UNIVERSIDADE REGIONAL DE BLUMENAU

CENTRO DE CIÊNCIAS EXATAS E NATURAIS

CURSO DE CIÊNCIAS DA COMPUTAÇÃO

(Bacharelado)

SISTEMAS DE INFORMAÇÃO EXECUTIVA BASEADO EM UM *DATA MINING* UTILIZANDO A TÉCNICA DE ÁRVORES DE DECISÃO

TRABALHO DE CONCLUSÃO DE CURSO SUBMETIDO À UNIVERSIDADE REGIONAL DE BLUMENAU PARA A OBTENÇÃO DOS CRÉDITOS NA DISCIPLINA COM NOME EQUIVALENTE NO CURSO DE CIÊNCIAS DA COMPUTAÇÃO — BACHARELADO

GEANDRO LUIS COMPOLT

BLUMENAU, NOVEMBRO/1999

1999/2-18

SISTEMAS DE INFORMAÇÃO EXECUTIVA BASEADO EM UM *DATA MINING* UTILIZANDO A TÉCNICA DE ÁRVORES DE DECISÃO

GEANDRO LUIS COMPOLT

ESTE TRABALHO DE CONCLUSÃO DE CURSO, FOI JULGADO ADEQUADO PARA OBTENÇÃO DOS CRÉDITOS NA DISCIPLINA DE TRABALHO DE CONCLUSÃO DE CURSO OBRIGATÓRIA PARA OBTENÇÃO DO TÍTULO DE:

BACHAREL EM CIÊNCIAS DA COMPUTAÇÃO

_____

Prof. Oscar Dalfovo — Orientador na FURB

_____

Prof. José Roque Voltolini da Silva — Coordenador do TCC

BANCA EXAMINADORA

_____

Prof. Oscar Dalfovo — Orientador na FURB

_____

Prof. Maurício Capobianco Lopes

_____

Prof. Neide de Melo A. Silva

À meus pais, a meus filhos Letícia e Leandro e a todos que contribuíram direta ou indiretamente para a realização deste trabalho.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

KDD - *Knowledge Discovery in Databases*

MBR - *Memory-Based Reasoning*

OLAP - *On Line Analytic Processing*

OLTP - *On Line Transaction Processing*

SAD - Sistema de Apoio à Decisão

SAE - Sistema de Automação de Escritórios

SE - Sistema Especialista

SI - Sistema de Informação

SIE - Sistema de Informações Executivas

SIG - Sistema de Informações Gerenciais

SPT - Sistema de Processamento de Transações

RESUMO

O trabalho tem como objetivo principal gerar um modelo de classificação de dados utilizando técnicas de *Data Mining*, mais especificamente árvores de decisão. Para auxiliar esta tarefa foi implementado um protótipo que permite ao usuário definir um valor prioridade para cada atributo que fará parte do modelo de classificação. Para a elaboração do protótipo, foram analisadas as características de Sistemas de Informação, bem como as técnicas de *Data Mining* e montado uma base de dados fictícia com informações de condições que conduzem a concessão de crédito a fornecedores. Estas informações serão a base que será aplicada à classificação. Como conseqüência do desenvolvimento deste trabalho, verificou-se que a aplicação do *Data Mining* juntamente com as etapas do KDD foi muito eficiente. Foram realizados testes e foi possível desenvolver modelos de classificação onde colocou-se em prática o uso de árvores de decisão.

ABSTRACT

This project has as main goal to create a data classification patern using *Data Mining* technics, more precisely decision trees. To help this task it was geared up a prototype that allows to the user definy a priority value for each attribute that will make part of the classification patern. For the prototype building were analysed the Information Systems characteristics, the *Data Mining* technics as well and built an imaginary database with information and conditions that turn possible the suppliers credit concession - those information will be the base that will be applied to the classification. As a result of this project development, we verifyed that the *Data Mining* use together with other KDD stages was very efficient. Tests were simulated and it was possible to develop classification paterns where it was put in practice the use of decision trees.

1 INTRODUÇÃO

Devido a competição exaltada e a necessidade de cultivar lucros, as empresas estão transformando algumas das tecnologias de informação em ferramentas para obterem sucesso no gerenciamento dos seus negócios, utilizando os dados armazenados em banco de dados durante o decorrer do tempo a uma tomada de decisão. Toda esta informação pode ser usada para melhorar seus procedimentos, permitindo que a empresa detecte tendências e características disfarçadas, e reaja rapidamente a um evento que ainda pode estar por vir. Alguns exemplos disto são o crescimento dos mecanismos de leitura de preço nos supermercados, dos caixas eletrônicos, dos cartões de crédito, da televisão por assinatura, do *home shopping*, da transferência eletrônica de fundos.

Apesar da grande importância desses dados, a maioria das empresas são incapazes de aproveitar total e eficazmente o que está armazenado em seus arquivos. Esta informação valiosa está escondida sob uma montanha de dados, e não pode ser descoberta utilizando-se dos métodos convencionais; elas precisam de um significado. O *Data Mining* veio para apresentar um significado a esses dados.

O significado permite a análise dos dados observando modelos, estabelecendo mecanismos e tendo novas idéias para fazer previsões sobre o futuro. Conforme [HAR98], o *Data Mining*, do modo como é usado o termo, é a exploração e análise, por meios automáticos ou semi-automáticos, de grandes quantidades de dados para descobrir modelos e regras significativas.

A tecnologia utilizada no *Data Mining* utiliza da procura em grandes quantidade de dados armazenados procurando extrair padrões e relacionamentos que podem ser fundamentais para os negócios da empresa. O *Data Mining* trabalha com um conjunto de técnicas avançadas e princípios de inteligência artificial para identificar os padrões e associações que os dados refletem, com isso oferecendo conclusões que podem trazer valiosas vantagens a nível de mercado para as empresas.

Reconhecendo o *Data Mining* como uma forma de incorporar significado aos dados, propõe-se especificar e desenvolver um Sistema de Informação para efetuar classificação e segmentação utilizando as técnicas de *Data Mining*.

1.1 OBJETIVOS

O objetivo principal deste trabalho é elaborar um modelo de classificação e segmentação de dados afim de auxiliar o executivo na tomada de decisões em uma empresa, através de um Protótipo de Sistema de Informação baseado em Árvore de Decisão utilizando técnicas de *Data Mining*, mais especificamente para efetuar classificações e segmentações de dados.

Os objetivos específicos são:

a) estudar as tarefas e técnicas que o *Data Mining* incorpora;
b) demonstrar o potencial do *Data Mining* para classificação e segmentação de dados;
c) desenvolver um protótipo que demonstre a construção de modelos de classificação.

1.2 ORGANIZAÇÃO DO TEXTO

O primeiro capítulo define os objetivos do trabalho, apresentando a justificativa para seu desenvolvimento.

O segundo capítulo apresenta uma visão geral sobre os SI, do qual o trabalho propõe-se a utilizar, mostrando conceitos, tipos, problemas e utilidades dos mesmos.

O terceiro capítulo enfatiza os conceitos, técnicas e aplicações do *Data Mining*.

O quarto capítulo apresenta a análise, as características, o desenvolvimento e a utilização do modelo criado.

O quinto capítulo completa o trabalho, apresentando as conclusões, limitações e sugestões para serem implementadas e aprimoradas.

2 SISTEMAS DE INFORMAÇÃO

2.1 CONCEITOS

De acordo com [ALT92] atualmente todas as organizações possuem um sistema de informação com o propósito de a auxiliar no cumprimento da sua missão (razão pela qual a empresa destina-se). Esse sistema é normalmente composto de diversos sub-sistemas de natureza conceitual idêntica à daquele que integram, mas com características específicas quanto à sua finalidade e justificação, quanto ao tipo de tecnologias utilizadas e quanto ao nível dos processos ou natureza das pessoas que envolvem.

A designação Sistema de Informação (SI) é indistintamente utilizada para referir cada um dos diferentes sub-sistemas de informação. Estes sub-sistemas de informação envolvem inevitavelmente a utilização de computadores e correspondem à sua definição, também correntemente designados por "Sistemas de Informação Baseados em Computador", ou simplesmente aplicações.

De acordo com [OLI98] toda empresa tem informações que proporcionam a sustentação para as suas decisões. Entretanto, apenas algumas têm um sistema estruturado de informações gerenciais que possibilita otimizar o seu processo decisório. E as que estão neste estágio evolutivo seguramente possuem vantagem empresarial interessante.

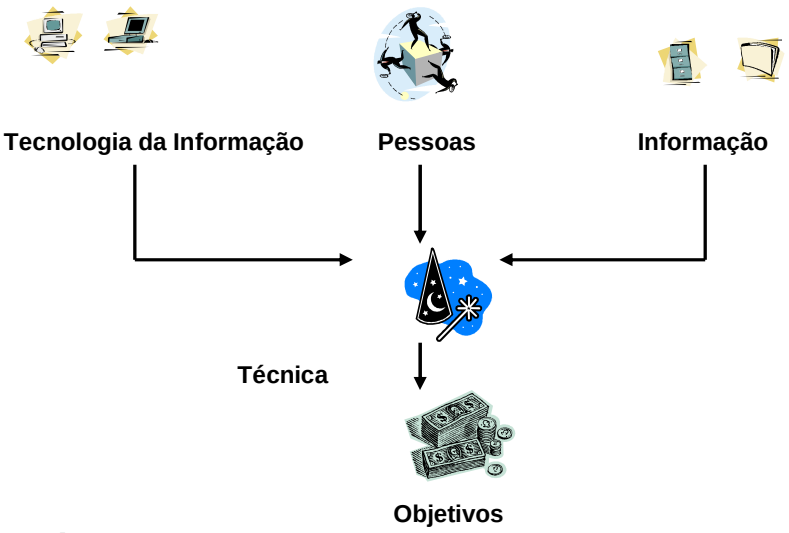
Um Sistema de Informação é um tipo especializado de sistema e pode ser definido de inúmeros modos. Um modo é dizer que sistemas de informação são conjuntos de elementos ou componentes inter-relacionados que coletam (entrada), manipulam e armazenam (processo), disseminam (saída) os dados e informações e fornecem um mecanismo de *feedback*. A entrada é a atividade de captar e reunir novos dados, o processamento envolve a conversão ou transformação dos dados em saídas úteis, e a saída envolve a produção de informação útil. O *feedback* é a saída que é usada para fazer ajustes ou modificações nas atividades de entrada ou processamento.

A informação tem papel importante nos Sistemas de Informação, pois é através das informações que dependerá o futuro da empresa. De nada adianta uma sobrecarga das

informações ou um sistema de banco de dados abarrotados de informações, pois esse acúmulo poderá levar a empresa à desinformação. Um Sistema de Informação deve apresentar informações claras, sem interferência de dados que não são importantes, e deve possuir um alto grau de precisão e rapidez para não perder sua razão de ser em momentos críticos. Além disso, a informação deve sempre chegar a quem tem necessidade dela. Sistemas de Informação tornaram-se hoje um elemento indispensável para dar apoio às operações e à tomada de decisões na empresa moderna.

De acordo com [PRA94], Sistemas de Informação são formados pela combinação estruturada de vários elementos, organizados da melhor maneira possível, visando atingir os objetivos da organização. São integrantes dos Sistemas de Informação: a informação (dados formatados, textos livres, imagens e sons), os recursos humanos (pessoas que coletam, armazenam, recuperam, processam, disseminam e utilizam as informações), as tecnologias de informação (o hardware e o software usados no suporte aos Sistemas de Informação) e as práticas de trabalho (métodos utilizados pelas pessoas no desempenho de suas atividades). Pode-se observar estes elementos na figura 1.

Figura 1 - Elementos de um Sistema de Informação



Fonte: [ALT92]

2.2 TIPOS DE SISTEMAS DE INFORMAÇÃO

De acordo com [ALT92] a utilização de diferentes critérios e das suas combinações, na classificação dos diversos tipos de SI, torna possível encontrar inúmeras propostas, de diferentes autores, sobre as características fundamentais de cada um desses tipos. São, contudo, mais frequentes e aceitas as classificações que utilizam como critérios:

a) o que os sistemas fazem (funções) e os componentes que o integram (atributos);
b) os níveis de gestão que prioritariamente servem;
c) a era que pertencem.

Os principais tipos de Sistemas de Informação, segundo [ALT92], estão identificados na tabela 1.

Tabela 1 - Tipos de Sistemas de Informação

Tipo de sistema	Definição
Sistema Processamento de Transações	Recolhe e mantém informação sobre transações e controla pequenas decisões que fazem parte das transações
Sistema de Informação de Gestão	Converte informação sobre transações em informação para a gestão da organização
Sistema de Apoio à Decisão	Ajuda os utilizadores na tomada de decisões não estruturáveis fornecendo-lhes informação, modelos e ferramentas para analisar a informação
Sistema de Informação para Executivos	Fornece aos gestores, de modo muito interativo e flexível, acesso a informação geral para a gestão da organização
Sistema Pericial	Suporta os profissionais do desenho, diagnóstico e avaliação de situações complexas que requerem conhecimento especializado em áreas bem definidas
Sistema de Automação de Escritório	Mantém as tarefas de comunicação e processamento de informação características de ambiente de escritório

Fonte: adaptado de [ALT92]

Para [EAR88] os tipos de SI são mais voltados para os níveis de gestão a que pertencem, de acordo com demonstrativo na tabela 2.

Tabela 2 – Tipos de Sistemas de Informação (níveis de gestão)

Nível de gestão	Tipo de sistema
Planejamento Estratégico	Sistema de Informação para Executivos
Controle de Gestão	Sistema de Apoio à Decisão
Controle Operacional	Sistema de Processamento de Transações

Fonte: adaptado de [EAR88]

A classificação dos diferentes tipos de SI pela identificação da era a que pertencem é uma das formas mais práticas (e mais útil em muitas situações), de o fazer. As eras são definidas de forma diferente por diferentes autores, mas todas elas têm uma evolução temporal de alguma característica fundamental da composição, justificação ou utilização dos diversos SI conforme demonstrativo na tabela 3.

Tabela 3 – Tipos de Sistemas de Informação (eras)

Foco de gestão	Era	Objetivo
Tecnologias da Informação	Sistema de Processamento de Dados	Automatização eficiente de processos básicos
Informação	Sistema de Informação de Gestão	Satisfação eficaz das necessidades de informação
---	Sistema de Informação Estratégica	Potencializar a competitividade da organização

Fonte: adaptado de [EAR88]

De todas estas classificações resulta inevitavelmente alguma confusão, quer ao nível das designações, quer ao nível dos próprios conceitos. Dos critérios utilizados, ou utilizáveis, para a classificação de SI, os níveis de gestão propostos por [ALT92] são sem dúvida o referencial de grande importância pelo efeito estruturador, isto pela vasta divulgação e aceitação de que é alvo.

Os principais tipos de Sistemas de Informação, segundo [ALT92], são os seguintes:

a) Sistema de Processamento de Transações (SPT): coletam e armazenam dados sobre transações e às vezes controlam decisões que são executadas como parte de uma transação. Uma transação é um evento empresarial que pode gerar ou modificar dados armazenados num Sistema de Informação. Ele foi o primeiro Sistema de Informação que surgiu e é freqüentemente encontrado. Por exemplo, quando paga-se uma conta com o Cartão de Crédito é o SPT que efetua a transação com a Central e valida o cartão. Enfim, ele grava as informações e assegura que as mesmas estão consistentes e disponíveis;
b) Sistema de Automação de Escritório (SAE): ajuda as pessoas a processar documentos e fornece ferramentas que tornam o trabalho no escritório mais eficiente e eficaz. Também pode definir a forma e o método para executar as tarefas diárias e dificilmente afeta as informações em si. Exemplos deste tipo de sistema são editores

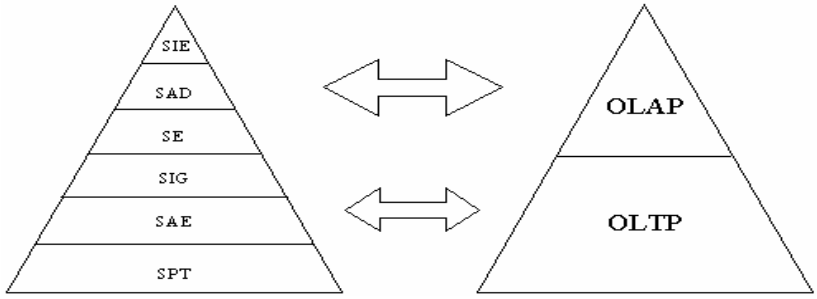
de texto, planilhas de cálculo, softwares para correio eletrônico e outros. Todas as pessoas que tem em sua função tarefas como redigir textos, enviar mensagens, criar apresentações são usuárias de Sistemas de Automação de Escritórios.

c) Sistema de Informação Gerencial (SIG): converte os dados de uma transação do SPT em informação para gerenciar a organização e monitorar o desempenho da mesma. Ele enfatiza a monitoração do desempenho da empresa para efetuar as devidas comparações com as suas metas. As pessoas que o utilizam são os gerentes e as que precisam monitorar seu próprio trabalho. Um exemplo disto são os relatórios que são tirados diariamente para acompanhar o Faturamento da empresa;
d) Sistemas Especialistas (SE): torna o conhecimento de especialistas disponível para outros, e ajuda a resolver problemas de áreas onde o conhecimento de especialistas é necessário. Ele pode guiar o processo de decisão e assegurar que os fatores chave serão considerados, e também pode ajudar uma empresa a tomar decisões consistentes. As pessoas que usam estes sistemas são aquelas que efetuam tarefas onde deveria existir um especialista. Um sistema especialista pode ser, por exemplo, um sistema onde médicos dizem os sintomas e é pesquisado em uma base de conhecimento os possíveis diagnósticos;
e) Sistema de Apoio à Decisão (SAD): ajuda as pessoas a tomar decisões, provendo informações, padrões, ou ferramentas para análise de informações. Ele pode prover métodos e formatos para porções de um processo de decisão. Os maiores usuários são os analistas, gerentes e outros profissionais. Os sistemas que disponibilizam gráficos 3D para comparativos são exemplos;
f) Sistema de Informações Executivas (SIE): fornece informações aos executivos de uma forma rápida e acessível, sem forçar os mesmos a pedir ajuda a especialistas em Análises de Informações. É utilizado para estruturar o planejamento da organização e o controle de processos, e pode eventualmente também ser utilizado para monitorar o desempenho da empresa. Um exemplo são os sistemas que fornecem comparativos simples e fáceis de Vendas x Estoque x Produção.

A evolução dos SI nos últimos anos transformou a forma de apresentação dos mesmos, antes existia uma pirâmide dividida em seis partes, na primeira camada os SPT, seguido do SAE, SIG, SE, SAD e o SIE.

Atualmente, segundo [MAC96] estas seis partes se transformaram em apenas duas, onde as linhas que separavam o segundo nível do sexto nível não fazem mais sentido. Estas duas camadas são a *On Line Transaction Processing* (OLTP) que fica na base da Pirâmide e a *On Line Analytic Processing* (*OLAP*) que fica no topo (figura 2).

Figura 2 – Evolução dos Sistemas de Informação



Fonte: [MAC96]

Conforme [MAC96], o motivo pelo qual houve a fusão entre estes grupos de sistemas reside nas mudanças por que passaram as organizações nos últimos anos. O SIE, por exemplo, voltava-se para a alta direção e tinha um aspecto mais informativo ao mesmo tempo que o SAD voltava-se para a gerência que tomava as decisões.

Atualmente, as modificações na forma de gestão das empresas levaram as pessoas do topo a tomar mais decisões. Do mesmo modo, os gerentes que antes tomavam a maior parte das decisões tiveram seu número reduzido, consequentemente reduzindo a hierarquia e os funcionários que antes só obedeciam ordens agora podem dar sugestões para a mudança de processos.

Outro aspecto que ajudou na mudança dos Sistemas de Informação diz respeito a própria evolução tecnológica da informática. Muitas das tarefas que antes eram executadas em mainframes agora são executadas através de redes de micros, operando de forma Cliente/Servidor. Esta estrutura facilitou a montagem de sistemas compartilhados voltados para um maior número de gerentes [MAC96].

Os sistemas baseados em OLTP são configurados e otimizados para prover respostas rápidas à transações individuais. Nestes sistemas, as transações devem ser realizadas rapidamente, e com grande confiança. Os dados são dinâmicos, mudando com grande freqüência. Já nos sistemas baseados em OLAP, a velocidade das transações não é relevante, pois o banco de dados pode armazenar os dados em forma estática, e são configurados e otimizados para suportar complexas decisões baseadas em dados históricos [OLI98].

De acordo com [DAL98], os SI estão sendo utilizados nas estruturas de decisões da empresa e se corretamente aplicado o seu desenvolvimento, trará, às mesmas, uma melhor utilização das suas informações. Desta forma, trarão certamente resultados positivos às empresas, caso contrário, tornam-se difíceis de serem implementados pelas mesmas, até mesmo pelo seu alto custo. Porém, é necessário antes de tudo saber ao certo onde se quer chegar e o que necessita-se obter dos SI, para que possam ser bem elaborados e desenvolvidos, tornando-se fundamentais e capacitados para a tomada de decisões da empresa.

3 *DATA MINING*

As expressões *Data Mining*, mineração de dados ou garimpagem de dados referem-se ao processo de extrair informações potencialmente úteis a partir de dados brutos que estão armazenados em um Data Warehouse ou nos bancos de dados dos diversos sistemas implantados nas empresas. A tecnologia utilizada no *Data Mining* utiliza da procura em grandes quantidade de dados armazenados procurando extrair padrões e relacionamentos que podem ser fundamentais para os negócios da empresa. O *Data Mining* trabalha com um conjunto de técnicas avançadas e princípios de inteligência artificial para identificar os padrões e associações que os dados refletem, com isso oferecendo conclusões que podem trazer valiosas vantagens a nível de mercado para a empresa.

O processo de descobrimento realizado pelo *Data Mining* pode ser utilizado a partir dos sistemas transacionais, porém, é muito mais eficiente utilizá-lo a partir de um Data Warehouse onde os dados já estão mais consistentes e íntegros, e habilitam descobertas mais abrangentes e precisas. O *Data Mining* oferece funções muito sofisticadas, porém as tecnologias estão totalmente embutidas no software, deixando os usuários totalmente isentos de conhecer técnicas estatísticas ou de inteligência artificial e permitindo ainda a exportação de dados para planilhas eletrônicas ou processadores de textos ou ainda outras ferramentas que servem de apoio á decisão.

Recentemente as organizações vem aumentando sua capacidade de gerar e armazenar informações, com o aumento do uso de banco de dados utilizado pelas mais diversas áreas, tais como: comercial, administrativa, científica, governamental e outras. Surgiu a necessidade de novas técnicas e ferramentas que possam de forma automática e inteligente gerar informações “escondidas” nas bases de dados.

De acordo com [BER97] o objetivo do *Data Mining* é descobrir o conhecimento, extraí-lo implicitamente sem que seja necessário conhecer a estrutura das informações do banco de dados sobre ele aplicado. Este processo é denominado de *Knowledge Discovery in Databases (KDD)* que será detalhado no item seguinte.

3.1 PROSPECÇÃO DE CONHECIMENTO

Prospecção de conhecimento em bases de dados (*Knowledge Discovery in Databases - KDD*) é um processo que envolve a automação da identificação e do reconhecimento de padrões em um banco de dados. Trata-se de uma pesquisa de fronteira, que começou a se expandir mais rapidamente nos últimos cinco anos. Sua principal característica é a extração não-trivial de informações a partir de uma base de dados de grande porte. Essas informações são necessariamente implícitas, previamente desconhecidas, e potencialmente úteis [FIG98].

Devido a essas características incomuns, todo o processo de KDD depende de uma nova geração de ferramentas e técnicas de análise de dados, e envolve diversas etapas. A principal, que forma o núcleo do processo, e que muitas vezes se confunde com ele, chama-se *Data Mining*, ou Mineração de Dados, também conhecido como processamento de padrões de dados, arqueologia de dados, ou colheita de informação (*information harvesting*).

O KDD compreende todo o processo de descoberta de dados, enquanto o *Data Mining* refere-se a aplicação de algoritmos para extração de padrões de dados, sem os passos adicionais do KDD e da análise dos resultados [AVI98].

3.2 AS ETAPAS DO PROCESSO DE KDD

O processo de KDD (figura 3) começa com o entendimento do domínio da aplicação e a relevância do conhecimento em relação às metas a serem atingidas. Em seguida, é feita a seleção dos conjuntos de dados a serem utilizados durante o processo do KDD, isto é, um agrupamento organizado de dados, que será o alvo da prospecção. A etapa da limpeza dos dados (*data cleaning*) vem a seguir, através de um pré-processamento dos dados, visando adequá-los aos algoritmos. Isso se faz através da integração de dados heterogêneos, eliminação de incompletude dos dados, repetição de tuplas, problemas de tipagem, etc. Essa etapa pode tomar até 80% do tempo necessário para todo o processo, devido às bem conhecidas dificuldades de integração de bases de dados heterogêneas [FAY96].

Figura 3 - As etapas do processo de KDD

Fonte: [FIG98]

Os dados pré-processados devem ainda passar por uma transformação que os armazena adequadamente, visando facilitar o uso das técnicas de *Data Mining*.

Prosseguindo no processo, chega-se à fase de *Data Mining* especificamente, que começa com a escolha dos algoritmos a serem aplicados. Essa escolha depende fundamentalmente do objetivo do processo de KDD: classificação, segmentação, agrupamento por afinidades, estimativas, etc. De modo geral, na fase de *Data Mining*, ferramentas especializadas procuram padrões nos dados. Essa busca pode ser efetuada automaticamente pelo sistema ou interativamente com um analista, responsável pela geração de hipóteses. Diversas ferramentas distintas, como redes neurais, indução de árvores de decisão, sistemas baseados em regras e programas estatísticos, tanto isoladamente quanto em combinação, podem ser então aplicadas ao problema. Em geral, o processo de busca é iterativo, de forma que os analistas revêem o resultado, formam um novo conjunto de questões para refinar a busca em um dado aspecto das descobertas, e realimentam o sistema com novos parâmetros. Ao final do processo, o sistema de *Data Mining* gera um relatório das descobertas, que passa então a ser interpretado pelos analistas de mineração. Somente após a interpretação das informações obtidas encontra-se o conhecimento.

Uma diferença significativa entre *Data Mining* e outras ferramentas de análise está na maneira como exploram as inter-relações entre os dados. As diversas ferramentas de análise disponíveis dispõem de um método baseado na verificação, isto é, o usuário constrói hipóteses sobre inter-relações específicas e então verifica ou refuta, através do sistema. Esse modelo torna-se dependente da intuição e habilidade do analista em propor hipóteses interessantes, em

manipular a complexidade do espaço de atributos, e em refinar a análise baseado nos resultados de consultas ao banco de dados potencialmente complexas. Já o processo de *Data Mining* fica responsável pela geração de hipóteses, garantindo mais rapidez, acurácia e completude aos resultados.

3.3 REQUISITOS DE UM *DATA MINING*

a) conhecimento de diferentes tipos de dados: Os bancos de dados possuem vários tipos de dados complexos, tais como: hipertextos, sons, imagens além dos tipos de dados tradicionais. Todavia o tratamento desses diversos tipo de dados, em relação às metas que se deseja atingir com o *Data Mining* podem trazer resultados irreais ou insatisfatórios. Por isso um *Data Mining* é construído para abranger tipo de dados específicos, procurando se dedicar a banco de dados relacionais, transacionais ou banco de dados de multimídia.

b) algoritmos eficientes e concisos é outro requisito básico para o *Data Mining* poder trazer resultados confiáveis e satisfatórios;

c) utilização, certificação e expressividade dos resultados obtidos;

d) forma de apresentação dos resultados obtidos pelo *Data Mining*: Diferentes tipos de conhecimento podem ser descobertos em um conjunto de dados, todavia, examinando de formas diferentes, pode-se visualizar os resultados sob diversos ângulos, dando ao usuário uma representação mais abrangente do problema.

e) interatividade durante o processo de mineração: Possibilitar ao usuário interativamente, definir dinamicamente a alteração no foco da pesquisa, restringir os resultados obtidos, navegar a outros níveis desejados, flexibilizar a visualização dos resultados em diversos níveis de abstração e de diferentes ângulos.

f) mineração de diferentes fontes de dados: Hoje com o crescimento das redes locais, metropolitanas e da própria internet, existe a possibilidade de mais facilmente ser acessadas fontes de dados remotas e distribuídas. Todavia para um *Data Mining* funcionar sob tais aspectos existe a necessidade de algoritmos de distribuição paralela e distribuída que é encontrada somente em alguns banco de dados.

g) proteção e segurança dos dados: Quando os dados de uma organização podem ser vistos sob diferentes ângulos e diferentes níveis de abstração é necessário que os dados sejam protegidos, para que pessoas não autorizadas venham a utilizá-los de forma inadequada, por isso o *Data Mining* deve ser desenvolvido de forma a prever tais invasões.

3.4 FUNÇÕES DO *DATA MINING*

O *Data Mining* pode desempenhar uma série limitada de tarefas dependendo das circunstâncias. Cada classe de aplicação em *Data Mining* tem como base um conjunto de algoritmos que serão usados na extração de relações relevantes dentro de uma massa de dados [HAR98]:

a) classificação;
b) estimativa;
c) agrupamento por afinidade;
d) previsão;
e) segmentação.

Cada uma destas propostas difere quanto à classe de problemas que o algoritmo será capaz de resolver.

3.4.1 CLASSIFICAÇÃO

Classificação é uma técnica que consiste no mapeamento ou pré-classificação de um conjunto pré-definido de classes. Em geral, algoritmos de classificação incluem árvores de decisão ou redes neurais.

Os algoritmos classificadores utilizam-se de exemplos para determinar um conjunto de parâmetros, codificados em um modelo, que será mais tarde utilizado para a discriminação do restante dos dados.

Uma vez que o algoritmo classificador foi desenvolvido de forma eficiente, ele será usado de forma preditiva para classificar novos registros naquelas mesmas classes pré-definidas.

3.4.2 ESTIMATIVA

Uma variação do problema de classificação envolve a geração de valores ao longo das dimensões dos dados: são os chamados algoritmos de estimativa. A estimativa lida com resultados contínuos, ao contrário da classificação que lida com resultados discretos. Fornecidos alguns dados, usa-se a estimativa para estipular um valor para alguma variável contínua desconhecida como receita, altura ou saldo de cartão de crédito.

Ao invés de um classificador binário determinar um risco “positivo” ou “negativo”, a técnica gera valores de “escore”, dentro de uma determinada margem. A abordagem de estimativa tem a grande vantagem de que os registros individuais podem ser agora ordenados por classificação, e as redes neurais são adequadas a esta tarefa.

Exemplos de estimativa incluem:

a) estimar o número de filhos numa família;
b) estimar a renda total de uma família;
c) estimar o valor em tempo de vida de um cliente.

3.4.3 AGRUPAMENTO POR AFINIDADE

Esta técnica identifica afinidades entre ítens de um subconjunto de dados. Essas afinidades são expressas na forma de regras: “72% de todos os registros que contém os itens A, B, e C também contém D e E”. A porcentagem de ocorrência (72 no caso) representa o fator de confiança da regra, e costuma ser usado para eliminar tendências fracas, mantendo apenas as regras mais fortes. Dependências funcionais podem ser vistas como regras de associação com fator de confiança igual a 100%.

Trata-se de um algoritmo tipicamente endereçado à análise de mercado, onde o objetivo é encontrar tendências dentro de um grande número de registros de compras, por exemplo, expressas como transações. Essas tendências podem ajudar a entender e explorar padrões de compra naturais, e pode ser usada para ajustar mostruários, modificar prateleiras ou propagandas, e introduzir atividades promocionais específicas. Um exemplo mais distinto, onde essa mesma

técnica pode ser utilizada, é o caso de um banco de dados escolar, relacionando alunos e disciplinas. Uma regra do tipo “84% dos alunos inscritos em ‘Introdução ao Unix’ também estão inscritos em ‘Programação em C’” pode ser usada pela direção ou secretaria para planejar o currículo anual, ou alocar recursos como salas de aula e professores.

3.4.4 PREVISÃO

A previsão é o mesmo que classificação ou estimativa, exceto pelo fato de que os registros são classificados de acordo com alguma atitude futura prevista. Em um trabalho de previsão, o único modo de confirmar a precisão da classificação é esperar para ver.

Essa tarefa é uma variante do problema de agrupamento por afinidades, onde as regras encontradas entre as relações podem ser usadas para identificar seqüências interessantes, que serão utilizadas para predizer acontecimentos subsequentes. Nesse caso, não apenas a coexistência de ítens dentro de cada transação é importante, mas também a ordem em que aparecem, e o intervalo entre elas. Seqüências podem ser úteis para identificar padrões temporais, por exemplo entre compras em uma loja, ou utilização de cartões de crédito, ou ainda tratamentos médicos.

Exemplos de tarefas de previsão:

a) previsão de quais clientes sairão nos próximos seis meses;
b) previsão da quantia de dinheiro que um cliente utilizará caso seja oferecido a ele um certo limite de cartão de crédito.

3.4.5 SEGMENTAÇÃO

A segmentação é um processo de agrupamento de uma população heterogênea em vários subgrupos ou *clusters* mais homogêneos. O que a distingue da classificação é que segmentação não depende de classes pré-determinadas. Essa segmentação é realizada automaticamente por algoritmos que identificam características em comum e particionam o espaço n-dimensional definido pelos atributos.

Os registros são agrupados de acordo com a semelhança e depende do usuário determinar qual o significado de cada segmento, caso exista algum. Muitas vezes a segmentação é uma das primeiras etapas dentro de um processo de *Data Mining*, já que identifica grupos de registros correlatos, que serão usados como ponto de partida para futuras explorações. O exemplo clássico é o de segmentação demográfica, que serve de início para uma determinação das características de um grupo social, visando desde hábitos de compras até utilização de meios de transporte.

3.5 TÉCNICAS DE *DATA MINING*

Muitas das técnicas usadas em ferramentas atuais de *Data Mining* se originaram na pesquisa em inteligência artificial da década de 80 e princípio da década de 90. Entretanto, somente agora essas técnicas passaram a ser utilizadas em sistemas de banco de dados de grande escala, devido a confluência de diversos fatores que aumentaram o valor líquido da informação, dentre os quais se destacam as relacionados nos itens a seguir.

3.5.1 REGRESSÃO LINEAR

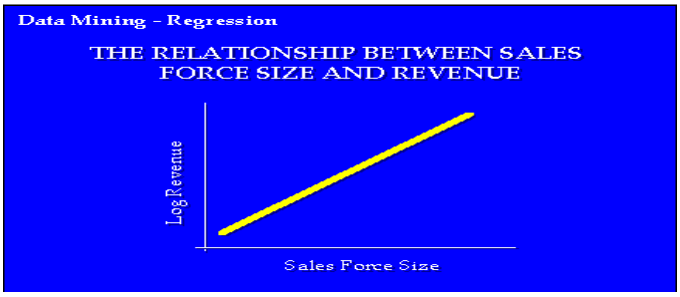
Regressão linear é um método que calcula o valor de uma variável através do valor de outra. Essa técnica é construída sobre um modelo em linha com a seguinte forma:

$$aX + bY + c = 0$$

Onde a, b, c são os parâmetros e X e Y são as variáveis. Para um dado valor de X, estima-se o valor de Y. Este tipo de modelo é um dos mais simples existentes.

Um exemplo de regressão linear é o gráfico representado pela figura 4, onde a linha tem inclinação para cima, isto significa que a variável independente que seria o valor das vendas tem um efeito positivo na variável dependente que seria a renda. Se a linha está se inclinando para baixo há um efeito negativo. Quanto mais acentuada a linha, maior é o efeito da variável independente sobre a variável dependente [GRO97].

Figura 4 – Regressão Linear



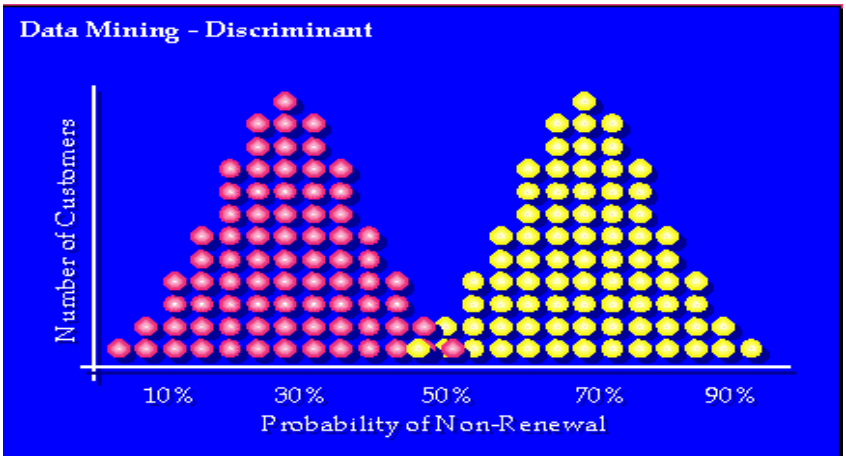
Fonte: [GRO97]

3.5.2 ANÁLISE DISCRIMINATÓRIA

Análise discriminatória é um método de classificação que mede a importância dos fatores que determinam os membros dentro de uma categoria. Por exemplo, podería-se querer testar os fatores que conduzem a não concessão de um empréstimo a partir das informações cadastrais de milhares de pessoas, usado como suporte a decisão no momento de conceder um empréstimo, o modelo deveria poder usar estes fatores para discriminar "prováveis a receber" e "os prováveis a não receber" o empréstimo.

A figura 5 descreve o resultado de uma "análise discriminatória bem sucedida ". O modelo provido pode achar fatores que separam os grupos que receberam daqueles que não receberam. Inevitavelmente, haverá um pouco de variabilidade nas pontuações dentro de cada um dos grupos como é mostrado pela distribuição de probabilidades. Uma "análise discriminatória bem sucedida" poderá minimizar a quantidade de sobreposições entre estas duas distribuições [GRO97].

Figura 5 – Análise Discriminatória



Fonte: [GRO97]

3.5.3 ANÁLISE DE GRUPO

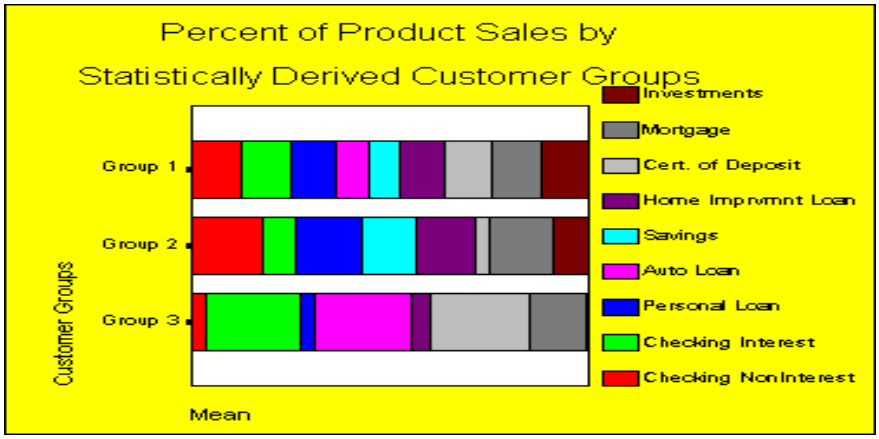
Análise de grupo é uma técnica de agrupamento de dados que constitui na construção de modelos para encontrar dados semelhantes, e estas reuniões por semelhança que são chamadas de grupos (*clusters*). É uma forma de *Data Mining* não-direcionado, onde a meta é encontrar similaridades não conhecidas anteriormente.

Por exemplo, suponha-se que um banco queira descobrir os segmentos de clientes baseando-se no tipo de conta que eles abrem. A análise de agrupamento feita sobre uma base de dados distinguiu três tipos de clientes como mostra a figura 6. As diferentes cores dos segmentos representam um resumo dos eventos e transações que são realizados de acordo com o tipo de cliente.

O primeiro grupo, que revelou percentagens quase iguais entre os diversos eventos, esses podem ser tratados como clientes gerais. O segundo grupo possui mais hipotecas, investimentos, empréstimos para compra de imóveis, e volume de depósitos, podendo ser tratados como clientes a longo prazo. E o terceiro utiliza mais conta corrente, poupança, e empréstimos pessoais, estes podendo ser chamados de clientes a curto prazo. Com base nestas informações o banco pode

adotar diferentes estratégias para tratar cada segmento, a fim de melhorar seus negócios [GRO97].

Figura 6 – Análise de Grupo



Fonte: [GRO97]

3.5.4 ANÁLISE DE VÍNCULOS

A análise de vínculos segue as relações entre registros para desenvolver modelos baseados em padrões nas relações. Esse é um aplicativo de construção de teoria gráfica de *Data Mining*. Esta técnica não é muito compatível com a tecnologia de banco de dados relacionais e sua maior área de aplicação é a área policial, onde pistas são ligadas entre si para solucionar os crimes. As poucas ferramentas que existem, enfocam mais a visualização de vínculos que a análise de padrões [HAR98].

3.5.5 MBR

O MBR (*Memory-Based Reasoning* – raciocínio baseado em memória) é uma técnica de *Data Mining* dirigida que usa exemplos conhecidos como modelo para fazer previsões sobre exemplos desconhecidos. O MBR procura os vizinhos mais próximos nos exemplos conhecidos e combina seus valores para atribuir valores de classificação ou de previsão [BER97].

Os elementos-chave no MBR são a função de distância usada para encontrar os vizinhos mais próximos e a função de combinação, que combina valores dos vizinhos mais próximos para fazer uma previsão. Uma vantagem do MBR é sua habilidade de aprender sobre novas classificações simplesmente introduzindo novos exemplos no banco de dados. Uma vez encontrada a função de distância e a função de combinação corretas tendem a permanecer muito estáveis, mesmo com a incorporação de novos exemplos para novas categorias nos dados conhecidos. Aliás, esta é uma característica que diferencia o MBR da maior parte das outras técnicas de *Data Mining*.

3.5.6 REDES NEURAIS ARTIFICIAIS

De acordo com [KRA99] as redes neurais são modelos que simulam a estrutura do cérebro humano, adaptados para o uso em computadores e são, provavelmente, a técnica de *Data Mining* mais utilizada. Elas aprendem com um conjunto de dados de treinamento, generalizando modelos para classificação e previsão. Esta técnica pode também ser aplicada ao *Data Mining* não-dirigido e às previsões em séries temporais.

Uma das principais vantagens na utilização desta técnica é a sua variedade de aplicação. Elas são interessantes porque detectam padrões nos dados de forma análoga ao pensamento humano. Mas existem duas desvantagens em seu uso:

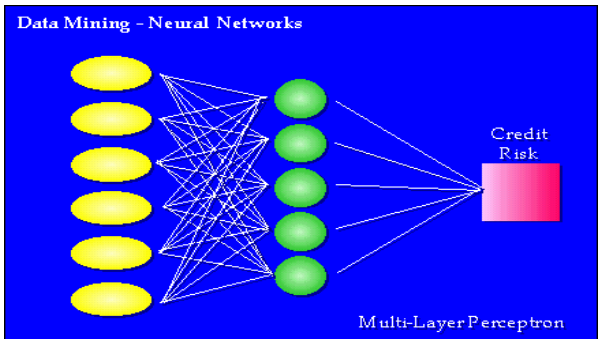
a) a dificuldade de interpretar os modelos produzidos por elas;
b) a sensibilidade ao formato dos dados que a alimentam, pois representações de dados diferentes podem produzir resultados diversos.

Uma rede neural artificial é composta por várias unidades de processamento, cujo funcionamento é bastante simples. Essas unidades, geralmente são conectadas por canais de comunicação que estão associados a determinado peso. As unidades fazem operações apenas sobre seus dados locais, que são entradas recebidas pelas suas conexões. O comportamento inteligente de uma rede neural artificial vem das interações entre as unidades de processamento da rede [LOE96].

A maioria dos modelos de redes neurais possui alguma regra de treinamento, onde os pesos de suas conexões são ajustados de acordo com os padrões apresentados. Em outras palavras, elas aprendem através de exemplos [LOE96].

Arquiteturas neurais são tipicamente organizadas em camadas, com unidades que podem estar conectadas às unidades da camada posterior conforme figura 7.

Figura 7 - Organização das camadas.



Fonte: [GRO97]

Usualmente as camadas são classificadas em três grupos:

a) Camada de Entrada: onde os padrões são apresentados à rede;
b) Camadas Intermediárias ou Escondidas: onde é feita a maior parte do processamento, através das conexões ponderadas; podem ser consideradas como extratoras de características;
c) Camada de Saída: onde o resultado final é concluído e apresentado.

3.5.7 ALGORITMOS GENÉTICOS

Os algoritmos genéticos aplicam a mecânica da genética e seleção natural à pesquisa usada para encontrar os melhores conjuntos de parâmetros que descrevem uma função de previsão. Eles são utilizados no *Data Mining* dirigido e são semelhantes à estatística, em que a forma do modelo precisa ser conhecida em profundidade. Os algoritmos genéticos usam os operadores seleção, cruzamento e mutação para desenvolver sucessivas gerações de soluções.

Com a evolução do algoritmo, somente os mais previsíveis sobrevivem, até as funções convergirem em uma solução ideal [BER97].

Esta técnica é apropriada para resolver os mesmos tipos de problemas que as outras técnicas de *Data Mining*, mas ela também pode ser usada para aprimorar MBRs e redes neurais.

3.5.8 ÁRVORES DE DECISÃO

De acordo com [QUI93] árvores de decisão expressam uma forma simples de lógica condicional buscando a representação de uma série de questões que estão escondidas sobre a base de dados. Em uma árvore de decisão existem dois tipo de atributos, o decisivo, que é aquele que contém o resultado ao qual deseja-se obter e os não decisivos que contém os valores que conduzem a uma decisão.

Através de uma fórmula matemática, denominada entropia, são realizados cálculos sobre os atributos não decisivos, denominados classes, onde é escolhido um nó inicial também chamado de raiz; a partir deste nó será realizado uma série de novos cálculos com o objetivo de decidir a estrutura de formação da árvore a ser gerada. Este processo é repetido até que todos os atributos a serem processados estejam perfeitamente classificados ou já se tenha processado todos os atributos.

Os três principais algoritmos conhecidos que implementam árvores de decisão, são ID3, (demonstrado na figura 9), C4.5 e PERT, sendo que os algoritmos C4.5 e PERT são um aperfeiçoamentos do algoritmo ID3 com alguns conceitos avançados de podagem (técnica de cortar nós da árvore que não são potencialmente úteis) e preocupação com a performance do mesmo em relação ao tempo de processamento.

O objetivo do algoritmo ID3 é gerar os valores categóricos de um atributo chamado *classe*, para isso utilizando-se de um método de classificação que tem o objetivo de realizar testes que são introduzidos na árvore, separando os casos de treino em subconjuntos. Cada subconjunto deve consistir de exemplos de uma única classe.

A distribuição de classes pode ser representada em forma de uma lista de probabilidades p(c1) .. p(cn), em que cada p_i indica a probabilidade de um exemplo pertencer à uma classe.

Os valores das funções que calculam essas probabilidades representam a informação necessária para classificar um caso e são chamados de entropia e *gain*, sendo calculados com as seguintes fórmulas demonstradas na figura 8.

Figura 8 – Fórmulas para calcular entropia e *gain*

Entropia(S) = ∑ -(1) p(I) log_2p(I) onde

Log_2 é o logaritmo de um número com base 2

p(I) é quantidade de ocorrências cada valor possível de uma classe dividido pela quantidade total da classe.

Gain (S,A) = Entropia(S) - ∑ ((|Sv|) / |S|) * Entropia(Sv)) onde

∑ é cada valor possível de todos os valores do atributo A

Sv é a quantidade de ocorrências de cada atributo definido por A

|Sv| é o número total de elementos definido por Sv

|S| é o número total de elementos da coleção.

Fonte: adaptado de [QUI93]

[B1] Comentário:

Figura 9 – Algoritmo ID3

```
      function ID3 (R: a set of non-goal attributes,
                    C: the goal attribute,
                     S: a training set) returns a decision tree;

      begin
        If S is empty,
          return a single node with value Failure;

        If S consists of records all with the same value for the goal
attribute,
          return a single node with that value;

        If R is empty, return a single node with as value
         the most frequent of the values of the goal attribute
         that are found in records of S;

         [note that then there will be errors, that is, records that
          will be improperly classified];

        Let D be the attribute with largest
           Gain(D,S) among attributes in R;

        Let {dj| j=1,2, .., m} be the values of attribute D;

        Let {Sj| j=1,2, .., m} be the subsets of S consisting
          respectively of records with value dj for attribute D;

        Return a tree with root labeled D and arcs labeled
         d1, d2, .., dm going respectively to the trees

        ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), .., ID3(R-{D}, C, Sm);

      end ID3;
```

Fonte: [QUI93]

4 DESENVOLVIMENTO DO SIE

Para o desenvolvimento do SIE, adotou-se a metodologia de análise estruturada. Esta, segundo [YOU90], é uma metodologia no qual tanto os analistas quanto os usuários sabem que o produto final da prototipação será o próprio sistema, já na sua forma aperfeiçoada.

4.1 ESPECIFICAÇÃO

De acordo com [YOU90] a atividade de desenvolvimento e análise de sistemas estruturada enfatiza que um sistema de processamento de dados envolve dados e processamento, e que não se pode construir um sistema com êxito sem a participação de ambos os componentes. O processamento de um sistema é, certamente, um aspecto importante para ser modelado e examinado. A modelagem de dados utilizando a técnica estruturada utiliza-se de ferramentas para descrever o processo de entradas em saídas e uma delas é o diagrama de fluxo de dados (DFD), um diagrama de fluxo de dados consiste em processos, depósitos de dados, fluxos e entidades.

a) processos: são representados por círculos no diagrama e apresentam as diversas funções individuais que o sistema executa. Essas funções são as responsáveis em transformar as entradas em saídas.

Figura 10 – Processos

Fonte: adaptado de [YOU90]

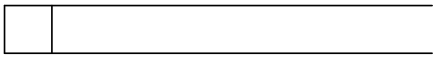
b) fluxos: são representados por setas direcionadas ou curvas. Elas são as conexões entre os processos , e representam a informação que os processos exigem como entrada e/ou informações que eles geram como saída.

Figura 11 – Fluxos

Fonte: adaptado de [YOU90]

c) depósitos de dados: são mostrados como duas linhas paralelas no diagrama. Eles mostram as coleções de dados que o sistema deverá contemplar. Quando a parte física do sistema será implantada eles serão traduzidos como arquivos ou tabelas em banco de dados.

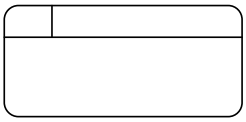
Figura 12 – Depósito de Dados



Fonte: adaptado de [YOU90]

d) entidades: mostram as entidades externas com as quais o sistema se comunica. As entidades são, tipicamente, indivíduos, grupos de pessoas, um departamento em uma empresa, sistemas ou organizações externas.

Figura 13 – Entidades

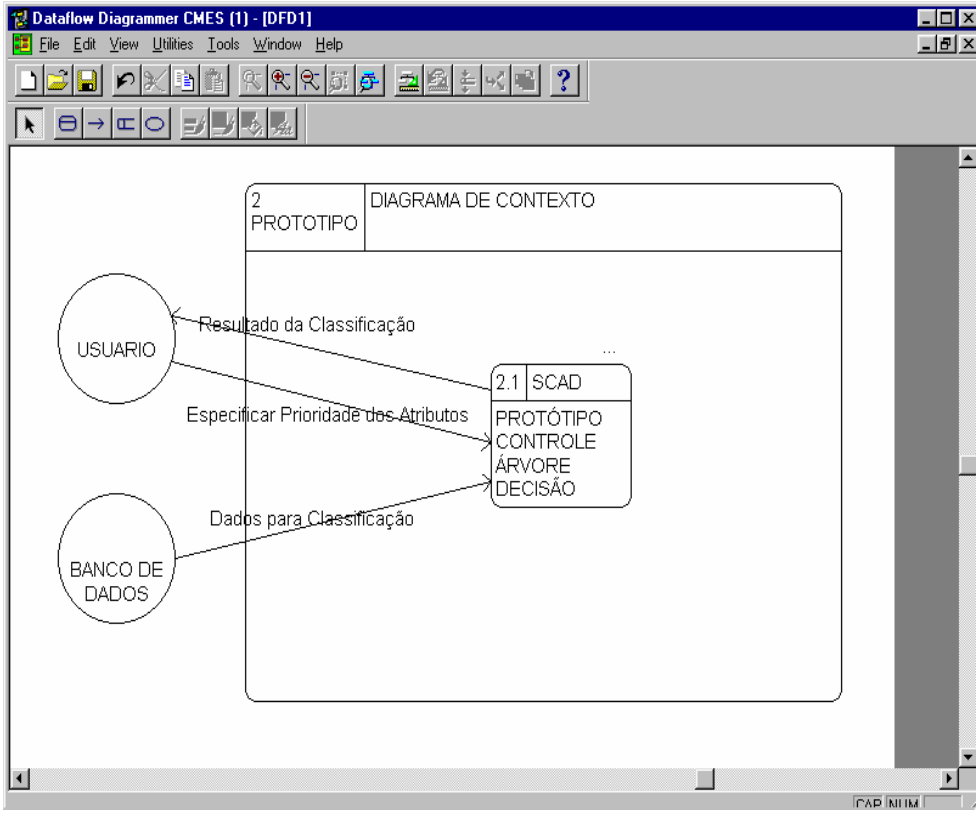


Fonte: adaptado de [YOU90]

As representações dos processos, fluxos, depósito de dados e entidades contidas diagrama de contexto e DFD nível 0 do protótipo são adaptadas de [YOU90] conforme demonstrado nas figuras 14 e 15.

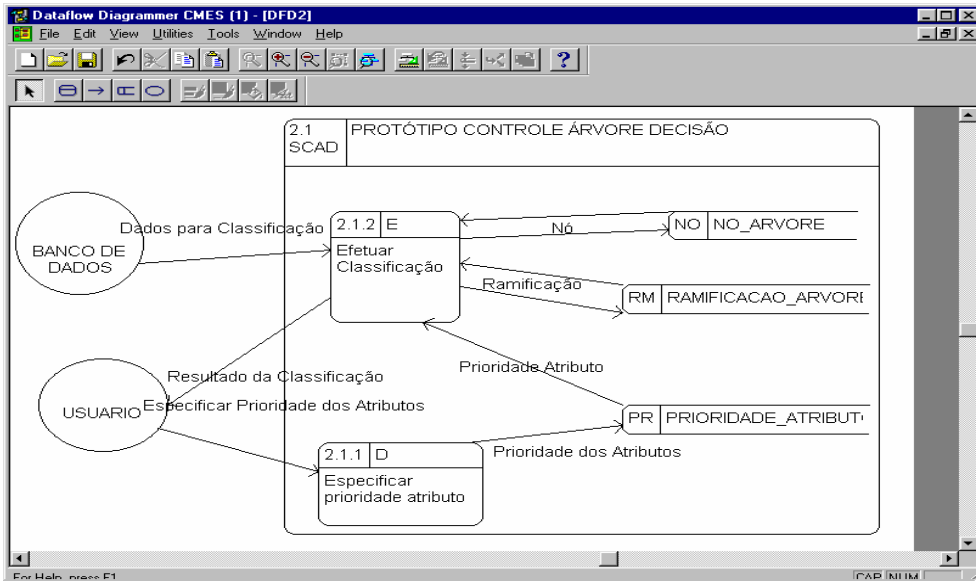
Para construção do diagrama de contexto, DFD nível 0 e do modelo entidade relacionamento deste protótipo foi utilizada a ferramenta case Oracle Designer 2000, através das ferramentas de análise de modelo de sistemas e a análise de modelo físico de dados.

Figura 14 – Diagrama de Contexto



O protótipo irá interagir com o usuário que fará a especificação da ordem de prioridade dos atributos, e, por fim, serão efetuadas as classificações sobre a base de dados que serão apresentadas ao usuário com o resultado da classificação obtida pelo processo de *Data Mining*.

Figura 15 – DFD nível 0



Descreve-se a seguir os processos e entidades do DFD nível 0:

a) especificar prioridade atributo: processo onde o usuário definirá a ordem de prioridade dos atributos a serem classificados. Este processo é importante para em determinados momentos do processo, quando, o valor da entropia for o mesmo para mais de um atributo, ser priorizado o atributo que possuir maior valor especificado pelo usuário;
b) efetuar previsão: este processo é caracterizado pela utilização da ordem dos atributos especificada pelo usuário e revogado para efetuar a classificação dos dados, isto é, executado o processo de *Data Mining* em si. Neste momento é obtido o resultado do processo de classificação que será mostrado ao usuário;
c) nó: entidade responsável pelo armazenamento dos dados referentes aos nós da árvore, gerados pelo processo de classificação;
d) ramificação: entidade responsável pelo armazenamento dos dados referentes as ligações existentes entre os nós;
e) prioridade atributo: entidade responsável pelo armazenamento dos dados referentes a ordem dos atributos especificada pelo usuário.

Embora o diagrama de fluxo de dados ofereça uma prática visão geral dos principais componentes funcionais do sistema, não fornece qualquer detalhe sobre esses componentes. Para mostrar os detalhes de como a informação é transformada, necessita-se de ferramentas como o dicionários de dados e o modelo entidade relacionamento (MER).

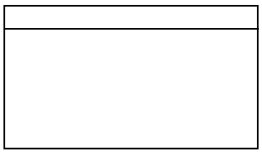
O dicionário de dados é uma listagem organizada de todos os elementos de dados pertinentes ao sistema, como definições precisas e rigorosas para que o analista de sistemas possa conhecer todas as entradas, saídas, componentes de depósitos e cálculos intermediários. O dicionário de dados define os elementos de dados da seguinte maneira.

a) descrevendo o significado dos fluxos e depósitos envolvidos no DFD;
b) descrevendo a composição da estrutura dos dados que se movimentam pelos fluxos;
c) descrevendo a estrutura dos depósitos de dados;
d) especificando os valores e unidades relevantes de partes elementares de informação;
e) descrevendo os detalhes de relacionamentos entre os depósitos de dados.

O modelo de entidade relacionamento é necessário por que a maioria dos sistemas a qual seu uso é justificado é bastante complexo. Não somente necessita-se saber, em detalhes, qual informação está contida nos depósitos de dados, mas também que relacionamentos existentes entre esses depósitos de dados. O MER possui dois componentes importantes:

a) entidades: são apresentadas por um quadro retangular no diagrama, representa uma coleção, conjunto, objetos do mundo real cujos membros desempenham um papel no sistema que está sendo desenvolvido.

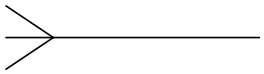
Figura 16 – Entidades



Fonte: adaptado de [YOU90]

b) relacionamentos: representados por losangos, representam um conjunto de conexões ou associações entre as entidades, isto é, de que forma e em que grau uma entidade está ligada com a outra.

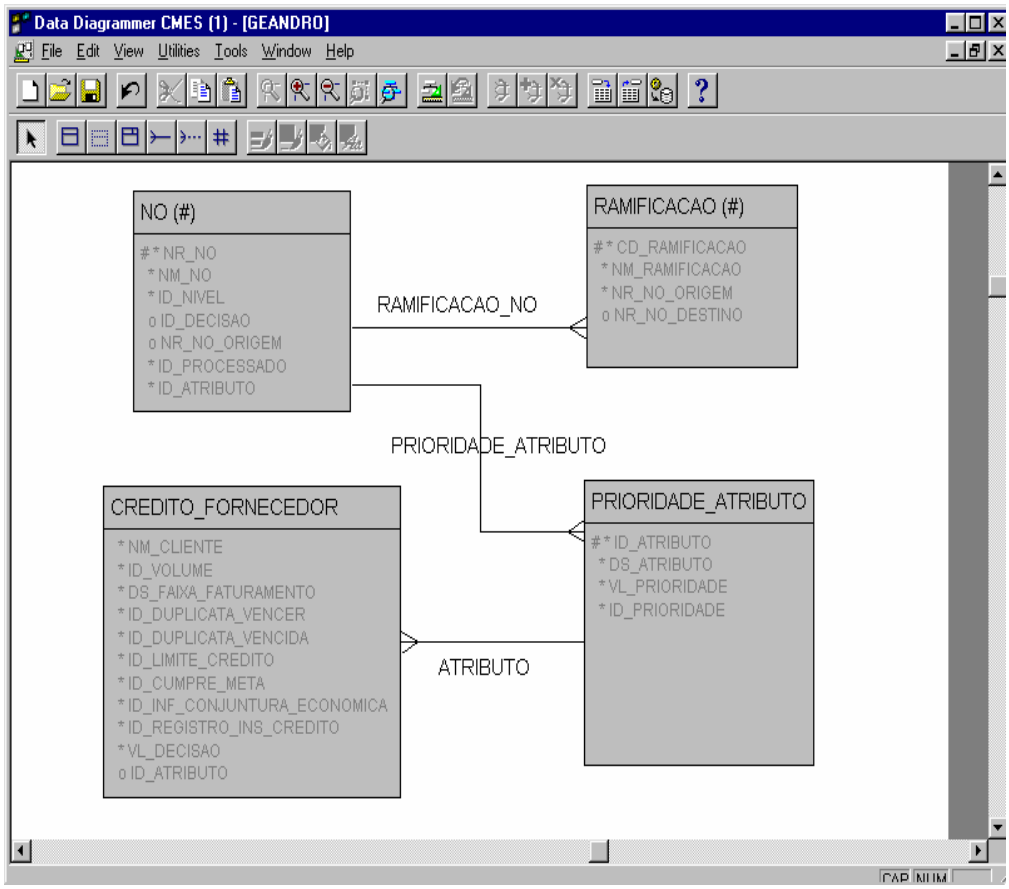
Figura 17 – Relacionamentos



Fonte: adaptado de [YOU90]

Na figura 18 demonstra-se a representação do modelo entidade relacionamento do protótipo.

Figura 18 – Modelo Entidade Relacionamento



As entidades “NO”, “RAMIFICACAO” e “PRIORIDADE_ATRIBUTO” surgiram a partir dos depósitos de dados contidos no DFD nível 0, e a tabela “CREDITO_FORNECEDOR” é a representação da entidade externa “BANCO DE DADOS” que será a base de dados utilizada para o processamento da árvore de decisão. A tabela 4 contém o detalhamento dessas entidades.

Tabela 4 - Descrição detalhada do modelo de dados.

ENTIDADE	ATRIBUTO	TIPO DE DADO	OPCIONAL ?	TAMANHO
NO	NR_NO	NUMÉRICO	NÃO	3
	NM_NO	ALFANUMÉRICO	NÃO	30
	ID_NIVEL	NUMÉRICO	NÃO	3
	ID_DECISAO	ALFANUMÉRICO	SIM	30
	NR_NO_ORIGEM	NUMÉRICO	SIM	3
	ID_PROCESSADO	NUMÉRICO	NÃO	1
	ID_ATRIBUTO	NUMÉRICO	NÃO	1
RAMIFICACAO	CD_RAMIFICACAO	NUMÉRICO	NÃO	3
	NM_RAMIFICACAO	ALFANUMÉRICO	NÃO	30
	NR_NO_ORIGEM	NUMÉRICO	NÃO	3
	NR_NO_DESTINO	NUMÉRICO	SIM	3
PRIORIDADE_ATRIBUTO	ID_ATRIBUTO	ALFANUMÉRICO	NÃO	30
	DS_ATRIBUTO	ALFANUMÉRICO	NÃO	30
	VL_PRIORIDADE	NUMÉRICO	NÃO	17,4
	ID_PRIORIDADE	NUMÉRICO	NÃO	1
CREDITO_FORNECEDOR	NM_CLIENTE	ALFANUMÉRICO	NÃO	50
	ID_VOLUME	ALFANUMÉRICO	NÃO	10
	DS_FAIXA_FATURAMENTO	ALFANUMÉRICO	NÃO	20
	ID_DUPLICATA_VENCER	ALFANUMÉRICO	NÃO	3
	ID_DUPLICATA_VENCIDA	ALFANUMÉRICO	NÃO	3
	ID_LIMITE_CREDITO	ALFANUMÉRICO	NÃO	10
	ID_CUMPRE_META	ALFANUMÉRICO	NÃO	3
	ID_INF_CONJUNTURA_ECONOMICA	ALFANUMÉRICO	NÃO	10
	ID_REGISTRO_INS_CREDITO	ALFANUMÉRICO	NÃO	3
	VL_DECISAO	ALFANUMÉRICO	NÃO	3
	ID_ATRIBUTO	NUMÉRICO	NÃO	1

4.2 BANCO DE DADOS

Foi utilizado o banco de dados ORACLE para gerenciar o armazenamento e controle dos dados necessários para especificar o protótipo. Este capítulo abrangerá as divisões que contém as funções realmente utilizadas no protótipo, evitando entrar em detalhes sobre a parte de administração e controle que o ORACLE realiza.

4.2.1 CONCEITO

De acordo com [CER95] o ORACLE é um Sistema Gerenciador de Banco de Dados relacional, isto é, utiliza o conceito de álgebra relacional, baseando-se em uma coleção de tabelas de duas dimensões (linhas e colunas) .Uma definição bastante difundida sobre banco de dados relacional é a seguinte:

"Um banco de dados relacional é uma coleção de dados organizados e integrados armazenados em forma de tabelas interligadas através de chaves primárias e estrangeiras, que constituem uma representação natural dos dados, sem imposição de restrições ou modificações, de forma a ser adequada a qualquer computador, podendo ser utilizada por todas as aplicações relevantes sem duplicação de dados, e sem a necessidade de serem definidos em programas, pois utiliza as definições existentes nas bases de dados, através do dicionário de dados ativo e dinâmico." [YOU90]

4.2.2 LINGUAGEM

De acordo com [CER95] o banco de dados ORACLE possui uma linguagem de interpretação e execução da maioria das funções que ele suporta, chamada de Structured Query Language (SQL) , sendo ela a responsável pela entrada e saída das instruções executadas sobre o banco de dados.

O SQL ORACLE é dividido em três principais divisões de comandos, que são chamadas de Data Manipulation Language (DML), Data Definition Language (DDL) e Data Control Language (DCL).

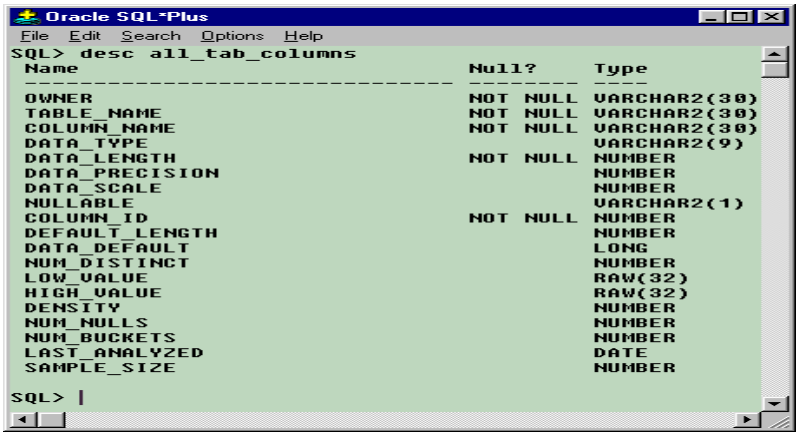
c) data manipulation language: é o conjunto de comandos utilizados para selecionar, inserir, atualizar, excluir, ordenar, agrupar, restringir, contar, unir registros e também comandos necessários para salvar e desfazer as alterações efetuadas sobre a base de dados, sendo eles: SELECT, INSERT, UPDATE, DELETE, ORDER BY, GROUP BY, WHERE, COUNT, UNION, COMMIT e ROLLBACK.
d) data definition language: é o conjunto de comandos utilizados para criar, alterar, excluir e renomear a estrutura das tabelas, sendo eles os comandos: CREATE, ALTER, DROP e RENAME.
e) data control language: é o conjunto de comandos utilizados para especificar o direito de acesso aos dados por parte dos usuários, sendo eles: GRANT, REVOKE e LOCK.

4.2.3 DICIONÁRIO DE DADOS ORACLE

O dicionário de dados é uma das mais importantes partes de qualquer sistema gerenciador de banco de dados, pois nele existe toda a referência sobre a estrutura das tabelas, sobre os direitos de acesso às tabelas, sobre os relacionamentos entre as tabelas, e sobre todos os mecanismos de controle utilizados pelo SGBD.

O acesso ao dicionário de dados ORACLE é feito através de comandos SELECTS nas tabelas e visões do dicionário. A manutenção destas tabelas e visões é de total responsabilidade do SGBD, que as altera à medida que a estrutura do banco de dados é alterada. Por exemplo, se fosse necessário acessar o dicionário de dados para descobrir quais as colunas e tipo de dados de uma coluna de uma determinada tabela, bastaria somente selecionar a visão (forma predeterminada de visualizar dados de uma ou mais tabelas como se fosse uma única tabela) ALL_TAB_COLUMNS, que contém a estrutura de todas as tabelas, visões e colunas existentes no banco de dados, conforme demonstrado na figura 19.

Figura 19 – Descrição da visão ALL_TAB_COLUMNS

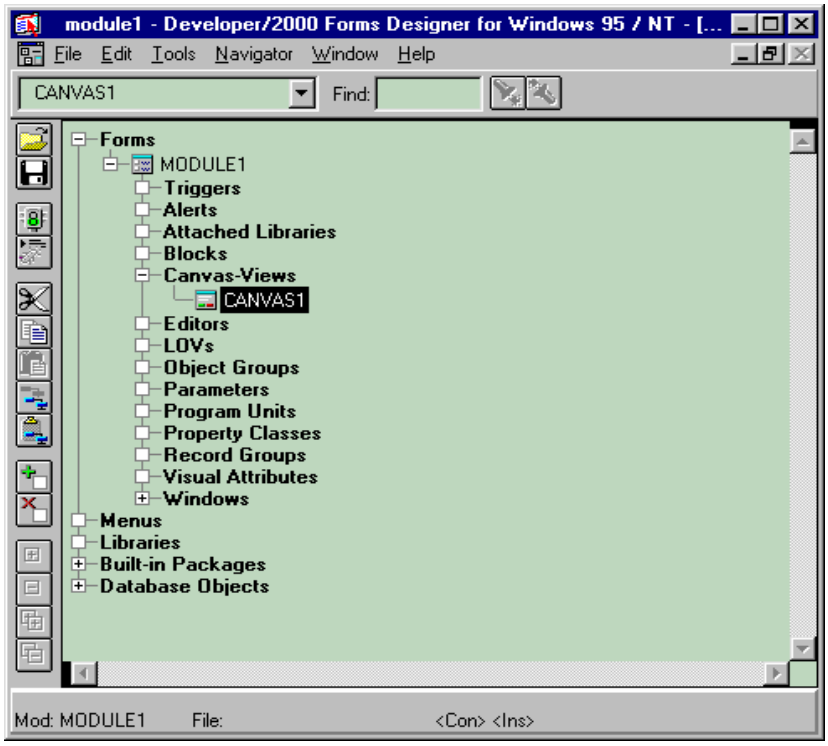


4.3 AMBIENTE VISUAL

A ferramenta gráfica utilizada para implementação do protótipo é ORACLE FORMS (tela principal demostrada na figura 20), que de acordo com [DAY95] é uma sofisticada ferramenta de desenvolvimento que simplifica a construção portável de aplicações gráficas baseadas em telas. Sua aplicação pode incorporar: botões, radio groups, list boxes, ícones, imagens, menus, listas de valores, suporte a PL/SQL, suporte a chamada de packages (bibliotecas escritas em PL/SQL), suporte a Object Link and Embeending(OLE) entre outras funções.

Com o ORACLE FORMS é possível gerar aplicações complexas sem a necessidade de escrever um grande volume de linhas de código, já que as funções básicas de acesso ao banco de dados e funções de controle de objetos padrão já estão previamente construídas pela ferramenta que utiliza da programação orientada a eventos, sendo necessário apenas incorporar as regras inerentes ao negócio.

Figura 20 – Tela principal do ORACLE FORMS



As principais funções e objetos do ORACLE FORMS são

a) *triggers*: código executado a partir do acontecimento de um evento, utilizado principalmente para validação de campos e tratamento de regras de negócio;
b) *alerts*: caixas de diálogo utilizada para formulação de perguntas ao usuário;
c) *attached libraries*: bibliotecas com implementações de funções externas;
d) *blocks*: referência a uma tabela ou visão do banco de dados;
e) *canvas-views*: desenho das telas que vão compor o sistema;
f) *lovs*: lista de valores e domínios para os campos;
g) *program-units*: procedures e functions de uso genérico pelo sistema;
h) *visual attributes*: características visuais dos campos e telas.

4.4 A TÉCNICA ÁRVORE DE DECISÃO

A coleção de dados representada pela tabela 5 servirá de exemplo para demonstração do funcionamento do algoritmo proposto neste protótipo, onde:

A1 = ID_VOLUME (Volume requerido de empréstimo).

A2 = DS_FAIXA_FATURAMENTO (Faixa de faturamento fornecedor).

A3 = ID_DUPLICATA_VENCER (Existência de duplicatas a vencer).

A4 = ID_DUPLICATA_VENCIDA (Existência de duplicatas vencidas).

A5 = ID_LIMITE_CREDITO (Limite de crédito do fornecedor).

A6 = ID_CUMPRE_META (Cumprimento de metas estabelecidas).

A7 = ID_INF_CONJUNTURA_ECONOMICA (Influência da conjuntura econômica).

A8 = ID_REGISTRO_INS_CREDITO (Registro em instituição de controle de crédito).

AD = VL_DECISAO (Atributo decisivo).

Tabela 5 – Informações sobre fornecedores

FORNECEDOR	A1	A2	A3	A4	A5	A6	A7	A8	AD
ATAMIRANDO DA FONSECA	ALTO	0 – 100000	NÃO	NÃO	ALTO	SIM	NÃO	NÃO	NAO
CHARLES LONGO	BAIXO	100000 – 500000	SIM	SIM	ALTO	NAO	SIM	SIM	NAO
CLAUDIO TAFFAREL	MEDIO	ACIMA DE 500000	SIM	NÃO	BAIXO	NAO	NAO	NAO	NAO
DANIEL CARLOS SANTANA	ALTO	0 – 100000	SIM	SIM	ALTO	SIM	SIM	SIM	SIM
ELIO LEITE	BAIXO	ACIMA DE 500000	NAO	NÃO	MEDIO	NÃO	NAO	NAO	SIM
VALDO DE OLIVEIRA	MEDIO	0 – 100000	SIM	SIM	ALTO	SIM	SIM	SIM	SIM

De acordo com [QUI93] o processo de formação da árvore decisão começa com a definição de que atributo será o nó inicial de árvore (também chamado de nó raiz) para isso deve-se calcular a entropia (conceito utilizado para determinar o fator de incidência de cada atributo não decisivo em relação ao decisivo) do atributo decisivo da coleção de dados, determinada pela fórmula apresentada no capítulo 3.5.8, conforme tabela 5 o atributo decisivo é chamado de “AD”. O processo de cálculo da entropia começa com a seleção distinta dos valores do atributo decisivo. Conforme tabela 5 os valores distintos para o atributo decisivo são (SIM,NÃO). Então calcula-se a quantidade de vezes que cada um desses valores ocorrem dentro da coleção.

Quantidade de ocorrências para (SIM,NÃO): (3,3)

Quantidade total de ocorrências : 6

Entropia = $\sum$ -(1) p(I) $\log_2$p(I)

[B2] Comentário:

I = Quantidade de ocorrências para valor distinto do atributo dividido pela quantidade total de ocorrências.

S = Coleção de dados (Tabela 5).

Log_2 = Logaritmo de base 2.

Entropia(S) = - (3/6) * $\log_2$(3/6) - (3/6) * $\log_2$(3/6)

Entropia(S) = 1

Após apurada a entropia do atributo decisivo deve-se calcular o valor do *Gain* para cada atributo não decisivo (conforme tabela 5 seriam os atributos A1, A2, A3, A4, A5, A6, A7 ,e A8) determinado pela fórmula abaixo também apresentada no capítulo 3.4.8, o atributo não decisivo que possuir o maior *Gain* será considerado o atributo inicial da árvore.

Gain (S,A) = Entropia(S) - $\sum$ ((|Sv|) / |S|) * Entropia(Sv))

Calculo do valor de *Gain* para o atributo A5:

Para calcular o valor de *Gain* deve-se selecionar os valores distintos de cada atributo não decisivo e contar a quantidade de vezes que cada um desses valores ocorrem em relação ao atributo decisivo, calculando-se também a sua entropia, conforme demonstração abaixo:

Valores distintos para Limite Crédito: (BAIXO,MEDIO,ALTO)

Quantidade de ocorrências para (BAIXO,MEDIO,ALTO) : (1,1,4)

Entropia(BAIXO) = BAIXOsim = 0 , BAIXOnão = 1

Entropia(BAIXO) = 0

Entropia(MEDIO) = MEDIOsim = 1 , MEDIOnão = 0

Entropia(MEDIO) = 0

Entropia(ALTO) = ALTOsim = 2 , ALTOnão = 2

Entropia(ALTO) = 1

Gain (S,A5) = 1 – ((2 / 6) * Entropia(S_{baixo}) + (2 / 6) * Entropia(S_{medio})
+ (4 / 6) * Entropia(S_{alto})

Gain(S, A5) = 1 – ((2/6) * 0)) + ((2/6) * 0) + (4/6) * 1)))

Gain(S, A5) = 0.333

O mesmo processo é realizado para os demais atributos onde obtêm-se o seguinte resultado:

Gain(S, A1) = 0

Gain(S, A2) = 0.207

Gain(S, A3) = 0

Gain(S, A4) = 0.081

Gain(S, A6) = 0.081

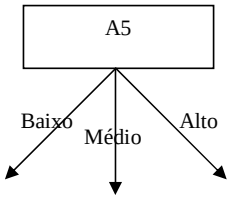
Gain(S, A7) = 0.081

Gain(S, A8) = 0.081

O atributo “A5” possui o maior valor de *Gain* logo ele será o atributo usado com o nó inicial da árvore.

Determinado o nó inicial da árvore, o próximo passo é definir a primeira ramificação que a árvore vai sofrer. Para isso deve-se selecionar os diferentes valores possíveis para o atributo considerado como o nó inicial da árvore, conforme tabela 5 o atributo A5 possui três valores distintos (Baixo,Médio,Alto) e para cada um desses valores deve-se criar uma ramificação.

Figura 21 – Primeira ramificação da árvore



O próximo passo é saber qual é o próximo nó a ser gerado para cada uma das novas ramificações existentes. Para isto deve-se agora considerar cada subconjunto gerado pelo valor dos atributos do nó raiz, conforme tabelas 6,7, e 8.

Tabela 6 - Subconjunto gerado pelo atributo A5 valor "BAIXO"

FORNECEDOR	A1	A2	A3	A4	A5	A6	A7	A8	AD
CLAUDIO TAFFAREL	MEDIO	ACIMA DE 500000	SIM	NÃO	BAIXO	NAO	NAO	NÃO	NAO

Tabela 7 - Subconjunto gerado pelo atributo A5 valor "MEDIO"

FORNECEDOR	A1	A2	A3	A4	A5	A6	A7	A8	AD
ELIO LEITE	BAIXO	ACIMA DE 500000	NAO	NÃO	MEDIO	NÃO	NAO	NAO	SIM

Tabela 8 - Subconjunto gerado pelo atributo A5 valor "ALTO"

FORNECEDOR	A1	A2	A3	A4	A5	A6	A7	A8	AD
ATAMIRANDO DA FONSECA	ALTO	0 – 100000	NÃO	NÃO	ALTO	SIM	NÃO	NÃO	NAO
CHARLES LONGO	BAIXO	100000 - 500000	SIM	SIM	ALTO	NAO	SIM	SIM	NAO
DANIEL CARLOS SANTANA	ALTO	0 – 100000	SIM	SIM	ALTO	SIM	SIM	SIM	SIM
VALDO DE OLIVEIRA	MEDIO	0 – 100000	SIM	SIM	ALTO	SIM	SIM	SIM	SIM

Seguindo o mesmo processo realizado pelo nó raiz deve-se calcular a entropia de cada umas das ramificações geradas pelo nó que foi gerado.

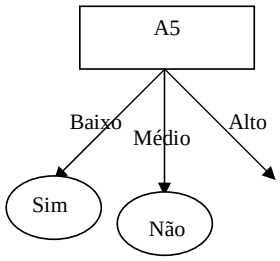
Entropia (Baixo) = 0

Entropia (Medio) = 0

Entropia (Alto) = 1

No algoritmo ID3 as ramificações que possuem valor de entropia igual a zero já estão perfeitamente classificadas, isto é, existe apenas um valor distinto de decisão para a mesma, logo já pode-se finalizar a ramificação com um nó decisão, atribuindo-se ao nó o valor distinto gerado pela sua ramificação, conforme demonstrado na figura 22.

Figura 22 – Geração dos nós decisão



Para a ramificação que ainda não esteja perfeitamente classificada deve determinar qual o próximo atributo à ser conectado a ramificação. Efetuando os mesmo cálculos sobre a coleção de dados representados pelas tabelas 6, 7, e 8 chega-se aos seguintes valor de Gain para os atributos restantes.

Gain(S, A1) = 0.5

Gain(S, A2) = 0.3112

Gain(S, A3) = 0.3112

Gain(S, A4) = 0.3112

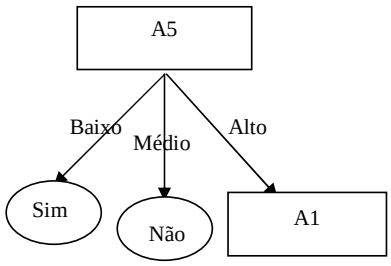
Gain(S, A5) = 0.3112

Gain(S, A7) = 0.3112

Gain(S, A8) = 0.3112

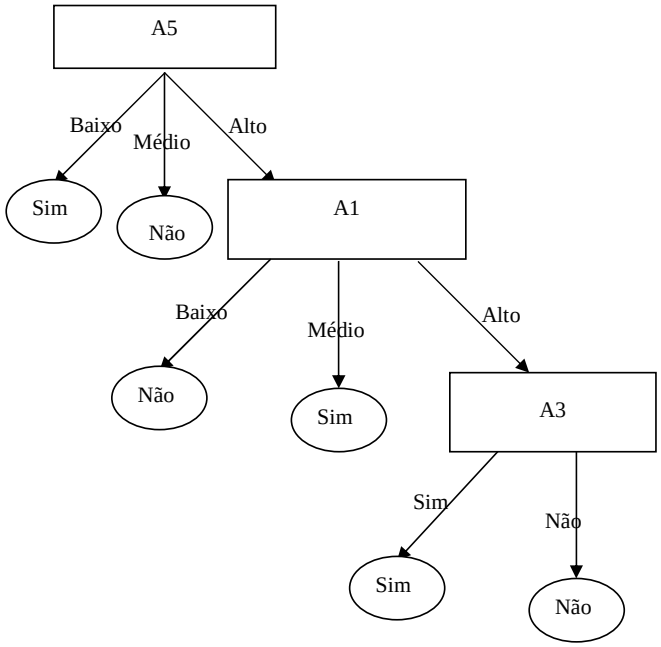
De acordo com os valores calculados o atributo com maior valor de Gain é o atributo A1. Caso exista mais de um atributo com o mesmo valor de Gain, o protótipo dará preferência ao atributo que tiver o maior valor de prioridade, que deve ser informada pelo usuário.

Figura 23 – Geração do próximo nó para a ramificação



Este processo deve ser repetido até que todos os atributos estejam perfeitamente classificados ou todos os atributos já tenham sido processados. Na figura 24 é demonstrado como ficará a árvore de decisão após o seu processamento completo.

Figura 24 – Árvore após seu processamento completo



Através dos dados contidos na tabela 5 chegou-se a representação gráfica da árvore de decisão, conforme demonstrada pela figura 24, onde, nem todos os atributos envolvidos estão presentes, isso ocorre devido a forma com que os dados estão dispostos na tabela 5, isto é, conforme os dos dados sofram modificações, a árvore poderá ganhar mais níveis e consequentemente mais nós.

4.5 DESENVOLVIMENTO DO PROTÓTIPO

Levando em conta os objetivos propostos por este trabalho, construiu-se um Sistema de Informação Executiva que fosse flexível e de fácil utilização. Aproveitando a flexibilidade da linguagem escolhida, resolveu-se utilizar a filosofia *Data Mining*, mais especificamente a técnica de árvores de decisão que foi desenvolvida para a mesma baseado em [QUI93].

Figura 25 – Tela de abertura do protótipo

4.5.1 SELEÇÃO DOS DADOS

Os dados utilizados neste protótipo consiste em simulação de um conjunto de situações que determinam se uma empresa deve ou não conceder empréstimos a seus fornecedores e parceiros responsáveis pelo fornecimento da matéria-prima necessária para a boa liquidez de seus negócios.

Foram definidos oito atributos para permitir que os dados pudessem ser convenientemente processados pelo protótipo. A característica principal desses atributos é o fato dos mesmos terem

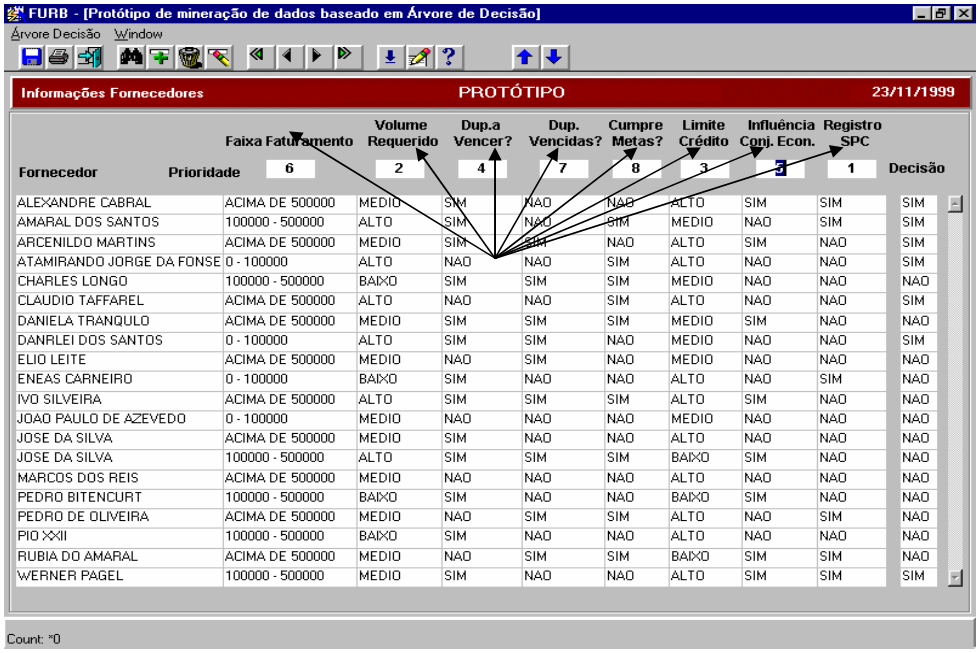
um domínio fixo, como por exemplo o atributo “FAIXA FATURAMENTO” que tem domínio {0-100000,101000 a 500000,ACIMA DE 500000}.

4.5.2 DOMÍNIO DA APLICAÇÃO

Esta etapa do KDD é muito importante na aplicação do *Data Mining*, pois é onde o usuário deve analisar qual o nível de prioridade que ele deseja atribuir aos atributos que farão parte do processo de classificação e apuração dos resultados.

Nesta etapa o usuário deve apenas informar para cada atributo não decisivo, valores de 1 a 8, de acordo com a importância que ele dá aos mesmos, levando-se em conta o seu conhecimento das regras de negócio. É baseado nesta prioridade que o protótipo priorizará os atributos em relação ao cálculo da entropia. Isso deve ser feito simplesmente informando o valor da prioridade nos campos localizados acima de cada atributo conforme demonstrado na figura 26.

Figura 26 – Informando a prioridade



4.5.2.1 PRÉ-PROCESSAMENTO E LIMPEZA

A etapa de pré-processamento visa adequar as informações aos algoritmos de *Data Mining*. Os algoritmos de *Data Mining* na maior parte das vezes requerem os dados formatados para o seu processamento.

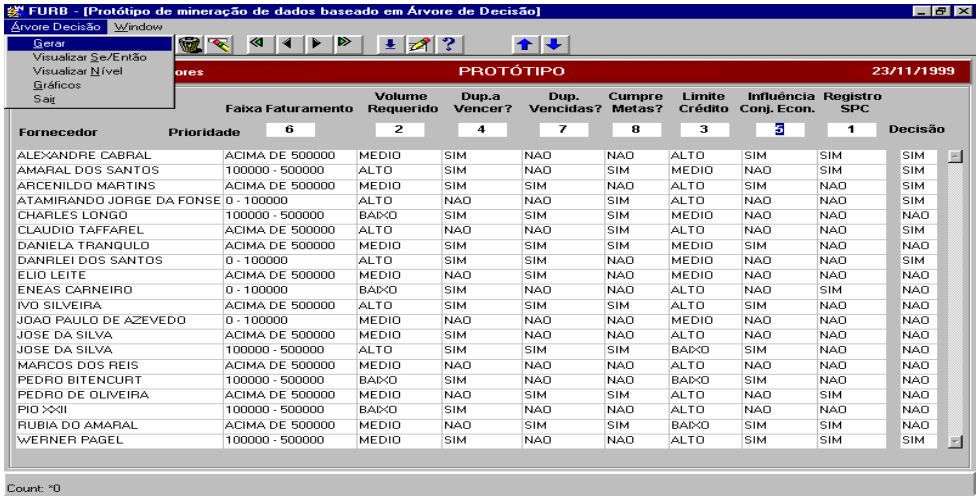
Este protótipo não utilizou a tarefa de pré-processamento e limpeza pois a base da dados foi construída de forma que os dados já atendessem as necessidades do processo de *Data Mining*.

4.5.2.2 *DATA MINING*

O *Data Mining* é a etapa onde o algoritmo, com base nos atributos definidos, deverá descobrir intuitivamente regras que forneçam diagnósticos, isto é, descobertas automáticas de conhecimento que é um dos seus objetivos principais.

Para executar o processamento da árvore de decisão o usuário deve clicar na opção "Árvore de Decisão" no menu principal e depois clicar em "Gerar" conforme demonstrado na figura 27.

Figura 27 – Execução do processo de *Data Mining*



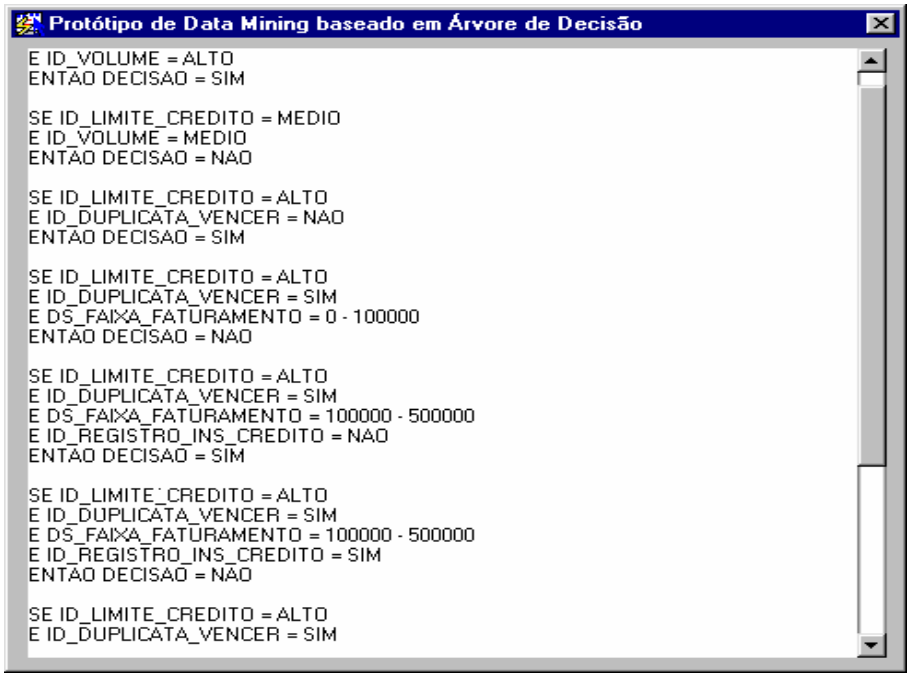
4.5.2.3 INTERPRETAÇÃO DO CONHECIMENTO

Após a etapa de *Data Mining* estar concluída e a classificação de dados estar estabelecida existe o processo de interpretação do conhecimento encontrado. Neste protótipo existem duas formas de visualização desse conhecimento, sendo elas, estrutura se/então e por nível.

4.5.2.3.1 VISUALIZAÇÃO SE/ENTÃO

Esta é a forma mais clássica de visualização do resultado obtido pelo *Data Mining*. Ela consiste em simples estruturas condicionais se/então obtidas a partir do processamento realizado pelo algoritmo. Para se chegar a essa representação deve-se navegar por todos os nós da árvore obtendo para cada ramificação o valor da decisão correspondente [QUI93], conforme representação na figura 28.

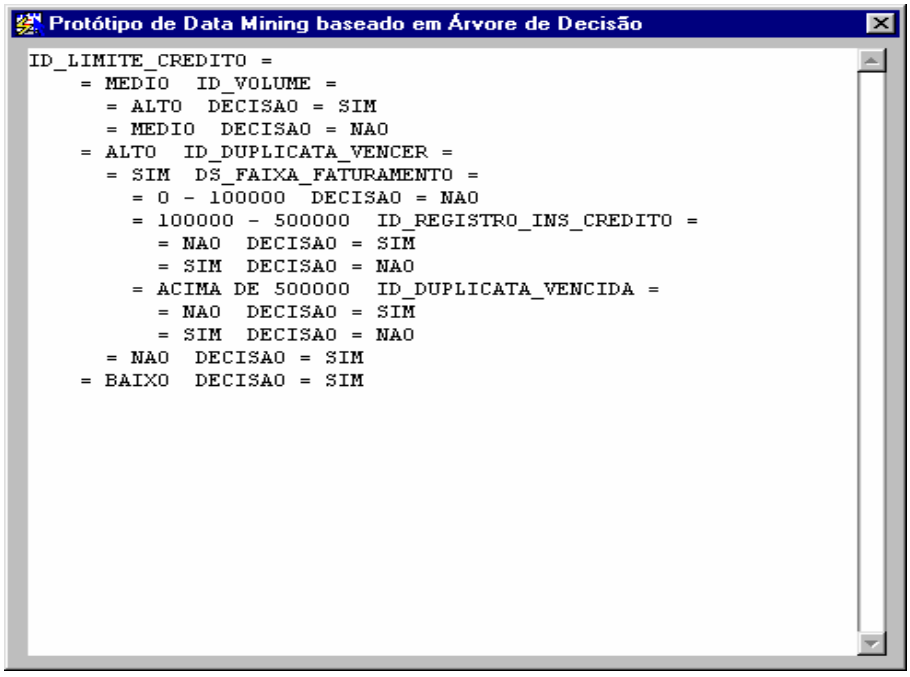
Figura 28 – Visualização Se/Então



4.5.2.3.2 VISUALIZAÇÃO POR NÍVEL

Esta é a forma utilizada para visualizar os diversos níveis da árvore gerados pelo algoritmo. Esta forma demonstra claramente a hierarquia existente entre os elementos pertencentes aos diversos nós da árvore conforme representação na figura 29.

Figura 29– Visualização por nível



5 CONCLUSÕES E SUGESTÕES

Este capítulo apresenta as conclusões, limitações e sugestões referentes ao trabalho desenvolvido.

5.1 CONCLUSÃO

Partindo da necessidade de se extrair conhecimento através de interpretação de dados foi estudada a tecnologia de *Data Mining*. Foram estudados suas funções, suas técnicas, e as etapas que levem a descoberta do conhecimento que é o objetivo principal do *Data Mining*.

Neste trabalho foi enfatizado o uso de *Data Mining* com Árvores de Decisão empregado em um Sistema de Informação Executiva para modelos de classificação e segmentação de dados. Tendo isso como base, verificou-se que a utilização do *Data Mining*, juntamente com as etapas de KDD se mostrou bastante eficiente.

Foram realizados testes com o modelo de dados construído para a execução do processo de *Data Mining* onde o protótipo mostrou ser eficiente para a definição de modelos de classificação e segmentação de dados.

Durante a construção do modelo, foram utilizadas algumas etapas/fases da metodologia de análise estruturada, as quais auxiliaram em muito no desenvolvimento do projeto. As ferramentas ORACLE FORMS e ORACLE GRAPHICS ajudaram muito pela facilidade de aprendizado que ela proporciona e sobre o fácil acesso aos dados que as mesmas proporcionam, já que foram construídas especialmente para o banco de dados ORACLE .

Considera-se que o objetivo principal do trabalho, o desenvolvimento de um SIE para efetuar classificações e segmentações de dados utilizando *Data Mining* foi atingido.

5.2 LIMITAÇÕES

O protótipo construído apresenta as seguintes limitações:

a) a fonte de dados que o protótipo utiliza para processamento é fixa, desta forma não permitindo ao usuário mudar a fonte de dados ou alterar o conjunto de atributos a serem processados;
b) os atributos envolvidos no processo de classificação possuem domínio fixo.

5.3 SUGESTÕES

Sugere-se o estudo do *Data Mining* aplicando outras tarefas e técnicas para a tomada de decisões, como o uso de outras técnicas.

Em relação a fonte de dados poderia ser implementado uma opção onde o usuário especificasse uma fonte de dados variável.

Um outro item importante na questão da origem dos dados que poderia ser implementado, seria um acesso a dados que fosse além do ORACLE. Sugere-se implementar acesso também à outros bancos como Microsoft SQL Server, Sybase Server, Informix, etc.

REFERÊNCIAS BIBLIOGRÁFICAS

[ALT92] ALTER, Steven. **Information systems: a management perspective.** USA : Addison-Wesley Publishing, 1992.

[AVI98] ÁVILA, Bráulio Coelho. ***Data Mining*.** VI Escola Regional de Informática da SBC – Regional Sul. Blumenau, 1998. p. 87-106.

[BER97] BERRY, Michael J. A.; LINOFF, Gordon. ***Data Mining* techniques.** USA : Wiley Computer Publishing, 1997.

[BRY95] BRYAN, Richard. **Oracle PL/SQL programming**. Dallas : Books Inc,1995.

[CER95] CERÍCOLA, Vicent Oswald. ORACLE - **Banco de dados relacional Distribuído**. Ferramentas para desenvolvimento. São Paulo : Makron – McGraw-Hill, 1995.

[DAL98] DALFOVO, Oscar. **Desenho de um modelo de sistemas de informação.** Blumenau, 1998. Dissertação (mestrado em Administração de Negócios) Centro de Ciências Sociais e Aplicadas, FURB.

[DAY95] DAY, Simon. **ORACLE education – D2D data desing using designer/2000.** Makron – McGraw-Hill, 1995.

[EAR88] EARL, M. J., **Exploiting IT for Strategic Advantage - A framework of frameworks.** Oxford Institute of Information Management, 1988.

[FAY96] FAYYAD, Usama M... [et all]. **Advances in knowledge discovery and *Data Mining*.** Mento Park : AAAI : MIT, 1996.

[FIG98] FIGUEIRA, Rafael Medeiros Andrade. **Miner: um software de inferência de dependências funcionais.** Rio de Janeiro, 1998. Trabalho de Conclusão de Curso – Instituto de Matemática, Universidade Federal do Rio de Janeiro.

[GRO97] GROTH, Robert. ***Data Mining*: a hands-on approach for business professionals conceitos e soluções**. New Jersey : Prentice Hall, 1997.

[HAR98] HARRISON, Thomas H. **Intranet data warehouse.** São Paulo : Berkeley Brasil, 1998.

[INM97] INMON, William H. **Como construir o data warehouse.** Rio de Janeiro : Campus, 1997.

[KRA99] KRAMER, Ricardo. **Sistema de Apoio à Decisão para previsões genéricas utilizando técnicas de *Data Mining*.** Blumenau, 1999. Trabalho de Conclusão de Curso – Centro de Ciências Exatas e Naturais, Universidade Regional de Blumenau.

[LOE96] LOESCH, Claudio; SARI, Solange Teresinha. **Redes neurais artificiais : fundamentos e modelos.** Blumenau: FURB, 1996.

[MAC96] MACHADO, Carlos. Como dar o tiro certo na hora de decidir. **Exame Informática.** São Paulo, v. 11, n. 120, p. 27-29, mar. 1996.

[OLI98] OLIVEIRA, Adelize Generini de. **Data warehouse: conceitos e soluções.** Florianópolis : Advanced, 1998.

[PRA94] PRATES, Maurício. **Conceituação de Sistemas de Informação do ponto de vista do Gerenciamento**. Revista do Instituto de Informática, PUC-CAMP, Março/Setembro, 1994.

[QUI93] QUINLAN, Ross and KAUFMANN, Morgan. **Programs for Machine Learning.** USA : McGraw-Hill, 1993.

[WES98] WESTPHAL, Christopher and BLAXTON, Teresa. ***Data Mining* solutions.** Canadá : John Wiley & Sons Inc, 1998.

[YOU90] YOURDON, Edward. **Análise Estruturada Moderna.** Rio de Janeiro : Campus, 1997.