

UNIVERSIDADE REGIONAL DE BLUMENAU
CENTRO DE CIÊNCIAS EXATAS E NATURAIS
CURSO DE CIÊNCIAS DA COMPUTAÇÃO
(Bacharelado)

**A TÉCNICA DE *KNOWLEDGE DISCOVERY IN DATABASES*
(KDD) APLICADA NAS OCORRÊNCIAS ATENDIDAS PELA
POLÍCIA MILITAR**

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

EMERSON TENFEN

BLUMENAU, JUNHO/2003

2003/1-21

**A TÉCNICA DE *KNOWLEDGE DISCOVERY IN DATABASES*
(KDD) APLICADA NAS OCORRÊNCIAS ATENDIDAS PELA
POLÍCIA MILITAR**

EMERSON TENFEN

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. Jomi Fred Hubner — Orientador na FURB

Prof. José Roque Voltolini da Silva — Coordenador do TCC

BANCA EXAMINADORA

Prof. Jomi Fred Hubner – Orientador na FURB

Prof. Maurício Capobianco Lopes

Prof. Dr. Oscar Dalfovo

AGRADECIMENTOS

Ao meu orientador Jomi Fred Hubner pelo apoio e motivação nas horas em que achei que não conseguiria.

Ao meu pai Alfredo e minha mãe Albertina, pelo empenho e dedicação com que conduziram minha educação básica, fazendo com que me tornasse o que sou hoje, mesmo que algumas vezes sem condições para tal.

Às minhas irmãs Ivone, Ilze, Clarice, Simone e ao meu irmão Aldo pela cumplicidade que demonstram e pela alegria que me trazem até hoje.

À minha esposa Glaucia pela sensibilidade e compreensão com que me conduziu durante a realização deste trabalho, e por todos os momentos felizes vividos até hoje e que estão por vir.

A todas as pessoas que, de alguma forma contribuíram para a realização deste trabalho, entre eles: amigos, professores, colaboradores, monitores, colegas de trabalho e outros, ainda que não estejam aqui relacionados.

RESUMO

Este trabalho visa gerar um modelo de classificação de dados utilizando técnicas de mineração de dados, mais especificamente árvores de decisão. Para a elaboração do aplicativo, foram analisados os processos de descobrimento de conhecimento em banco de dados, bem como técnicas de mineração de dados e montado um banco de dados fornecido pela Polícia Militar. Estas informações serão a base que será aplicada à classificação e ao algoritmo de árvore de decisão. Foram realizados testes e foi possível desenvolver modelos de classificação, nos quais colocou-se em prática a técnica de árvores de decisão.

ABSTRACT

This research has like main objective to generate a classification model using data mining techniques, more specifically decision trees. For the elaboration of the prototype, the process standard recognition of informations in data base were analyzed, as well as techniques data mining and building a data base supplied by Military Police. These information will be the base that will be applied to the classification and the decision tree algorithm. Tests were accomplished and it was possible to develop classification models, us which it was placed in practices the use of decision trees.

LISTA DE FIGURAS

FIGURA 1 - FICHA DE OCORRÊNCIA DA POLÍCIA MILITAR.....	13
FIGURA 2 - PROCESSO DE KDD.....	16
FIGURA 3 - ÁRVORE DE DECISÃO GERADA A PARTIR DOS DADOS DA TABELA 3	22
FIGURA 4 - CLIENTES DIVIDIDOS EM SEGMENTOS QUE POSSUEM SEMELHANÇAS	24
FIGURA 5 – GRÁFICO DE DEMONSTRAÇÃO DA ENTROPIA	26
FIGURA 6 - NODO GERADO PELO PONTO DE REFÊNCIA = 70.5.....	33
FIGURA 7 - DIAGRAMA DE CONTEXTO	35
FIGURA 8 - DFD NÍVEL 0.....	36
FIGURA 9 - MODELO ENTIDADE RELACIONAMENTO.....	37
FIGURA 10 - TELA DE APRESENTAÇÃO	39
FIGURA 11 - TELA PRINCIPAL DO APLICATIVO.....	40
FIGURA 12 - TELA DE ESCOLHA DOS ATRIBUTOS.....	41
FIGURA 13 – TELA PRINCIPAL DO PROGRAMA PARA CRIAR BASE DE DADOS DO APLICATIVO.....	42
FIGURA 14 - TELA DE GERAÇÃO DE NOVA ÁRVORE	44
FIGURA 15 - TELA DE VISUALIZAÇÃO SE/ENTÃO.....	45

LISTA DE TABELAS

TABELA 1 - PROFISSIONAIS ENVOLVIDOS NO PROCESSO DE KDD	16
TABELA 2 - CONJUNTO DE VARIÁVEIS SUBMETIDO AO PROCESSO DE REDUÇÃO	18
TABELA 3 - CONJUNTO DE EXEMPLOS DE ESCOLHA DE LINGUAGENS DE PROGRAMAÇÃO	21
TABELA 4 - ENTRADA DE DADOS PARA A DESCOBERTA DE REGRAS DE ASSOCIAÇÃO	23
TABELA 5 - CONJUNTO DE TREINO PARA CÁLCULO DA ENTROPIA E GANHO DE INFORMAÇÃO	28
TABELA 6 - PROBLEMA DE JOGAR TÊNIS COM ATRIBUTO NUMÉRICO	31
TABELA 7 - DESCRIÇÃO DETALHADA DO MODELO DE DADOS	38
TABELA 8 - DADOS ORIGINAIS PARA O ATRIBUTO "NATUREZA"	42
TABELA 9 - ETAPA DE PRÉ-PROCESSAMENTO REALIZADA NO ATRIBUTO "TURNO"	43

LISTA DE QUADROS

QUADRO 1 - PASSOS PARA A CONSTRUÇÃO DE UMA ÁRVORE DE DECISÃO	25
QUADRO 2 - FÓRMULA DO CÁLCULO DA ENTROPIA	27
QUADRO 3 - CÁLCULO DA ENTROPIA	27
QUADRO 4 - FÓRMULA DO CÁLCULO DO GANHO DE INFORMAÇÃO.....	28
QUADRO 5 - CÁLCULO DO GANHO DE INFORMAÇÃO PARA OS ATRIBUTOS TEMPO E VENTO....	29
QUADRO 6 - PSEUDO-CÓDIGO DO ALGORITMO C4.5	30
QUADRO 7 - PARTIÇÕES DO ATRIBUTO TEMPERATURA.....	31
QUADRO 8 - CÁLCULO DA ENTROPIA PARA VALORES MENORES E VALORES MAIORES QUE O PONTO DE REFERÊNCIA	32
QUADRO 9 - CÁLCULO DA ENTROPIA PARA O ATRIBUTO TEMPERATURA.....	32
QUADRO 10 - CÁLCULO DO GANHO DE INFORMAÇÃO PARA O PONTO DE REFERÊNCIA	32

LISTA DE SIGLAS E ABREVIATURAS

KDD *Knowledge Discovery in Databases*

MD Mineração de Dados

COPOM Centro de Operações da Polícia Militar

SUMÁRIO

AGRADECIMENTOS	3
RESUMO	4
ABSTRACT	5
LISTA DE FIGURAS	6
LISTA DE TABELAS	7
LISTA DE QUADROS	8
LISTA DE SIGLAS E ABREVIATURAS	9
1 INTRODUÇÃO.....	12
1.1 OBJETIVOS DO TRABALHO.....	14
1.2 ESTRUTURA DO TRABALHO.....	14
2 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS (KDD)	15
2.1 PROCESSOS DE KDD	15
2.1.1 DESENVOLVIMENTO E ENTENDIMENTO DO DOMÍNIO DA APLICAÇÃO ...	16
2.1.2 CRIAÇÃO DE UM GRUPO DE DADOS ALVO	17
2.1.3 LIMPEZA DOS DADOS	17
2.1.4 REDUÇÃO E PROJEÇÃO DOS DADOS.....	18
2.1.5 MINERAÇÃO	18
2.1.6 INTERPRETAÇÃO.....	19
2.1.7 CONSOLIDAÇÃO DO CONHECIMENTO DESCOBERTO	19
3 MINERAÇÃO DE DADOS (MD)	20
3.1 TAREFAS DE MINERAÇÃO DE DADOS.....	20
3.1.1 CLASSIFICAÇÃO.....	21
3.1.2 ESTIMATIVA.....	22

3.1.3 ASSOCIAÇÃO	23
3.1.4 SEGMENTAÇÃO.....	24
4 ÁRVORES DE DECISÃO	25
4.1 ENTROPIA.....	26
4.2 GANHO DE INFORMAÇÃO	27
4.3 ALGORITMO C4.5	29
4.3.1 ATRIBUTOS NUMÉRICOS	31
5 DESENVOLVIMENTO DO APLICATIVO.....	34
5.1 ESPECIFICAÇÃO.....	34
5.2 IMPLEMENTAÇÃO	38
5.2.1 DOMÍNIO DA APLICAÇÃO	39
5.2.2 CRIAÇÃO DE UM GRUPO DE DADOS ALVO.....	41
5.2.3 LIMPEZA DOS DADOS	42
5.2.4 REDUÇÃO E PROJEÇÃO DOS DADOS.....	43
5.2.5 MINERAÇÃO DE DADOS.....	43
5.2.6 INTERPRETAÇÃO DO CONHECIMENTO.....	44
6 CONCLUSÕES	46
6.1 LIMITAÇÕES.....	47
6.2 SUGESTÕES	47
REFERÊNCIAS	48

1 INTRODUÇÃO

A violência, que até algum tempo atrás, deixava as pessoas surpresas, por tratar-se de um fato isolado e raro, hoje tornou-se costumeiro e banal entre todos. No jornal A Notícia de janeiro de 2003, a seguinte notícia, apesar de grave, faz parte do dia a dia de qualquer pessoa que lê jornais, assiste ao noticiário da TV ou ouve rádio:

“O empresário Vilmar Braz, de 30 anos, foi baleado no pé por dois assaltantes que entraram na sua panificadora, no bairro Morro do Meio, em Joinville, na noite de sábado. Braz e sua família acabaram reféns dos bandidos por quase uma hora.” (BRAGA, 2003, p. A9).

Segundo Maldonado (2002, p. 1), define que

“as raízes e as expressões da violência são múltiplas e a escalada da violência nas últimas décadas, em grande número de países, tem atingido proporções consideradas epidêmicas. A questão do controle e da prevenção da violência passou a ser vista como um problema de saúde pública, demandando intervenções em vários níveis.”

Uma das instituições reconhecidas no combate à violência é a Polícia Militar, que atua de forma preventiva e repressiva. A prática de um delito pode ser repensada pelos delinquentes pelo simples fato de haverem dois policiais fardados próximo ou circulando pelo local. Quando a prevenção falha e acontece algum delito, a Polícia Militar é chamada para atender a ocorrência.

Todas as ocorrências atendidas pela Polícia Militar são registradas e depois de resolvidas são armazenados dados referentes à ocorrência em uma base de dados. A figura 1 apresenta uma ficha de ocorrência que é preenchida pela guarnição de serviço responsável pela ocorrência.

FIGURA 1 - FICHA DE OCORRÊNCIA DA POLÍCIA MILITAR

ESTADO DE SANTA CATARINA POLÍCIA MILITAR 10º BATALHÃO - BLUMENAU					Nº 19215
N.º OCORRÊNCIA	NATUREZA	DATA	HORA	VIATURA	
1 - GUARNIÇÃO DE SERVIÇO					
2 - CONDUZIDOS					
NOME:		Nº	SEXO	BAIRRO	IDADE
END:					
NOME:		Nº	SEXO	BAIRRO	IDADE
END:					
NOME:		Nº	SEXO	BAIRRO	IDADE
END:					
3 - VÍTIMAS / TESTEMUNHAS					
NOME:		Nº	SEXO	BAIRRO	IDADE
END:					
NOME:		Nº	SEXO	BAIRRO	IDADE
END:					
NOME:		Nº	SEXO	BAIRRO	IDADE
END:					
4 - DADOS DA OCORRÊNCIA ATENDIDA					
END:	R.				Nº
BAIRRO:			PTO REFERÊNCIA:		
5 - DESTINO DOS CONDUZIDOS					
DP	HOSPITAL	CIP	RESIDÊNCIA	GMT	CASA KOLPING
					OUTROS (CITAR LOCAL)
6 - CONDIÇÕES FÍSICAS DOS CONDUZIDOS					
CONFORME ACIC Nº					
7 - OBJETOS DE VALOR E MATERIAIS RECOLHIDOS DOS CONDUZIDOS					
CARTEIRA COM DOCUMENTOS:					
CI	CNH	T.E.	C.T.	C.N.	CPF
					CARTÃO DE CRÉDITO (CITAR BANCO E QUANTIDADE)
<small>CI - CARTeira DE IDENTIDADE; CNH - CARTeira NACIONAL DE HABILITACÃO; T.E. - TITULO DE ELEITOR; C.T. - CARTeira DE TITULO DE CIDADÃO; C.N. - CARTeira DE NACIMENTO; CPF - CARTeira DE PESSOA FISICA</small>					
DINHEIRO (DISCRIMINAR VALOR)			TALÃO DE CHEQUES (NOME DO BANCO E Nº DE FOLHAS)		
RELÓGIO DE PULSO (MARCA E SITUAÇÃO)			CORRENTES/PULSEIRAS (TIPO E QUANTIDADE)		
OUTROS OBJETOS: (CITAR)					
ARMAS BRANCAS		ARMAS DE FOGO		TÓXICOS (TIPO E QUANTIDADE)	
8 - HISTÓRICO DA OCORRÊNCIA ATENDIDA (DE FORMA CLARA, PRECISA E LEGÍVEL)					
→					
→					
→					
→					
→					
→					
9 - ASSINATURAS					
CMT DA GUARNIÇÃO PM			AUTORIDADE OU RESPONSÁVEL		
MATRÍCULA:			MATRÍCULA:		
NOME LEGÍVEL:			NOME LEGÍVEL:		
<small>1ª Via - 10º BPM</small>			<small>2ª Via - Recipiente</small>		
<small>ASSINATURA DO CONDUZIDO</small>					

Fonte: 10º Batalhão de Polícia Militar (2003).

Na base de dados do Centro de Operações da Polícia Militar (COPOM), ficam armazenados dados como o endereço em que aconteceu a ocorrência, data, horário e natureza da ocorrência.

Tendo em vista os aspectos acima citados, propõe-se desenvolver um aplicativo que, a partir dos dados lidos em ocorrências atendidas pela Polícia Militar, utilize-se da técnica de *Knowledge Discovery in Databases* (KDD), reconhecendo padrões. O resultado obtido pelo aplicativo será utilizado pela Polícia Militar para auxiliar no controle e distribuição de viaturas pela área de atuação do 10º Batalhão de Polícia Militar.

O tema *Data Mining* já foi abordado em outros TCC's como Compolt (1999) e Nardelli (2000), porém a diferença deste trabalho é o domínio da aplicação e o algoritmo a ser usado no processo de *Data Mining*. Este trabalho propõe-se a implementar o algoritmo C4.5 (RUGGIERI, 2000) que é uma extensão do algoritmo ID3, implementado pelos TCC's anteriores.

1.1 OBJETIVOS DO TRABALHO

O objetivo deste trabalho é desenvolver um aplicativo, utilizando técnicas de KDD para o reconhecimento de padrões a partir de dados de ocorrências atendidas pela Polícia Militar.

Os objetivos específicos do trabalho são:

- a) seguir de forma geral os processos de KDD e detalhar o processo de *Data Mining*;
- b) analisar os dados armazenados na base de dados do COPOM;
- c) gerar uma árvore de decisão a partir dos dados analisados;
- d) implementar neste aplicativo o algoritmo C4.5.

1.2 ESTRUTURA DO TRABALHO

O trabalho está organizado conforme descrito a seguir.

O capítulo dois apresenta uma visão geral de KDD e seus processos.

O capítulo três enfatiza os conceitos, detalhando as tarefas de *Data Mining*.

O capítulo quatro é dedicado a árvores de decisão e especificamente ao algoritmo C4.5, o qual será detalhado neste capítulo.

O capítulo cinco apresenta a especificação, implementação e o funcionamento do aplicativo do ambiente de aprendizado de máquina.

Como finalização, o capítulo seis é utilizado para a apresentação das conclusões gerais, originadas durante o estudo e confecção do trabalho e apresentação de algumas possíveis extensões para futuros trabalhos correlatos.

2 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS (KDD)

Com o surgimento dos bancos de dados e sistemas gerenciadores de bancos de dados, as empresas e instituições públicas viram a possibilidade de armazenar de uma maneira mais prática todos os dados relativos às ações e transações efetuadas por elas. Esta facilidade em armazenar dados fez com que as empresas, em um curto espaço de tempo, obtivessem uma grande quantidade de dados gravados em seus computadores. Assim, diretores e pesquisadores começaram a se perguntar o que farão com a grande quantidade de dados armazenada.

Conforme Quoniam (2002, p. 20),

“a capacidade de armazenamento em banco de dados, assim como sua utilização, vem crescendo na mesma proporção dos avanços em novas tecnologias de informação e comunicação. A atividade de extrair informações relevantes, por conseguinte, está se tornando bastante complexa. Este processo de ‘garimpagem’ é chamado de *Knowledge Discovery in Databases* – Descoberta de Conhecimento em Bases de Dados (KDD).”

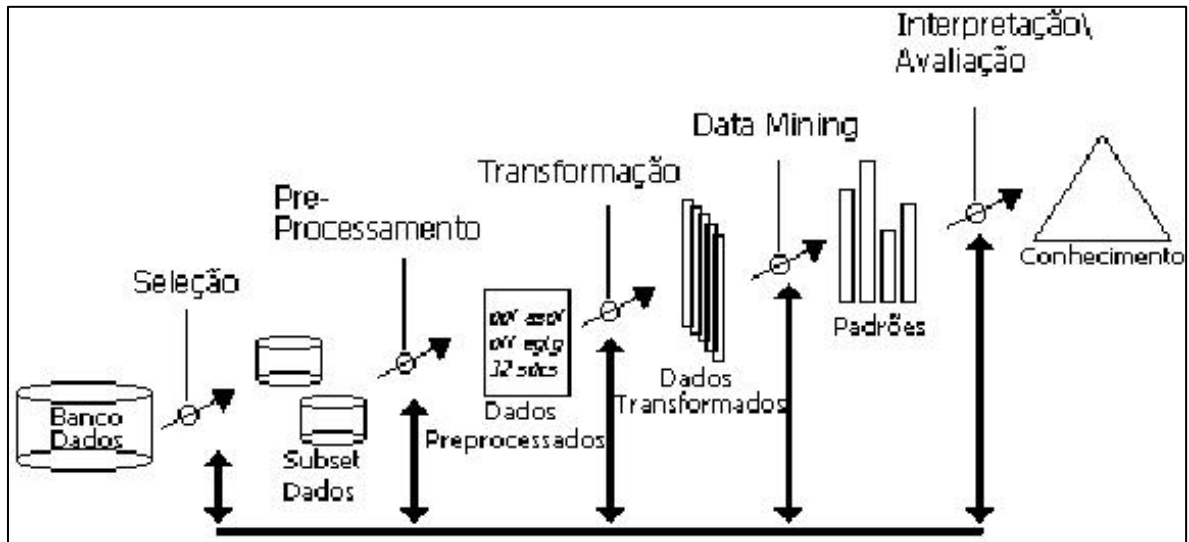
Conforme Fayyad (1996, p. 6), a definição de KDD é “o processo não trivial de identificação de padrões em dados que sejam válidos, novos, potencialmente úteis e compreensíveis”.

Iniciada nos anos 70, a metodologia KDD englobava recursos de estatística, reconhecimento de padrões, máquinas de aprendizado e métodos de visualização. Esta técnica obtinha formas de regras para dar suporte de análise de dados e descobrir princípios que estavam embutidos nos dados.

2.1 PROCESSOS DE KDD

Segundo Liebstein (2001), o processo de KDD destina-se encontrar e interpretar padrões nos dados através da repetição de algoritmos e da análise de resultados. A figura 2 mostra o processo de KDD.

FIGURA 2 - PROCESSO DE KDD



Fonte: Fayyad (1996).

Cada fase deste processo pode possuir uma interseção com os demais. Deste modo, os resultados produzidos numa fase podem ser utilizados para melhorar os resultados das próximas fases, o que torna o processo de KDD repetitivo. Outra característica dos processos de KDD é a sua interatividade, pois a cada fase pode haver uma interação entre os profissionais envolvidos no processo (FAYYAD, 1996). A tabela 1 mostra os profissionais envolvidos nos processos de KDD:

TABELA 1 - PROFISSIONAIS ENVOLVIDOS NO PROCESSO DE KDD

PROFISSIONAIS	FUNÇÕES
Nível decisório	São diretores, gerentes, pessoas de nível gerencial e executivo. Tomam decisões baseados no resultado da pesquisa.
Analista do conhecimento	Domina as técnicas e ferramentas de <i>data mining</i> . Deve ser capaz de conversar com os usuários de nível decisório.
Pessoal de tecnologia da informação	Responsáveis pela programação, manutenção, limpeza de bases de dados (BD) e administração das BD.

Fonte: Adaptado de Liebsstein (2001).

2.1.1 DESENVOLVIMENTO E ENTENDIMENTO DO DOMÍNIO DA APLICAÇÃO

A definição do problema é fundamental para o processo de KDD. Essa definição requer um entendimento perfeito do problema existente e um objetivo bem especificado, ou

seja, aquilo que se deseja obter ou extrair deve estar bem explícito no projeto. Para isso, pode ocorrer, durante este passo do processo, uma interação entre o desenvolvedor e um especialista na área de conhecimento a que se irá aplicar a técnica de KDD (TWO CROWS CORPORATION, 1998).

2.1.2 CRIAÇÃO DE UM GRUPO DE DADOS ALVO

O primeiro passo consiste em selecionar um conjunto de dados, um subconjunto de variáveis ou uma amostragem dos dados sobre os quais o processo será executado. Dependendo do objetivo da tarefa, uma informação como a idade do cliente pode ser mais valiosa do que o nome da rua onde ele mora e vice-versa. Por isso, para que se tenha qualidade nos resultados a serem obtidos pelo processo de KDD, é necessário investir nesta etapa (RODRIGUES, 2000).

2.1.3 LIMPEZA DOS DADOS

Após definir um grupo de dados alvo, o segundo passo é a realização de operações básicas, como coletar informações necessárias para modelar ou tratar os ruídos, decidir estratégias para tratar campos sem valor, remover ruídos e valores desconhecidos. Também são tratadas questões ligadas ao sistema gerenciador de bancos de dados (SGBD), tais como tipos de dados e esquemas. A execução desta etapa corrige a base de dados, eliminando consultas desnecessárias que seriam executadas pelo algoritmo de *data mining*. Um exemplo simples de limpeza dos dados seria atribuir valores mínimos e valores máximos para um determinado atributo, {0..10}. Qualquer valor fora destes limites seria desconsiderado.(RODRIGUES, 2000)

2.1.4 REDUÇÃO E PROJEÇÃO DOS DADOS

No terceiro passo do processo os dados são “ajustados”. Aham-se características úteis para representar os dados, dependendo do objetivo da tarefa. Usa-se redução de dimensionalidade ou métodos de transformação para reduzir o número efetivo de variáveis ou achar representações variadas para os dados. A tabela 2 mostra um exemplo de um dado que foi submetido à redução (BARAZETTI, 2001).

TABELA 2 - CONJUNTO DE VARIÁVEIS SUBMETIDO AO PROCESSO DE REDUÇÃO

HORÁRIO	TURNOS
01:30	MADRUGADA
03:15	
08:45	MANHÃ
10:20	
13:50	TARDE
17:15	
19:30	NOITE
22:20	

Fonte: Adaptado de Rodrigues (2000).

2.1.5 MINERAÇÃO

O quarto passo é a etapa de descoberta de conhecimento, onde são escolhidas e processadas as tarefas e os algoritmos de aprendizagem de máquina e de reconhecimento de padrões. Esta etapa do processo de KDD divide-se ainda em três fases detalhadas a seguir:

- a) definição da tarefa de *data mining*: definir o objetivo do modelo a ser gerado por uma técnica de *data mining*, considerando o objetivo geral de KDD. As tarefas podem ser de classificação, estimativa, associação ou segmentação. Estas tarefas de *data mining* serão detalhadas no capítulo 3.
- b) definição do algoritmo de *data mining*: nesta etapa são selecionados os métodos a serem utilizados para pesquisa de padrões, bem como os modelos e parâmetros que podem ser apropriados. A escolha do algoritmo está diretamente ligada à definição

da tarefa de *data mining*. Cada tarefa de *data mining* pode consistir na aplicação de vários algoritmos.

- c) mineração: é a etapa de aplicação do algoritmo de *data mining* sobre os dados selecionados, buscando os padrões de interesse da aplicação. Nesta etapa do processo, podem ocorrer constantes ajustes dos parâmetros para refinamento do modelo.

2.1.6 INTERPRETAÇÃO

No quinto passo, depois que os dados já foram submetidos ao algoritmo de mineração, gerando padrões, deve-se interpretar os padrões descobertos.

Esta interpretação é feita através de visualizações, removendo padrões redundantes ou irrelevantes e traduzindo os padrões importantes em condições que podem ser facilmente compreendidas pelos usuários finais. Esta etapa pode ser, possivelmente, um ponto de retorno para quaisquer dos passos anteriores, pois os resultados da etapa de mineração não têm efeito algum até que sejam validados (BARAZETTI, 2001).

2.1.7 CONSOLIDAÇÃO DO CONHECIMENTO DESCOBERTO

O sexto passo consiste na incorporação do conhecimento descoberto na etapa anterior, ou simplesmente na documentação deste conhecimento e envio às partes interessadas. O conhecimento servirá para auxiliar na tomada de decisões e planejamentos futuros (BARAZETTI, 2001).

3 MINERAÇÃO DE DADOS (MD)

Todas as técnicas que permitem extrair conhecimento de uma massa de dados que, de outra maneira, permaneceria escondido nas grandes bases de dados, chama-se *data mining* (QUONIAM, 2002).

Considerada o núcleo do processo de KDD, a DM, ou mineração de dados, é a etapa que consiste na aplicação dos algoritmos em grandes volumes de dados, a fim de descobrir informações úteis que normalmente não estão visíveis.

Facilmente o processo de KDD é confundido com DM, porém enquanto o KDD compreende todo o processo de descoberta de conhecimento, a DM refere-se à aplicação dos algoritmos de mineração de dados, sem os passos adicionais de KDD e a análise dos resultados.

Conforme Groth (1997) os negócios profissionais procuram modelos de DM que possam suprir suas necessidades. Primeiramente, um profissional de negócios requer que o modelo seja legível, ou de fácil compreensão para ele. Em segundo lugar, o modelo deve ter um bom desempenho, ou seja, atender a necessidade de tempo do profissional. E o mais importante para o profissional é que o modelo deve ser da maior exatidão possível, trazendo-lhe confiança para tomar as decisões a partir do conhecimento descoberto.

3.1 TAREFAS DE MINERAÇÃO DE DADOS

A técnica de DM 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, podendo exercer tarefas de classificação, estimativa, associação e segmentação, entre outras (HARMON, 1988)

Em geral, as tarefas de *data mining* podem ser classificadas em duas categorias:

- a) descritivas: descrevem o conjunto de dados de uma maneira concisa e resumida e descobre propriedades gerais interessantes nestes dados;

- b) preditivas: constroem um modelo ou um conjunto de modelos, realiza inferências sobre o conjunto de dados disponíveis e tenta prever o comportamento de novos conjuntos de dados.

3.1.1 CLASSIFICAÇÃO

Classificação é uma tarefa de DM que consiste em levantar as características de um objeto e associar essas características a classes pré-determinadas. Os algoritmos que constituem a tarefa de classificação se utilizam de árvores de decisão ou redes neurais, começando por um treinamento a partir de transações-exemplo (BARAZETTI, 2001).

Os exemplos levantados permitem determinar um conjunto de parâmetros, reunidos em um modelo que no decorrer do processo serão utilizados para discriminar o restante dos dados. A tabela 3 mostra um conjunto de exemplos de escolha de linguagem de programação para desenvolvimento de aplicações levando em consideração certas características do sistema.

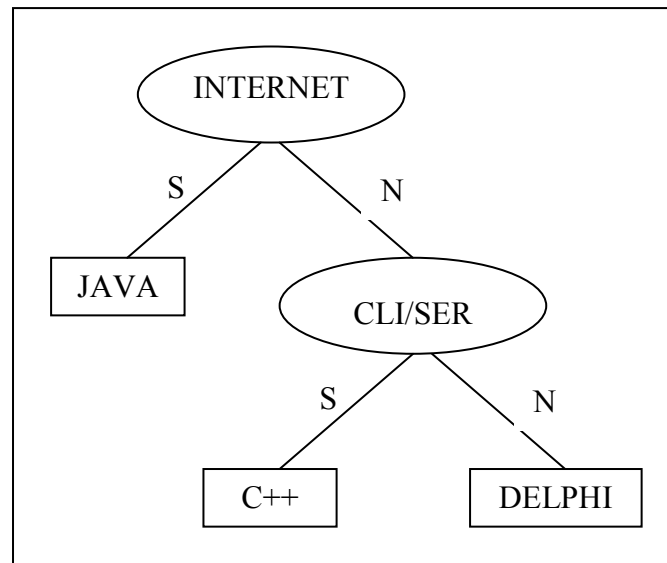
TABELA 3 - CONJUNTO DE EXEMPLOS DE ESCOLHA DE LINGUAGENS DE PROGRAMAÇÃO

Aplicação	Comercial	Distribuída	Cli/Ser	Internet	Matemática	Tempo Real	Linguagem
1	S	N	N	N	S	N	Delphi
2	S	S	S	N	S	S	C++
3	S	S	N	S	S	N	Java
4	N	N	N	S	N	S	Java
5	N	N	N	N	S	S	C++
6	N	S	N	S	N	N	Java

Fonte: Hubner (2002).

Submetendo estes exemplos a um algoritmo classificador, por exemplo, um algoritmo de geração de árvore de decisão, seria gerada a árvore mostrada na figura 3.

FIGURA 3 - ÁRVORE DE DECISÃO GERADA A PARTIR DOS DADOS DA TABELA 3



Fonte: Hubner (2002).

De posse destes dados, um desenvolvedor de software ao analisar a árvore, tomaria a decisão de escolher a linguagem de programação que mais se adaptaria ao propósito da aplicação.

3.1.2 ESTIMATIVA

Enquanto o problema de classificação lida com valores discretos (sim ou não), o problema de estimativa consiste na determinação de valores contínuos. Através de dados de entrada, utiliza-se a estimativa para gerar o valor de alguma variável contínua, como a renda, altura ou saldo de cartão de crédito de um indivíduo.

Um exemplo de utilização do modelo de estimativa poderia ser, por exemplo, uma empresa de cartão de crédito interessada em anunciar uma promoção de equipamentos de esqui. Para isso, a empresa deveria classificar seus clientes em dois conjuntos: praticante de esqui e não praticante de esqui. Uma alternativa a essa abordagem é a construção de um modelo que associe a cada cliente um valor entre 0 e 1 para a probabilidade de uso de equipamentos de esqui. Desta forma, estima-se a probabilidade do cliente praticar o esporte (RODRIGUES, 2000).

A vantagem dos modelos de estimativa está na possibilidade de ordenar os dados. Com os dados ordenados pela probabilidade da prática de esqui, a empresa enviaria o material

promocional somente para os clientes com maior probabilidade de praticar o esporte, obtendo resultados satisfatórios com menores gastos.

Algumas aplicações da técnica de estimativa são:

- a) estimativa do número de crianças em uma família;
- b) estimativa do total de pessoas em uma família;
- c) estimativa da probabilidade de resposta para mala direta.

3.1.3 ASSOCIAÇÃO

A técnica de associação tem por objetivo encontrar relacionamentos ou padrões freqüentes entre conjuntos de dados, tais como “90% de pessoas que compram pão também compram leite”. Atualmente, grandes empresas de varejo utilizam-se das regras de associação, pois permitem que o processo de *marketing* seja dirigido, além de servirem para reorganização de *layout* das lojas, onde itens podem ser agrupados de modo a induzir a venda de artigos relacionados (CERVO, 2002).

Um exemplo de entrada de dados para descoberta de regras de associação pode ser vista na tabela 4. Como se pode observar, a relação entre os itens café, pão e manteiga é de 100%, ou seja, nesta amostra de vendas efetuadas, todos os clientes que compraram café também compraram pão e manteiga.

TABELA 4 - ENTRADA DE DADOS PARA A DESCOBERTA DE REGRAS DE ASSOCIAÇÃO

ID	LEITE	CAFÉ	CERVEJA	PÃO	MANTEIGA	ARROZ	FEIJÃO
1	Não	Sim	Não	Sim	Sim	Não	Não
2	Sim	Não	Sim	Sim	Sim	Não	Não
3	Não	Sim	Não	Sim	Sim	Não	Não
4	Sim	Sim	Não	Sim	Sim	Não	Não

Fonte: Rodrigues (2000).

3.1.4 SEGMENTAÇÃO

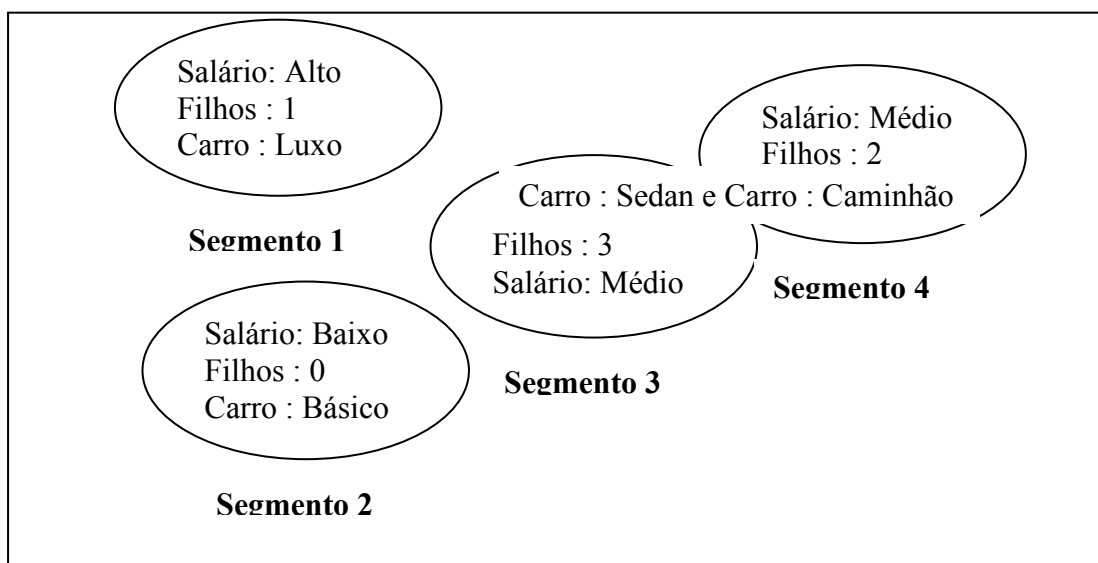
Nesta técnica, o algoritmo deve criar conjuntos de dados com valores semelhantes a partir de um banco de dados heterogêneo. Uma vez criados os conjuntos, pode-se aplicar um algoritmo de classificação, criando assim regras para os mesmos. Na classificação, os dados são subdivididos e colocados em alguma das classes definidas. Em um problema de segmentação, por outro lado, os registros são agrupados com base em similaridades. Muitas vezes a segmentação é uma das primeiras etapas dentro de um processo de DM, já que identifica grupos de registros correlatos que serão usados como ponto de partida para futuras explorações. (RODRIGUES, 2000).

Como aplicações de segmentação, pode-se citar:

- determinação do número de região de vendas;
- determinação de grupos de consumidores potenciais;
- agrupamento das casas de uma área de acordo com sua categoria, área construída e localização geográfica.

A figura 4 mostra quatro segmentos extraídos de um banco de dados de clientes. No primeiro segmento, pode-se perceber um cliente que possui alto salário e carro de luxo, diferentemente do segmento 2 que possui um salário baixo e carro popular. Os segmentos 3 e 4 são quase um mesmo segmento, diferenciado apenas pelo item filhos.

FIGURA 4 - CLIENTES DIVIDIDOS EM SEGMENTOS QUE POSSUEM SEMELHANÇAS



Fonte: Rodrigues (2000).

4 ÁRVORES DE DECISÃO

Segundo Berry (1998) árvores de decisão são ferramentas analíticas usadas para descobrir automaticamente regras e relacionamentos entre dados, subdividindo a informação em subconjuntos. Amplamente utilizadas em algoritmos de classificação, as árvores de decisão são representações simples do conhecimento e um meio eficiente de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados.

Em uma árvore de decisão existem dois tipos de atributo: o atributo-alvo, que é aquele que contém o resultado ao qual se quer chegar, e os outros atributos que contêm os valores que conduzem a uma decisão.

Segundo Gama (2002) e Mitchell (1997) a idéia base da construção de um árvore de decisão deve seguir os seguintes passos, conforme quadro:

QUADRO 1 - PASSOS PARA A CONSTRUÇÃO DE UMA ÁRVORE DE DECISÃO

- 1 Escolher um atributo;
- 2 Estender a árvore adicionando um ramo para cada valor do atributo;
- 3 Passar os exemplos para as folhas (tendo em conta o valor do atributo escolhido);
- 4 Para cada folha:
 - 4.1 Se todos os exemplos são da mesma classe, associar essa classe à folha;
 - 4.2 Senão repetir os passos de 1 a 4.

Fonte: Gama (2002).

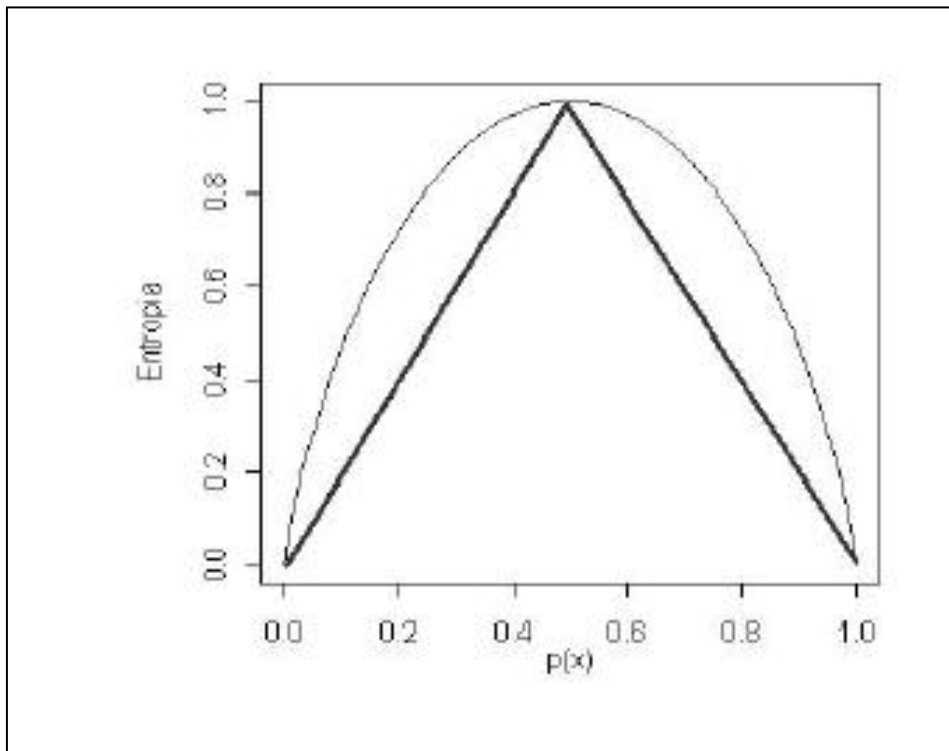
Uma árvore de decisão tem a função de particionar recursivamente um conjunto de treinamento, até que cada subconjunto obtido deste particionamento contenha casos de uma única classe. Para atingir esta meta, a técnica de árvores de decisão examina e compara a distribuição de classes durante a construção da árvore. A distribuição de classes pode ser representada em forma de uma lista de probabilidades $p(c_1) \dots p(c_n)$, em que cada p_i indica a probabilidade de um exemplo pertencer a uma classe. Os valores das funções que calculam essas probabilidades representam a informação necessária para classificar um caso e são chamados de entropia e ganho de informação, que serão detalhados a seguir. O resultado

obtido, após a construção de uma árvore de decisão, são dados organizados de maneira compacta, que são utilizados para classificar novos casos.

4.1 ENTROPIA

Segundo Antunes (2002), a entropia é a medida da impureza do conjunto de dados. Assumindo o valor máximo (1) quando existem tantos elementos positivos como negativos, e o valor mínimo (0) quando todos os elementos são da mesma classe, como se pode ver na figura 5.

FIGURA 5 – GRÁFICO DE DEMONSTRAÇÃO DA ENTROPIA



Fonte: Gama (2002).

No contexto das árvores de decisão, a entropia é usada para estimar a aleatoriedade da variável a prever: o atributo alvo. A construção de uma árvore de decisão é guiada pelo objetivo de diminuir a entropia, ou seja, a aleatoriedade do atributo alvo (GAMA, 2002).

O quadro 2 mostra a fórmula de cálculo da entropia, onde S é o conjunto de treino, C é o número de valores para o atributo alvo e p_i é a proporção de exemplos com atributo alvo (número de exemplos de S_i / número total de exemplos)

QUADRO 2 - FÓRMULA DO CÁLCULO DA ENTROPIA

$$\text{Entropia (S)} = \sum_{i=1}^C - p_i * \log_2 (p_i)$$

Fonte: Antunes (2002).

Tendo como base o quadro 2, pode-se dar valores às variáveis e efetuar o cálculo da entropia. O resultado é mostrado no quadro 3.

QUADRO 3 - CÁLCULO DA ENTROPIA

$$\begin{aligned} S &= \{1,2,\dots,14\}; \\ C &= \{2\} - (\text{Jogar=Sim e Jogar=Não}) \\ P_1(\text{Sim}) &= 9/14 \\ P_2(\text{Não}) &= 5/14 \\ \\ \text{Entropia}(\{1,2,\dots,14\}) &= (-9/14 * \log_2(9/14)) + (-5/14 * \log_2(5/14)) \\ &= 0,41 + 0,53 \\ &= 0,94 \end{aligned}$$

Fonte: Adaptado de Hubner (2002).

4.2 GANHO DE INFORMAÇÃO

Para se conseguir gerar uma árvore de decisão com uma alta taxa de predição é necessário fazer a escolha correta dos atributos que serão usados como teste no agrupamento dos casos. Estes testes devem gerar uma árvore com o menor número possível de subconjuntos, ou seja, o ideal é escolher os testes de modo que a árvore final seja a menor possível.

Dado um conjunto de exemplos, que atributo escolher para teste? O ganho de informação é a redução esperada no valor da entropia, devido à ordenação do conjunto de treino segundo os valores do atributo (ANTUNES, 2002). O quadro 4 mostra a fórmula de cálculo do ganho de informação:

QUADRO 4 - FÓRMULA DO CÁLCULO DO GANHO DE INFORMAÇÃO

$$\text{Ganho de Informação (S,A)} = \text{Entropia (S)} - \sum_{V \text{ 0 valores(A)}} \frac{|S_v|}{|S|} \text{Entropia (S}_v\text{)}$$

Fonte: Antunes (2002).

Com o conjunto de treino mostrado na tabela 5, pode-se calcular o valor da entropia e do ganho de informação.

TABELA 5 - CONJUNTO DE TREINO PARA CÁLCULO DA ENTROPIA E GANHO DE INFORMAÇÃO

	Tempo	Temperatura	Umidade	Vento	Jogar
1.	ensolarado	alta	alta	fraco	não
2.	ensolarado	alta	alta	forte	não
3.	nublado	alta	alta	fraco	sim
4.	chuva	média	alta	fraco	sim
5.	chuva	baixa	normal	fraco	sim
6.	chuva	baixa	normal	forte	não
7.	nublado	baixa	normal	forte	sim
8.	ensolarado	média	alta	fraco	não
9.	ensolarado	baixa	normal	fraco	sim
10.	chuva	média	normal	fraco	sim
11.	ensolarado	média	normal	forte	sim
12.	nublado	média	alta	forte	sim
13.	nublado	alta	normal	fraco	sim
14.	chuva	média	alta	forte	não

Fonte: Hubner (2002).

O conjunto de treino é formado pelo atributo alvo (Jogar), usado para efetuar o cálculo da entropia e pelos atributos restantes (Tempo, Temperatura, Umidade, Vento) que serão usados para demonstrar o cálculo do ganho de informação.

Para o cálculo do ganho de informação, toma-se como base o quadro 4 e substitui-se o valor das variáveis. O quadro 5 mostra o cálculo do ganho de informação para os atributos tempo e vento.

QUADRO 5 - CÁLCULO DO GANHO DE INFORMAÇÃO PARA OS ATRIBUTOS TEMPO E VENTO

$$\begin{aligned}
 \text{Ganho}(\{1,2,\dots,14\}, \text{Tempo}) &= 0,94 - [(5/14 * \text{Entropia}(\text{ensolarado})) + (4/14 * \\
 &\quad \text{Entropia}(\text{nublado})) + (5/14 * \text{Entropia}(\text{chuva}))] \\
 &= 0,94 - [(5/14 * 0,97) + (4/14 * 0) + (5/14 * 0,97)] \\
 &= 0,246 \\
 \\
 \text{Ganho}(\{1,2,\dots,14\}, \text{Vento}) &= 0,94 - [(6/14 * \text{Entropia}(\text{forte})) + (8/14 * \\
 &\quad \text{Entropia}(\text{fraco}))] \\
 &= 0,94 - [(5/14 * 1) + (8/14 * 0,81)] \\
 &= 0,048
 \end{aligned}$$

Fonte: Adaptado de Hubner (2002).

Entre os atributos tempo e vento, o que deve ser escolhido é o tempo por possuir o maior ganho de informação.

4.3 ALGORITMO C4.5

Segundo Ruggieri (2000) os algoritmos de classificação despertam um considerável interesse em pesquisadores na área de *machine learning* e *data mining*. Dentre estes algoritmos, o C4.5 destaca-se por ser o resultado de pesquisas de evolução do algoritmo ID3 e por ser considerado o mais rápido em relação aos algoritmos de memória-principal para *machine learning* e *data mining*. Dentre as evoluções que o C4.5 possui pode-se citar a possibilidade de tratar os atributos numéricos. O quadro 6 apresenta um pseudo-código do algoritmo C4.5:

QUADRO 6 - PSEUDO-CÓDIGO DO ALGORITMO C4.5

```

FormaArvore(T)
  (1) CalculaEntropia(T);
  (2) Se todos os exemplos são de uma mesma classe
      Retorna uma folha com o nome da classe;
      Cria um nodo Raiz;
  (3) Para cada atributo A
      Calcula Ganho de informação(A);
  (4) Raiz = Atributo com maior ganho de informação;
  (5) Se Atributo = Continuo
      EncontraPontoReferencia;
  (6) Para Cada T' possível em T
  (7)   Se T' é vazio
      Retorna folha com valor mais comum
      Else
  (8)   AdicionaArvore = FormaArvore(T');

```

Fonte: Ruggieri (2000).

O algoritmo C4.5 constrói a árvore de decisão com a estratégia dividir para conquistar. O algoritmo inicia com a raiz da árvore. Em seguida, é executado o algoritmo mostrado no quadro 6. No passo 1, calcula-se a Entropia do conjunto de treino. Se todos os exemplos do conjunto de treino pertencem a uma mesma classe, o retorno é uma folha (passo 2). Cria-se um novo nodo de decisão. No passo 3, para cada atributo do conjunto de treino, calcula-se o ganho de informação. O nodo receberá o atributo que possuir o maior ganho de informação (passo 4). No passo 5 é verificada a possibilidade do atributo ser contínuo. Caso seja contínuo, deve-se dividir o atributo em dois conjuntos, exemplos onde o valor do atributo < ponto de referência e exemplos onde o valor do atributo >= ponto de referência. Conforme Aurora (2001) para calcular o ponto de referência deve-se ordenar o conjunto de treino. Para cada valor V_i é calculado $(V_i + V_{i+1})/2$ como sendo o ponto de referência que será utilizado no teste para fazer a divisão do atributo em dois subconjuntos. Baseado nesses subconjuntos é então calculado o ganho de informação. Sendo m o conjunto de valores do atributo A deve-se calcular $m-1$ possibilidades para o ponto de referência e seu ganho. O ponto de referência que apresentar o maior ganho de informação será o representante do atributo A . Seguindo o algoritmo, para cada subconjunto de T (passo 6), verifica-se se T' é vazio (passo 7), caso positivo, o filho deste nodo é uma folha, caso contrário, inicia-se a função tendo como parâmetro o T' ou seja, o subconjunto de T (passo 8).

4.3.1 ATRIBUTOS NUMÉRICOS

Segundo Gama (2002), para entender melhor o cálculo do atributo numérico, pode-se tomar como exemplo a decisão de jogar tênis ou não, acrescentando o atributo numérico temperatura. Neste exemplo, o ponto de referência foi fixado arbitrariamente em 70.5. Em uma classificação real, para cada valor único (V_i) do atributo numérico, o ponto de referência é igual ao valor(V_i) + próximo valor (V_{i+1}) / 2. Em seguida, calcula-se o ganho de informação para o ponto de referência. O ponto de referência que obtiver o maior ganho de informação será o escolhido para dividir os atributos numéricos em dois subconjuntos, conforme o exemplo a seguir. A tabela 6 mostra o atributo numérico temperatura em graus Farenheite o atributo alvo jogar tênis.

TABELA 6 - PROBLEMA DE JOGAR TÊNIS COM ATRIBUTO NUMÉRICO

TEMPERATURA	JOGA
64	Sim
65	Não
68	Sim
69	Sim
70	Sim
71	Não
72	Não
72	Sim
75	Sim
75	Sim
80	Não
81	Sim
83	Sim
85	Não

Fonte: Gama (2002).

Considerando o ponto de referência da temperatura igual a 70.5, os dados seriam divididos em dois subconjuntos, exemplos onde a temperatura é menor que 70.5 e exemplos onde a temperatura é maior que 70.5. Para calcular o ganho de informação deste ponto de referência, os dados são divididos em partições, mostradas no quadro 7.

QUADRO 7 - PARTIÇÕES DO ATRIBUTO TEMPERATURA

partição(sim temperatura < 70.5) = 4/5
partição(não temperatura < 70.5) = 1/5
partição(sim temperatura > 70.5) = 5/9
partição(não temperatura > 70.5) = 4/9

Fonte: Adaptado de Gama (2002).

Calcula-se a entropia para valores menores que o ponto de referência e para valores maiores que o ponto de referência, conforme o quadro 8.

QUADRO 8 - CÁLCULO DA ENTROPIA PARA VALORES MENORES E VALORES MAIORES QUE O PONTO DE REFERÊNCIA

$$\begin{aligned} &\text{Entropia (joga | temperatura < 70.5)} \\ &-4/5 \log_2 4/5 - 1/5 \log_2 1/5 = 0,721 \end{aligned}$$

$$\begin{aligned} &\text{Entropia (joga | temperatura > 70.5)} \\ &-5/9 \log_2 5/9 - 4/9 \log_2 4/9 = 0,991 \end{aligned}$$

Fonte: Adaptado de Gama (2002).

Calcula-se então a entropia para os quatorze valores do atributo temperatura, dos quais cinco são abaixo do ponto de referência e nove são acima do ponto de referência, como apresentado no quadro 9.

QUADRO 9 - CÁLCULO DA ENTROPIA PARA O ATRIBUTO TEMPERATURA

$$\begin{aligned} &\text{Entropia (Temperatura)} \\ &5/14 * 0,721 + 9/14 * 0,991 = 0,895 \end{aligned}$$

Fonte: Adaptado de Gama (2002).

Por fim, calcula-se o ganho de informação para o ponto de referência, subtraindo a entropia do atributo temperatura da entropia dos exemplos, como mostra o quadro 10.

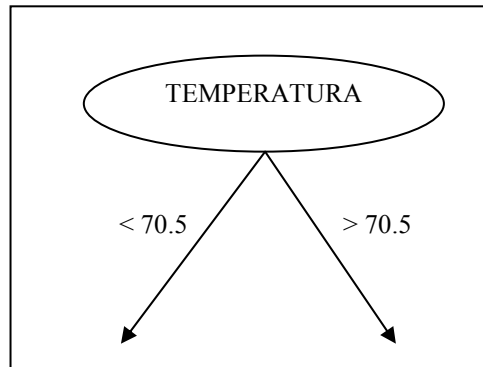
QUADRO 10 - CÁLCULO DO GANHO DE INFORMAÇÃO PARA O PONTO DE REFERÊNCIA

$$\begin{aligned} &\text{Ganho(ponto de referência)} \\ &0,940 - 0,895 = 0,045 \end{aligned}$$

Fonte: Adaptado de Gama (2002).

O ganho de informação para o ponto de referência = 70.5 é de 0,045. O ponto de referência que obtiver o maior ganho de informação será usado para dividir o conjunto de exemplos em dois subconjuntos. Se o ponto de referência = 70.5 obtivesse o maior ganho de informação, teríamos um dos nodos da árvore como mostra a figura 6.

FIGURA 6 - NODO GERADO PELO PONTO DE REFÊNCIA = 70.5



Fonte: Adaptado de Gama (2002).

5 DESENVOLVIMENTO DO APLICATIVO

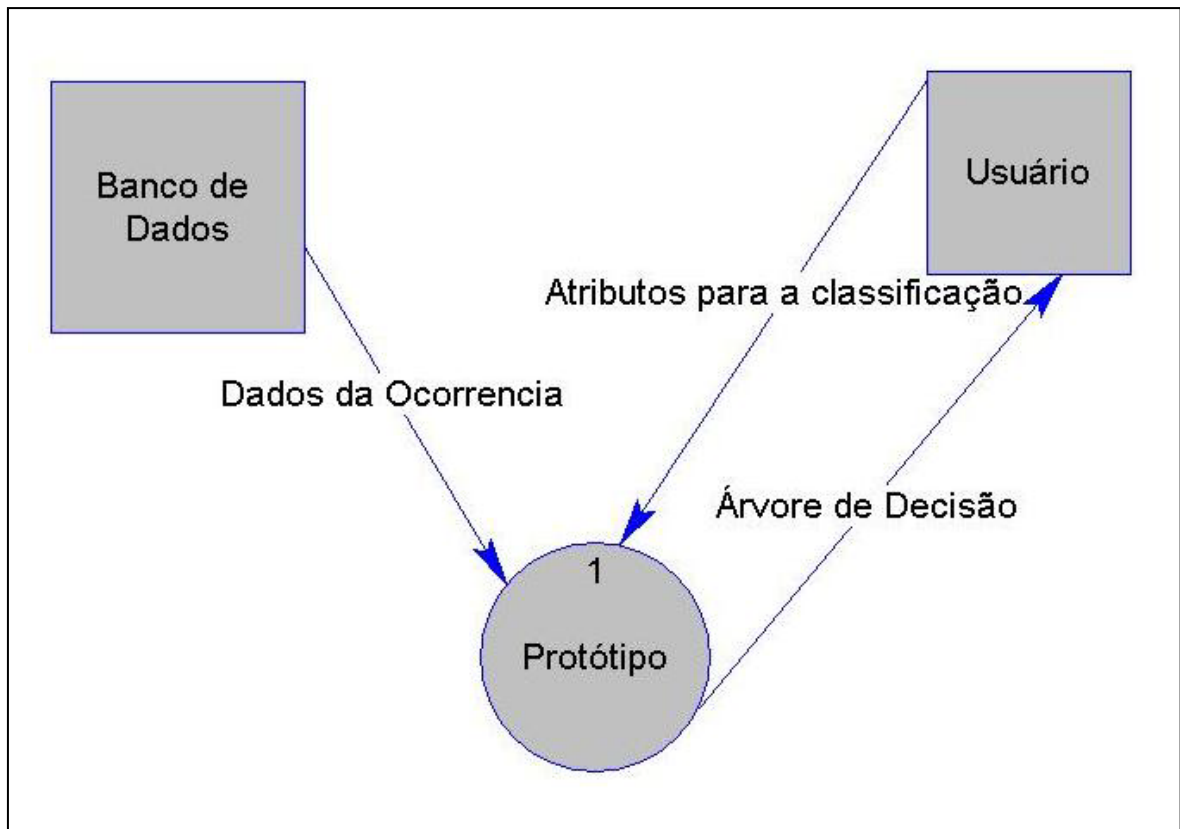
Levando em conta os objetivos propostos por este trabalho, construiu-se uma aplicação que fosse flexível e de fácil utilização. Utilizou-se a filosofia *Data Mining* com as etapas do processo KDD, a técnica de árvores de decisão e a implementação do algoritmo C4.5.

5.1 ESPECIFICAÇÃO

De acordo com Yourdon (1990) 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. A modelagem de dados utilizando a técnica estruturada utiliza-se de ferramentas para descrever o processo de entradas e saídas e uma delas é o diagrama de fluxo de dados (DFD), que consiste em processos, depósitos de dados, fluxos de dados e entidades.

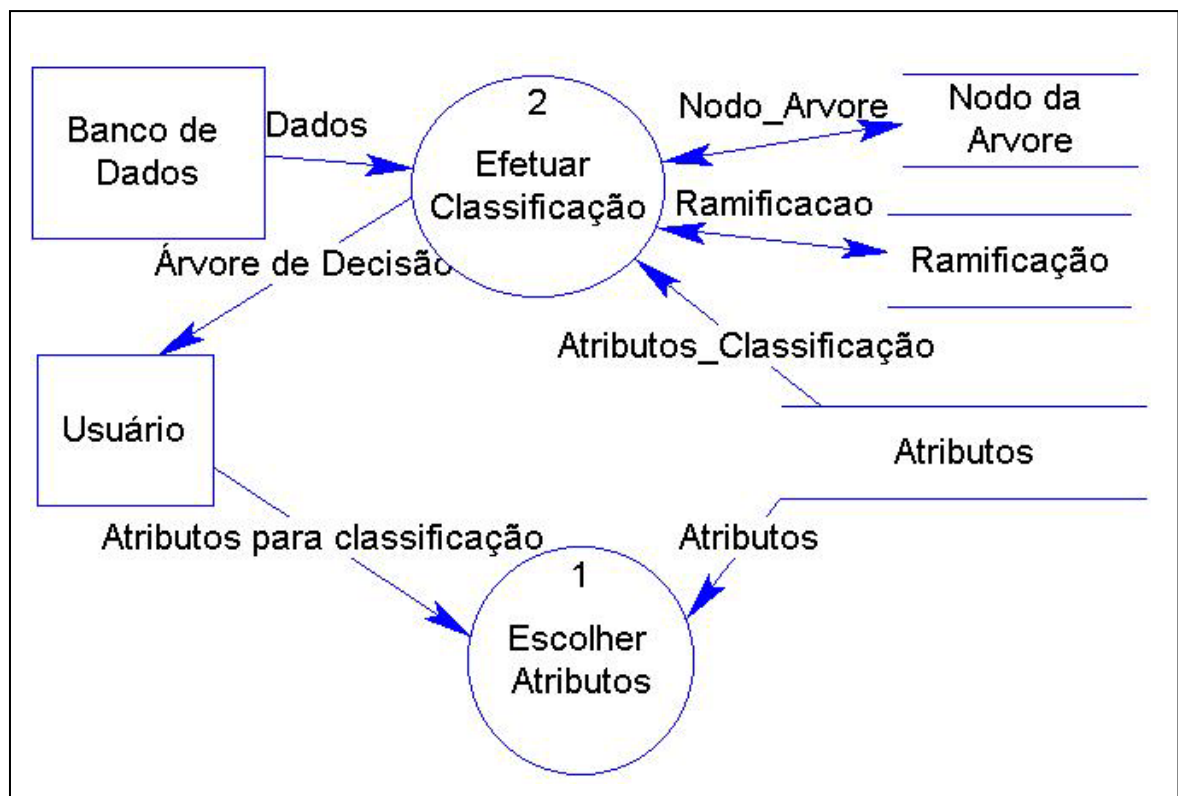
No desenvolvimento deste aplicativo utilizou-se para a construção do diagrama de contexto e DFD nível 0 a ferramenta *case* Power Designer - *ProcessAnalyst* e no desenvolvimento do modelo entidade-relacionamento utilizou-se o Power Designer - *DataArchitect*. A figura 7 mostra o diagrama de contexto do aplicativo.

FIGURA 7 - DIAGRAMA DE CONTEXTO



O aplicativo irá interagir com o usuário que fará a escolha dos atributos que serão processados pelo algoritmo de mineração de dados, gerando uma árvore de decisão que será apresentada ao usuário em forma de visualização se/então.

FIGURA 8 - DFD NÍVEL 0



Descreve-se a seguir os processos do DFD nível 0, mostrado na figura 8:

- a) escolher atributos: processo onde o usuário definirá quais atributos serão usados para gerar a classificação;
- b) efetuar classificação: este processo é caracterizado pela utilização da indicação dos atributos especificados pelo usuário. De posse da informação de quais são os atributos, começa a ser feita a classificação com o cálculo da entropia e o ganho de cada atributo.

A seguir estão descritos os depósitos do DFD nível 0:

- a) nodo: entidade responsável pelo armazenamento dos dados referentes aos nós da árvore, gerados pelo processo de classificação;
- b) ramificação: entidade responsável pelo armazenamento dos dados referentes às ligações existentes entre os nós.

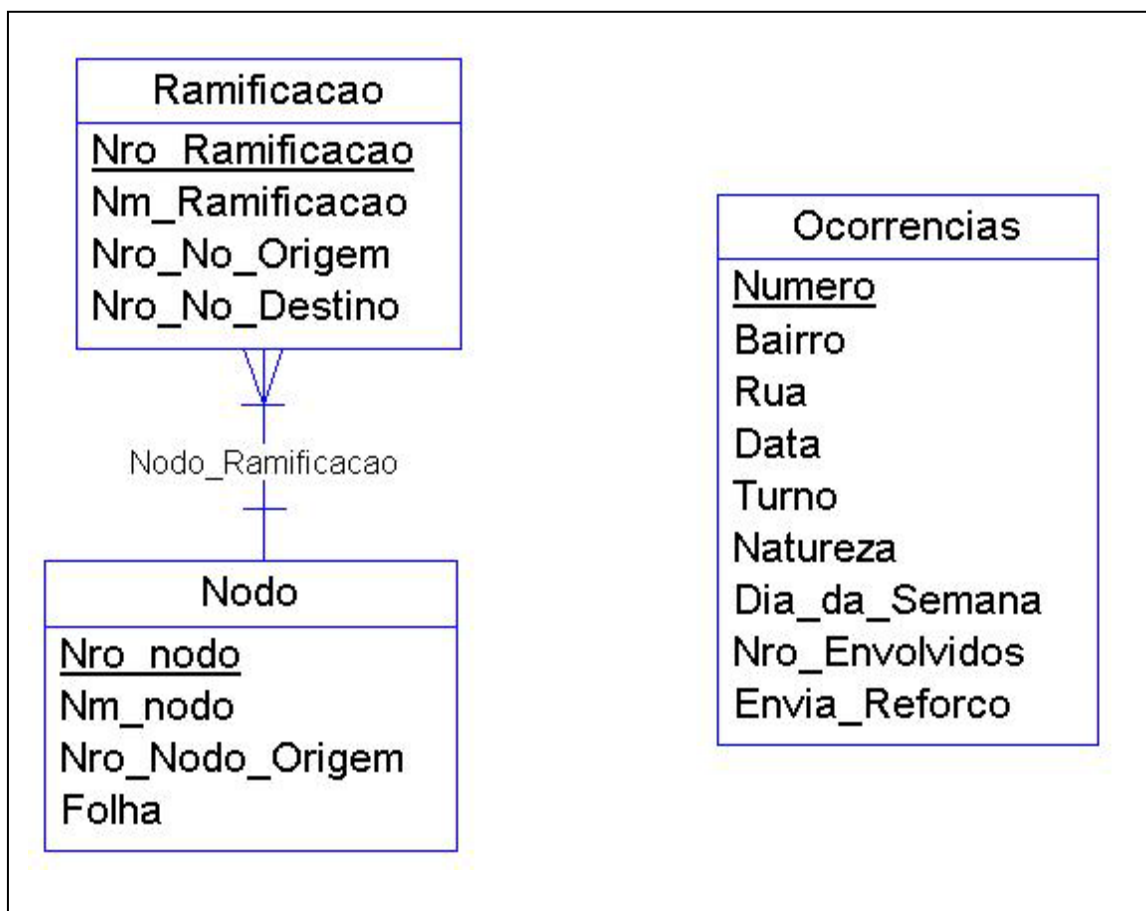
Embora o diagrama de fluxo de dados ofereça uma prática visão geral dos principais componentes funcionais do sistema, ele não fornece qualquer detalhe sobre esses componentes.

Para mostrar os detalhes de como a informação é transformada, necessita-se de ferramentas como o modelo entidade relacionamento (MER).

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 existem entre esses depósitos de dados.

A figura 9 mostra o modelo entidade relacionamento do aplicativo.

FIGURA 9 - MODELO ENTIDADE RELACIONAMENTO



As entidades "Nodo" e "Ramificação" surgiram a partir dos depósitos de dados contidos no DFD nível 0, e a entidade Ocorrências é a representação da entidade externa "Banco de Dados" que será utilizada para o processamento da árvore de decisão.

Optou-se por armazenar a árvore de decisão em tabelas de banco de dados, por motivo de capacidade de armazenamento, uma vez que a árvore gerada pode alcançar um número alto de nodos e ramos, atingindo assim a capacidade de memória do computador. Outro motivo é que armazenada em banco de dados, toda vez que o aplicativo for reiniciado, a árvore

permanecerá armazenada no banco de dados, o que não acontece com a memória do computador.

A tabela 7 contém o detalhamento das entidades envolvidas no aplicativo.

TABELA 7 - DESCRIÇÃO DETALHADA DO MODELO DE DADOS

ENTIDADE	ATRIBUTO	TIPO DE DADOS	TAMANHO	DESCRIÇÃO
OCORRENCIAS	Bairro	ALFANUMÉRICO	30	nome do bairro onde houve a ocorrência
	Data	DATA	-	data da ocorrência
	Dia_da_Semana	ALFANUMÉRICO	15	dia da semana em que houve a ocorrência
	Natureza	ALFANUMÉRICO	4	delito (crime) da ocorrência
	Nro_Envolvidos	NUMÉRICO	1	número de pessoas envolvidas na ocorrência
	Numero	NUMÉRICO	1	número da ocorrência (número que identifica cada ocorrência)
	Rua	ALFANUMÉRICO	40	rua onde houve a ocorrência
	Turno	ALFANUMÉRICO	9	compreende o turno em que houve a ocorrência (madrugada, manhã, tarde, noite)
NODO	Nro_nodo	NUMÉRICO	1	identifica cada nodo da árvore
	Nm_nodo	ALFANUMÉRICO	50	nome do nodo
	Folha	ALFANUMÉRICO	1	Indica se o nodo é uma folha ou não
	Visualizado	ALFANUMÉRICO	1	Indica se o nodo já foi visualizado ou não
	Nro_Nodo_Origem	NUMÉRICO	1	de qual nodo originou
RAMIFICACAO	Nro_Ramificacao	NUMÉRICO	1	identifica cada ramificação da árvore
	Nm_Ramificacao	ALFANUMÉRICO	50	nome da ramificação
	Nro_No_Destino	NUMÉRICO	1	para qual nodo a ramificação aponta
	Nro_No_Origem	NUMÉRICO	1	de qual nodo esta ramificação vem

5.2 IMPLEMENTAÇÃO

A seguir serão apresentados o Banco de Dados usado para o desenvolvimento do aplicativo, linguagem de programação e o desenvolvimento do aplicativo conforme as etapas do processo de KDD.

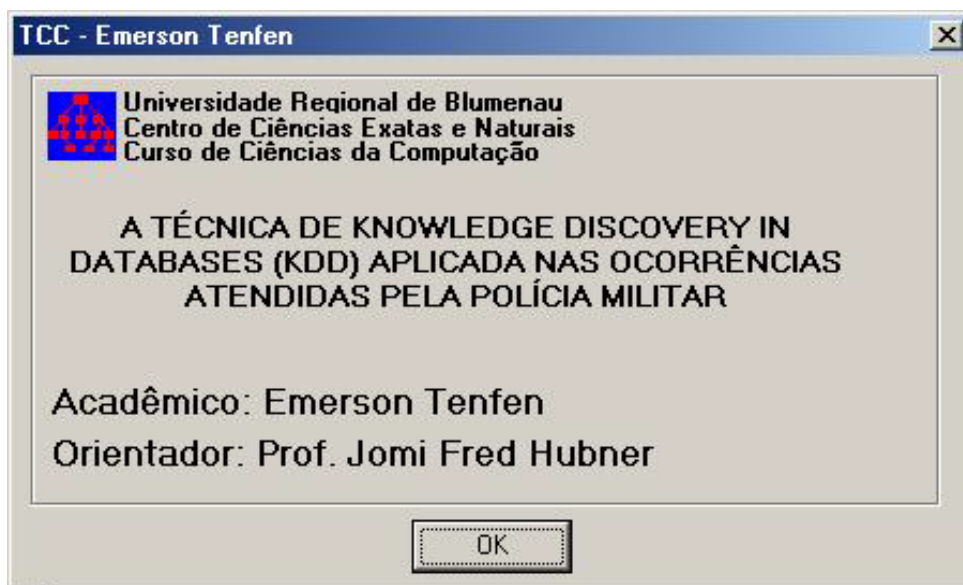
Foi utilizado o banco de dados Paradox 7.0 para o armazenamento dos dados da ocorrência, dos nodos e ramificações da árvore. Segundo Borland (1995), Paradox é um sistema completo de gerenciamento de bancos de dados relacional que pode ser usado como

um sistema autônomo em um computador simples ou como um sistema multiusuário em uma rede e que, mesmo sendo um programa de banco de dados eficiente e capaz de realizar tarefas complexa, fácil de se aprender e rápido de se ativar.

Para o desenvolvimento do aplicativo, utilizou-se a linguagem de programação *Object Pascal* no ambiente de desenvolvimento Delphi 5.0. Segundo Cantu (2000), Delphi é uma versão de desenvolvimento rápido de aplicativos do Turbo Pascal para Windows. O Delphi oferece uma interface melhorada e muitos recursos que facilitam o desenvolvimento de aplicativos.

A figura 10 mostra a tela de apresentação do aplicativo.

FIGURA 10 - TELA DE APRESENTAÇÃO



5.2.1 DOMÍNIO DA APLICAÇÃO

Esta etapa é muito importante no processo de KDD. O usuário que fizer uso do aplicativo desenvolvido deverá possuir um prévio conhecimento das rotinas de atendimento de ocorrências para poder indicar os atributos que serão usados pelo aplicativo para gerar a classificação.

Nesta etapa, o usuário deve apenas informar quais os atributos que serão submetidos ao algoritmo de classificação. Para ter acesso à tela de configuração, demonstrada na fig. 12,

o usuário deverá selecionar a opção Sistema-Configuração no Menu Principal conforme demonstra a figura 11.

FIGURA 11 - TELA PRINCIPAL DO APLICATIVO

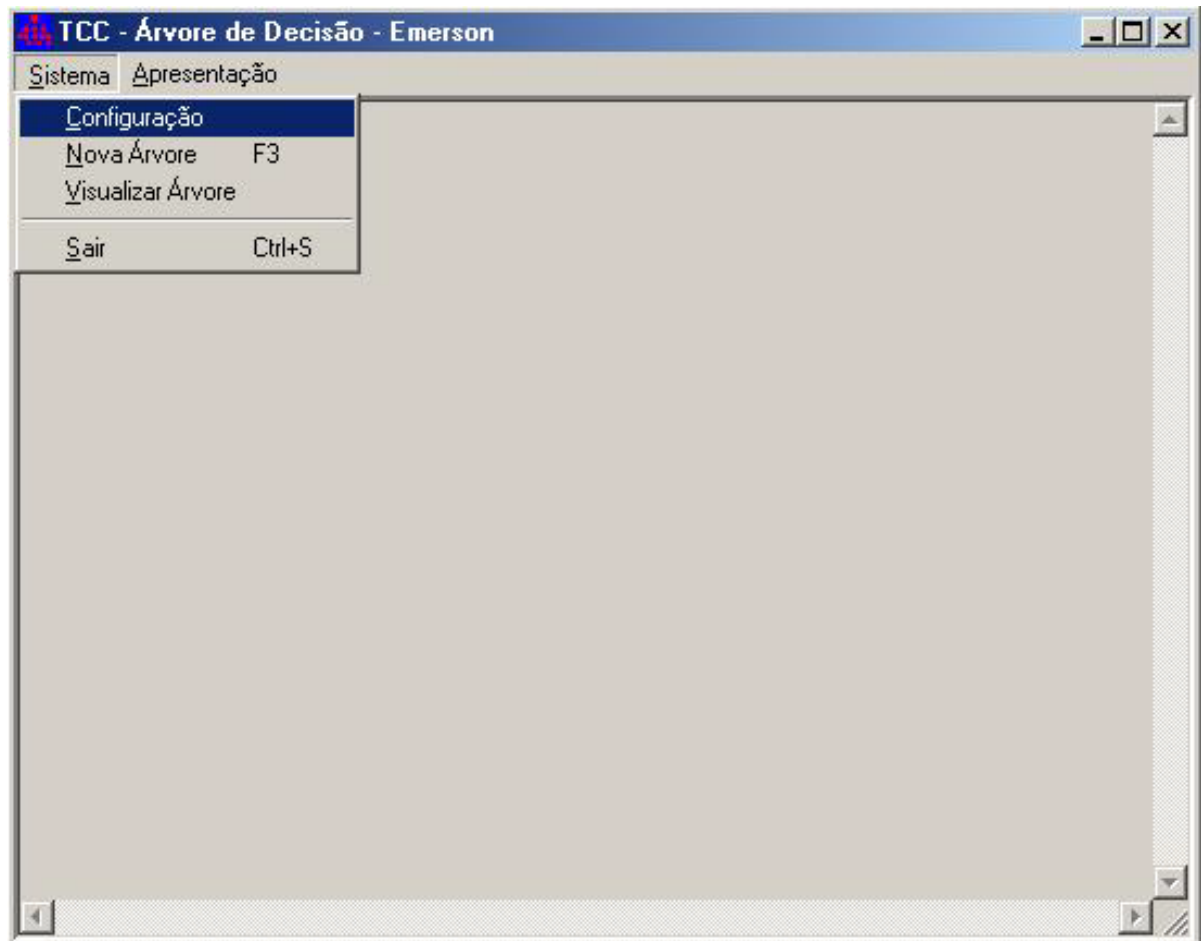


FIGURA 12 - TELA DE ESCOLHA DOS ATRIBUTOS

The screenshot shows a window titled "Configuração dos Dados" with a table of data and a configuration panel below it.

Numero	Bairro	Rua	Data
177259	DO ASILO	JOSE DEEKE	18/05/03
175815	DO ASILO	JOSE DEEKE	03/05/03
169067	DO ASILO	JOSE DEEKE	28/02/03
169694	DO ASILO	JOSE DEEKE	07/03/03
169698	DO ASILO	JOSE DEEKE	07/03/03
173224	PONTA AGUDA	JOSE ISIDORO CORREA	10/04/03
172860	PONTA AGUDA	JOSE ISIDORO CORREA	07/04/03
169506	PONTA AGUDA	JOSE ISIDORO CORREA	04/03/03
177091	ITOUPAVA NORTE	JOSE MANOEL DEPLA	17/05/03
173971	FORTALEZA	JOSE MANOEL GOUVEIA	17/04/03
166933	DA VELHA	JOSE MARTINS	06/02/03
174316	DA VELHA	JOSE REUTER	20/04/03
175845	DA VELHA	JOSE REUTER	04/05/03
175855	DA VELHA	JOSE REUTER	04/05/03

Below the table, the configuration panel includes:

- Campos da Tabela:** A list containing "Numero" and "Data".
- Atributos para Classificacao:** A list containing "Bairro", "Rua", "Turno", "Natureza", and "Dia_da_Semana".
- Atributo Continuo:** "Nro_Envolvidos".
- Atributo Alvo:** "Envia_Reforco".
- Total Registros:** A text box displaying "1638".
- Buttons:** "OK" (with a green checkmark) and "Sair" (with a red X).

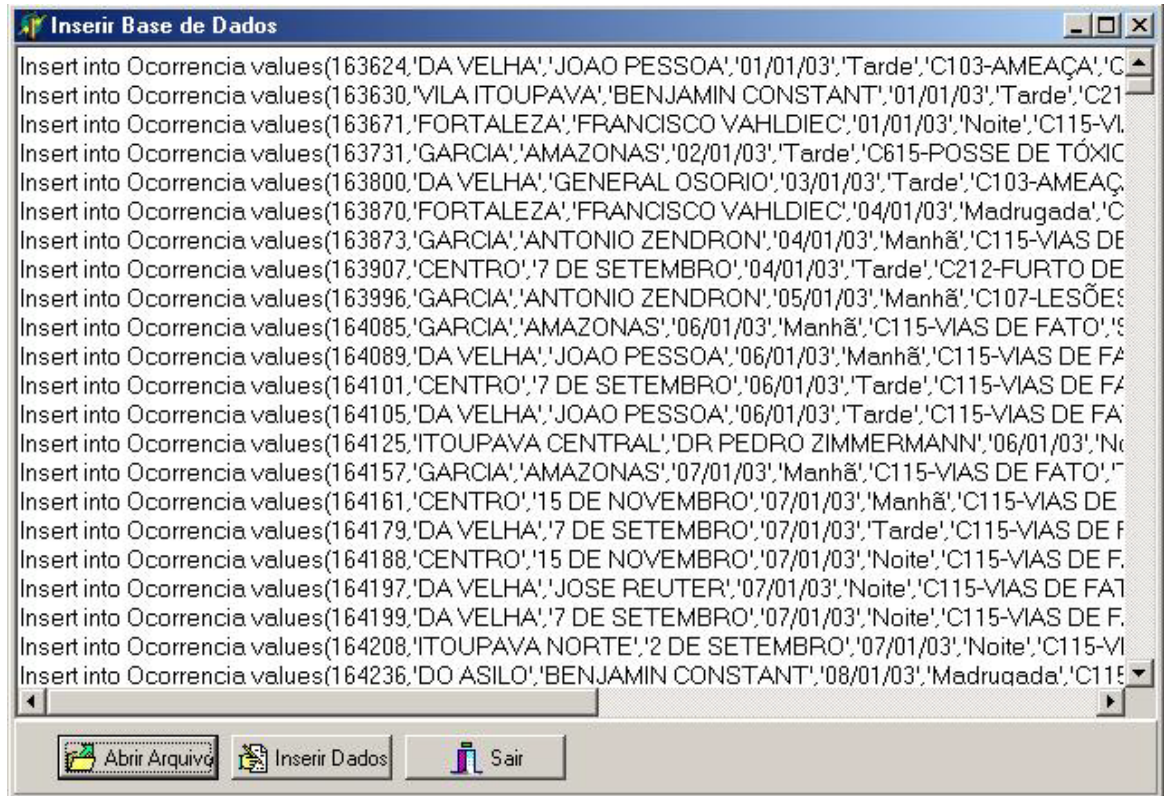
O atributo Nro_Envolvidos é um atributo com valores numéricos e neste aplicativo foi utilizado para comprovar a eficiência do algoritmo C4.5 no tratamento com atributos desta natureza. O atributo Envia_Reforco é o atributo alvo e é utilizado como resultado da classificação. Com base nos dados armazenados, o aplicativo será capaz de auxiliar um operador do COPOM a tomar a decisão de enviar reforço ou não no atendimento de uma ocorrência.

5.2.2 CRIAÇÃO DE UM GRUPO DE DADOS ALVO

Esta etapa consiste na criação da base de dados utilizada no modelo de classificação. Por uma questão de segurança, não foi possível trabalhar diretamente com a base de dados da Polícia Militar. Desta forma, os dados eram apresentados em forma de arquivos do Microsoft Excel, transformados posteriormente em arquivos texto. Foi desenvolvido um programa para,

a partir destes arquivos, inserir os dados no banco de dados do aplicativo. A tela do programa desenvolvido é mostrada na figura 13.

FIGURA 13 – TELA PRINCIPAL DO PROGRAMA PARA CRIAR BASE DE DADOS DO APLICATIVO



5.2.3 LIMPEZA DOS DADOS

A etapa de limpeza dos dados visa adequar os dados aos algoritmos de *Data Mining*, eliminando ruídos e valores desconhecidos. No atributo "Natureza", realizou-se a limpeza dos dados. A tabela 8 mostra os dados originais para o atributo "Natureza".

TABELA 8 - DADOS ORIGINAIS PARA O ATRIBUTO "NATUREZA"

GRUPOS DE NATUREZA
A000-Auxílios à Comunidade/Órgão Públicos
C000-Crimes e Contravenções
D000-Ocorrências Diversas
E000-Emergência/Traums/Acidentes
I000-Incêndios
N000-Contra o meio ambiente
P000 - Serviços/Atividades Profissionais
S000 - Serviços/Atividades Afins
Y000 – Trânsito

Como o objetivo do aplicativo seria fazer uma classificação para auxiliar no combate à violência, no atributo "Natureza" serão consideradas somente as naturezas do grupo C000-Crimes e Contravenções.

5.2.4 REDUÇÃO E PROJEÇÃO DOS DADOS

A etapa de redução dos dados consiste em “adequar” os dados. Esta etapa foi realizada no atributo "Turno". Nos dados originais constam as horas em que acontecem a ocorrência. A tabela 9 mostra a transformação do atributo "Turno" após a redução e projeção.

TABELA 9 - ETAPA DE PRÉ-PROCESSAMENTO REALIZADA NO ATRIBUTO "TURNO"

INTERVALO	TURNO
00:01	MADRUGADA
06:00	
06:01	MANHÃ
12:00	
12:01	TARDE
18:00	
18:00	NOITE
24:00	

5.2.5 MINERAÇÃO DE DADOS

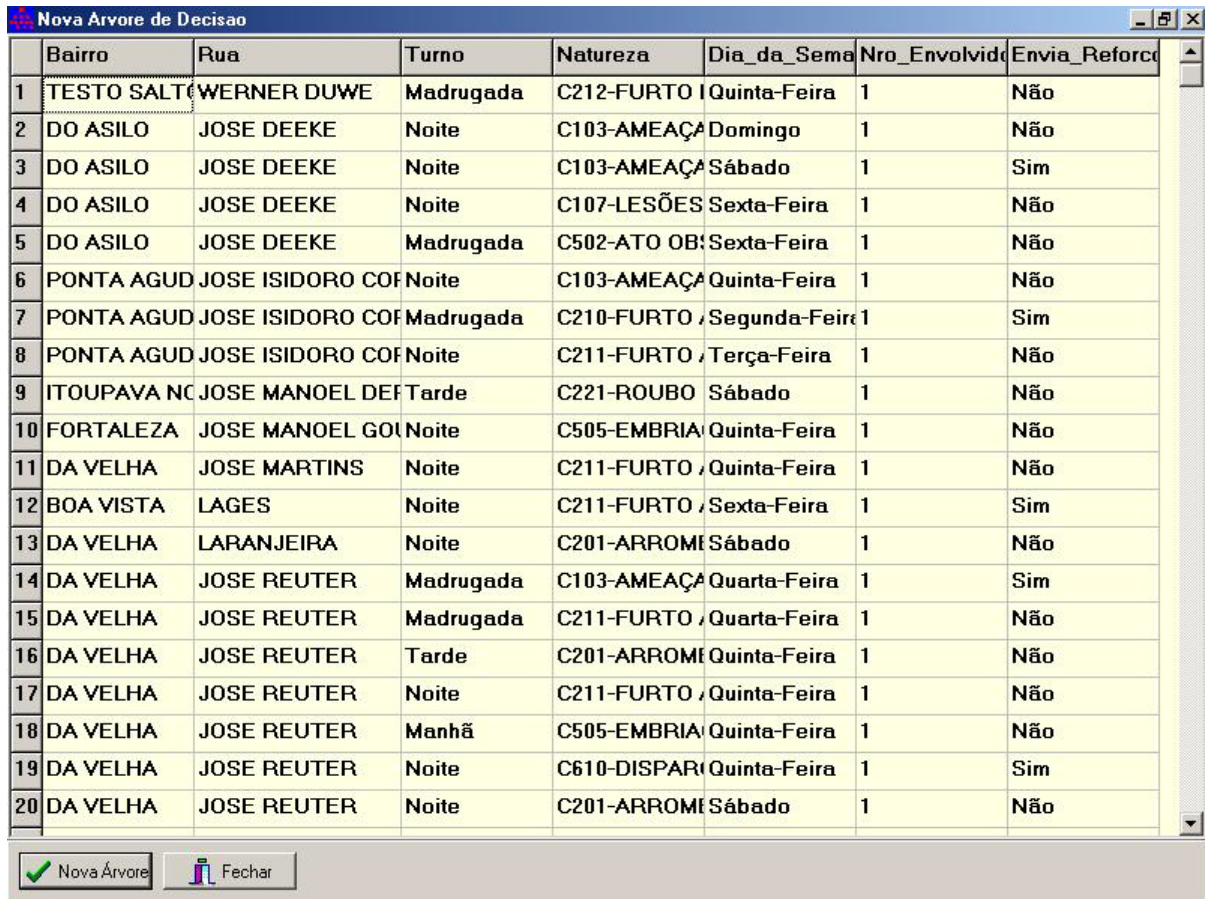
A etapa de mineração de dados é a etapa onde o algoritmo, com base nos atributos definidos, deverá descobrir intuitivamente regras que forneçam diagnósticos, isto é, descobertas automáticas de conhecimento que é um de seus objetivos principais.

Para este aplicativo foi definida como tarefa de MD a classificação e como algoritmo de MD o C4.5.

Para executar o processamento da árvore de decisão, o usuário deverá selecionar primeiramente a opção Sistema-Nova Árvore no menu principal como demonstrado na fig.

15, ou simplesmente pressionar a tecla F3, abrindo-se assim a tela de geração da nova árvore de decisão, demonstrada na figura 14.

FIGURA 14 - TELA DE GERAÇÃO DE NOVA ÁRVORE



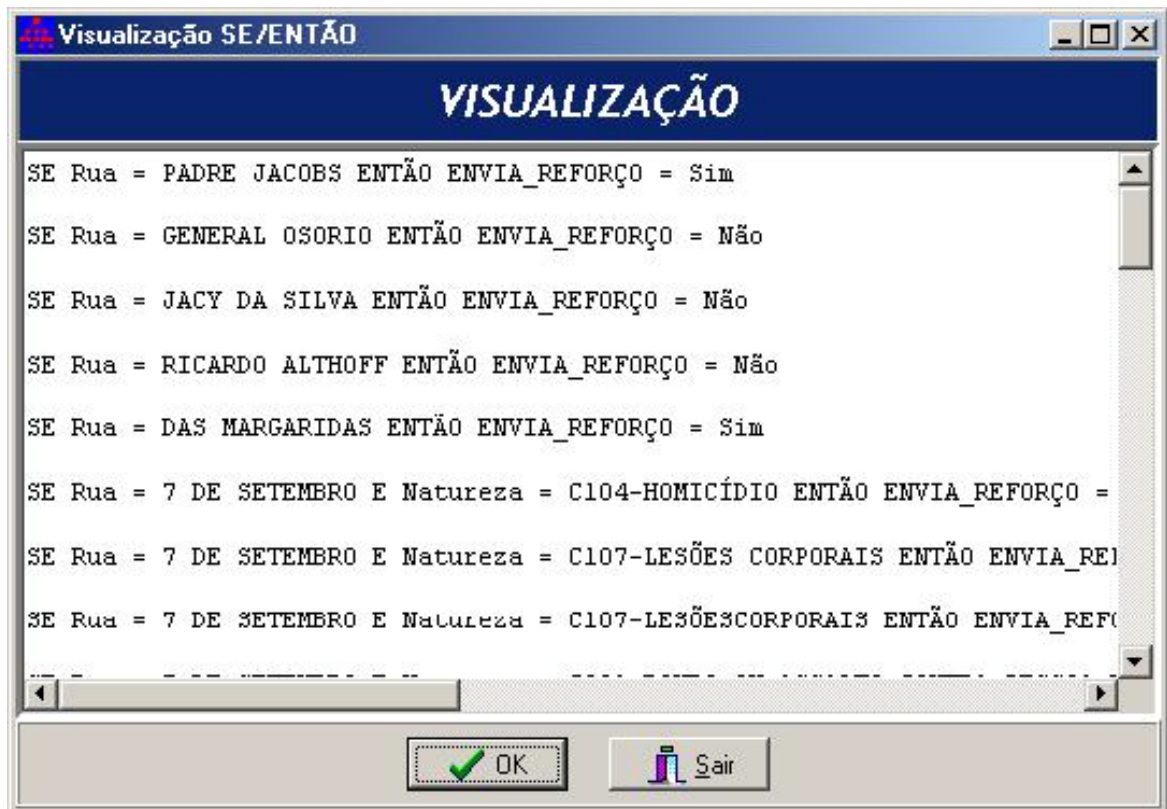
	Bairro	Rua	Turno	Natureza	Dia_da_Sema	Nro_Envolvido	Envvia_Reforco
1	TESTO SALT	WERNER DUWE	Madrugada	C212-FURTO	Quinta-Feira	1	Não
2	DO ASILO	JOSE DEEKE	Noite	C103-AMEAÇA	Domingo	1	Não
3	DO ASILO	JOSE DEEKE	Noite	C103-AMEAÇA	Sábado	1	Sim
4	DO ASILO	JOSE DEEKE	Noite	C107-LESÕES	Sexta-Feira	1	Não
5	DO ASILO	JOSE DEEKE	Madrugada	C502-ATO OB	Sexta-Feira	1	Não
6	PONTA AGUD	JOSE ISIDORO COF	Noite	C103-AMEAÇA	Quinta-Feira	1	Não
7	PONTA AGUD	JOSE ISIDORO COF	Madrugada	C210-FURTO	Segunda-Feira	1	Sim
8	PONTA AGUD	JOSE ISIDORO COF	Noite	C211-FURTO	Terça-Feira	1	Não
9	ITOUPAVA NC	JOSE MANOEL DEF	Tarde	C221-ROUBO	Sábado	1	Não
10	FORTALEZA	JOSE MANOEL GOI	Noite	C505-EMBRIA	Quinta-Feira	1	Não
11	DA VELHA	JOSE MARTINS	Noite	C211-FURTO	Quinta-Feira	1	Não
12	BOA VISTA	LAGES	Noite	C211-FURTO	Sexta-Feira	1	Sim
13	DA VELHA	LARANJEIRA	Noite	C201-ARROM	Sábado	1	Não
14	DA VELHA	JOSE REUTER	Madrugada	C103-AMEAÇA	Quarta-Feira	1	Sim
15	DA VELHA	JOSE REUTER	Madrugada	C211-FURTO	Quarta-Feira	1	Não
16	DA VELHA	JOSE REUTER	Tarde	C201-ARROM	Quinta-Feira	1	Não
17	DA VELHA	JOSE REUTER	Noite	C211-FURTO	Quinta-Feira	1	Não
18	DA VELHA	JOSE REUTER	Manhã	C505-EMBRIA	Quinta-Feira	1	Não
19	DA VELHA	JOSE REUTER	Noite	C610-DISPARI	Quinta-Feira	1	Sim
20	DA VELHA	JOSE REUTER	Noite	C201-ARROM	Sábado	1	Não

Quando o usuário clicar no botão Nova Árvore, será iniciado o processamento do algoritmo de construção da árvore de decisão.

5.2.6 INTERPRETAÇÃO DO CONHECIMENTO

Nesta etapa os padrões gerados pelo algoritmo de *Data Mining* serão interpretados através de visualizações. Este aplicativo possui a forma de visualização com estrutura se/então. A figura 15 apresenta a tela de visualização se/então.

FIGURA 15 - TELA DE VISUALIZAÇÃO SE/ENTÃO



6 CONCLUSÕES

Com o objetivo de se extrair conhecimento por meio da interpretação de dados, aplicou-se a tecnologia de KDD. Foram estudados os processos de desenvolvimento, com ênfase ao processo de mineração de dados que leva ao descobrimento de padrões que é o objetivo principal deste trabalho.

Para desenvolver o processo de mineração de dados, utilizou-se neste aplicativo a técnica de árvore de decisão com o algoritmo C4.5 para gerar uma classificação a partir dos dados extraídos de ocorrências atendidas pela Polícia Militar. Tendo isso como base, foi possível verificar que a utilização de *Data Mining*, como um dos processos de KDD, mostrou-se bastante eficiente.

Foram realizados testes com o aplicativo construído para a execução do processo de descobrimento de padrões nos quais o aplicativo mostrou-se eficiente para a definição de modelos de classificação de dados. Contudo, reconhecendo que, alguns atributos escolhidos podem vir a comprometer a classificação, pois estão fortemente ligados, como o atributo "Rua" e o atributo "Bairro", por uma rua sempre pertencer ao mesmo bairro.

Anteriormente, os dados das ocorrências atendidas pela Polícia Militar eram apenas coletados e armazenados, sem uma utilidade definida. A partir da utilização do aplicativo desenvolvido, os dados passam a ter um propósito, sendo utilizados para auxiliar na tomada de decisão dos operadores do COPOM na distribuição de viaturas pela área de atuação do 10º Batalhão de Polícia Militar.

O aplicativo permite que os operadores passem a obter conclusões de caráter lógico, o que antes eram de caráter dedutivo. Sem este recurso, os operadores distribuía as viaturas aleatoriamente, de acordo com a experiência adquirida ao longo do tempo de serviço. Com o aplicativo desenvolvido, a distribuição das viaturas ocorre de acordo com os resultados obtidos no modelo de classificação.

Por fim, considera-se que o objetivo principal do trabalho, "desenvolver um aplicativo, utilizando técnicas de KDD para o reconhecimento de padrões a partir de dados de ocorrências atendidas pela Polícia Militar" foi atingido.

6.1 LIMITAÇÕES

O aplicativo construído apresenta as seguintes limitações:

- a) a fonte de dados que o aplicativo utiliza para processamento é fixa, desta forma não permitindo ao usuário mudar a fonte de dados;
- b) os atributos envolvidos no processo de classificação possuem domínio fixo;
- c) o atributo alvo que o aplicativo utiliza para realizar a classificação dos dados é fixo.

6.2 SUGESTÕES

Sugere-se o estudo de *Data Mining* aplicando em outras áreas para a tomada de decisão, como o uso de outras técnicas.

Sugere-se também o desenvolvimento de um aplicativo com uma fonte de dados variável, para que o usuário possa escolher a fonte de dados que deseja submeter ao modelo de classificação.

Sugere-se também o desenvolvimento de uma aplicação que tenha uma opção de visualização gráfica da árvore de decisão.

Outro ponto importante seria o desenvolvimento de um aplicativo que possibilite ao usuário a escolha do atributo alvo.

REFERÊNCIAS

- 10º BATALHÃO DE POLÍCIA MILITAR. **Ficha de Ocorrência**. Material destinado aos Policiais Militares para preenchimento dos dados relativos às ocorrências atendidas. 2003. 01 f.
- ALVIM, J.E. Carreira. **Violência: frente e verso**, [S.l.], [2002?]. Disponível em: <<http://www.trf2.gov.br/emarf/artigoviolenca.html>>. Acesso em: 03 jan. 2003.
- ANTUNES, M. Claudia. **Árvores de decisão**, [S.l.], [2002?]. Disponível em: <http://mega.ist.utl.pt/~ic-apr/documentos/aulas/aula2_arvores_decisao.pdf>. Acesso em: 22 abr. 2003.
- BARAZETTI, Mônica Cristina. **Data Mining**, [S.l.], [2001?]. Disponível em: <<http://www.pr.gov.br/batebyte/edicoes/1999/bb90/estagiario.htm>>. Acesso em: 04 abr. 2003.
- BERRY, Michael. **Data Mining solutions: methods and tools for solving real-world problems**. USA: John Wiley & Sons, Inc., 1998.
- BORLAND. **Paradox para windows: primeiros passos**. USA: Borland International, 1995.
- BRAGA, Marco Aurélio. **Família mantida refém em panificadora no Norte**. Jornal A Notícia, Joinville, 27 jan. 2003. Polícia, p. A9.
- CANTU, Marco. **Dominando o Delphi 5: a bíblia**. São Paulo: Makron Books, 2000.
- CERVO, Leonardo Vieira; GEYER, Cláudio Fernando Resin; BRUSSO, Marcos José. **Ferramenta para descoberta de regras de associação em bancos de dados relacionais na área da saúde**, [S.l.], [2002?]. Disponível em: <http://www.sis.org.ar/tlibres/D/d_6.PDF>. Acesso em: 23 abr. 2003.
- COMPOLT, Geandro Luis. **Sistemas de informação executiva baseado em um Data Mining utilizando a técnica de árvores de decisão**. 1999. Trabalho de Conclusão de Curso. (Graduação em Ciências da Computação) – Departamento de Sistemas e Computação, FURB, Blumenau.

FAYYAD, M. Usama. et al. *Advances in knowledge discovery and data mining*. Califórnia: Massachusetts Institute of Technology, 1996.

GAMA, João. **Árvores de decisão**, [S.l.], [2002]. Disponível em:
<http://www.liacc.up.pt/~jgama/Aulas_ECD/arv.pdf>. Acesso em: 22 abr. 2003.

GROTH, Robert. *Data Mining: a hands-on approach for business professionals*. New Jersey: Prentice Hall, 1997.

HARMON, Paul. **Sistemas Especialistas**. Rio de Janeiro: Campus, 1988.

HUBNER, Jomi Fred. **Árvores de Decisão**. Material distribuído pelo professor da disciplina de Inteligência Artificial ministrada no I semestre de 2002 no curso de Ciências da Computação da Universidade Regional de Blumenau, 2002. 01 f.

LIEBSTEIN, Lourdes H.. **Data Mining** – teoria e prática, Porto Alegre, [2001?]. Disponível em: < http://www.inf.ufrgs.br/~clesio/cmp151/cmp15120021/artigo_lourdes.pdf >. Acesso em: 31 mar. 2003.

MALDONADO, T. Maria. **Construindo a paz: caminhos da prevenção da violência**, [S.l.], [2002?]. Disponível em:
<http://www.bapera.com.br/REVISTA/Psicoterapia/Construindo_a_paz.htm>. Acesso em: 09 jan. 2003.

MITCHELL, M. Tom. *Machine learning*. Portland: WCB/McGraw-Hill Companies, 1997.

NARDELLI, Bianca. **Protótipo de um sistema de informação gerencial aplicado a central de informação aos alunos da FURB utilizando Data Mining**. 2000. Trabalho de Conclusão de Curso. (Graduação em Ciências da Computação) – Departamento de Sistemas e Computação, FURB, Blumenau.

QUONIAM, Luc et al. **Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil**, [S.l.], [2002?]. Disponível em:
<<http://www.ibict.br/cionline/300201/3020104.pdf>>. Acesso em: 09 jan. 2003.

RODRIGUES, Alexandre Medeiros. **Técnicas de data mining classificadas do ponto de vista do usuário**. Rio de Janeiro: UFRJ, 2000. Tese.

RUGGIERI, Salvatore. *Efficient C4.5*, [S.l.], [2000]. Disponível em:
<<http://www.kdd.di.unip.it>>. Acesso em: 11 abr. 2003.

TWO CROWS CORPORATION. *Introduction to data mining and knowledge discovery*, 2^o
ed., [1998]. Disponível em: <<http://www.geocities.com/vienna/9128/indm2e.pdf>>. Acesso
em: 15 abr. 2003.

YOURDON, Edward. **Análise estruturada moderna**. Rio de Janeiro: Campus, 1990.