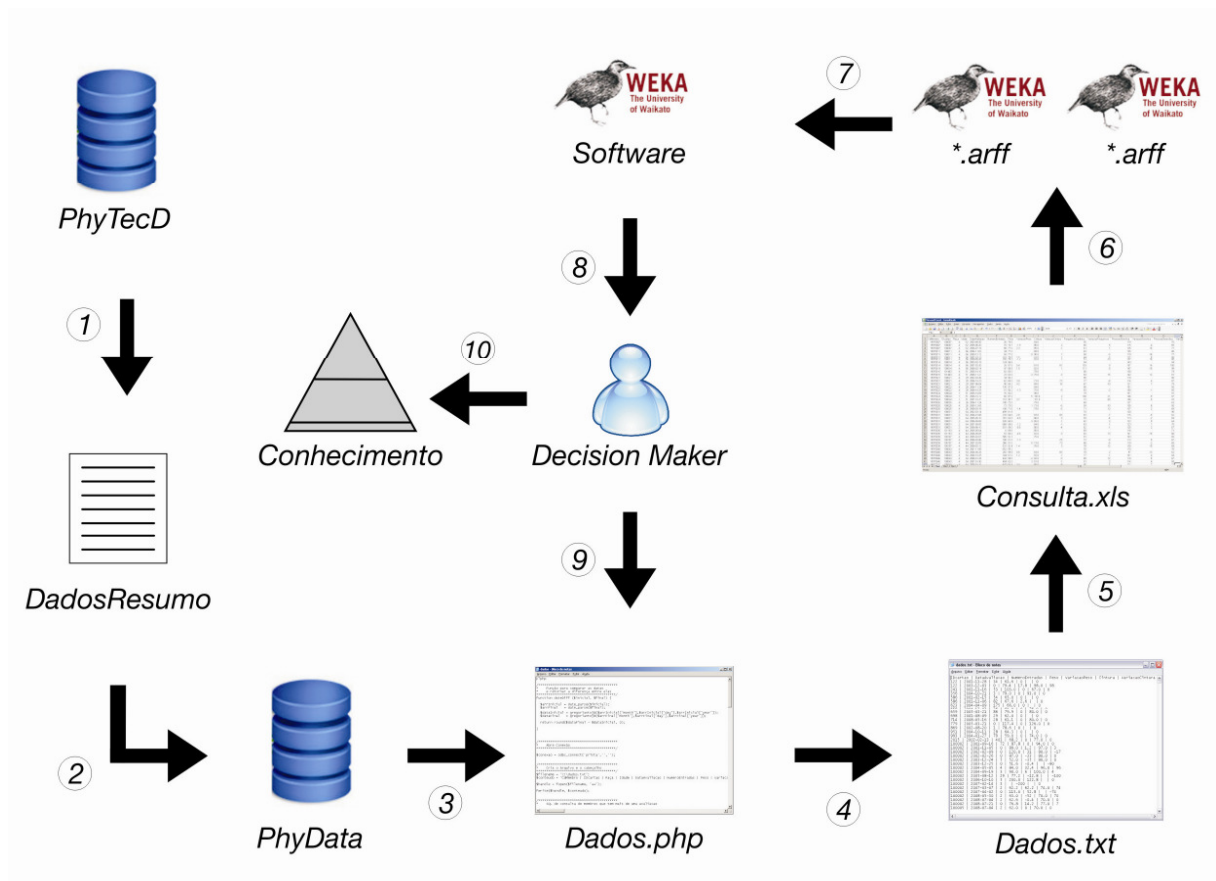


Figura 4.1 – Metodologia Utilizada para Realização do Trabalho

Fonte: do autor.

#### 4.1 Pré-Processamento

Conforme citado no capítulo 1 deste trabalho, o Pré-Processamento possui grande relevância no processo de descoberta de conhecimento. Fazem parte dessa etapa a seleção,

limpeza e transformação dos dados. Para uma melhor adequação dos dados e aplicação de algoritmos de mineração neles, foi necessário realizar todas essas etapas para a realização deste trabalho. A seguir serão relatados todos os procedimentos realizados na execução dessas etapas.

#### **4.1.1 Seleção dos Dados**

Após analisar a base, foi possível identificar alguns atributos que poderiam ser estudados mais profundamente, a fim de transformá-los em possíveis perfis de alunos da organização. A partir disso, foi necessário definir a maneira como esses atributos iriam ser extraídos do banco, tornando-se dados analisáveis.

A linguagem escolhida para realizar essa extração foi a PHP, uma vez que essa linguagem é mais familiar ao autor deste trabalho. Após a programação dessa consulta e aplicação dela na base, foi gerado um arquivo em formato de arquivo texto (.txt) com os dados extraídos, com o objetivo de transformá-lo e analisar possíveis resultados no Weka.



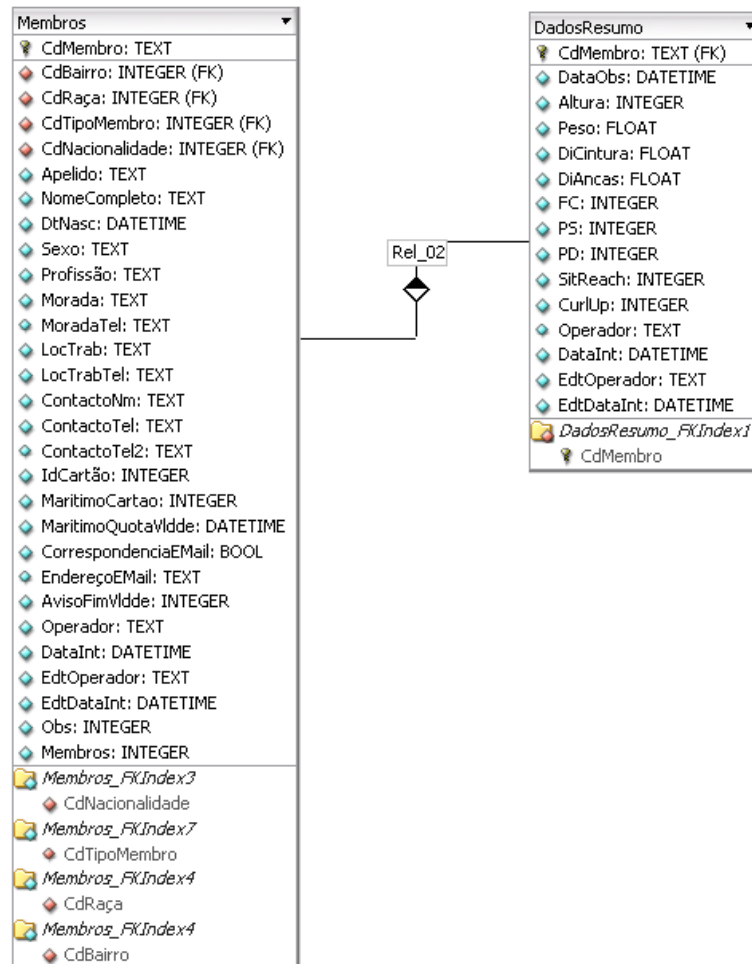


Figura 4.2 – Relacionamento das Tabelas e Atributos

Fonte: do autor.

A figura 4.2 demonstra a relação das tabelas utilizadas para a extração dos dados e os atributos presentes nelas. Na tabela DadosResumo, o atributo FC significa a frequência cardíaca da pessoa em repouso, o atributo PS é a pressão sistólica do aluno e o PD se refere a pressão diastólica. A seguir, a figura 4.3 demonstra o início do primeiro arquivo de texto gerado pela consulta.

IDCartao	DataAvaliacao	NumeroEntradas	Peso	VariacaoPeso	Cintura	VariacaoCintura
122	2001-11-28	34	61.6	0	0	
122	2003-12-03	0	79.4	17.8	98.0	98
241	2001-11-16	53	103.0	0	67.0	0
356	2004-10-21	3	79.0	0	93.0	0
586	2001-02-15	54	65.0	0	0	0
586	2001-11-06	82	67.6	2.6	0	0
623	2004-04-09	175	66.0	0	0	0
658	2004-12-02	26	61.0	0	76.0	0
659	2005-03-21	86	79.9	0	80.0	0
698	2001-08-09	29	62.0	0	0	0
714	2008-05-16	28	61.1	0	80.0	0
779	2005-03-21	0	117.4	0	126.0	0
869	2002-08-20	1	78.6	0	0	0
951	2004-10-11	28	64.3	0	0	0
993	2004-01-27	79	59.0	0	74.0	0
2015	2002-02-13	403	68.2	0	77.0	0
100002	2001-09-16	72	87.8	0	94.0	0
100002	2001-11-05	0	89.0	1.2	97.0	3
100002	2002-02-09	0	120.0	31	80.0	-17
100002	2002-02-26	0	87.0	-33	80.0	0
100002	2003-12-24	5	52.0	-35	80.0	0
100002	2003-12-25	0	51.6	-0.4	0	-80
100002	2004-05-05	4	84.0	32.4	96.0	96
100002	2004-09-19	3	90.0	6	100.0	4
100002	2005-08-12	29	77.2	-12.8	0	-100
100002	2006-10-10	5	200.0	122.8	0	0
100002	2007-02-14	3	0	-200	0	0
100002	2007-03-07	2	62.2	62.2	74.0	74
100002	2007-04-02	0	115.0	52.8	0	-74
100002	2008-05-30	2	63.0	-52	70.0	70
100002	2008-07-04	2	62.6	-0.4	70.0	0
100002	2008-07-21	0	76.8	14.2	77.0	7
100005	2008-07-04	2	62.0	0	70.0	0

Figura 4.3 - Primeiro Arquivo de Texto Gerado

Fonte: do autor.

Este primeiro arquivo possui os seguintes atributos:

- IdCartao: código do aluno dentro da academia;
- DataAvaliacao: data em que foram feitas as avaliações de peso e medidas realizadas nos alunos;
- NumeroEntradas: quantidade de vezes que a pessoa frequentou a academia entre uma avaliação e outra. Essa quantidade não está gravada no banco de dados, e foi gerada pela consulta PHP que conta quantas vezes que o aluno entrou na academia entre a avaliação atual e a anterior. O registro de entrada do aluno na academia fica armazenado na tabela Entradas da base PhyData;
- Peso: peso em quilogramas do aluno na avaliação em questão;

- VariacaoPeso: variação do peso do aluno entre uma avaliação e outra. Essa variação não está gravada no banco de dados, sendo gerada através da consulta PHP criada, que subtrai o valor do peso da avaliação anterior do peso atual do aluno;

- Cintura: medida da cintura do aluno, em centímetros, na avaliação em questão;

- VariacaoCintura: variação da medida da cintura do aluno entre as avaliações. Essa variação não está gravada no banco de dados, sendo gerada através da consulta PHP criada utilizando a mesma fórmula que calcula a variação do peso. É feita a subtração da medida da cintura do aluno na avaliação anterior da medida atual.

O arquivo de texto gerado foi importado para uma planilha do Microsoft Excel, a fim de tornar a visualização dos dados mais fácil. A partir desta consulta, os 6.485 registros obtidos foram analisados e tornou-se visível que seria necessário obter maiores informações sobre os alunos da academia para tentar atingir o objetivo de criar perfis dos mesmos.

Após esta constatação, a consulta PHP foi modificada, sendo que, foram incluídos outros dados presentes na base, além daqueles já citados anteriormente. Sendo assim, o segundo arquivo texto gerado possui os seguintes dados, além das informações já presentes do primeiro arquivo:

- Raça: código que determina a raça do aluno;

- Idade: idade do aluno. Essa informação não está gravada no banco de dados, sendo gerada através da data de nascimento da pessoa, que é transformada em idade através de uma função adicionada ao *script* da consulta;

- FrequenciaCardiaca: frequência cardíaca do aluno no momento da avaliação;

- VariacaoFrequencia: variação da frequência cardíaca do aluno entre uma avaliação e outra. Para o cálculo dessa variação, foi subtraído o valor referente à frequência cardíaca da pessoa na avaliação anterior da frequência na avaliação atual;

- PressaoSistolica: pressão sistólica do aluno no momento da avaliação;

- VariacaoSistolica: variação da pressão sistólica do aluno entre uma avaliação e outra. Essa variação não está gravada no banco de dados, sendo gerada através da consulta PHP, utilizando a mesma fórmula que calcula a variação da frequência cardíaca;

- *PressaoDiastolica*: pressão diastólica do aluno no momento da avaliação;
- *VariacaoDiastolica*: variação da pressão diastólica do aluno entre uma avaliação e outra. Essa variação não está gravada no banco de dados, sendo gerada através da consulta PHP criada;
- *DiasEntreAvaliaco*: quantidade de dias que passou entre uma avaliação e outra. Esse valor é calculado através de uma função adicionada ao *script* da consulta, que compara a data de avaliação anterior com a data atual e calcula a quantidade de dias transcorridos entre essas avaliações.

Além desses dados, este arquivo possui também todas as faixas de variação geradas na fase de transformação dos dados, que será explicada mais adiante. Assim como no primeiro arquivo gerado, este arquivo texto foi importado para uma planilha do Excel, a fim de facilitar a visualização dos dados e tornar a manipulação dos mesmos mais simples.

#### **4.1.2 Limpeza dos Dados**

Na etapa de limpeza, alguns dados inconsistentes foram retirados, visando evitar possíveis problemas e falhas por parte do algoritmo de árvores de decisão. Para tanto, foram desconsiderados atributos e até mesmo registros que não seguiam os padrões pretendidos para análise do *software* Weka.

Um exemplo que pode ser citado é o atributo Raça da tabela Membros. Alguns alunos não possuem sua raça indicada, sendo que nesses casos o valor presente nesse campo é nulo. Sendo assim, os arquivos gerados que utilizassem o atributo Raça iriam possuir valores nulos entre os registros. Visto que o algoritmo de árvore de decisão não interpreta valores nulos, foi necessário considerar somente os alunos que possuem sua raça indicada.

Para isso, foi criado um script que foi incluído na consulta PHP, que considera somente os membros que possuem no atributo Raça um valor diferente de nulo. Após isso, os arquivos foram gerados e aplicados no Weka.

Outro exemplo da limpeza realizada se refere à primeira avaliação do aluno. Neste momento não é possível calcular a variação do peso da pessoa, por exemplo, pois não há um peso anterior para realizar a comparação. Sendo assim, os valores da primeira avaliação não

são adicionados ao arquivo texto, mas ficam registrados para ser realizado o cálculo da variação do peso da segunda avaliação em diante.

### **4.1.3 Transformação dos Dados**

Concluída a etapa de limpeza dos dados viu-se a necessidade de padronizar alguns registros, transformando-os em faixas de valores a fim de possibilitar que o algoritmo de mineração de dados os analise de maneira mais eficaz. Assim, alguns dados considerados relevantes para o estudo na etapa de seleção foram transformados.

Para melhor entendimento, pode-se citar um exemplo desse processo de transformação. A variação de peso dos alunos entre uma avaliação e outra possui valores entre -35 e 35 quilos, levando em conta que a variação negativa significa que o aluno perdeu peso e a variação positiva significa que o aluno ganhou peso. Caso o valor dessa variação seja zero, significa que a pessoa continuou com o mesmo peso entre uma avaliação e outra.

Esse valor é calculado comparando o peso do aluno na data de avaliação atual, com o peso dele na data de avaliação anterior. Existe ainda o valor da variação nulo, o que significa que aquela avaliação é a primeira que o aluno fez, pois não há data anterior para ser comparada com a data de avaliação atual.

Num primeiro momento os valores de cada faixa de variação foram escolhidos aleatoriamente. Os valores das faixas de variação do peso, da cintura, da frequência cardíaca, da pressão sistólica e da pressão diastólica são os mesmos, sendo que a única diferença é o nome da faixa de variação. Foram criadas também faixas para a quantidade de vezes que o usuário frequentou a academia entre as avaliações, sendo que essa informação está presente em todos os arquivos testados. Sendo assim, foram criadas faixas de variação de peso de acordo com a tabela 4.1.

Tabela 4.1 – Faixa de Variação do Peso

Nome da Faixa	Valor Inicial	Valor Final
FNaoVariaPeso	<i>Null</i>	<i>Null</i>
FVariaPeso0		< -30
FVariaPeso1	>= -30	<= -15
FVariaPeso2	> -15	<= -10
FVariaPeso3	> -10	<= -5
FVariaPeso4	>-5	<= -2
FVariaPeso5	> -2	<= 0
FVariaPeso6	> 0	<= 2
FVariaPeso7	> 2	<= 5
FVariaPeso8	> 5	<= 10
FVariaPeso9	> 10	<= 15
FVariaPeso10	> 15	<= 30
FVariaPeso11	>30	

Fonte: do autor.

Após analisar alguns dados, viu-se a necessidade de eliminar a faixa FNãoVariaPeso, pois essa faixa equivale a primeira avaliação do aluno, quando ainda não é possível realizar uma comparação entre o peso da avaliação anterior e a atual. Sendo assim, passaram a ser considerados somente os dados referentes à segunda avaliação em diante.

Outro caso de transformação que convém ser citado refere-se à idade do membro. Na tabela Membros da base PhyData está armazenada a data de nascimento do aluno. Essa data

foi transformada em idade, através de uma função criada na consulta PHP e após isso foi necessário criar faixas etárias, a fim de facilitar a análise após utilização do algoritmo de mineração de dados e também porque a entrada do algoritmo utilizado exige isso. Sendo assim, foram criadas faixas etárias conforme demonstra a tabela 4.2.

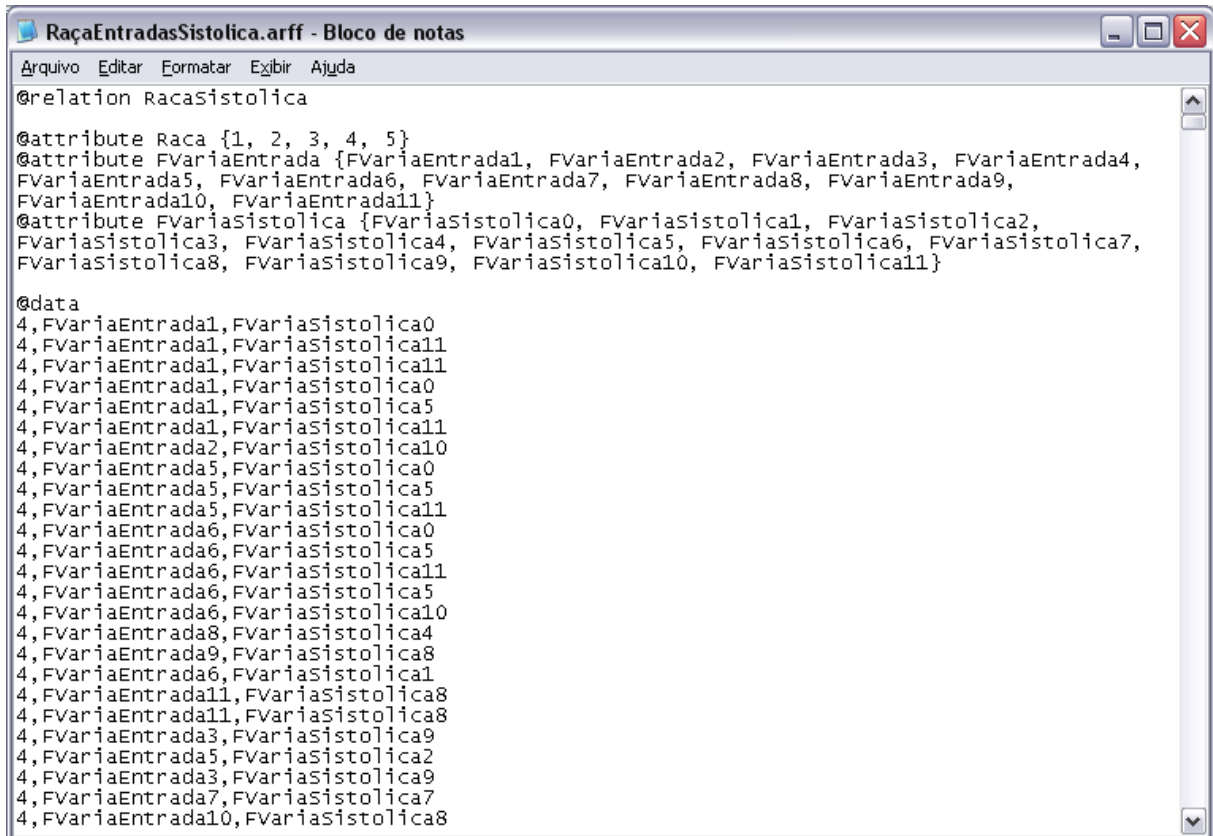
Tabela 4.2 – Faixas Etárias

Nome da Faixa	Valor Inicial	Valor Final
F1	0	9
F2	10	19
F3	20	29
F4	30	39
F5	40	49
F6	50	59
F7	60	69
F8	70	79
F9	80	89
F10	90	99

Fonte: do autor.

Visto que o arquivo Excel possui todos os registros selecionados na consulta, foram criados manualmente vários arquivos Excel com atributos que poderiam estar relacionados e seriam propícios para criação de perfis de alunos. Após isso, os arquivos Excel foram salvos no formato CSV (separado por vírgulas). Quando o arquivo CSV é salvo, surgem espaços em branco entre os registros, esses espaços devem ser retirados. É necessário também acrescentar um cabeçalho, necessário para aplicação do algoritmo presente no WEKA. Por fim, esses arquivos são salvos no formato arff, e então é aplicado o algoritmo de árvore de decisão através do *software* WEKA

Este cabeçalho que é inserido nos arquivos possui primeiramente a *tag @relation*, que identifica o nome da relação que será estudada pelo WEKA. Em seguida, é criada a *tag @attribute*, que identifica os atributos presentes no arquivo, que são seguidos pelos possíveis valores presentes no mesmo. Por último, é inserida a *tag @data*, que informa ao *software* o ponto onde iniciam os dados que devem ser analisados. A figura 4.4 demonstra o cabeçalho e alguns registros de um dos arquivos gerados para análise.



```

RaçaEntradasSistolica.arff - Bloco de notas
Arquivo Editar Formatar Exibir Ajuda
@relation RacaSistolica

@attribute Raca {1, 2, 3, 4, 5}
@attribute FVariaEntrada {FVariaEntrada1, FVariaEntrada2, FVariaEntrada3, FVariaEntrada4,
FVariaEntrada5, FVariaEntrada6, FVariaEntrada7, FVariaEntrada8, FVariaEntrada9,
FVariaEntrada10, FVariaEntrada11}
@attribute FVariaSistolica {FVariasistolica0, FVariasistolica1, FVariasistolica2,
FVariasistolica3, FVariasistolica4, FVariasistolica5, FVariasistolica6, FVariasistolica7,
FVariasistolica8, FVariasistolica9, FVariasistolica10, FVariasistolica11}

@data
4,FVariaEntrada1,FVariasistolica0
4,FVariaEntrada1,FVariasistolica11
4,FVariaEntrada1,FVariasistolica11
4,FVariaEntrada1,FVariasistolica0
4,FVariaEntrada1,FVariasistolica5
4,FVariaEntrada1,FVariasistolica11
4,FVariaEntrada2,FVariasistolica10
4,FVariaEntrada5,FVariasistolica0
4,FVariaEntrada5,FVariasistolica5
4,FVariaEntrada5,FVariasistolica11
4,FVariaEntrada6,FVariasistolica0
4,FVariaEntrada6,FVariasistolica5
4,FVariaEntrada6,FVariasistolica11
4,FVariaEntrada6,FVariasistolica5
4,FVariaEntrada6,FVariasistolica10
4,FVariaEntrada8,FVariasistolica4
4,FVariaEntrada9,FVariasistolica8
4,FVariaEntrada6,FVariasistolica1
4,FVariaEntrada11,FVariasistolica8
4,FVariaEntrada11,FVariasistolica8
4,FVariaEntrada3,FVariasistolica9
4,FVariaEntrada5,FVariasistolica2
4,FVariaEntrada3,FVariasistolica9
4,FVariaEntrada7,FVariasistolica7
4,FVariaEntrada10,FVariasistolica8

```

Figura 4.4 – Exemplo de Cabeçalho e Registros de um Arquivo arff

Fonte: do autor.

Como citado anteriormente, os arquivos foram criados com os atributos que, quando relacionados, podem identificar perfis de alunos da academia. Esses registros foram escolhidos levando em conta dados técnicos referentes às avaliações dos alunos e considerando também características raciais e etárias. Para tanto, cada arquivo possui três atributos, de acordo com a tabela 4.3 que descreve cada um deles.



Tabela 4.3 – Primeiros Arquivos Gerados e seus Atributos

Nome do arquivo	Atributos
FaixaEtariaEntradasCardiaca.arff	FaixaEtaria, FVariaEntrada, FVariaFrequencia
FaixaEtariaEntradasCintura.arff	FaixaEtaria, FVariaEntrada, FVariaCintura
FaixaEtariaEntradasDiastolica.arff	FaixaEtaria, FVariaEntrada, FVariaDiastolica
FaixaEtariaEntradasPeso.arff	FaixaEtaria, FVariaEntrada, FVariaPeso
FaixaEtariaEntradasSistolica.arff	FaixaEtaria, FVariaEntrada, FVariaSistolica
RaçaEntradasCardiaca.arff	Raça, FVariaEntrada, FVariaFrequencia
RaçaEntradasCintura.arff	Raça, FVariaEntrada, FVariaCintura
RaçaEntradasDiastolica.arff	Raça, FVariaEntrada, FVariaDiastolica
RaçaEntradasPeso.arff	Raça, FVariaEntrada, FVariaPeso
RaçaEntradasSistolica.arff	Raça, FVariaEntrada, FVariaSistolica

Fonte: do autor.

O arquivo `FaixaEtariaEntradasCardiaca.arff`, por exemplo, possui os dados referentes a faixa etária do aluno, a faixa que possui a quantidade de vezes que o aluno entrou na academia entre uma avaliação e outra e a faixa de variação da frequência cardíaca do aluno entre as avaliações. Os demais arquivos foram gerados seguindo a mesma linha de raciocínio, somente alterando alguns atributos.

Após criar todos os arquivos visando empregá-los na etapa de mineração de dados, os mesmos foram utilizados para criar dois arquivos distintos: um para treinamento e outro para teste. Para isso foram criadas macros no Microsoft Excel, conforme Stahnke (2008).

Essas macros primeiramente separam o conteúdo presente na planilha Excel original em duas outras planilhas, equivalentes a 70% e 30% do conteúdo total. Após isso, são criados dois arquivos no formato `arff`, um para treinamento e outro para teste, respectivamente. A tabela 4.4 apresenta todos os arquivos gerados dessa forma.

Tabela 4.4 – Arquivos de Treinamento e Testes Gerados e seus Atributos

Nome do arquivo	Atributos
FaixaEtariaEntradasCardiacaTreinamento.arff	FaixaEtaria, FVariaEntrada, FVariaFrequencia
FaixaEtariaEntradasCardiacaTeste.arff	FaixaEtaria, FVariaEntrada, FVariaFrequencia
FaixaEtariaEntradasCinturaTreinamento.arff	FaixaEtaria, FVariaEntrada, FVariaCintura
FaixaEtariaEntradasCinturaTeste.arff	FaixaEtaria, FVariaEntrada, FVariaCintura
FaixaEtariaEntradasDiastolicaTreinamento.arff	FaixaEtaria, FVariaEntrada, FVariaDiastolica
FaixaEtariaEntradasDiastolicaTeste.arff	FaixaEtaria, FVariaEntrada, FVariaDiastolica
FaixaEtariaEntradasPesoTreinamento.arff	FaixaEtaria, FVariaEntrada, FVariaPeso
FaixaEtariaEntradasPesoTeste.arff	FaixaEtaria, FVariaEntrada, FVariaPeso
FaixaEtariaEntradasSistolicaTreinamento.arff	FaixaEtaria, FVariaEntrada, FVariaSistolica
FaixaEtariaEntradasSistolicaTeste.arff	FaixaEtaria, FVariaEntrada, FVariaSistolica
RaçaEntradasCardiacaTreinamento.arff	Raça, FVariaEntrada, FVariaFrequencia
RaçaEntradasCardiacaTeste.arff	Raça, FVariaEntrada, FVariaFrequencia
RaçaEntradasCinturaTreinamento.arff	Raça, FVariaEntrada, FVariaCintura
RaçaEntradasCinturaTeste.arff	Raça, FVariaEntrada, FVariaCintura
RaçaEntradasDiastolicaTreinamento.arff	Raça, FVariaEntrada, FVariaDiastolica
RaçaEntradasDiastolicaTeste.arff	Raça, FVariaEntrada, FVariaDiastolica
RaçaEntradasPesoTreinamento.arff	Raça, FVariaEntrada, FVariaPeso
RaçaEntradasPesoTeste.arff	Raça, FVariaEntrada, FVariaPeso
RaçaEntradasSistolicaTreinamento.arff	Raça, FVariaEntrada, FVariaSistolica
RaçaEntradasSistolicaTeste.arff	Raça, FVariaEntrada, FVariaSistolica

Fonte: do autor.

Realizando alguns testes com os arquivos gerados, encontrou-se a importância de uma conversa com alguma pessoa da área de Educação Física, que pudesse fornecer algumas informações e dar sugestões de valores e intervalos para as faixas de variação. Assim, foi realizada uma entrevista com o proprietário da academia, que possui doutorado na área e

também com um Bacharel em Enfermagem, atuante na profissão e com experiência em medidas de pressão e frequência cardíaca.

A faixa de variação do peso, por exemplo, foi dividida em intervalo de 2 kg sendo desconsiderados valores maiores que 20 kg e menores do que -20 kg. As faixas de variação da cintura foram reformuladas, recebendo novos intervalos de dados, variando 5 centímetros como demonstra a tabela 4.5.

Tabela 4.5 – Faixas de Variação da Cintura Reformuladas

Nome da Faixa	Valor Inicial	Valor Final
FVariaCintura1	$\geq -30$	$\leq -25$
FVariaCintura2	$> -25$	$\leq -20$
FVariaCintura3	$> -20$	$\leq -15$
FVariaCintura4	$> -15$	$\leq -10$
FVariaCintura5	$> -10$	$\leq -5$
FVariaCintura6	$> -5$	$\leq 0$
FVariaCintura7	$> 0$	$\leq 5$
FVariaCintura8	$> 5$	$\leq 10$
FVariaCintura9	$> 10$	$\leq 15$
FVariaCintura10	$> 15$	$\leq 20$
FVariaCintura11	$> 20$	$\leq 25$
FVariaCintura12	$> 25$	$\leq 30$

Fonte: do autor.

Como pode se perceber, os valores maiores que 30 cm e menores que -30 cm foram desconsiderados, pois se entende que uma variação tão alta assim pode ser considerada como

ruído. Com essas doze faixas de variação de cintura criadas de acordo com as indicações dos profissionais da área, pretende-se obter melhores resultados após a aplicação do algoritmo de árvore de decisão nos registros. Cabe salientar também que as faixas etárias foram reformuladas após essa entrevista, sendo que seus valores ficaram de acordo com a tabela 4.6.

Tabela 4.6 – Faixas Etárias Reformuladas

Nome da Faixa	Valor Inicial	Valor Final
Adolescente	0	19
AdultoJovem	20	39
MeiaIdade	40	65
TerceiraIdade	66	100

Fonte: do autor.

Dessa forma, as informações fornecidas nestas entrevistas contribuíram para que a consulta PHP fosse modificada, sendo atribuídos os novos valores para as faixas de variação. As faixas de variação de peso, cintura, frequência cardíaca, pressão sistólica e pressão diastólica foram reformuladas.

## 4.2 Mineração de Dados

Na etapa de Mineração de Dados, todos os arquivos gerados são minerados pelo algoritmo J48. Esse algoritmo gera uma árvore de decisão durante a fase de treinamento, além de ilustrar em forma de gráfico a relação entre os atributos presentes nos arquivos conforme demonstra a figura 4.5.

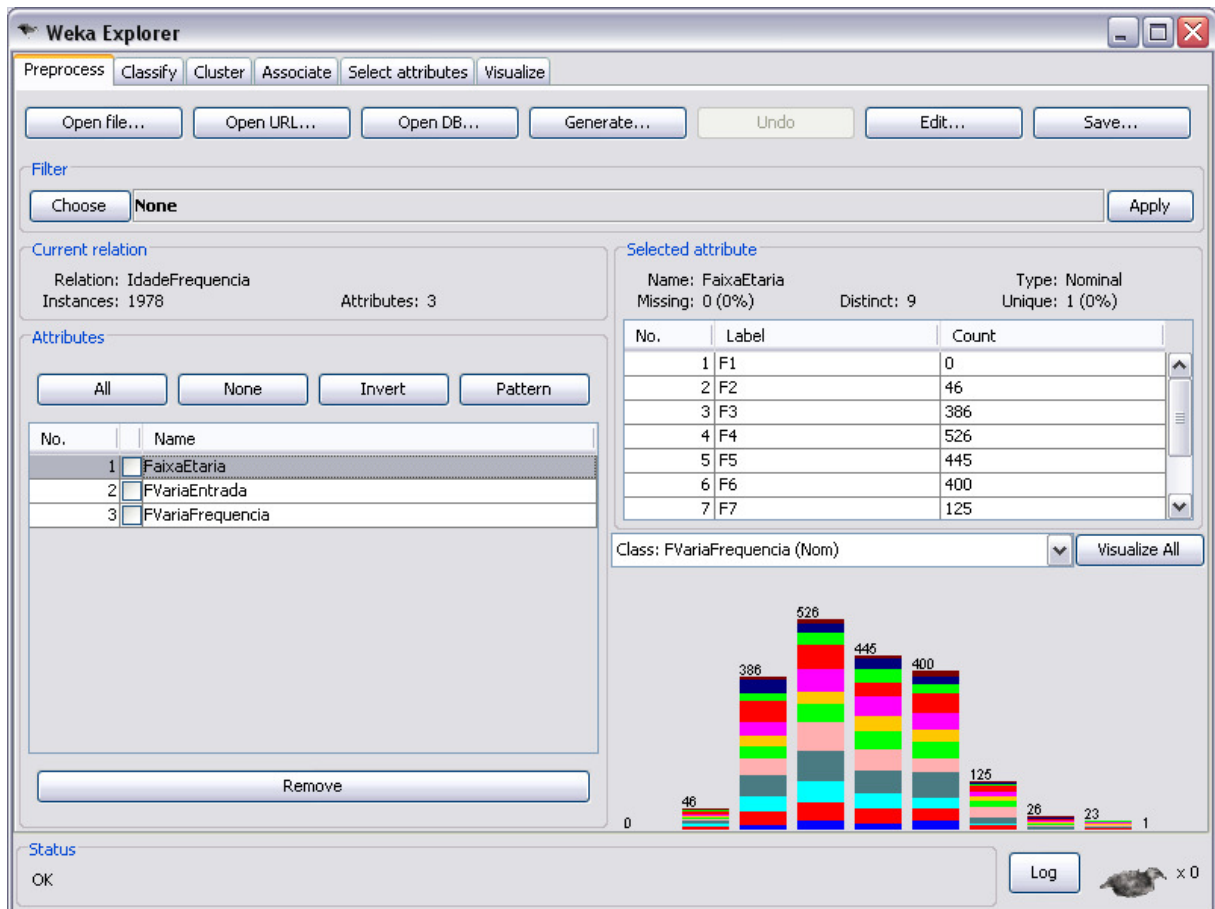


Figura 4.5 – Gráfico Gerado pelo *Software Weka*

Fonte: do autor.

Nessa figura, pode-se perceber também que é informado o número de registros e a quantidade de atributos presentes no arquivo. É informada também a quantidade de registros que se encaixam em cada faixa, podendo-se fazer a relação entre as mesmas para um melhor exame.

As barras do gráfico representam o atributo *FaixaEtaria* e as diferentes cores presentes nas barras representam o atributo *FVariaFrequencia*. O número que aparece acima da barra, significa a quantidade de registros que se encaixam em cada faixa etária. Dessa forma, é possível ver a relação entre os atributos *FaixaEtaria* e *FVariaFrequencia* e também a relação entre qualquer outro atributo presente nesse arquivo.

Após análise desses gráficos é possível realizar os testes de validação da árvore de decisão, estes testes podem ser realizados de várias maneiras. Neste trabalho isso foi feito de três formas: utilizando o mesmo arquivo para treinamento e testes; dividindo o arquivo em

70% e 30% do seu total, sendo que os 70% são utilizados para treinamento e geração da árvore de decisão e os 30% são utilizados para testar a árvore gerada; e utilizando o *Cross-Validation* com *Folds* 10, onde a quantidade de registros do arquivo é dividida em 10 partes, cada parte é testada separadamente e após isso é feita a média de acertos e erros de cada parte, formando o resultado final.

Essa árvore de decisão é resultado do treinamento do arquivo carregado, sendo que após esse treinamento é realizado o teste de acordo com os parâmetros indicados pelo *decision maker*. O único parâmetro alterado para a realização de testes neste trabalho foi o *confidence factor*, que é o fator de confiança utilizado para a geração da árvore de decisão. Quanto maior o valor colocado no fator de confiança, maior será a árvore de decisão gerada e menos precisa será a árvore gerada. Nos demais parâmetros foram mantidos os valores *default* do Weka. A figura 4.6 ilustra um exemplo de árvore de decisão gerada.

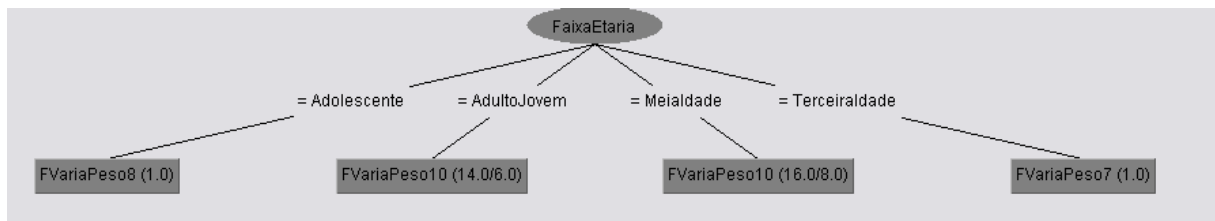


Figura 4.6 – Árvore de Decisão Gerada pelo Algoritmo J48

Fonte: do autor.

O objetivo principal do *decision maker* é fazer com que essa árvore de decisão fique cada vez menor e apresente resultados dos testes cada vez mais precisos. Para isso é necessário realizar modificações nos arquivos a fim de fazer com que esses resultados tragam algum benefício para a empresa em questão.

## 5 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Com base nos arquivos criados, estabeleceu-se o objetivo de encontrar alguma informação que possa ser útil à academia no sentido de poder entender melhor o desenvolvimento de seus alunos quanto à parte física. Sendo assim, todos os arquivos criados foram treinados e testados pelo algoritmo J48 do Weka com diversos parâmetros e seus resultados serão exibidos a seguir.

Todo o processo de descoberta de conhecimento se torna árduo, à medida que as etapas do processo precisam ser repetidas cada vez que não se encontra um resultado contundente. Neste trabalho isso não foi diferente: todos os arquivos criados foram testados, as faixas de variação foram redefinidas, a forma como o teste é realizado foi modificada, tudo isso com o objetivo de se obter o melhor resultado possível.

Ao realizar estes testes todas as árvores geradas foram analisadas e, à medida que os resultados exibidos não apresentaram um percentual de acerto relevante os arquivos eram ajustados. Assim foram realizadas novas tentativas, sendo que as árvores foram geradas novamente e os resultados foram analisados novamente.

Já que a academia possui diversas categorias de membros como, por exemplo, Membro Família, Membro Sênior e Membro Prata, um arquivo foi gerado para relacionar a raça da pessoa, a categoria de membro dela na academia e seu sexo. Contudo, foi constatado que 90% dos membros estão incluídos na mesma categoria de sócio, o que torna o resultado da mineração de dados tendencioso e pouco útil.

A seguir serão exibidos os resultados obtidos com os últimos testes gerados, com as faixas de variação recomendadas pelo proprietário da academia. Esses resultados serão comentados e as particularidades de cada árvore gerada neste último teste serão apresentadas.

## 5.1 Relação Entradas - Faixa Etária - Variação da Cintura

O arquivo `FaixaEtariaEntradaCinturaTreinamento.arff`, conforme citado anteriormente, possui a faixa etária do aluno, a faixa de variação de entradas na academia entre uma avaliação e outra, e a faixa de variação da cintura entre as avaliações. Este arquivo possui 70% do total de registros das avaliações presentes na base de dados, sendo que os outros 30% estão no arquivo `FaixaEtariaEntradaCinturaTeste.arff`.

Aplicando o algoritmo de árvore de decisão J48 no arquivo de treinamento e utilizando o arquivo de teste para testar a árvore com fator de confiança 0.4, diversas informações são demonstradas através da *decision tree* gerada. A figura 5.1 demonstra parte da árvore de decisão gerada utilizando estes arquivos. Foi possível apresentar somente parte da árvore, pois a árvore completa é muito grande, possuindo 35 folhas.

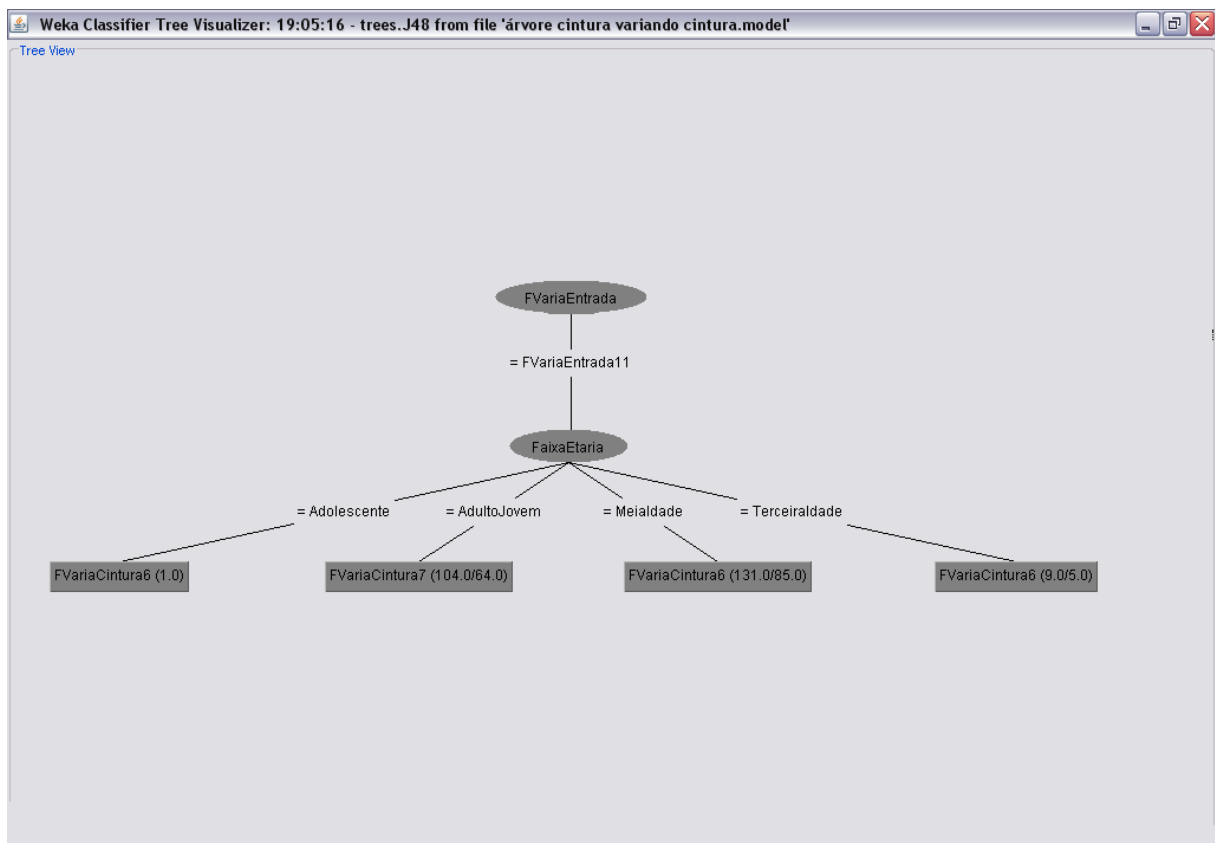


Figura 5.1 – Árvore de Decisão da Variação da Cintura

Fonte: do autor.

Foi utilizado o fator de confiança 0.4, pois fatores com valores menores do que esse geraram árvores com somente um nodo, o que torna o resultado da mineração inconsistente.



Através dessa imagem, é possível perceber que a maior parte dos registros se encontram nas faixas etárias AdultoJovem e MeiaIdade.

É possível perceber que os alunos que frequentam a academia entre 100 e 199 vezes entre as avaliações e pertencem à faixa etária MeiaIdade se encaixam na faixa de variação de cintura FVariaCintura6. Portanto, o valor de perda de cintura do aluno ficou entre -5 e 0 centímetros, ou seja, uma variação pequena mas que já demonstra que a assiduidade da pessoa pode fazer com que sua cintura diminua ou pelo menos continue do mesmo tamanho. Dos 131 registros que eram esperados nessa faixa de variação de cintura, o teste do algoritmo errou 85, possuindo 35% de acertos.

Já com essa mesma variação de entradas na academia, porém com faixa etária AdultoJovem, se encaixaram na variação de cintura FVariaCintura7, o que equivale a uma variação de 0 a 5 centímetros de cintura. Dos 104 registros esperados para esta faixa, o algoritmo acertou 40, o que corresponde a 38% de acerto.

Cabe salientar que esses valores foram obtidos utilizando somente uma parte da árvore gerada. O acerto obtido com a árvore inteira foi de 33%, resultado relativamente baixo. Isto pode significar que o algoritmo empregado não traz os melhores resultados com o tipo de dados utilizados, ou que pode ser realizada mais uma reformulação nas faixas para tentar obter um resultado melhor.

## **5.2 Relação Entradas - Variação Frequência - Faixa Etária**

Neste relacionamento, o arquivo criado possui a faixa de variação de entradas do aluno na academia entre uma avaliação e outra, a faixa etária do mesmo e a faixa de variação da frequência cardíaca em repouso, entre as avaliações. Foi criado um arquivo para treinamento e outro para testes, como citado na seção anterior, e gerada a árvore de decisão do relacionamento desses atributos.

O fator de confiança utilizado para esse teste foi 0.25, pois quando este é diminuído as árvores ficam mais genéricas, chegando ao ponto de possuir apenas um nodo. No total, a árvore gerada possui 31 folhas, o que demonstra que ela é muito grande. Sendo assim, foi retirada uma parte dela para análise, como demonstra a figura 5.2.

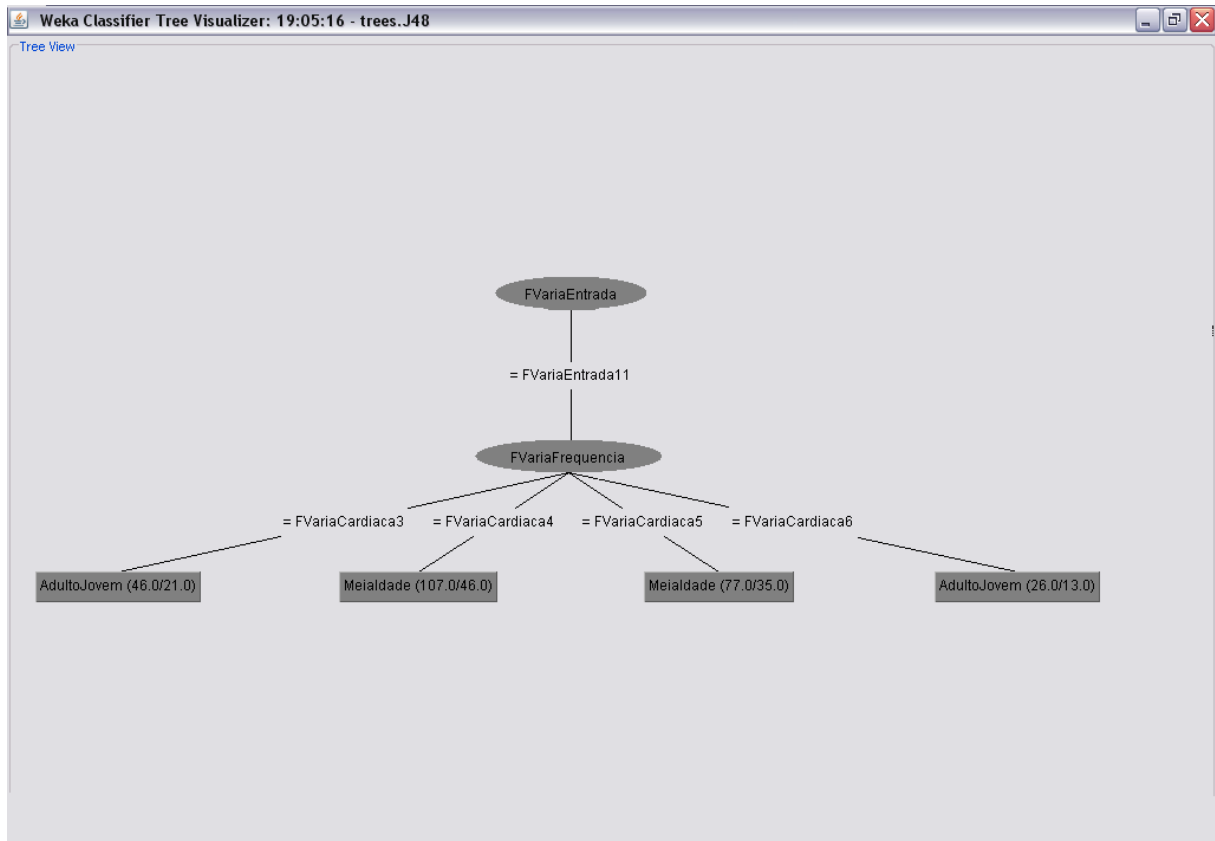


Figura 5.2 – Árvore de Decisão da Variação da Frequência Cardíaca  
Fonte: do autor.

Pode-se perceber pela imagem que somente as faixas de variação intermediárias da frequência cardíaca são exibidas, pois é onde se concentra a maioria dos registros, ou seja, quanto mais alta a variação menos registros a faixa possui. Da mesma forma que aconteceu com a cintura, a maior parte dos registros se encaixa nas faixas de variação com valores menores, pois estão dentro de um padrão de variação normal.

Quando um aluno se encaixa na faixa de variação de entrada FVariaEntrada11, que corresponde de 100 a 199 entradas na academia entre uma avaliação e outra, e quando a frequência cardíaca dele ficou na FVariaCardiaca4, que é uma variação entre -10 e 0 batimentos por minuto, dos 107 registros esperados para a faixa etária MeiaIdade o algoritmo acertou 61, com 57% de acerto. Já na faixa de variação cardíaca FVariaCardiaca5, onde houve uma variação entre 0 e 10 batimentos por minuto, dos 77 registros esperados na faixa etária MeiaIdade, o algoritmo acertou 42, ou seja, houve 54% de acerto.

De uma maneira geral, o teste deste arquivo trouxe 55% de acerto e 45% de erro. Quando a árvore de decisão é testada, o algoritmo gera uma matriz de confusão onde são exibidos acertos e erros do teste. A matriz de confusão deste teste é exibida na tabela 5.1.

Tabela 5.1 – Matriz de Confusão Gerada pelo Teste da Freqüência

	Adolescente	AdultoJovem	MeiaIdade	TerceiraIdade
Adolescente	0	19	0	0
AdultoJovem	0	215	69	0
MeiaIdade	0	175	139	0
TerceiraIdade	0	9	14	0

Fonte: do autor.

Nessa matriz, o valor 215, por exemplo, significa que 215 registros foram classificados corretamente na faixa etária AdultoJovem, já 69 registros que deveriam ser classificados nessa faixa foram classificados de maneira incorreta na faixa etária MeiaIdade. Tudo isso tem o objetivo de mostrar de diversas maneiras como o resultado da árvore de decisão foi gerado.

### 5.3 Relação Variação Peso - Entradas - Faixa Etária

Para este relacionamento, o arquivo possui a faixa de variação do peso do aluno entre as avaliações, a quantidade de entradas na academia entre uma avaliação e outra e a faixa etária do aluno. A metodologia utilizada foi a mesma dos outros testes, ou seja, foi utilizado um arquivo para treinamento, com 70% dos registros, e outro arquivo para testes, com os 30% restantes.

O fator de confiança utilizado para este teste foi de 0.45, pois quando valores menores do que esse eram inseridos, as árvores geradas ficavam muito genéricas, possuindo somente um nodo. A árvore possui 52 folhas, porém, para análise de resultados foi utilizada somente uma parte dela, conforme demonstra a figura 5.3.

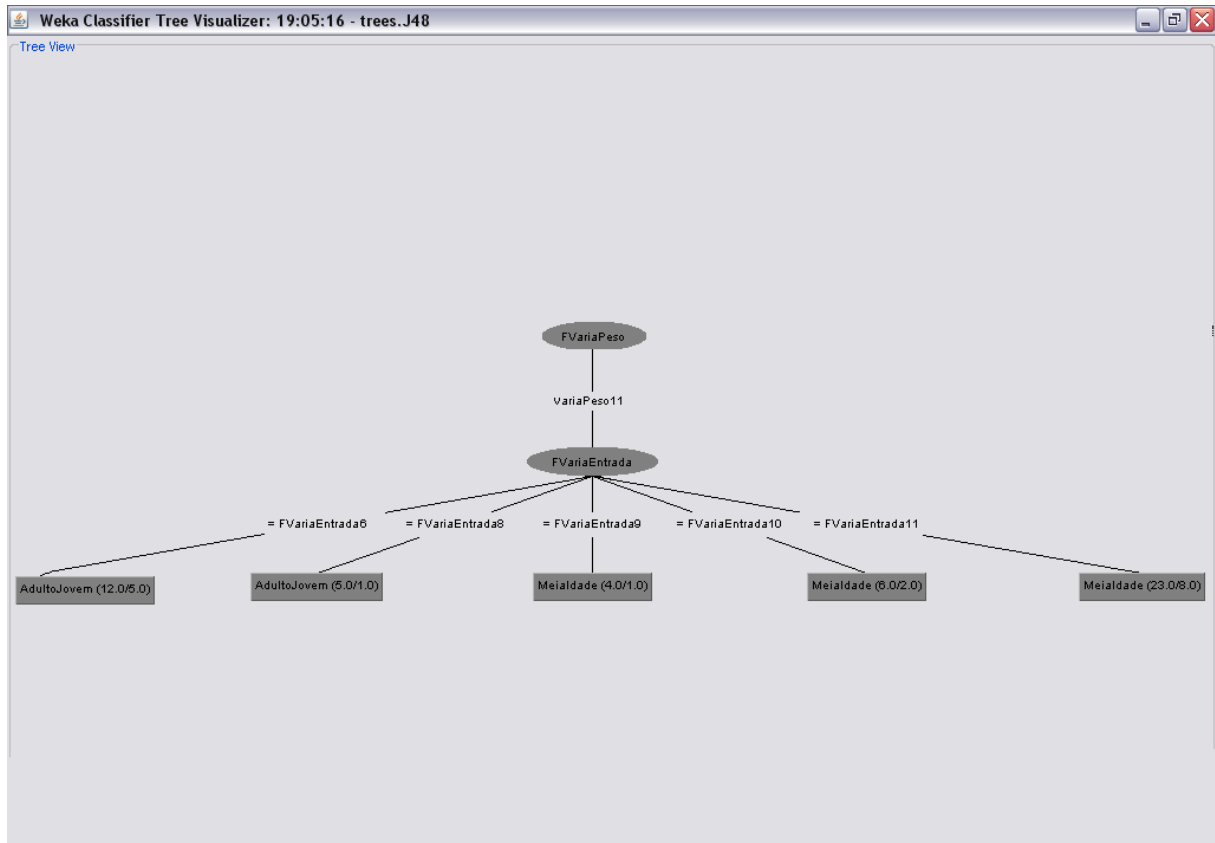


Figura 5.3 – Árvore de Decisão da Variação do Peso  
Fonte: do autor.

Na faixa de variação de peso *VariaPeso11*, onde a variação está entre 0 e 2 quilogramas, e faixa de variação de entrada *FVariaEntrada11*, onde o aluno frequentou a academia entre 100 e 199 vezes entre as avaliações e possui faixa etária *Meialdade*, dos 23 registros esperados o algoritmo acertou 15 o que representa 65% de acerto. Quando a faixa de variação de entrada foi *FVariaEntrada6*, onde o aluno frequentou entre 50 e 59 vezes a academia, e faixa etária *AdultoJovem*, dos 12 registros esperados o algoritmo acertou 7 o que representa 58% de acerto.

Pode-se chegar à conclusão através dessas regras que existe uma grande possibilidade de o aluno que frequenta a academia entre 100 e 199 vezes entre as avaliações e possui de 40 a 65 anos, possua uma variação de 0 a 2 quilos, o que representa que ele pode até mesmo ganhar peso frequentando a academia. A matriz de confusão gerada por esse teste está representada na tabela 5.2

Tabela 5.2 – Matriz de Confusão Gerada pelo Teste do Peso

	Adolescente	AdultoJovem	MeiaIdade	TerceiraIdade
Adolescente	0	6	1	0
AdultoJovem	1	56	43	0
MeiaIdade	0	37	72	1
TerceiraIdade	0	3	6	0

Fonte: do autor.

Nessa matriz, o valor 56, por exemplo, significa que 56 registros foram classificados corretamente na faixa etária AdultoJovem, já 43 registros que deveriam ser classificados nessa faixa foram classificados de maneira incorreta na faixa etária MeiaIdade sendo que neste teste o acerto total foi de 57%.

#### 5.4 Relação Entradas - Variação Pressão Sistólica - Faixa Etária

Neste teste foram utilizados os seguintes atributos: quantidade de entradas do aluno na academia entre as avaliações; variação da pressão sistólica do aluno entre uma avaliação e outra; e sua faixa etária. Foram utilizados dois arquivos, um para treinamento e um para teste da árvore de decisão

O fator de confiança utilizado foi 0.25 e de uma maneira geral o teste apresentou 56% de acerto e 44% de erro. A árvore gerada possui diversas ramificações e para realizar essa análise foi utilizada apenas uma parte dela, conforme demonstra a figura 5.4.

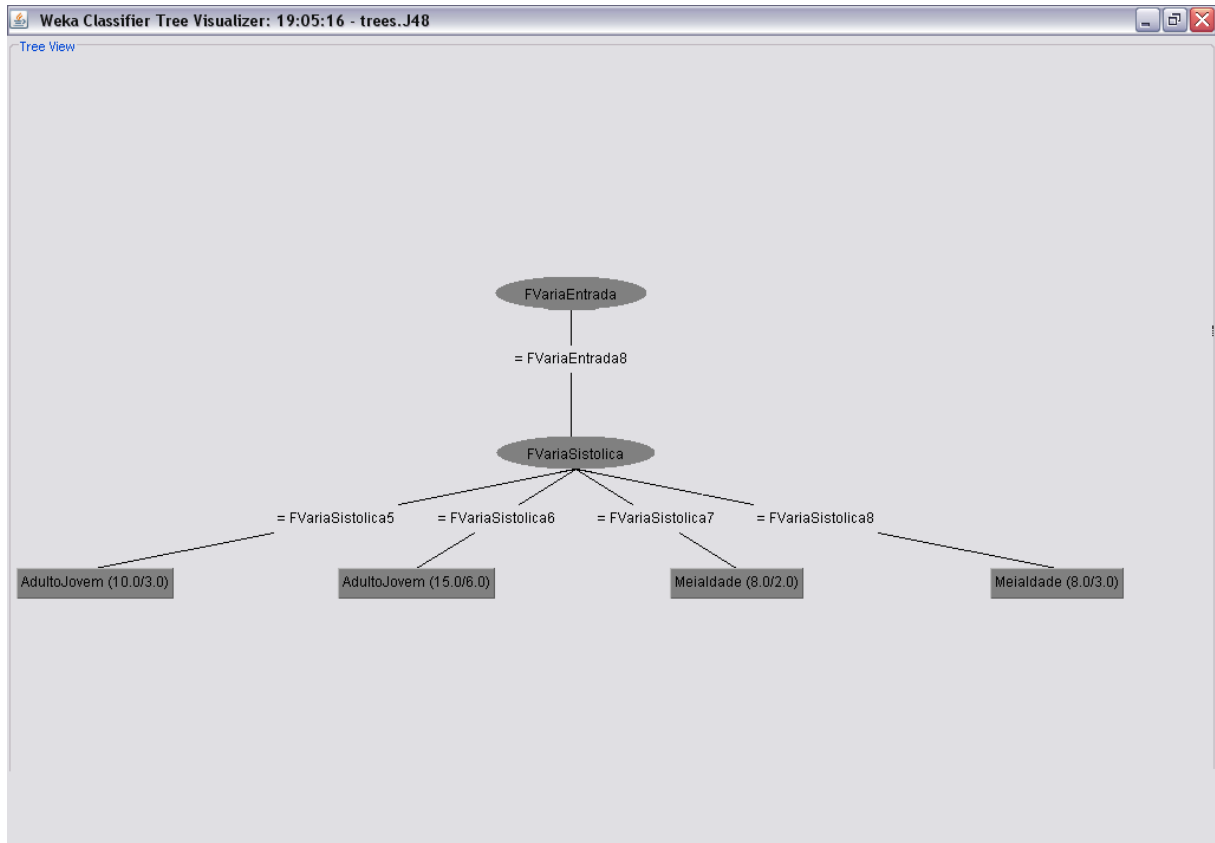


Figura 5.4 – Árvore de Decisão da Variação da Pressão Sistólica

Fonte: do autor.

Através dessa imagem, é possível perceber que quando a faixa de variação de entrada é FVariaEntrada8, onde o aluno frequenta a academia entre 70 e 79 vezes entre as avaliações, e ainda sua variação da pressão sistólica fica na FVariaSistolica5, onde a pressão variou entre -10 e -5, dos 10 registros esperados para a faixa etária AdultoJovem o algoritmo acertou 7, o que representa 70% de acerto. Já na faixa de variação de pressão FVariaSistolica6, onde a pressão variou entre -5 e 0, dos 15 registros esperados na faixa etária AdultoJovem, o algoritmo acertou 9, o que totaliza 60% de acerto.

De uma maneira geral, o teste da árvore de decisão obteve 56% de acertos, um valor considerado bastante razoável. Pode-se perceber que novamente apenas as faixas intermediárias foram apresentadas e analisadas, pois nas faixas com um valor mais alto de variação se encontram a menor parte dos registros.

### 5.5 Relação Entradas - Raça - Variação Cintura

Os testes realizados relacionando a raça do aluno com os demais atributos não resultaram em informações relevantes para a academia. Este teste foi realizado com fator de confiança de 0.25, pois fatores com valores menores que esse gerava árvores de apenas um nodo.

Visto que o atributo raça está gravado na tabela Membros do banco de dados no formato de número e está relacionada à tabela Raça, que possui a descrição delas, a tabela 5.3 demonstra qual raça corresponde a cada número que será demonstrado na figura que será exibida logo depois.

Tabela 5.3 – Códigos e Raças presentes na Base de Dados

Código da Raça	Descrição da Raça
1	Negra
2	Mestiça
3	Indiana
4	Caucasiana
5	Asiática

Fonte: do autor.

Através dessa tabela, pode-se perceber que a raça caucasiana corresponde à raça branca, conforme definido no Brasil. A maior parte dos membros da academia são da raça negra e caucasiana. A figura 5.5 representa apenas uma parte da árvore de decisão gerada pelo algoritmo J48 do Weka.

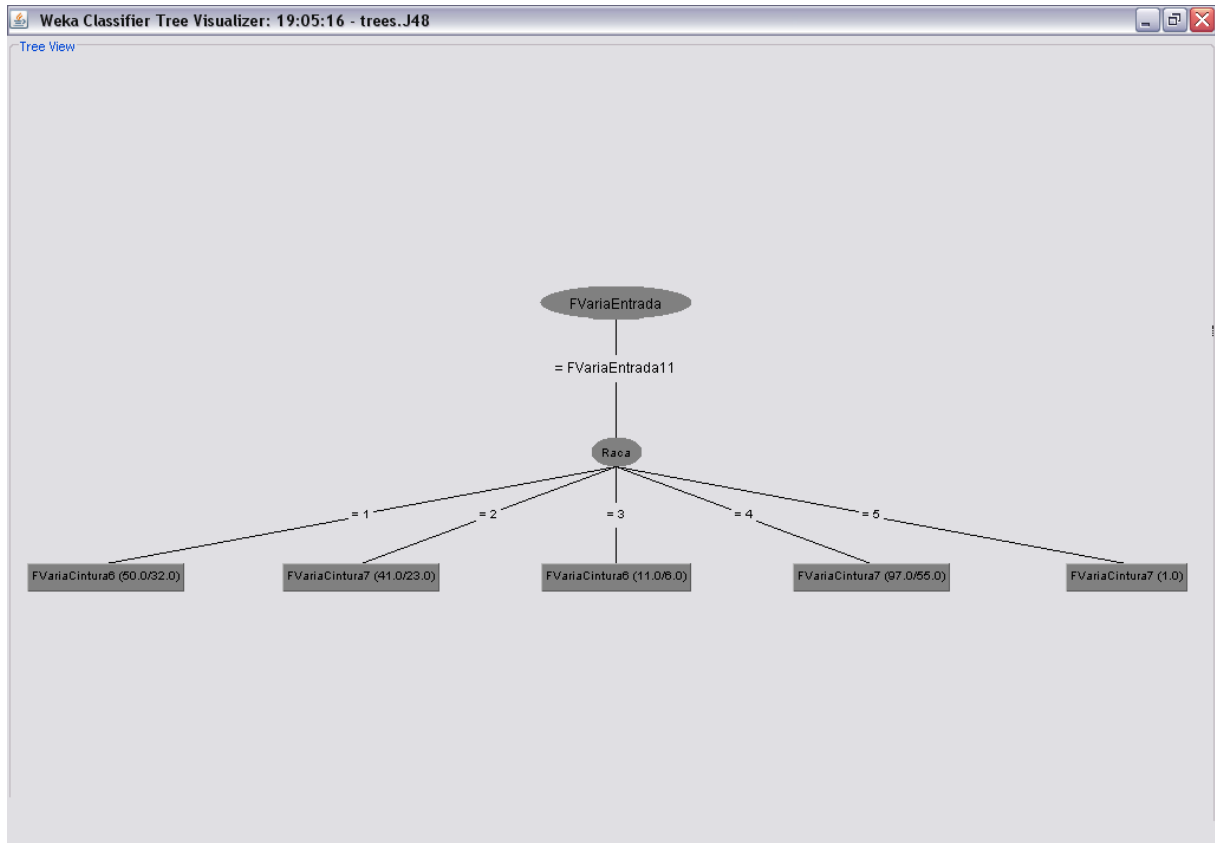


Figura 5.5 – Árvore de Decisão da Raça  
Fonte: do autor.

Quando o aluno frequenta a academia entre 100 e 199 vezes entre as avaliações, é da raça negra e faixa de variação da cintura está na FVariaCintura6, que corresponde a uma variação de -5 a 0 centímetros entre as avaliações, dos 50 registros esperados o teste da árvore acertou 18, o que corresponde a 36% de acerto. Quando a raça é caucasiana e a variação da cintura ficou na faixa FVariaCintura7, que corresponde uma variação entre 0 e 5 centímetros, dos 97 registros esperados o teste acertou 42, o que corresponde a 43% de acerto.

No total a árvore gerada possui 37 folhas, com um acerto total de 31%. Nos demais testes relacionados a raça, o total de acerto ficou nessa média, portanto não serão apresentados neste trabalho.

Ao longo do trabalho, foram criados 56 arquivos no formato arff, com diferentes faixas de variação e que trouxeram os mais diferentes tipos de resultados. De todos os testes realizados, os que trouxeram os melhores resultados foram citados nesse capítulo. Simulações utilizando *Cross-Validation* com *Folds* 10 e utilizando o mesmo arquivo para treinamento e testes foram realizadas e não trouxeram resultados mais significativos do que os apresentados.



## CONCLUSÃO

Pode-se afirmar que muitas vezes os dados presentes nas bases de dados das organizações não são aproveitados da melhor maneira possível. Assim, torna-se fundamental a aplicação de técnicas de mineração de dados nessas bases a fim de transformar informações desconhecidas em conhecimento útil e lucrativo para as empresas. A partir dessas técnicas de mineração de dados é possível ter um conhecimento maior do que realmente existe na base de dados, fazendo com que ações possam ser tomadas por parte dos homens de negócios na tentativa de aumentar o potencial da empresa.

A aplicação de técnicas de mineração de dados na base da academia de musculação proporcionou um estudo focado no desempenho dos alunos de acordo com sua frequência e suas medidas que são retiradas nas avaliações periódicas que a academia realiza. Também foi possível relacionar alguns atributos referentes aos alunos, na tentativa de compreender se realmente a quantidade de vezes que o aluno comparece à academia entre as avaliações influencia na sua forma física.

Buscou-se também obter dados estatísticos sobre os clientes, além de encontrar alguma relação das atividades dos alunos com a sua frequência na academia, evolução das medidas, de acordo com faixa etária, raça, entre outras informações. O uso do algoritmo J48 foi relativamente eficiente quando aplicado nos relacionamentos com frequência cardíaca e peso, por exemplo, pelo fato de a faixa de variação ter sido definida de forma mais acertada. Porém, esse mesmo algoritmo não trouxe resultados relevantes quando aplicado à variação da cintura e relacionamentos com a raça da pessoa, uma vez que os dados se tornaram muito heterogêneos para as faixas definidas.

Todas as informações obtidas com este estudo foram transmitidas para o proprietário da academia que, após analisá-las, irá decidir a questão da sua aplicação ou não. Com essas

informações, podem-se definir quantas vezes que o aluno deve ir à academia por semana para obter um melhor resultado focado na perda de peso por exemplo.

Durante o desenvolvimento do estudo encontrou-se algumas dificuldades, como entrar em contato com o proprietário da academia, já que ele reside em Moçambique e o fuso horário é diferente. Em virtude disso, houve carência de um especialista e de um domínio com alguns dados técnicos, como frequência cardíaca e pressão sanguínea, o que também dificultou realização do estudo, uma vez que esses dados eram altamente importantes no contexto que foi definido para a realização do trabalho.

Outra dificuldade encontrada foi a falta de experiência no que diz respeito à mineração de dados em sua totalidade. Isso fez com que houvesse uma demora excessiva para a preparação dos dados e até mesmo a transformação deles para então aplicá-los no Weka. A análise dos resultados na ferramenta foi aprimorada durante o estudo e no final do trabalho pode-se afirmar que foram dominadas pelo seu autor.

Como trabalhos futuros podem ser citados a aplicação de outras técnicas de mineração de dados, ou até mesmo outros algoritmos na tentativa de se obter melhores resultados; a utilização de outra base da academia que possui as informações referentes aos exercícios que os alunos executam; a utilização da base com os dados financeiros da academia para fazer uma análise das vendas e das aquisições. Pode-se também criar um processo automatizado em que, através de uma interface, o analista informe os dados que quer relacionar e a consulta gerasse automaticamente o arquivo para ser colocado no Weka, Tudo isso poderia auxiliar na tomada de decisões na academia e ser um diferencial em relação às concorrentes.

## REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, Leandro Maciel; PADILHA, Thereza Patrícia P.; OLIVEIRA, Fernando Luiz de; PREVIERO, Conceição Aparecida. **Uma Ferramenta para Extração de Padrões**. Revista Eletrônica de Iniciação Científica, v. 3, 2003. Disponível em: <<http://www.sbc.org.br/reic/edicoes/2003e4/cientificos/UmaFerramentaParaExtracaoDePadroes.pdf>>. Acesso em 20 Ago. 2008.

ARAÚJO, Anderson Viçoso de. **Árvore de Decisão Fuzzy na mineração de imagens do sistema Footscanage**. Curitiba, PR: 2006. Dissertação (Mestrado) – Programa de Pós-Graduação em Informática, Universidade Federal do Paraná, 2006.

BORGES, Helyane Bronoski. **Redução de Dimensionalidade de Atributos em Bases de Dados de Expressão Gênica**. Curitiba, PR: 2006. 123 p. Dissertação (Mestrado) – Programa de Pós Graduação em Informática. Pontifícia Universidade Católica do Paraná, 2006.

BRAGA, Antônio de Pádua; LAUDERMIR, Teresa Bernarda; CARVALHO, André Carlos Ponce de Leon Ferreira. **Redes neurais artificiais: teoria e aplicações**. Livros Técnicos e Científicos Editora S.A., 2000.

CARVALHO, Juliano Varella de. **Reconhecimento de Caracteres Manuscritos Utilizando Regras de Associação**. Campina Grande, PB: 2000. Dissertação (Mestrado) - Centro de Ciências e Tecnologia, Universidade Federal da Paraíba, 2000.

CARVALHO, Luis Alfredo Vidal de. **DataMining : A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração**. Rio de Janeiro: Ciência Moderna, 2005.

CARVALHO, Luis Marcelo Tavares de; FIGUEIREDO, Symone Maria de Mello. **Avaliação da exatidão do mapeamento da cobertura da terra em Capixaba, Acre utilizando classificação por árvore de decisão**. Cerne, Lavras, v. 12, n. 1, p. 38 – 47, 2006.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery: An overview. In: FAYYAD et al. **Advances in Knowledge Discovery and Data Mining**. G. Cambridge-Mass:AAAI/MIT Press, 1996.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data Mining: Um Guia Prático: Conceitos, Técnicas, Ferramentas, Orientações e Aplicações**. Rio de Janeiro, RJ: Elsevier, 2005.

GONCHOROSKI, Sidnei Pereira. **Utilização de Técnicas de KDD em um Call Center Ativo**. Novo Hamburgo, RS: 2007. 119 p. Monografia (Bacharelado em Ciência da Computação) – Instituto de Ciências Exatas e Tecnológicas, Centro Universitário Feevale, 2007.

HAYKIN, Simon S. **Redes neurais: princípios e prática**. 2. ed. Porto Alegre, RS: Bookman, 2001. 900 p.

JERONIMO, Paulo Marcelo. **Estudo sobre: Data Mining : Data Warehouse : Cases - Data Warehouse**. Novo Hamburgo, RS: 2001. 73 p. Monografia (Bacharelado em Ciência da Computação) – Instituto de Ciências Exatas e Tecnológicas, Centro Universitário Feevale, 2001.

JUNIOR, Adelir José Schuler; PEREZ, Anderson Luiz Fernandes. **Análise do perfil do usuário de serviços de telefonia utilizando técnicas de mineração de dados**. Revista Eletrônica de Sistemas de Informação, Florianópolis, p. 1 - 8, 01 jun. 2006.

KRANZ, Paulo Henrique. **Business Intelligence: Estudo Aplicado em Cooperativa Médica**. Novo Hamburgo, RS: 2004. 103 p. Monografia (Bacharelado em Ciência da Computação) – Instituto de Ciências Exatas e Tecnológicas, Centro Universitário Feevale, 2004.

LANDWEHR, Niels; HALL, Mark; FRANK, Eibe. **Logistic model trees**. Proceedings of the 14th European Conference on Machine Learning, p. 241-252, 2003.

MARTINHAGO, Sergio. **Descoberta de Conhecimento sobre o processo seletivo da UFPR**. Curitiba, PR: 2005. 114 p. Dissertação (Mestrado) – Departamento de Matemática, Universidade Federal do Paraná, 2005.

MARTINS, Flávia de Toledo et al. **Uso de árvore de decisão para predição da prevalência de esquistossomose no Estado de Minas Gerais, Brasil**. In: Simpósio Brasileiro de Sensoriamento Remoto, 13., 2007, Florianópolis. Anais... Florianópolis: INPE, 2007. p. 2841-2848.

OLIVEIRA, Ivana Corrêa de. **Aplicação de Data Mining na Busca de um Modelo de Prevenção da Mortalidade Infantil**. Florianópolis, SC: 2001. Dissertação (Mestrado) – Engenharia e Sistemas, Universidade Federal de Santa Catarina, 2001.

PASSINI, Sílvia Regina Reginato; TOLEDO, Carlos Miguel Tobar. **Mineração de Dados para Detecção de Fraudes em Ligações de Água**. XI SEMINCO – SEMINÁRIO DE COMPUTAÇÃO, 2002, Blumenau, SC. Anais do XI Seminco. Blumenau, SC: s.n., 2002. p. 229- 242.

REZENDE, Solange Oliveira. **Mineração de Dados**. In: XXV Congresso da Sociedade Brasileira de Computação, 2005. Anais do XXV Congresso da Sociedade Brasileira de Computação. São Leopoldo: SBC, 2005. p. 397-433.

SANTOS, Daiana Pereira dos. **Mineração em Notas Fiscais de entrada de uma empresa calçadista**. Novo Hamburgo, RS: 2008. 93 p. Monografia (Bacharelado em Ciência da Computação) – Instituto de Ciências Exatas e Tecnológicas, Centro Universitário Feevale, 2008.

SOUSA, Mauro Sérgio Ribeiro de. **Mineração de Dados: Uma Implementação Fortemente Acoplada a um Sistema Gerenciador de Banco de Dados Paralelo**. Rio de Janeiro, RJ: 1998. 75 p. Dissertação (Mestrado) – Programa de Pós Graduação de Engenharia. Universidade Federal do Rio de Janeiro, 1998.

STAHNKE, Fernando Rafael. **Uso de *data mining* no mercado financeiro**. Novo Hamburgo, RS: 2008. 121 p. Monografia (Bacharelado em Ciência da Computação) – Instituto de Ciências Exatas e Tecnológicas, Centro Universitário Feevale, 2008.