SISTEMA INTELIGENTE PARA AVALIAÇÃO DE RISCOS
EM VIAS DE TRANSPORTE TERRESTRE

Custódio Gouvêa Lopes da Motta

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA CIVIL.

Aprovada por:

_____
Prof. Nelson Francisco Favilla Ebecken, D.Sc

_____
Prof. Alexandre Gonçalves Evsukoff, Dr.

_____
Prof. Carlos Cristiano Hasenclever Borges, D.Sc

RIO DE JANEIRO, RJ - BRASIL
JUNHO DE 2004

MOTTA, CUSTÓDIO GOUVÊA LOPES DA

Sistema Inteligente para Avaliação de Riscos em Vias de Transporte Terrestre [Rio de Janeiro] 2004.

IX, 120 p. 29,7 cm (COPPE/UFRJ, M. Sc., Engenharia Civil, 2004)

Tese - Universidade Federal do Rio de Janeiro, COPPE

1. Mineração de Dados

2. Geotecnia

3. Classificador Bayesiano

I. COPPE/UFRJ II. Título (série)

Aos meus pais,

Augusto e Gicélia, por tudo que eles me ensinaram

e à minha querida avó Mercedes, que do alto dos seus 98 anos,

não deixava de demonstrar seu carinho, sempre perguntando:

"Amanhã, você vai à sua aula de mestrado no Rio?".

AGRADECIMENTOS

Ao Professor Nelson Francisco Favilla Ebecken pela amizade, ensinamentos e orientação precisa e objetiva.

Aos Professores Alexandre Gonçalves Evsukoff e Carlos Cristiano Hasenclever Borges por aceitarem participar da Banca de avaliação desta tese.

Aos engenheiros Luiz Ernesto Bernardino Alves Filho e Raul Bomilcar do Amaral pelo apoio fundamental no desenvolvimento deste trabalho.

Ao Professor Antonio Carlos Salgado Guimarães do Laboratório Nacional de Computação Científica (LNCC/CNPq), pelos ensinamentos da linguagem C++ que muito contribuíram nas implementações desta tese.

Ao meu filho Lucas e à minha mulher Rachel pelo incentivo permanente e compreensão pelas horas extras dispensadas a este trabalho.

A funcionária do Programa de Engenharia Civil, Estela Sampaio, pela competência e presteza de seus apoios.

A meus colegas da UFJF, em especial aos Professores Clícia (DCC), Hélio (DCC), Raul (DCC), Rubens(DCC) e sua esposa Sônia (História) e, ainda, Wilhelm(Matemática), por todo o incentivo dispensado.

A meus parentes, especialmente meus irmãos Marta, Margô e Bráulio que sempre mostraram sua preocupação e incentivo para a conclusão deste trabalho.

A meus amigos Francisco (Chiquinho) e Carlos Carreira pelo incentivo paralelo, que tornou possível a continuidade e conclusão desta Tese.

A meus colegas de mestrado Gilberto, Hélcio e Ruy, que juntos constituímos uma grande turma, tão especial, que em termos de DM, ela poderia ser caracterizada como “inclassificável”.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M. Sc.)

SISTEMA INTELIGENTE PARA AVALIAÇÃO DE RISCOS EM VIAS DE TRANSPORTE TERRESTRE

Custódio Gouvêa Lopes da Motta

Junho/2004

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

O presente trabalho tem por objetivo desenvolver um sistema inteligente, usando uma atividade preditiva de mineração de dados, implementando um algoritmo de classificação para solução do problema de avaliação de riscos em vias de transporte terrestre. É feita uma análise dos métodos de classificação e selecionado o classificador Bayesiano simples como o algoritmo adotado. Os programas de treinamento e classificação foram submetidos a diversos testes para a verificação de sua acurácia de predição e seu desempenho. Finalmente, os programas foram aplicados numa base de dados real para predição dos riscos de acidentes geotécnicos num trecho da via, possibilitando a tomada de decisões em relação a diversos aspectos da conservação e manutenção de sua infra-estrutura.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M. Sc.)

INTELLIGENT SYSTEM FOR EVALUATION OF RISKS IN ROADS OF TERRESTRIAL TRANSPORT

Custódio Gouvêa Lopes da Motta

June/2004

Advisor: Nelson Francisco Favilla Ebecken

Department: Civil Engineering

This work intends to develop an intelligent system using Data Mining predictive activity, implementing a classification algorithm for solving problems of risk evaluation in roads for terrestrial transport. Several methods of classification were analyzed and the Simple Bayesian Classifier was selected as the algorithm to be used. The trainning and classification programs were tested in order to acertain their prediction accuracy and performance. Finally these programs were applied to a real data base in order to predict the risks of geotechnical acidents on the road, enabling decisions regarding the conservation and maintenance of its infrastructure to be made.

ÍNDICE

FIGURAS

TABELAS

LISTAGENS

CAPÍTULO 1
INTRODUÇÃO

A descoberta de conhecimento em bases de dados, também chamada de KDD (*Knowledge Discovery in Databases*) pode ser definida como o processo de identificação de padrões embutidos nos dados. Além disso, os padrões identificados devem ser válidos, novos, potencialmente úteis e compreensíveis (FAYYAD, PIATETSKY-SHAPIRO & SMITH, 1996[1]).

As pesquisas relativas a este processo ganharam rápido crescimento a partir da última década, motivadas pela evolução da tecnologia que vem permitindo a coleta, o armazenamento e o gerenciamento de quantidades cada vez maiores de dados (FAYYAD, PIATETSKY-SHAPIRO & SMITH, 1995, 1996[2]).

Outro motivador deste crescimento é a ampliação das áreas de aplicações de KDD. Como exemplos de áreas de aplicações, podem ser citadas (CUROTTO, 2003): bancária (aprovação de crédito), ciências e medicina (descoberta de hipóteses, predição, classificação, diagnóstico), comercialização (segmentação, localização de consumidores, identificação de hábitos de consumo), engenharia (simulação e análise, reconhecimento de padrões, processamento de sinais e planejamento), financeira (apoio para investimentos, controle de carteira de ações), gerencial (tomadas de decisão, gerenciamento de documentos), *Internet* (ferramentas de busca, navegação, extração de dados), manufatura (modelagem e controle de processos, controle de qualidade, alocação de recursos) e segurança (detecção de bombas, icebergs e fraudes).

O processo de descoberta de conhecimento em base de dados envolve diversas etapas, destacando-se a seguinte seqüência (FAYYAD, PIATETSKY-SHAPIRO & SMITH, 1996[1]):

1. Consolidação de dados: onde os dados são obtidos a partir de diferentes fontes (arquivos texto, planilhas ou bases de dados) e consolidados numa única fonte.

2. Seleção e pré-processamento: nesta etapa, diversas transformações podem ser aplicadas sobre os dados, como reduzir o número de exemplos, de

atributos ou de intervalos de atributos, normalizar valores etc., de forma a obter, no final, um conjunto de dados preparados para utilização dos algoritmos de mineração.

3. Mineração de dados ou DM (*Data Mining*): é a etapa de extração de padrões propriamente dita, onde, primeiramente, é feita a escolha da tarefa de mineração conforme os objetivos desejáveis para a solução procurada, isto é, conforme o tipo de conhecimento que se espera extrair dos dados. A Figura 1.1 ilustra as tarefas de mineração organizadas em atividades preditivas e descritivas.

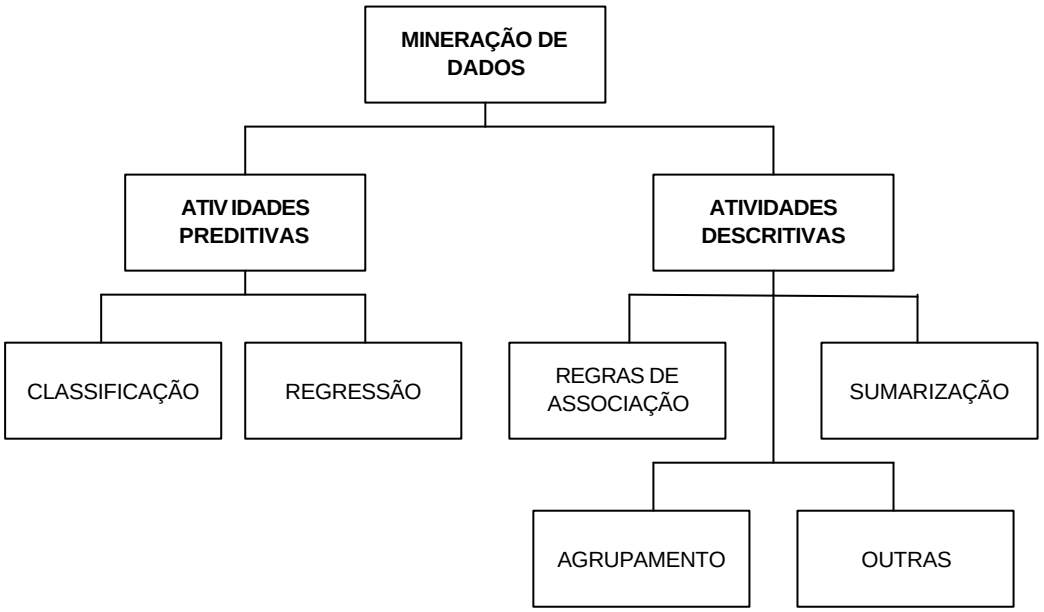

Figura 1.1. Tarefas de Mineração de Dados (REZENDE, PUGLIESI, MELANDA & DE PAULA, 2003)

As atividades preditivas buscam identificar a classe de uma nova amostra de dados, a partir do conhecimento adquirido de um conjunto de amostras com classes conhecidas. Já as atividades descritivas trabalham com um conjunto de dados que não possuem uma classe determinada, buscando identificar padrões de comportamento comuns nestes dados.

Em seguida, é escolhido o algoritmo que atenda a tarefa de mineração eleita e que possa representar satisfatoriamente os padrões a serem encontrados. Os

algoritmos de mineração mais comuns são: Algoritmos Estatísticos, Algoritmos Genéticos, Árvores de Decisão, Regras de Decisão, Redes Neurais Artificiais, Algoritmos de Agrupamento, Lógica *Fuzzy*.

A mineração de dados é na verdade uma atividade interdisciplinar pela diversidade de tecnologias que podem estar envolvidas. A Figura 1.2 sintetiza os assuntos envolvidos com DM.

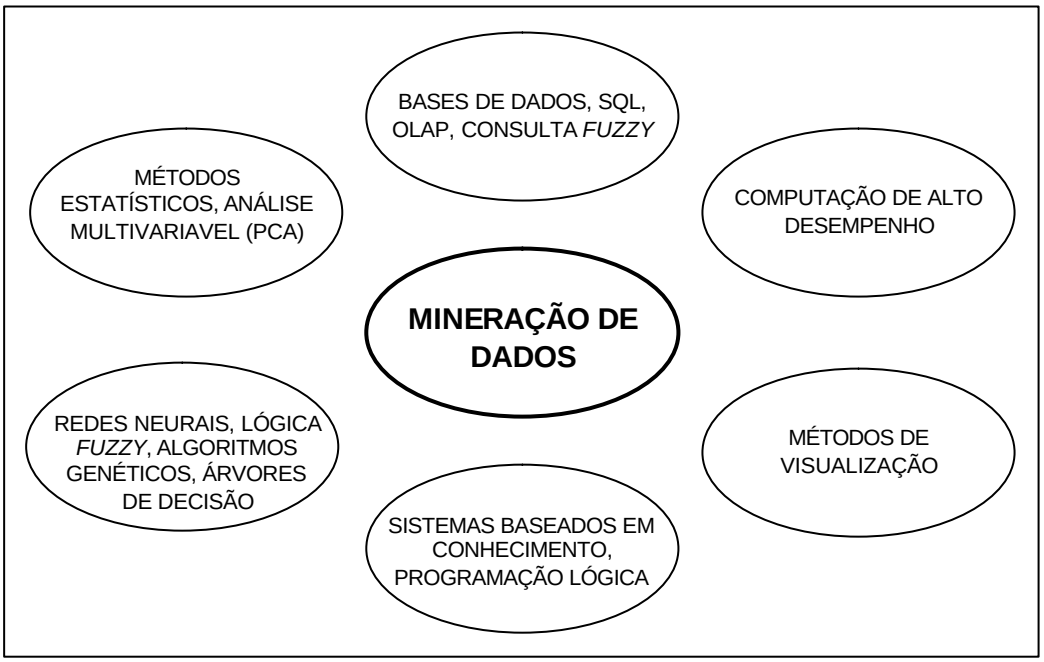

Figura 1.2. Assuntos envolvidos com Mineração de Dados (Han & Kamber, 2001)

4. Avaliação e interpretação: nesta etapa são avaliados o desempenho e a qualidade das regras extraídas, bem como verificada a facilidade de interpretação dessas regras.

Deve-se destacar que o processo de KDD ocupa uma posição no ciclo de solução do problema, não se esgotando por si só. Este ciclo é também conhecido como ciclo virtuoso e é apresentado na Figura 1.3.

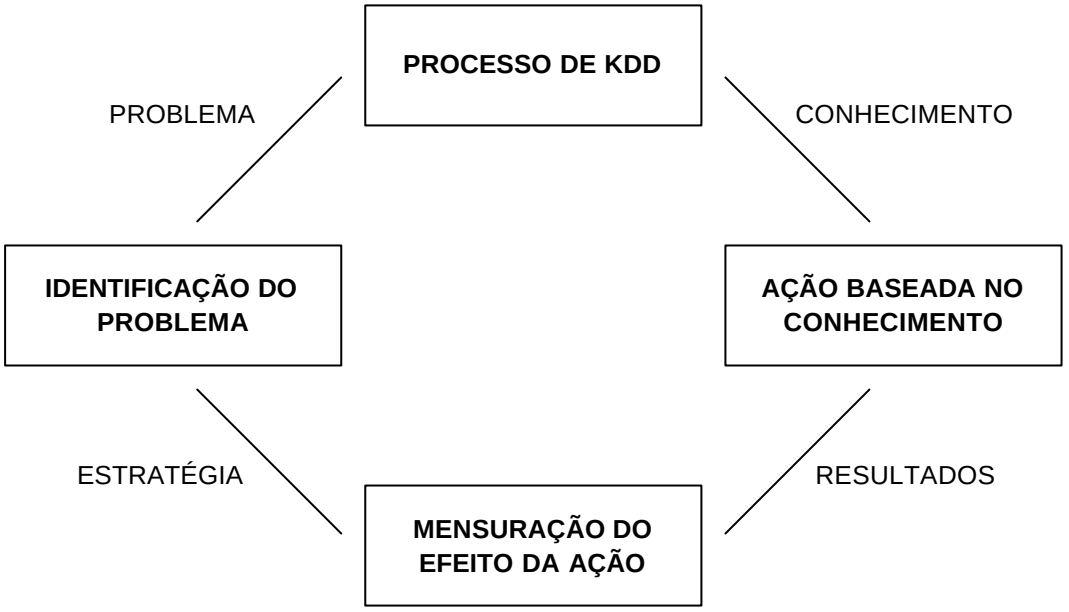

Figura 1.3. Ciclo Virtuoso (SILVER, 1998)

A utilização do conhecimento obtido no processo de KDD é realizada através de um sistema inteligente ou de um ser humano como forma de apoio à tomada de decisão.

Entende-se como inteligente um sistema computacional que possui habilidades inteligentes e sabe como elas modelam tarefas específicas. Entre essas habilidades, está a de usar conhecimento para resolver problemas (REZENDE, 2003).

O objetivo principal desta tese é desenvolver um sistema inteligente, usando uma atividade preditiva de DM, implementando um algoritmo de classificação para solução do problema de avaliação de riscos em vias de transporte terrestre.

No Capítulo 2, é apresentado um estudo da atividade preditiva de classificação de dados e os seus métodos mais usuais.

Os algoritmos de treinamento e predição do classificador CBS (Classificador Bayesiano Simples), as suas implementações e experimentos computacionais realizados para avaliar o comportamento dessas implementações são detalhados no Capítulo 3.

O Capítulo 4 apresenta um problema de geotecnia, onde o classificador CBS é utilizado sobre uma base de dados real, como sistema inteligente para avaliação de riscos.

As conclusões do trabalho desenvolvido são descritas no Capítulo 5.

Finalizando, são listadas todas as referências citadas no texto, como também as referências complementares usadas como base de formação para a composição desta tese.

Como complemento, três anexos são incluídos para detalhar: os algoritmos do classificador CBS (ANEXO I); as suas implementações (ANEXO II), e os experimentos computacionais realizadas em diversas bases de dados (ANEXO III).

CAPÍTULO 2
CLASSIFICAÇÃO DE DADOS

As atividades preditivas de mineração de dados são formas de análises em bases de dados usadas para extrair padrões que descrevem tendências futuras de dados. Essas atividades podem ser divididas em classificação e regressão. A diferença básica é que enquanto a classificação prediz valores discretos (classes), a regressão modela funções contínuas.

Como o problema proposto no Capítulo 4 trata da predição de valores discretos, será apresentada a seguir, a atividade de classificação de dados, onde, inicialmente, será descrito o funcionamento do processo de uma forma geral, os critérios para avaliação dos métodos e alguns cuidados na preparação dos dados. Em seguida, serão apresentados os três métodos considerados mais usuais: indução de árvore de decisão, redes neurais artificiais e classificação Bayesiana.

2.1 O Processo de Classificação

O processo de classificação de dados é realizado em dois passos. O primeiro, conhecido como treinamento ou aprendizado, caracteriza-se pela construção de um modelo que descreve um conjunto predeterminado de classes de dados. Essa construção é feita analisando as amostras de uma base de dados, onde as amostras são descritas por atributos e cada uma delas pertence a uma classe predefinida, identificada por um dos atributos, chamado atributo rótulo da classe ou, simplesmente, classe. O conjunto de amostras usadas neste passo é o conjunto de treinamento, dados de treinamento ou amostras de treinamento.

As formas mais comuns de representar o conhecimento (ou padrões) aprendido na fase de treinamento são regras de classificação, árvores de decisão ou formulações matemáticas. Este conhecimento pode ser usado para predizer as classes de amostras desconhecidas futuras, bem como pode permitir um melhor entendimento dos conteúdos da base de dados.

No segundo passo, o modelo construído é testado, isto é, o modelo é usado para classificação de um novo conjunto de amostras, independentes daquelas usadas no treinamento, chamado conjunto de teste, dados de teste ou amostras de teste. Como este conjunto também possui as classes conhecidas, após a classificação, pode-se calcular o percentual de acertos, comparando as classes preditas pelo modelo com as classes esperadas (ou conhecidas). Este percentual é conhecido como acurácia ou precisão do modelo para o conjunto de teste em questão.

Se a acurácia for considerada aceitável, o modelo pode ser usado na classificação de amostras desconhecidas futuras, ou seja, amostras cuja classe não é conhecida.

As Figuras 2.1 e 2.2 exemplificam o treinamento e o teste, respectivamente, usando um algoritmo de geração de regras de classificação (Han & Kamber, 2001).

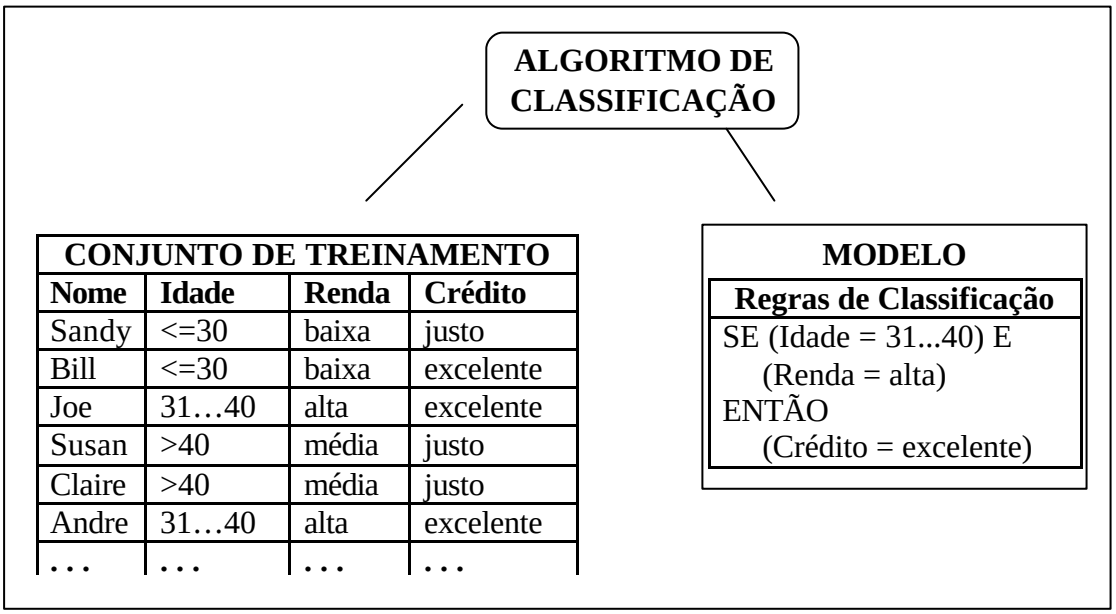

Figura 2.1. Treinamento

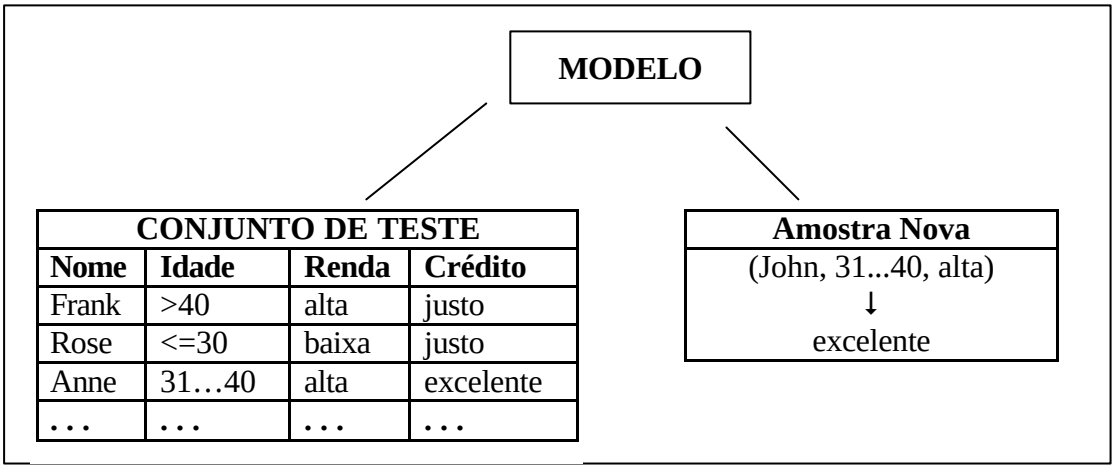

Figura 2.2. Teste e Classificação

2.2 Avaliação dos Métodos de Classificação

A partir da identificação da necessidade de resolver um problema de classificação, deve-se escolher um dos diversos métodos existentes. Para isso, pode-se comparar esses métodos conforme os seguintes critérios (Han & Kamber, 2001):

- Acurácia de Predição: é a habilidade do modelo predizer corretamente a classe de amostras desconhecidas.
- Desempenho: critério relativo aos custos computacionais envolvidos na geração e na utilização do modelo.
- Robustez: é a habilidade do modelo fazer predições corretas em amostras com atributos faltando ou com ruídos.
- Escalabilidade: é a habilidade de construir um modelo eficiente a partir de grandes quantidades de dados.
- Interpretabilidade: é a habilidade de tornar compreensível o conhecimento gerado pelo modelo.

Estes critérios serão detalhados para cada método de classificação apresentado ao longo desse capítulo.

2.3 Preparação dos Dados para Classificação

Visando melhorar a acurácia, o desempenho e a escalabilidade do modelo, pode-se executar um pré-processamento sobre os dados, de forma a prepará-los para a classificação. Essa preparação envolve as seguintes tarefas:

- Limpeza: são técnicas que devem ser usadas para garantir a qualidade dos dados. As mais comuns são eliminação de erros gerados no processo de coleta (erros de digitação ou de leitura por sensores), tratamento de atributos faltando e eliminação ou redução de ruídos.

- Análise de relevância: é uma análise realizada sobre os atributos das amostras de treinamento para identificar e excluir atributos irrelevantes ou redundantes, que em nada contribuem no processo de classificação. A diminuição do tamanho das amostras com essa exclusão concorre para melhorar o desempenho e a escalabilidade do modelo.

- Transformação: as transformações mais comuns aplicáveis aos dados de treinamento são: resumo, onde um conjunto de atributos é agrupado para formar resumos; discretização, onde dados contínuos são transformados em discretos da forma baixo, médio e alto, por exemplo; transformação de tipo, para que o dado fique numa forma mais apropriada para classificação, e normalização, aplicada sobre dados contínuos para colocá-los em intervalos determinados de valores, por exemplo, entre -1 e 1.

2.4 Classificação por Indução de Árvore de Decisão

Árvore de decisão é um classificador simbólico representado como estrutura de árvore, onde cada nó interno indica o teste em um atributo, cada ramo representa um resultado do teste, e os nós terminais representam classes ou distribuições de classe. O topo da árvore representa a raiz e os nós terminais, as folhas.

Para classificar uma amostra desconhecida, os valores de seus atributos são testados ao longo da árvore, traçando um caminhamento da raiz até um nó folha que prediz a classe da amostra.

2.4.1 Indução de Árvore de Decisão

O algoritmo básico para indução de árvore de decisão é um algoritmo que constrói árvores de decisão recursivamente, de cima para baixo, através de divisão e conquista.

A idéia geral do algoritmo de aprendizado de árvore de decisão é testar o primeiro atributo mais importante chamado atributo divisor ou de teste, isto é, aquele que faz a maior diferença para a classificação da amostra. A intenção é encontrar a classificação correta com o menor número de testes, pois quanto menor a árvore, melhor o desempenho da classificação.

A Figura 2.3 compara a importância dos atributos TEMPO e TEMPERATURA, utilizando as amostras de treinamento da Tabela 2.1 a seguir.

	Atributos				**Classe**
DIA	**TEMPO**	**TEMPERATURA**	**UMIDADE**	**VENTO**	**JOGA-TÊNIS**
1	Sol	Quente	Alta	Fraco	Não
2	Sol	Quente	Alta	Forte	Não
3	Nublado	Quente	Alta	Fraco	Sim
4	Chuva	Moderado	Alta	Fraco	Sim
5	Chuva	Frio	Normal	Fraco	Sim
6	Chuva	Frio	Normal	Forte	Não
7	Nublado	Frio	Normal	Forte	Sim
8	Sol	Moderado	Alta	Fraco	Não
9	Sol	Frio	Normal	Fraco	Sim
10	Chuva	Moderado	Normal	Fraco	Sim
11	Sol	Moderado	Normal	Forte	Sim
12	Nublado	Moderado	Alta	Forte	Sim
13	Nublado	Quente	Normal	Fraco	Sim
14	Chuva	Moderado	Alta	Forte	Não

Tabela 2.1. Situações Favoráveis ou não para Jogar Tênis (MITCHELL, 1997).

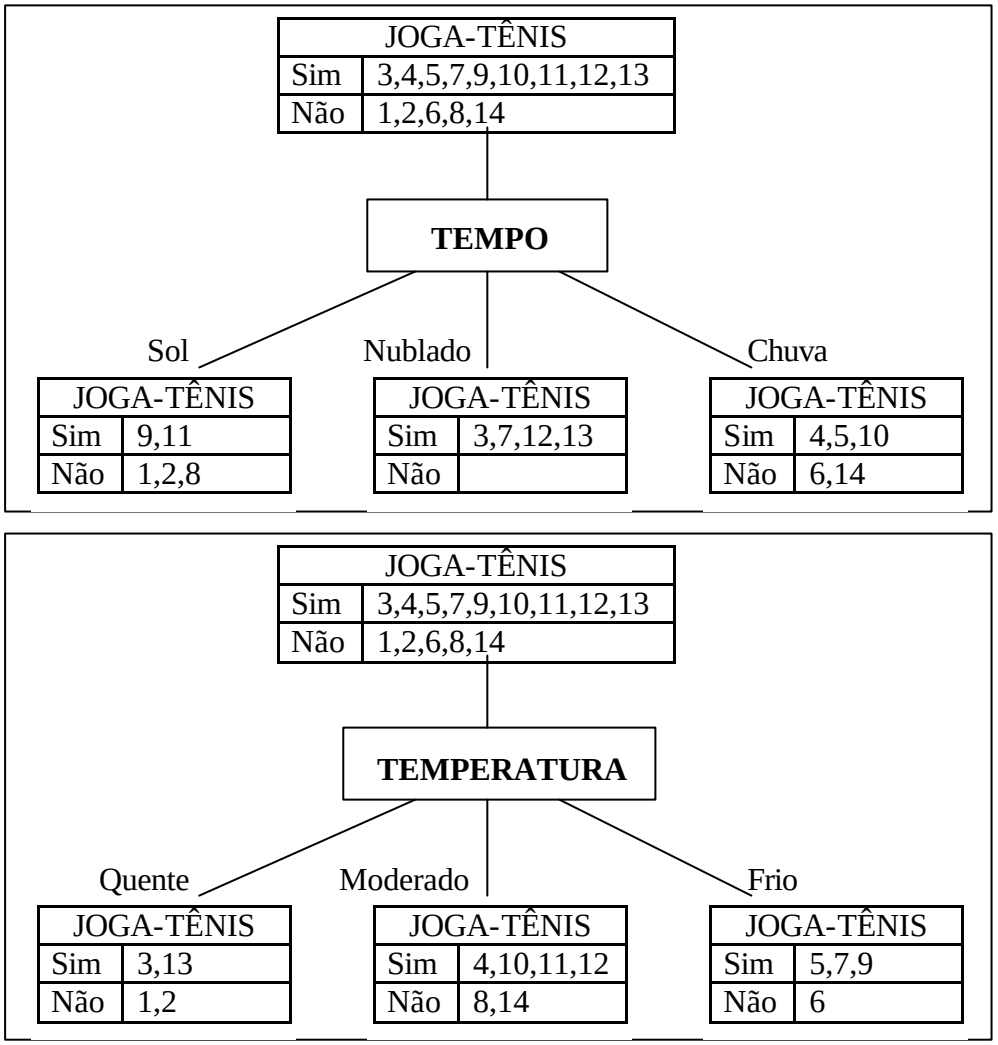

Figura 2.3. Comparação da importância dos atributos TEMPO e TEMPERATURA.

Observa-se que o atributo TEMPO é mais importante que TEMPERATURA, uma vez que se TEMPO = Nublado, então a classificação já alcança a folha JOGA-TÊNIS = Sim. Se o valor de TEMPO for Sol ou Chuva, serão necessários outros testes.

Após a definição do primeiro atributo, cada saída ou ramo se torna, recursivamente, um novo problema de aprendizagem de árvore decisória, com menos amostras e menos atributos. O processo recursivo é finalizado quando ocorrer uma das seguintes situações(Han & Kamber, 2001):

- Se houver somente um ramo de saída, isto é, se existir somente uma classe para todas as amostras, então o nó se torna uma folha rotulada com essa classe.

- Se não resta mais nenhum atributo para dividir amostras de classes distintas, então o nó se torna uma folha rotulada com a classe mais comum entre essas amostras.

- Não há nenhuma amostra para um determinado ramo de saída. Neste caso, é criada uma folha rotulada com a classe da maioria das amostras.

Uma versão de um algoritmo chamado ID3 é apresentada na Listagem 2.1 a seguir.

Descrição:
Chamada: GERA_ÁRVORE (*amostras*, *list_atrib*).
Objetivo: Gerar uma árvore de decisão a partir de um conjunto de dados de treinamento.
Entrada: *amostras*: conjunto de amostras de treinamento, representadas por atributos discretos
list_atrib: conjunto de atributos candidatos a teste.
Saída: Uma árvore de decisão.

Algoritmo:

```
início
    crie um nó N;
    se amostras são todas da classe C
        então retorne N como nó folha rotulada com a classe C;
    senão
        se list_atrib estiver vazia
            então retorne N como nó folha rotulada com a classe mais comum
                  entre as amostras;
        senão
            selecione o atributo de teste, isto é, o atributo da list_atrib com o
            maior ganho de informação;
            rotule o nó N com o atributo de teste;
            para cada valor possível a_i do atributo de teste faça
                inclua um ramo a partir do nó N, com a condição "atributo de teste
                = a_i";
                atribua a s_i o subconjunto de amostras contido em amostras que
                possuem o atributo de teste = a_i;
                se s_i estiver vazia
                    então inclua uma folha rotulada com a classe mais comum entre
                          as amostras;
                senão
                    exclua o atributo de teste de list_atrib;
                    inclua o nó retornado por GERA_ÁRVORE (s_i, list_atrib);
fim.
```

Listagem 2.1. Algoritmo de Criação de Árvore de Decisão (Han & Kamber, 2001).

O problema central do algoritmo ID3 consiste na escolha do atributo de teste entre todos os envolvidos com o nó em questão. Para solução desse problema, o algoritmo usa um método estatístico, selecionando o atributo que possui o maior ganho de informação (*Ganho*), isto é, aquele que melhor classifica o conjunto de treinamento.

Este método tem por objetivo minimizar o número de testes necessários para classificar uma amostra, onde o valor do *Ganho* de um atributo é calculado da seguinte forma (Han & Kamber, 2001):

Seja S um conjunto de s amostras de dados. Supondo que S possui m classes distintas C_i (i = 1, ..., m) e sendo s_i o número de amostras de S com classe igual a C_i, a informação necessária para classificar uma determinada amostra é dada por:

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i), \qquad (2.1)$$

onde p_i é a probabilidade que uma amostra qualquer pertença a classe C_i e é calculada como s_i / s. O uso da função log na base 2, representa o número de bits necessários para codificar a informação.

Se um atributo A possui v valores distintos possíveis, $\{a_1, a_2, ..., a_v\}$, ele pode ser usado para particionar S em v subconjuntos, $\{S_1, S_2, ..., S_v\}$, onde S_j contém todas as amostras de S com o atributo A igual a a_j. Se A for selecionado como o atributo de teste, então estes subconjuntos seriam distribuídos pelos v ramos descendentes do nó rotulado com A.

Considerando s_{ij} o número de amostras do subconjunto S_j com classe igual a C_i, pode-se calcular a entropia (expectativa de informação baseada na divisão de S no teste do atributo A) como:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj}). \qquad (2.2)$$

O termo $(s_{1j} + \cdots + s_{mj})/s$ age como o peso do j-ésimo subconjunto e corresponde ao número de amostras deste subconjunto que possuem o atributo A igual a a_j, dividido pelo número total de amostras de S.

Para um determinado subconjunto S_j, pode-se calcular:

$$I(s_{1j}, s_{2j}, \ldots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2 (p_{ij}), \quad (2.3)$$

onde, p_{ij} é a probabilidade que uma amostra de S_j pertença a classe C_i e é calculada como s_{ij} / $|S_j|$.

Finalmente, calcula-se o ganho de informação como:

$$Ganho(A) = I(s_1, s_2, \ldots, s_m) - E(A). \quad (2.4)$$

Para exemplificar a indução de uma árvore de decisão, será utilizado o conjunto de treinamento da Tabela 2.1, onde se observa que o atributo de classe, JOGA-TÊNIS, possui dois valores distintos: Sim e Não (m = 2). Do total de 14 (S) amostras, 9 (s_1) são da classe Sim e 5 (s_2) da classe Não.

A informação necessária para classificar uma determinada amostra é calculada pela equação (2.1) e resulta em:

$$I(s_1, s_2) = I(9,5) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0{,}940$$

Em seguida, para calcular a entropia de cada atributo, deve-se, primeiramente, quantificar a distribuição de amostras por classe, considerando cada instância de cada atributo (s_{ij}) e calcular a informação esperada para cada uma destas distribuições, usando a equação (2.3). A Tabela 2.2 a seguir apresenta esses valores.

Atributo	Instância	Classe	Amostras	Informação
TEMPO	Sol	Sim	$s_{11} = 2$	$I(s_{11}, s_{21}) = 0,971$
		Não	$s_{21} = 3$	
	Nublado	Sim	$s_{12} = 4$	$I(s_{12}, s_{22}) = 0$
		Não	$s_{22} = 0$	
	Chuva	Sim	$s_{13} = 3$	$I(s_{13}, s_{23}) = 0,971$
		Não	$s_{23} = 2$	
TEMPERATURA	Quente	Sim	$s_{11} = 2$	$I(s_{11}, s_{21}) = 1$
		Não	$s_{21} = 2$	
	Moderado	Sim	$s_{12} = 4$	$I(s_{12}, s_{22}) = 0,918$
		Não	$s_{22} = 2$	
	Frio	Sim	$s_{13} = 3$	$I(s_{13}, s_{23}) = 0,811$
		Não	$s_{23} = 1$	
UMIDADE	Alta	Sim	$s_{11} = 3$	$I(s_{11}, s_{21}) = 0,985$
		Não	$s_{21} = 4$	
	Normal	Sim	$s_{12} = 6$	$I(s_{12}, s_{22}) = 0,592$
		Não	$s_{22} = 1$	
VENTO	Forte	Sim	$s_{11} = 3$	$I(s_{11}, s_{21}) = 1$
		Não	$s_{21} = 3$	
	Fraco	Sim	$s_{12} = 6$	$I(s_{12}, s_{22}) = 0,811$
		Não	$s_{22} = 2$	

Tabela 2.2. Distribuição de Amostras por Classe e por Instância de Atributo.

Para calcular a entropia, usa-se a equação (2.2), obtendo:

E (TEMPO) = 0,694

E (TEMPERATURA) = 0,911

E (UMIDADE) = 0,789

E (VENTO) = 0,892

Finalmente, aplicando a equação (2.4), chega-se ao *Ganho* de cada atributo:

Ganho (TEMPO) = 0,940 – 0,694 = 0,246

Ganho (TEMPERATURA) = 0,940 – 0,911 = 0,029

Ganho (UMIDADE) = 0,940 – 0,789 = 0,151

Ganho (VENTO) = 0,940 – 0,892 = 0,048

O atributo TEMPO é escolhido como divisor do nó raiz da árvore, por ser o que possui o maior ganho de informação. É criado um ramo para cada instância de TEMPO e as amostras são divididas adequadamente, como mostrado em Figura 2.3. Para TEMPO igual a Sol ou Chuva, deve-se repetir o processo nas amostras respectivas, criando novos nós e assim sucessivamente, até gerar a árvore apresentada na Figura 2.4.

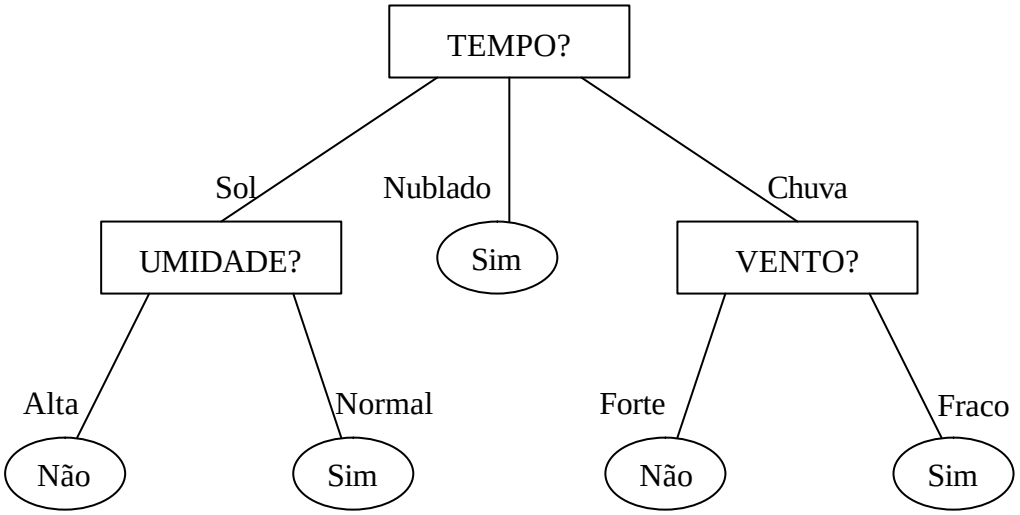

Figura 2.4. Árvore de Decisão para classificar JOGA-TÊNIS.

2.4.2 Outras Considerações sobre Árvores de Decisão

Evolução:

Os algoritmos de indução de árvore de decisão tem sido usados para classificação em uma grande variedade de domínios de aplicação, principalmente, por se tratar de um método que trabalha bem em relação aos cinco critérios de avaliação de classificadores: acurácia, desempenho, robustez, escalabilidade e interpretabilidade. Algumas versões são bastante conhecidas, como o ID3, o C4.5 e o CART.

Um dos primeiros a se tornar conhecido foi o algoritmo ID3, que utiliza uma pequena quantidade de amostras no treinamento e a árvore gerada pode processar, eficientemente, um grande conjunto de amostras desconhecidas. É um algoritmo com heurística míope, isto é, selecionado um atributo para teste em um determinado nível da árvore, jamais ocorrerá de reconsiderar a escolha. Caso exista algum atributo contínuo,

ele deverá ser dividido em intervalos, uma vez que o ID3 só trabalha com atributos discretos.

O algoritmo C4.5 é uma evolução do ID3 e apresenta um grande número de complementos, destacando-se: tratamento de amostras com valores de atributos faltando; trabalha com atributos contínuos, avaliando e dividindo-o em intervalos que particionam o conjunto de treinamento de forma a maximizar o *Ganho*, e usa uma técnica chamada Regras C4.5 que exclui (poda) nós antecedentes redundantes na classificação.

O algoritmo CART trabalha de forma semelhante ao C4.5, porém montando uma árvore de decisão binária, com formas distintas de tratamento de valores faltando, de atributos contínuos e de mecanismo de poda.

Critérios de Poda em Árvores de Decisão

Na construção de uma árvore de decisão, alguns de seus ramos podem conter anomalias causadas principalmente por ruídos nos dados de treinamento. Este tipo de problema representa uma classificação bastante específica, uma forma de discriminação ou memorização excessiva e é denominado overfitting. A utilização de métodos que usam medidas estatísticas para identificar e excluir ramos menos seguros, isto é, podar a árvore, é uma forma de controlar este problema, resultando em uma classificação mais rápida e na melhoria da habilidade de classificar amostras de teste corretamente.

Os critérios mais comuns de poda são a pré-poda e a pós-poda.

A pré-poda é realizada durante o treinamento e consiste em interromper o processo de divisão do nó em função da avaliação de um conjunto de medidas, transformando o nó em folha rotulada com a classe majoritária.

Entre as medidas mais usadas estão o número mínimo de amostras, a quantidade mínima de ganho de informação e a utilização de uma técnica de validação cruzada (cross-validation) onde o desempenho da árvore é verificado a cada divisão com um conjunto de teste.

Já a pós-poda é executada somente após a finalização do processo de construção da árvore, sendo aplicado recursivamente, de baixo para cima. Uma forma de pós-poda, por exemplo, consiste em eliminar os ramos de um nó intermediário, transformando-o em uma folha rotulada com a classe mais freqüente de seu conjunto de treinamento. Isso é feito progressivamente para cada nó intermediário, gerando um conjunto de árvores. Em seguida, um conjunto de teste independente é usado para determinar a precisão de cada uma, sendo escolhida a árvore de decisão que apresentar o melhor resultado.

Regras de Decisão

Após o treinamento da árvore de decisão, o conhecimento adquirido pode ser também representado sob a forma de um conjunto de regras de classificação ou de decisão, melhorando a legibilidade por humanos (MITCHELL, 1997).

Cada nó folha possui uma regra associada e para obtê-la, basta relacionar o conjunto de decisões tomadas, percorrendo o caminho desde a raiz da árvore até a folha.

Uma regra é apresentada no formato “SE <condição> ENTÃO <classe>”, onde a <condição> é formada pela conjunção das decisões tomadas por cada atributo dos nós intermediários ao longo do caminho e <classe> é a instância da classe rotulada na folha em questão.

Da árvore representada na Figura 2.4, por exemplo, pode-se extrair o seguinte conjunto de regras de classificação:

SE (TEMPO = Sol) E (UMIDADE = Alta) ENTÃO (JOGA-TÊNIS = Não)

SE (TEMPO = Sol) E (UMIDADE = Normal) ENTÃO (JOGA-TÊNIS = Sim)

SE (TEMPO = Nublado) ENTÃO (JOGA-TÊNIS = Sim)

SE (TEMPO = Chuva) E (VENTO = Forte) ENTÃO (JOGA-TÊNIS = Não)

SE (TEMPO = Chuva) E (VENTO = Fraco) ENTÃO (JOGA-TÊNIS = Sim)

2.5 Classificação com Redes Neurais Artificiais

O estudo de redes neurais artificiais, ou simplesmente redes neurais, surgiu a partir do conhecimento dos conceitos básicos das redes neurais biológicas. Em outras palavras, uma rede neural busca simular ou modelar a forma como o cérebro realiza uma determinada tarefa, através de seus neurônios.

Tecnicamente, uma rede neural é um sistema paralelo distribuído, composto por unidades de processamento simples, chamadas elementos processadores, nodos ou simplesmente neurônios, que têm o objetivo de calcular determinadas funções matemáticas. Os neurônios são organizados em uma ou mais camadas, interligadas por um grande número de conexões, geralmente unidirecionais, havendo um peso (peso sináptico) associado a cada conexão. Estes pesos armazenam o conhecimento representado pelo sistema e servem para ponderar as entradas recebidas por cada neurônio.

Durante a fase de treinamento, os pesos das conexões da rede vão sendo ajustados de forma que o conhecimento extraído dos dados possa ser representado internamente.

Assim, pode-se observar a semelhança entre o cérebro e uma rede neural em dois aspectos (HAYKIN, 2001):

- Aquisição de conhecimento a partir do ambiente através de um processo de aprendizagem (treinamento).

- O conhecimento adquirido é armazenado nas conexões entre os neurônios.

Embora as redes neurais necessitem da definição de muitos parâmetros como a sua estrutura e valores iniciais dos pesos, além de longos tempos de treinamento, elas são bastante empregadas na solução de problemas de classificação. Isso porque elas possuem grandes vantagens como boa capacidade de generalização (aprender com um conjunto reduzido de amostras e predizer coerentemente a classe de amostras desconhecidas) e alta tolerância a dados com ruídos. Além disto, diversos algoritmos tem sido desenvolvidos para extração de regras de classificação de redes neurais, melhorando a sua interpretabilidade.

2.5.1 Modelo de Neurônio

A partir do funcionamento básico de um neurônio biológico, foram descritos modelos para sua representação. A Figura 2.5 a seguir mostra um modelo de neurônio, onde se destacam um conjunto de conexões, um combinador linear e uma função de ativação.

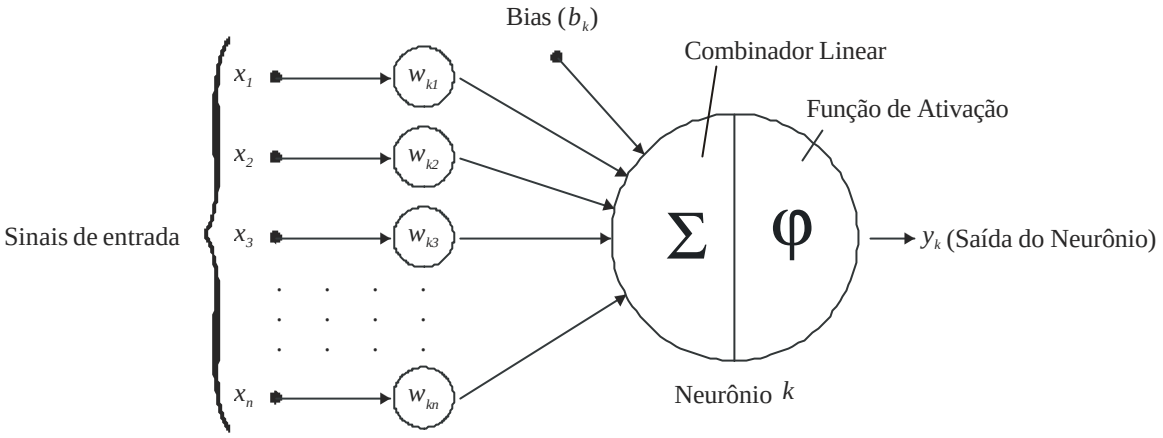

Figura 2.5. Modelo de Neurônio.

Cada conexão j possui um valor de entrada x_j (ou sinal de entrada) e um peso w_{kj}, onde o primeiro índice do peso identifica o neurônio k e o segundo, a conexão.

O combinador linear calcula a entrada líquida (u_k) do neurônio como o somatório de todas as entradas multiplicadas pelos pesos respectivos, mais o valor do bias b_k. O bias é uma espécie de excitador ou inibidor e tem o efeito de aumentar ou diminuir a entrada líquida da unidade, dependendo se o seu valor for positivo ou negativo, respectivamente.

Em seguida, a função de ativação f é aplicada sobre a entrada líquida gerando o valor de saída y_k do neurônio. Essa função é também conhecida como função restritiva, já que limita o intervalo possível da saída a um valor finito. As funções logística e tangente hiperbólica, que fornecem resultados no intervalo entre 0 e 1 e entre -1 e 1, respectivamente, são bastante usadas como funções de ativação de neurônios.

Matematicamente, isto é representado pelas duas equações abaixo.

$$u_k = \sum_{j=1}^{n} w_{kj} x_j + b_k \tag{2.5}$$

$$y_k = j(u_k) \tag{2.6}$$

2.5.2 Arquitetura de uma Rede Neural

Existem várias formas de organizar a estrutura de uma rede neural. A quantidade de camadas, a quantidade de neurônios em cada camada e a forma de conexão entre estes neurônios na rede, devem ser definidas antes do treinamento e dependem do problema que se deseja resolver.

No caso de um problema de classificação em que as soluções não são linearmente separáveis, a arquitetura da rede deverá prever, no mínimo, duas camadas, além da camada de entrada. Este tipo de rede é conhecida como rede de múltiplas camadas (ou *multilayer*), onde a primeira camada é a camada de entrada, a última é a camada de saída e as intermediárias são chamadas camadas ocultas ou escondidas.

A Figura 2.6 ilustra uma rede neural de múltiplas camadas (no caso duas) e alimentação adiante (feedforward), isto é, o fluxo de informações é sempre no sentido da camada de entrada para a de saída, não retornando a nenhum neurônio de camadas anteriores. Ela é totalmente conectada, já que cada neurônio provê entrada para cada um dos neurônios da camada seguinte.

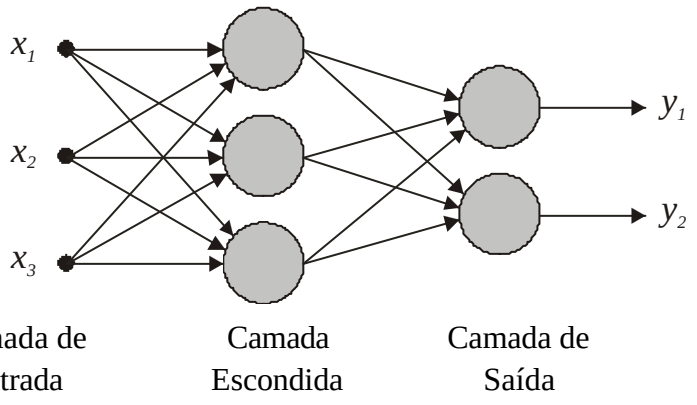

Figura 2.6. Rede Neural de Múltiplas Camadas e Alimentação Adiante.

As entradas da rede são aos valores dos atributos de uma amostra de treinamento que são entregues, simultaneamente, aos neurônios da camada de entrada, que, por sua vez, distribuem os sinais recebidos (valores dos atributos) aos neurônios da segunda camada (primeira camada escondida). Deve-se observar que, distintamente de todos os demais neurônios da rede, os neurônios da camada de entrada não são unidades processadoras como o modelo da Figura 2.5, mas, simplesmente, unidades distribuidoras dos valores de entrada para a primeira camada escondida. As saídas dos neurônios de cada camada escondida são as entradas de outra camada escondida (se existir) e assim por diante, até chegar à última, chamada camada de saída, que emite a predição para a amostra em questão. O número de camadas escondidas é arbitrário, embora na prática, normalmente seja usada somente uma.

Quanto ao número de neurônios por camada, pode-se definir:

- Para a camada de entrada: Um neurônio para cada atributo contínuo, sendo recomendável a normalização destes atributos na preparação de dados para melhorar a velocidade do treinamento. Se o atributo for discreto, usa-se um neurônio por instância e o atributo é codificado em binário. Como exemplo, para o atributo TEMPO com domínio {Sol, Nublado, Chuva} são usados três neurônios de entrada N_1, N_2 e N_3. Se na amostra de treinamento TEMPO = Chuva, serão fornecidos os valores 0, 0 e 1, respectivamente, para N_1, N_2 e N_3.

- Para uma camada escondida: Neste caso, não há nenhuma forma padrão sobre o número de neurônios que deve ser usado em cada camada escondida, devendo-se considerar nesta definição que, quanto menor o número de neurônios, melhor o desempenho da rede.

- Para a camada de saída: Havendo somente duas classes possíveis, pode-se usar um neurônio que fornecerá valor 0 para representar uma classe e 1 para a outra. No caso de mais de duas classes, usa-se um neurônio por classe, sendo que aquele que fornecer resultado igual a um, identifica a classe respectiva e os demais fornecerão saída igual a 0.

A definição da rede é um processo de tentativa e erro no sentido de alcançar uma precisão aceitável nos resultados de seu treinamento. Em algumas situações, pode

acontecer de não se atingir uma acurácia desejável no treinamento, sendo necessária a repetição de todo o processo com uma arquitetura diferente para a rede e/ou com alterações nos valores dos pesos iniciais e bias.

2.5.3 Algoritmo de Treinamento

O algoritmo de retropropagação dos erros, mais conhecido como *backpropagation*, foi desenvolvido na década de 80, se transformando no algoritmo de treinamento mais usado para redes de múltiplas camadas e com alimentação adiante.

Ele executa um processamento iterativo em um conjunto de amostras de treinamento, comparando para cada amostra, o valor predito pela rede com o valor conhecido da classe da amostra. A partir desta comparação, é calculado o erro quadrático médio que é retropropagado pela rede, no sentido da camada de saída até a camada de entrada, modificando os pesos das conexões, de forma a minimizar o erro. Em geral, os pesos convergem para valores que tornam o erro aceitável e o treinamento é finalizado. Uma versão do algoritmo é apresentada na Listagem 2.2.

Descrição:
Algoritmo: *Backpropagation*.
Objetivo: Treinar uma rede neural para classificação.
Entrada: *amostras*: conjunto de amostras de treinamento.
rede: uma rede neural multi-camadas e alimentação adiante.
Saída: Uma rede neural treinada para classificar amostras desconhecidas.

Algoritmo:

```
início
  inicialize os parâmetros (pesos e bias) da rede;
  enquanto condição de finalização ainda não atendida faça
    para cada amostra de treinamento X de amostras faça

      /* fase de alimentação da rede ou forward */
      aplique X à camada de entrada da rede;
      para cada neurônio k de uma camada escondida ou de saída faça
        calcule a entrada u_k, a partir dos valores da camada anterior;
        calcule a saída y_k;
      para cada neurônio k da camada de saída faça
        calcule o erro produzido e_k;

      /* fase de retropropagação do erro ou backward */
      para cada neurônio k de uma camada escondida, da última até a
          primeira camada escondida faça
        calcule o erro e_k , a partir dos valores da camada posterior;
      para cada peso w_kj da rede faça
        calcule o incremento do peso ? w_kj;
        atualize o peso w_kj;
      para cada bias b_k da rede faça
        calcule o incremento da bias ? b_k;
        atualize a bias b_k;

fim.
```

Listagem 2.2. Algoritmo *Backpropagation* (HAN & KAMBER, 2001).

Do algoritmo de treinamento devem ser destacadas as seguintes ações:

- Inicialização dos parâmetros: os pesos e bias da rede são inicializados com valores pequenos e aleatórios, geralmente no intervalo entre -1.0 e 1.0, ou entre -0.5 e 0.5.

- Fase de alimentação da rede ou forward: uma amostra de treinamento *X* é fornecida à camada de entrada da rede e se propaga até que os neurônios da camada de saída forneçam os seus resultados. Durante este percurso, cada neurônio das camadas escondidas e da de saída emite o seu sinal de saída específico, usando uma função de ativação. Para uso deste algoritmo, a

função de ativação dos neurônios tem que ser diferenciável, de forma a permitir a retropropagação do erro (HAYKIN, 2001). Considerando que a presente versão do algoritmo utiliza a função logística, o valor da saída de um neurônio escondido *k* seria:

$$y_k = j(u_k) = \frac{1}{1+e^{-u_k}} \tag{2.7}$$

onde u_k é calculado pela equação (2.5).

O erro e_k produzido pelo neurônio *k* de uma camada de saída é calculado da seguinte forma:

$$e_k = y_k(1-y_k)(d_k - y_k) \tag{2.8}$$

onde, y_k é o valor de saída predito pelo neurônio *k*, d_k é o valor desejado ou esperado para esta saída, baseado na classe conhecida da amostra de treinamento *X* e a expressão $y_k (1 – y_k)$ é a derivada da função logística.

- Fase de retropropagação do erro ou backward: para calcular o erro de um neurônio *k* de uma camada escondida, basta multiplicar a derivada de sua saída pelo somatório dos erros dos *n* neurônios conectados a ele na camada seguinte, multiplicados pelos pesos das respectivas conexões, isto é:

$$e_k = y_k(1-y_k)\sum_{i=1}^{n} e_i w_{ik} \tag{2.9}$$

São então calculados os incrementos $?w_{kj}$ e $?b_k$, que representam a correção dos erros retropropagados da amostra *X* e, finalmente, esses incrementos atualizam os pesos correspondentes w_{kj} e o valor do bias b_k do neurônio *k*, ou seja:

$$\Delta w_{kj} = he_k x_j \tag{2.10}$$

$$\Delta b_k = he_k \tag{2.11}$$

$$w_{kj} = w_{kj} + \Delta w_{kj} \tag{2.12}$$

$$b_k = b_k + \Delta b_k \tag{2.13}$$

A interpretação gráfica do algoritmo *backpropagation* é que a escolha dos valores iniciais dos pesos e dos bias representa a escolha de um ponto inicial sobre a superfície de erro da rede. Dependendo do lugar em que este ponto se localiza, a rede pode tanto convergir para a resposta ótima num tempo razoável, quanto demorar muito para encontrar o ponto de valor mínimo do erro durante o treinamento. A ocorrência de regiões de depressão (mínimos locais) ou platôs (valores estáveis) na superfície influencia significativamente o desempenho do treinamento. A taxa de aprendizagem *?* usada nas equações (2.10) e (2.11) é uma constante que tem seu valor geralmente entre 0 e 1, serve para evitar essas regiões. É aconselhável usar um valor intermediário para a taxa de aprendizagem, pois um valor alto faz com que os pesos caminhem rapidamente para um valor ótimo, porém com instabilidade, enquanto que um valor baixo, compromete sensivelmente o desempenho.

Em outras palavras, o algoritmo backpropagation usa o método do gradiente descendente para procurar um conjunto de pesos que podem modelar um determinado problema de classificação, minimizando a distância quadrática média entre a predição de classe da rede e classe real das amostras de treinamento.

Na versão da Listagem 2.2, os pesos e bias são atualizados logo após a apresentação de cada amostra. Alternativamente, é usual atualizar os incrementos acumulados somente depois que todas as amostras tiverem sido submetidas ao treinamento. Esta estratégia é chamada de treinamento por época, onde a repetição do processo para todo o conjunto de treinamento é uma época.

- Condições de finalização do treinamento: quando o erro quadrático médio tiver seu valor minimizado, isto é, ficar menor que um valor pré-determinado (tolerância) ou quando se completar um determinado número de épocas, o treinamento é finalizado.

2.6 Classificação Bayesiana

Classificadores Bayesianos são classificadores estatísticos que, em resumo, calculam as probabilidades de que uma determinada amostra pertença a cada uma das classes possíveis, predizendo para a amostra, a classe mais provável, isto é, a classe que obteve a maior dessas probabilidades.

Em outras palavras, dada uma amostra desconhecida X com valores de seus atributos, respectivamente, iguais a $x_1, x_2, \ldots, x_n$ e sabendo que existem m classes possíveis $C_1, C_2, \ldots, C_m$, são calculadas m probabilidades $P(C_i \mid X)$, $i = 1, 2, \ldots, m$. Cada um dos valores $P(C_i \mid X)$ representa a probabilidade de que a amostra X pertença a uma classe C_i específica, considerando que se conhece os valores dos atributos de X. $P(C_i \mid X)$ é chamada probabilidade de C_i condicionada a X ou probabilidade posterior, a posteriori, condicional, ou ainda, mais detalhadamente, probabilidade de que ocorra a classe C_i, dado que se conhece os valores dos atributos de X.

O classificador determina, então, qual é o valor máximo entre os calculados para $P(C_i \mid X)$, predizendo a classe C_i correspondente, para a amostra X. A classe Ci para a qual $P(C_i \mid X)$ é máxima é chamada hipótese posterior máxima.

Este método se baseia na teoria de decisão de Bayes e deu origem ao desenvolvimento dos chamados classificadores Bayesianos.

2.6.1 Teoria de Decisão de Bayes

De um conjunto de treinamento com S amostras, distribuídas em, por exemplo, duas classes distintas C_1 e C_2, sendo que S_1 amostras pertencem à classe C_1 e S_2 à C_2, pode-se calcular as probabilidades anteriores ou a priori de cada classe como:

$$P(C_i) = \frac{S_i}{S}, \; i = 1, 2. \tag{2.14}$$

Outras quantidades estatísticas que podem ser calculadas a partir do conjunto de treinamento, são as probabilidades posteriores de uma amostra desconhecida X, condicionadas a C_i, $P(X|C_i)$, i = 1, 2, que descrevem a distribuição das instâncias dos atributos da amostra X, em cada classe (THEODORIDIS & KOUTROUMBAS, 1999).

As regras de classificação de Bayes neste exemplo, determinam:

se $P(C_1 \mid X) > P(C_2 \mid X)$, então X é classificado para C_1 (2.15)

se $P(C_2 \mid X) > P(C_1 \mid X)$, então X é classificado para C_2

Usando o teorema de Bayes para calcular as probabilidades de C_i condicionadas a X, tem-se:

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)},\ i = 1, 2. \quad (2.16)$$

Embora $P(X)$ também possa ser calculada a partir do conjunto de treinamento, ela é constante para todas as classes. Então, somente $P(X|C_i)P(C_i)$ precisa ser maximizado e, desta forma, as regras (2.15) passam a ser:

se $P(X \mid C_1)P(C_1) > P(X \mid C_2)P(C_2)$, então X é classificado para C_1 (2.17)

se $P(X \mid C_2)P(C_2) < P(X \mid C_1)P(C_1)$, então X é classificado para C_2

2.6.2 Classificador Bayesiano Simples

O classificador Bayesiano simples (Simple Bayesian Classifier), também conhecido como classificador Bayesiano ingênuo (Naive Bayesian Classifier) pode ser comparável em desempenho com classificadores que usam árvores de decisão ou redes neurais, apresentando também precisão alta e boa escalabilidade (HAN & KAMBER, 2001).

Ele é chamado de ingênuos pelo fato de que o efeito do valor de um atributo sobre uma determinada classe é considerado independente dos valores dos outros atributos. Esta suposição é chamada independência condicional de classe e simplificam sensivelmente as computações envolvidas.

Existem outros classificadores que usam a teoria de decisão de Bayes e que consideram a dependência entre subconjuntos de atributos, conhecidos como redes Bayesianas.

O funcionamento do classificador Bayesiano simples poderia ser assim descrito (HAN & KAMBER, 2001):

Dado um conjunto de treinamento com S amostras, sendo que cada amostra é representada por um vetor característico n-dimensional (x_1, x_2, . . . , x_n), correspondendo aos valores específicos de cada um dos n atributos A_1, A_2, . . . , A_n, considerando que existam m classes C_1, C_2, . . . , C_m e que o número de amostras de treinamento por classe seja, respectivamente, igual a S_1, S_2, . . . , S_m e dada uma amostra desconhecida X, o classificador indicará que ela pertence à classe que tem a maior probabilidade posterior condicionada a X.

Generalizando o uso as regras formuladas em (2.15), o classificador atribui a amostra desconhecida X para a classe C_i, se e somente se:

$$P(C_i \mid X) > P(C_j \mid X) \text{ para } 1 \le j \le m, j \ne i$$

o que corresponde a generalizar as regras formuladas em (2.17), isto é, classifica X para a classe C_i, se e somente se:

$$P(X \mid C_i)P(C_i) > P(X \mid C_j)P(C_j) \text{ para } 1 \le j \le m, j \ne i \tag{2.18}$$

Usando o conjunto de treinamento, os valores de P(Ci) são calculados como:

$$P(C_i) = \frac{S_i}{S}, \; i = 1, \ldots, m \tag{2.19}$$

Os valores de $P(X \mid C_i)$ também são obtidos do conjunto de treinamento e para reduzir o custo computacional neste cálculo, o classificador Bayesiano simples considera a independência condicional de classe. Desta forma:

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) \tag{2.20}$$

onde as probabilidades $P(x_k \mid C_i)$ são calculadas considerando:

- Se o atributo A_k é discreto, então:

$$P(x_k \mid C_i) = \frac{S_{ik}}{Si} \tag{2.21}$$

onde S_{ik} é o número de amostras de treinamento da classe C_i que possuem o valor x_k para A_k.

- Se o atributo A_k é contínuo, então é adotada uma distribuição Gaussiana para o atributo, sendo:

$$P(x_k \mid C_i) = g(x_k, m_{C_i}, s_{C_i}) = \frac{1}{\sqrt{2p}\, s_{C_i}} e^{-\frac{(x_k - m_{C_i})^2}{2 s_{C_i}^2}} \tag{2.22}$$

onde $g(x_k, m_{Ci}, s_{Ci})$ é a função Gaussiana (ou normal) do atributo A_k e m_{Ci} e s_{Ci} são a média e o desvio padrão, respectivamente, dos valores do atributo A_k das amostras de treinamento da classe C_i.

Finalmente, substituindo os valores calculados nas regras de classificação (2.18), pode-se determinar a classe C_i (mais provável) da amostra X.

Exemplificando o funcionamento do classificador Bayesiano simples, seja o conjunto de treinamento apresentado na Tabela 2.1 e a amostra desconhecida X com os seguintes valores de atributos: TEMPO = Sol, TEMPERATURA = Moderada, UMIDADE = Normal e VENTO = Fraco.

Para classificá-la, é necessário maximizar $P(X \mid C_i)P(C_i)$, para i = 1, 2, onde C_1 corresponde a JOGA-TÊNIS = Sim e C_2 a JOGA-TÊNIS = Não.

Do conjunto de treinamento, pode-se obter:

- Número total de amostras: S = 14
- Número de amostras com classe C_1: S_1 = 9
- Número de amostras com classe C_2: S_2 = 5

Calculando as probabilidades anteriores de cada classe usando a equação (2.19), tem-se:

- $P(C_1)$ = P(JOGA-TÊNIS = Sim) = 9/14 = 0.643
- $P(C_2)$ = P(JOGA-TÊNIS = Não) = 5/14 = 0.357

Como todos os atributos são discretos, para calcular $P(X \mid C_i)$ é necessário calcular as probabilidade de cada atributo condicionadas às classes, usando a equação (2.21):

- $P(x_1|C_1)$ = P(TEMPO = Sol | JOGA-TÊNIS = Sim) = 2/9 = 0.222
- $P(x_1|C_2)$ = P (TEMPO = Sol | JOGA-TÊNIS = Não) = 3/5 = 0.600
- $P(x_2|C_1)$ = P (TEMPERATURA = Moderada | JOGA-TÊNIS = Sim) = 4/9 = 0.444
- $P(x_2|C_2)$ = P (TEMPERATURA = Moderada | JOGA-TÊNIS = Não) = 2/5 = 0.400
- $P(x_3|C_1)$ = P (UMIDADE = Normal | JOGA-TÊNIS = Sim) = 6/9 = 0.667
- $P(x_3|C_2)$ = P (UMIDADE = Normal | JOGA-TÊNIS = Não) = 1/5 = 0.200
- $P(x_4|C_1)$ = P (VENTO = Fraco | JOGA-TÊNIS = Sim) = 6/9 = 0.667
- $P(x_4|C_2)$ = P (VENTO = Fraco | JOGA-TÊNIS = Não) = 2/5 = 0.400

Usando a equação (2.20):

- $P(X \mid C_1)$ = $P(X$ | JOGA-TÊNIS = Sim) = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
- $P(X \mid C_2)$ = $P(X$ | JOGA-TÊNIS = Não) = 0.600 x 0.400 x 0.200 x 0.400 = 0.019

Para usar as regras de classificação, calcula-se:

- $P(X \mid C_1)P(C_1)$:

  $P(X \mid$ JOGA-TÊNIS = Sim) P(JOGA-TÊNIS = Sim) = 0.044 x 0.643 = 0.028

- $P(X \mid C_2)P(C_2)$:

  $P(X \mid$ JOGA-TÊNIS = Não) P(JOGA-TÊNIS = Não) = 0.019 x 0.357 = 0.007

Portanto, o classificador Bayesiano simples prediz a classe JOGA-TÊNIS = Sim para a amostra X.

CAPÍTULO 3
IMPLEMENTAÇÃO DO CLASSIFICADOR

A solução do problema proposto no Capítulo 4 é o desenvolvimento de um classificador, a ser acoplado internamente num sistema chamado *InfraSystem*. O principal objetivo do *InfraSystem* é apoiar a tomada de decisões em relação a diversos aspectos da conservação e manutenção da infra-estrutura de uma via de transporte terrestre.

O acoplamento do classificador transforma o *InfraSystem* em um sistema inteligente, capaz de assimilar o conhecimento a partir de um conjunto de treinamento, gerado em vistorias na via por especialistas em geotecnia e, posteriormente, simular o comportamento desses especialistas, classificando novas situações que vierem a ocorrer.

A seguir, é apresentada uma discussão sobre a escolha do método de classificação a ser empregado neste caso, bem como os algoritmos de treinamento e predição do classificador, as suas implementações e os experimentos computacionais realizados para avaliar essas implementações.

3.1 Escolha do Método de Classificação

O estado da arte disponibiliza diversos métodos de classificação que poderiam ser empregados para solucionar o problema em questão. Limitando aos três métodos apresentados no Capítulo anterior para facilitar a análise, verifica-se que existem pequenas vantagens e desvantagens entre eles.

Tendo em vista a forma de trabalho que se deseja realizar, o método a ser adotado deve atender bem, principalmente, as qualidades de acurácia de predição e desempenho. Os outros critérios não são tão essenciais neste caso, pelas seguintes razões:

- Robustez: O *InfraSystem* possui três níveis de filtragem (ou crítica) na entrada de dados, o que impossibilita a existência de amostras com atributos errados, faltando ou com ruídos.

- Escalabilidade: Cada base de dados é montada a partir da vistoria em um trecho da via de transporte com características próprias (serra, planalto, baixada etc.), não constituindo, portanto, grandes quantidades de dados.

- Interpretabilidade: O usuário típico do sistema pertence à equipe de administração e planejamento da via, interessado, exclusivamente, no resultado da classificação e não no conhecimento gerado pelo modelo. Sendo assim, é desejável que o classificador funcione de maneira totalmente transparente para o usuário.

A partir desta análise, descartou-se o uso de rede neural, pois a robustez, que é um de seus pontos fortes, não é essencial e o seu treinamento poderia prejudicar a transparência desejada.

Finalmente, na dificuldade de escolher entre árvore de decisão e classificador Bayesiano, foi considerado que um classificador Bayesiano simples levava ligeira vantagem, por possuir um esforço de programação um pouco menor, desempenho de treinamento possivelmente melhor e maior facilidade no tratamento de atributos contínuos, sendo, portanto, o escolhido.

3.2 Algoritmos do Classificador

O classificador CBS é um classificador Bayesiano simples, com consideração de atributos de dados discretos e contínuos e múltiplas classes. Ele foi desenvolvido através de dois algoritmos, sendo o primeiro de treinamento e o segundo de predição.

A seguir são descritas as características gerais e apresentada uma versão em alto nível de cada algoritmo (Listagens 3.1 e 3.2). A descrição das variáveis usadas nos algoritmos é mostrada na Tabela 3.1. A versão detalhada de cada um foi incluída no ANEXO I.

3.2.1 Algoritmo de Treinamento

Nome: TREINA.

Objetivo: Realizar o treinamento.

Entrada: Arquivo de treinamento (XTREINA) contendo os parâmetros N, ND, M e MAXAD (descrição na Tabela 3.1) e todas as amostras de treinamento, onde cada uma deve estar organizada seqüencialmente com todos os atributos discretos (XD), todos os atributos contínuos (XC) e a sua classe (C).

Saída: Arquivo TREINADO contendo os parâmetros N, ND, M e MAXAD e o conhecimento adquirido, representado pelas matrizes PCi, PXCD, MEDIA, DESVIO e CTE (descrição na Tabela 3.1).

```
início

    N, M, ND, NC, MAXAD, C, S: int;
    XD: vet [1 . . ND] int;
    XC: vet [1 . . NC] int;
    Si: vet [1 . . M] int;
    PXCD: mat [1 . . M, 1 . . MAXAD, 1 . . ND] real;
    MEDIA, DESVIO, CTE: mat [1 . . M, 1 . . NC] real;
    PCi: vet [1 . . M] real;

    leia (N, ND, M, MAXAD);
    NC = N – ND;
    repita
            "ler vetores de atributos XD e XC";
            leia (C);
            "acumular S e Si";
            "preparar cálculo de PXCD";
    até "fim de arquivo de TREINAMENTO";
    "calcular PCi";
    "calcular PXCD, MEDIA, DESVIO e CTE";
    "gravar N, ND, M, MAXAD, PCi, PXCD, MEDIA, DESVIO e CTE";

fim.
```

Listagem 3.1. Algoritmo TREINA – Versão em Alto Nível.

A Tabela 3.1 contendo a descrição de todas as variáveis dos algoritmos de treinamento e de predição é apresentada a seguir.

Variável	Descrição	Tipo
N	Número total de atributos	Int
M	Número total de classes	Int
ND	Número de atributos discretos	Int
NC	Número de atributos contínuos	Int
MAXAD	Número máximo de valores distintos que um atributo discreto qualquer pode assumir	Int
XD	Primeira parte do vetor característico contendo somente os atributos discretos	vet [1 . . ND] int
XC	Segunda parte do vetor característico contendo somente os atributos contínuos	vet [1 . . NC] real
C	Classe de um vetor característico dado para treinamento	Int
S	Número total de amostras de treinamento	Int
Si	Número de amostras de treinamento da classe C_i	vet [1 . . M] int
PXCD	Probabilidade que ocorra um valor x_k para um atributo discreto A_k, numa amostra X de classe C_i	mat [1 . . M, 1 . . MAXAD, 1 . . ND] real
MEDIA	Média aritmética de todos os valores de cada atributo contínuo A_k, pertencentes às amostras de treinamento de classe C_i	mat [1 . . M, 1 . . NC] real
DESVIO	Desvio padrão de todos os valores de cada atributo contínuo A_k pertencentes às amostras de treinamento de classe *Ci*	mat [1 . . M, 1 . . NC] real
CTE	Constantes utilizadas na distribuição normal, relativas a cada atributo contínuo A_k pertencentes às amostras de treinamento de classe C_i	mat [1 . . M, 1 . . NC] real
PCi	Probabilidade *a priori* da classe *Ci*	vet [1 . . M] real
PXCi	Probabilidade que ocorra um valor x_k para cada atributo A_k, numa amostra X de classe C_i	vet [1 . . M] real
PCiX	Probabilidade que a amostra desconhecida X pertença à classe C_i, também chamada de probabilidade *a posteriori*	vet [1 . . M] real
CLASSE	Classe predita para uma amostra desconhecida X	Int

Tabela 3.1. Descrição de Variáveis.

3.2.2 Algoritmo de Predição

Nome: CLASS.

Objetivo: Realizar a classificação de amostras desconhecidas.

Entrada: Arquivo TREINADO contendo os parâmetros N, ND, M e MAXAD e o conhecimento adquirido, representado pelas matrizes PCi, PXCD, MEDIA, DESVIO e CTE.

Arquivo XTESTE contendo as amostras desconhecidas, onde cada uma deve estar organizada seqüencialmente com todos os atributos discretos (XD) e todos os atributos contínuos (XC).

Saída: Arquivo CLASSIFICADO contendo as CLASSEs preditas, respectivas a cada amostra desconhecida.

```
início

    N, M, ND, NC, MAXAD, CLASSE: int;
    XD: vet [1 . . ND] int;
    XC: vet [1 . . NC] int;
    PXCD: mat [1 . . M, 1 . . MAXAD, 1 . . ND] real;
    MEDIA, DESVIO, CTE: mat [1 . . M, 1 . . NC] real;
    PCi, PXCi, PCiX: vet [1 . . M] real;

    "ler N, ND, M, MAXAD, PCi, PXCD, MEDIA, DESVIO e CTE do arquivo de
    treinamento";
    NC = N – ND;
    repita
            "ler vetor característico de teste";
            "calcular PXCi";
            "calcular PCiX e CLASSE";
            grave (CLASSE);
    até "fim de arquivo XTESTE";

fim.
```

Listagem 3.2. Algoritmo CLASS – Versão em Alto Nível.

3.3 Implementações

O trabalho de implementação computacional foi realizado em um micro computador IBM® PC compatível, equipado com um processador Intel® Pentium III 850 MHz; 128 megabytes de memória RAM; 10 gigabytes de disco rígido; sistema operacional Microsoft® Windows® 2000 Server, Service Pack 4.

Para a implementação foi utilizado o ambiente de desenvolvimento Borland® C++, versão 5.0 e para visualização das tabelas geradas nos experimentos realizados, o Microsoft® Excel 2002.

As listagens dos programas TREINA.CPP e CLASS.CPP são apresentadas no ANEXO II e correspondem aos algoritmos com os respectivos nomes. Foram mantidos, também, os nomes das variáveis dos algoritmos, para facilitar o entendimento.

Todo o processo de desenvolvimento teve como base a obtenção de um produto final de uso geral e com o melhor desempenho possível. Neste sentido, deve-se destacar:

- Todos os arquivos de entrada e saída de dados são do tipo texto.

- A organização dos atributos nas amostras de treinamento e de teste nos arquivos de entrada XTREINA e XTESTE obedece a uma seqüência que contem todos os atributos discretos (vetor XD) e todos os atributos contínuos (vetor XC), evitando a necessidade de informar um vetor parâmetro que identifica se cada atributo é discreto ou contínuo e também, um número considerável de comparações.

- Os valores dos atributos discretos e das classes são usados diretamente como índices de matrizes. Portanto, na preparação de dados dos arquivos de entrada XTREINA e XTESTE, esses valores devem ser transformados para valores inteiros a partir de 1. A organização de todos os arquivos de entrada e saída, também são apresentadas no ANEXO II.

- A linguagem C++, considerada de médio nível, gera programas executáveis comprovadamente mais rápidos que as linguagens de alto nível. Além disso,

procurou-se utilizar todos os seus recursos que valorizassem essa característica. Dois exemplos claros desses recursos utilizados são:

1. As variáveis de controle de todas as iterações *for* foram declaradas como *register*, causando a alocação das variáveis diretamente em registradores da unidade aritmética e lógica do processador, eliminando o tempo de carregamento e minimizando o tempo de incremento (variáveis I e K na Listagem 3.3).

2. Padronização adotada para manipulação de matrizes: uma matriz para ser manipulada é entregue como parâmetro para uma função, que a recebe localmente como um ponteiro, causando a redução de uma dimensão na matriz e, conseqüentemente, diminuindo o tempo de cálculo de endereço de cada elemento a ser manipulado. Na função CALCPXCi do programa CLASS na Listagem 3.3 abaixo, o parâmetro de entrada MEDIA, por exemplo, é uma matriz bi-dimensional, recebida pelo ponteiro local (*PtMEDIA). Na verdade, ele aponta para uma linha da matriz MEDIA, necessitando de somente um índice de coluna para identificar um elemento desta linha, no caso (*PtMEDIA)[K]. Para apontar para a próxima linha, basta somar 1 no ponteiro, ou seja, PtMEDIA++.

```
//*** calcula P (X | Ci) ***
void CALCPXCi (int ND, int NC, int M, int XD[], double XC[],
                    double PXCD[][TMAXAD][TND], double MEDIA[][TNC],
                    double DESVIO[][TNC], double CTE[][TNC], double PXCi[])
        {
        register int I, K;
        int *PtXD;
        double *PtXC, (*PtPXCD)[TMAXAD][TND]=PXCD,
        (*PtMEDIA)[TNC]=MEDIA,
                (*PtDESVIO)[TNC]=DESVIO, (*PtCTE)[TNC]=CTE, *PtPXCi =
                PXCi;

        for (I = 0; I < M; I++, PtPXCi++, PtPXCD++, PtCTE++, PtMEDIA++,
        PtDESVIO++)
            {
            *PtPXCi = 1.0;
            for (K = 0, PtXD = XD; K < ND; K++, PtXD++)
                {
                *PtPXCi = *PtPXCi * (*PtPXCD)[*PtXD - 1][K];
                }
            for (K = 0, PtXC = XC; K < NC; K++, PtXC++)
                {
                if ((*PtDESVIO)[K])
                  *PtPXCi = *PtPXCi * (*PtCTE)[K] * exp(-quad(*PtXC -
                  (*PtMEDIA)[K]) /          (2 * quad((*PtDESVIO)[K])));
                else
                  *PtPXCi = 0.0;
                }
            }
        }  //*** Fim CALCPXCi ***
```

Listagem 3.3. Função CALCPXCi do Programa CLASS.CPP.

3.4 Experimentos Computacionais

Um procedimento comum para avaliar a acurácia de classificadores é a validação cruzada. Para utilizá-lo, divide-se a base de dados em N conjuntos com um número aproximadamente igual de amostras, com distribuição uniforme aleatória de classes. Cada conjunto constitui um conjunto de teste e os (N – 1) conjuntos restantes formam o conjunto de treinamento. Assim, N testes são realizados com a base de dados, sendo que as amostras de cada conjunto de teste não pertencem aos correspondentes conjuntos de treinamento. A acurácia do classificador para esta base de dados é a média das acurácias dos N testes.

Para realizar este procedimento com o classificador CBS, foi criado um programa na linguagem C++ que analisa a base de dados e verifica a possibilidade de dividi-la em 3 a 10 conjuntos de teste, dependendo do número de amostras por classe, executando, no mínimo, uma validação cruzada (com 3 conjuntos de teste, qualquer que seja a distribuição de amostras por classe na base de dados) e no máximo, oito (variando o número de conjuntos de teste de 3 a 10, respectivamente). Para exemplificar, seja uma base de dados que contem somente 7 amostras de uma determinada classe. Neste caso, o programa executará cinco validações cruzadas, com o número de conjuntos de teste variando de 3 a 7, respectivamente.

No final, o programa de validação cruzada apresenta um resumo de todos os testes realizados com a base de dados e considera como acurácia a média de todos eles.

Para ilustrar, será comentado a seguir, o resultado da execução da validação cruzada para a base de dados da flor Íris.

Primeiramente, o programa apresenta o número mínimo e máximo em que foi possível dividir a base de dados em conjuntos de teste e o número total de testes realizados (Tabela 3.2).

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Iris	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

Tabela 3.2. Cabeçalho do Resultado da Validação Cruzada.

Em seguida, são detalhados os resultados de cada teste em cada validação cruzada, isto é, para cada validação cruzada são informados o número de conjuntos de teste e o número de amostras de teste por conjunto e, para cada teste desta validação, os valores absoluto e percentual de acertos (Tabela 3.3).

TESTES REALIZADOS

Conjuntos de Teste		3		N° de Amostras / Conjunto	48
Teste		1	2	3	
Acertos	Abs.	46	46	42	
	Per.	95.83	95.83	87.50	

Conjuntos de Teste		4		N° de Amostras / Conjunto	36
Teste		1	2	3	4
Acertos	Abs.	34	34	34	33
	Per.	94.44	94.44	94.44	91.67

Conjuntos de Teste		5		N° de Amostras / Conjunto	30
Teste		1	2	3	4	5
Acertos	Abs.	29	27	30	26	29
	Per.	96.67	90.00	100.00	86.67	96.67

Conjuntos de Teste		6		N° de Amostras / Conjunto	24
Teste		1	2	3	4	5	6
Acertos	Abs.	23	23	22	24	20	23
	Per.	95.83	95.83	91.67	100.00	83.33	95.83

Conjuntos de Teste		7		N° de Amostras / Conjunto	21
Teste		1	2	3	4	5	6	7
Acertos	Abs.	20	20	19	21	19	19	20
	Per.	95.24	95.24	90.48	100.00	90.48	90.48	95.24

Conjuntos de Teste		8		N° de Amostras / Conjunto	18
Teste		1	2	3	4	5	6	7	8
Acertos	Abs.	17	17	17	17	18	16	16	17
	Per.	94.44	94.44	94.44	94.44	100.00	88.89	88.89	94.44

Conjuntos de Teste		9		N° de Amostras / Conjunto	15
Teste		1	2	3	4	5	6	7	8	9
Acertos	Abs.	14	15	14	13	15	15	13	13	14
	Per.	93.33	100.00	93.33	86.67	100.00	100.00	86.67	86.67	93.33

Conjuntos de Teste		10		N° de Amostras / Conjunto	15	
Teste		1	2	3	4	5	6	7	8	9	10
Acertos	Abs.	14	15	14	13	15	15	13	13	14	15
	Per.	93.33	100.00	93.33	86.67	100.00	100.00	86.67	86.67	93.33	100.00

Tabela 3.3. Resultados de cada Teste em cada Validação Cruzada.

Na seqüência, um resumo (Tabela 3.4) representa em cada coluna uma validação cruzada com os seguintes resultados: número de conjuntos de teste, número de amostras por conjunto e os valores absoluto e percentual de acertos mínimo, máximo e médio da

validação. A última linha deste resumo contem as acurácias de cada validação cruzada. A última coluna contem os percentuais mínimo, máximo e médio de todos os testes realizados (no caso 52), sendo este último valor (acerto percentual médio total) a acurácia que o CBS obteve para a base de dados *Iris* (93.73%).

RESUMO - ACERTOS										
Conjs. de Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		48	36	30	24	21	18	15	15	
Mínimo	**Absoluto**	42	33	26	20	19	16	13	13	
	Percentual	87.50	91.67	86.67	83.33	90.48	88.89	86.67	86.67	**83.33**
Máximo	**Absoluto**	46	34	30	24	21	18	15	15	
	Percentual	95.83	94.44	100.00	100.00	100.00	100.00	100.00	100.00	**100.00**
Médio	**Absoluto**	44.67	33.75	28.2	22.5	19.71	16.88	14	14.1	
	Percentual	93.06	93.75	94.00	93.75	93.88	93.75	93.33	94.00	**93.73**

Tabela 3.4. Resumo de Acertos.

O programa calcula ainda, a matriz de confusão de todos os testes realizados, apresentando somente a pior e a melhor situação (Tabela 3.6), logo após a uma distribuição de amostras por classe (Tabela 3.5).

NÚMERO DE AMOSTRAS POR CLASSE					
Classe	1	2	3	**Total**	
Abs.	50	50	50	150	
Per.	33.33	33.33	33.33	100.00	

Tabela 3.5. Distribuição de Amostras por Classe.

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL		**83.33**
CONJUNTOS DE TESTE	6	
AMOSTRAS / CONJUNTO	24	
TESTE	5	
ACERTOS - ABS.	20	

MATRIZ DE CONFUSÃO - ABS.				
		PREDITO		
		1	**2**	**3**
ESPERADO	**1**	8	0	0
	2	0	6	2
	3	0	2	6

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL					100.00
CONJUNTOS DE TESTE	5				
AMOSTRAS / CONJUNTO	30				
TESTE	3				
ACERTOS - ABS.	30				

MATRIZ DE CONFUSÃO - ABS.				
		PREDITO		
		1	2	3
ESPERADO	1	10	0	0
	2	0	10	0
	3	0	0	10

Tabela 3.6. Matrizes de Confusão da Pior e da Melhor Situação.

Finalmente, foi criada uma matriz de confusão com valores percentuais médios de todos os testes (Tabela 3.7), que possibilita uma análise do comportamento geral do CBS para a base de dados Íris. A soma da diagonal principal é, evidentemente, a acurácia considerada e os valores situados fora desta diagonal, representam de que forma o classificador está errando. Nota-se também, que o CBS não errou em nenhum dos testes, a predição da classe 1, que é uma classe linearmente separável.

ACERTO MÉDIO PERCENTUAL					93.73
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL					
		PREDITO			
		1	2	3	
ESPERADO	1	33.33	0.00	0.00	
	2	0.00	30.52	2.81	
	3	0.00	3.46	29.87	

Tabela 3.7. Matrizes de Confusão – Valores Médios Percentuais.

Na verificação da acurácia de predição do classificador CBS, foram executadas validações cruzadas em vinte e uma bases de dados (conhecidas como bases de dados acadêmicas), disponíveis em ftp://ftp.ics.uci.edu/pub/machine-learning-databases para testes. Os resultados específicos obtidos nestes testes constituem o ANEXO III e um resumo é apresentado na Tabela 3.8.

VALIDAÇÃO CRUZADA - RESUMO									
Base de Dados							Validação Cruzada		
Identificação	Atributos				Classes	Amostras	% Acerto		Acurácia
	N	ND	NC	MAXAD	M		Mín.	Máx.	
Blood Testing	4	0	4	0	2	209	72.00	100.00	88.51
Breast Cancer 1	9	0	9	0	2	683	91.67	98.97	95.75
Breast Cancer 2	9	9	0	10	2	683	93.81	100.00	96.77
Credit Screening	15	9	6	14	2	653	70.31	92.59	81.08
Diabetes	8	0	8	0	2	768	67.37	81.65	74.37
Echocardiogram	8	1	7	2	2	62	50.00	100.00	74.41
Glass	9	0	9	0	7	214	32.14	79.17	46.91
Images	18	0	18	0	7	210	21.43	42.86	33.08
Iris	4	0	4	0	3	150	83.33	100.00	93.73
Mushroom	22	22	0	9	2	5644	63.69	100.00	96.41
Parity 3	3	3	0	2	2	80	5.00	61.54	32.59
Parity 4	4	4	0	2	2	160	12.50	50.00	34.71
Parity 5	5	5	0	2	2	320	18.75	55.00	36.10
Sleepdata1	8	0	8	0	6	468	53.85	93.85	73.36
Sleepdata2	8	0	8	0	6	397	14.73	70.83	60.21
Sonar	60	0	60	0	2	208	48.00	96.00	69.90
Spiral	2	0	2	0	2	92	0.00	80.00	21.72
Synthetic1	2	0	2	0	2	1250	82.61	94.93	88.63
Vowel	10	0	10	0	11	990	38.18	76.77	54.70
Wine	13	0	13	0	3	178	83.33	100.00	97.53
WNBA	2	0	2	0	3	120	54.55	100.00	78.94
Máximos	**60**	**22**	**60**	**14**	**11**				

sendo:

N	Número total de atributos
ND	Número de atributos discretos
NC	Número de atributos contínuos
MAXAD	Número máximo de valores distintos que um atributo discreto qualquer pode assumir
M	Número total de classes

Tabela 3.8. Validação Cruzada – Resumo.

Para melhor analisar os resultados deste resumo, deve-se destacar que:

- As bases de dados *Parity* 3, 4 e 5 para a determinação da classe da paridade de conjuntos de 3, 4 ou 5 bits, respectivamente, são enunciadas como de difícil classificação. O mesmo acontece com *Spiral* que fornece um conjunto de pares ordenados que pertencem a uma de duas espirais distintas, traçadas a partir da origem dos eixos do plano cartesiano. Os resultados obtidos para suas acurácias comprovam essa dificuldade.

- A ocorrência simultânea de maiores quantidades de atributos contínuos e de classes distintas com pequena quantidade de amostras, também representam dificuldade para o CBS. Este é o caso de *Glass* e *Images*. Já a base de dado *Vowel*, que pode ser enquadrada nesta situação, possui uma quantidade de amostras um pouco maior e, em conseqüência, melhor acurácia.

- As bases de dados *Sleepdata* 1 e 2, servem para classificar 6 estágios do sono de duas pessoas e, além de possuírem uma distribuição bem irregular de amostras por classe, também demonstram a influência do número de amostras no valor obtido para a acurácia.

- *Sonar* com 60 atributos contínuos e poucas amostras, obteve uma acurácia que pode ser considerada satisfatória.

- *Breast Cancer* 1 foi testado com 9 atributos contínuos, conforme enunciado originalmente. Como nas amostras, todos os atributos são valores inteiros entre 1 e 10, foi criada, a título de experiência, a base de dados *Breast Cancer* 2, que considera os atributos como discretos, o que melhorou ainda mais o valor da acurácia.

- Doze bases de dados alcançaram acurácia superior a 70,00%.

A partir dos resultados das validações cruzadas e considerando os detalhes apresentados anteriormente, pode-se concluir que o CBS possui acurácia de predição satisfatória.

Quanto ao desempenho, embora não tenha sido mensurado exatamente o tempo de execução do classificador, observou-se que o processamento é muito rápido. As bases de dados que consumiram mais tempo na validação cruzada foram *Mushroom* com 5644 amostras e *Sonar* com 60 atributos contínuos. Nestes casos, o programa de validação cruzada administrou 52 treinamentos e os testes correspondentes em menos de três segundos.

CAPÍTULO 4
O PROBLEMA DE AVALIAÇÃO DE RISCOS

A partir da década de 70, ocorreu uma brusca mudança nos procedimentos de conservação e manutenção das estradas brasileiras. Essas atividades, até então presentes no cotidiano do gerenciamento das estradas, ficaram relegadas a segundo plano, quando não totalmente ausentes. A conseqüência foi uma crescente degradação da malha rodoviária e ferroviária do país, passando a engenharia a atuar fundamentalmente nos moldes corretivos e emergenciais.

A constatação deste fato levou os engenheiros Luiz Ernesto Bernardino Alves Filho e Raul Bomilcar do Amaral, embasados, cada um, em mais de vinte anos de trabalhos em conservação, manutenção e monitoramento de infra-estrutura de estradas, a desenvolver uma metodologia que disponibilizasse à equipe de administração, uma visão completa e integrada de todo o trecho da via, apresentando, para cada local, um diagnóstico da situação atual e a formulação de propostas e recomendações para a reabilitação da infra-estrutura, bem como o custo de cada intervenção sugerida.

Com a introdução de diversos indicadores por local, a metodologia auxilia no gerenciamento das interferências, dos fenômenos adversos e, principalmente, dos riscos e possibilita configurar uma hierarquização dos problemas constatados, fornecendo subsídios fundamentais para a elaboração do planejamento estratégico de intervenções na estrada.

4.1 Aplicação da Metodologia

Inicialmente, é realizado o levantamento dos dados através de uma vistoria no trecho da via, com a participação de pelo menos um especialista em geotecnia.

Define-se como um LOCAL a ser cadastrado, cada pequeno segmento que apresente intervenções e/ou alterações, como, por exemplo, terraplenos e encostas naturais instáveis, obras de contenção ou de reforço para estabilização de maciços ou um conjunto de dispositivos de drenagem.

O cadastramento de cada local inclui um identificador único, sua localização no trecho (quilômetro, estaca etc.), sua caracterização e fenomenologia, avaliação dos riscos e proposições e recomendações de intervenções.

Os dados essenciais na caracterização de um local são:

- Terraplenos e encostas: identificação da tipologia, do estágio de intemperismo dos materiais superficiais e as dimensões básicas (inclinações e alturas aproximadas dos taludes, larguras de bermas e dimensões da plataforma).

- Obras de contenção existentes: classificação dos tipos de obras, concepção estrutural, materiais empregados na construção, sua localização no terrapleno, estado atual de conservação e as dimensões básicas.

- Drenagem existente: classificação, materiais empregados na construção, conformação, eventuais alterações estruturais e disfunções, e dimensões.

- Cobertura vegetal: classificação do tipo e estado do revestimento nos taludes.

- Interferências: localização de equipamentos urbanos, dispositivos e intervenções e/ou alterações, próximas a via, que possuam ou possam vir a ter interface com o terrapleno.

- Aspectos ambientais relevantes: localização e estágios de conservação e de implantação atuais, quando interferem na operação e na integridade da estrada.

No registro da fenomenologia de um local, são identificados e dimensionados os fenômenos existentes, que, do ponto de vista geotécnico, já interfiram ou possam vir a promover alguma interferência negativa no fluxo normal da estrada. Neste sentido, as evidências encontradas relacionadas ao início de processos instabilizadores ou de degradação de equipamentos, passam por uma análise técnica aprofundada, resultando em comentários das prováveis conseqüências decorrentes da permanência ou da evolução destas ocorrências.

Para melhor avaliar os riscos a que a via de transporte está sujeita, em função das alterações constatadas, foram definidos seis indicadores, chamados parâmetros (ou variáveis) de decisão, que, devidamente graduados, procuram retratar o nível de criticidade do local, em relação ao conjunto de todos os locais do trecho.

As recomendações e proposições de intervenções no local são estabelecidas e dimensionadas, com medidas para curto, médio e longo prazos, de forma a possibilitar ações desde atenuar situações potenciais de risco, até a solução completa dos problemas, contribuindo no aumento da segurança da via.

Após a vistoria, as informações de todos os locais constituem uma base de dados gerenciada por um sistema computacional, chamado *InfraSystem*. A arquitetura deste sistema utiliza uma estruturação modular das seguintes tarefas: cadastramento de usuários, geração do banco de dados, consulta e geração de relatórios, consulta e geração de estatísticas e gerenciamento das intervenções por local, tempo ou custo.

Um ponto importante da metodologia consiste no relacionamento do risco representado por cada local, permitindo uma visão completa e integrada de todo o trecho da via e possibilitando o planejamento das ações a serem desenvolvidas.

A medida deste risco é o nível de criticidade do local, sendo, portanto, o principal parâmetro para que se possa determinar a melhor seqüência das correções necessárias. O seu valor inicial é determinado pelo(s) especialista(s) no momento do levantamento dos dados de cada local e é baseado nos parâmetros de decisão.

Na dinâmica da administração da estrada, um local que sofreu uma correção (parcial ou total), passa por uma nova vistoria, tendo alterado os seus parâmetros de decisão e, conseqüentemente, o seu nível de criticidade. O que também pode acontecer é a ocorrência de fenômenos naturais (uma tromba d'água, por exemplo) resultando no agravamento de problemas existentes ou na necessidade do cadastramento de novos locais e na determinação do nível de criticidade de cada um.

É desejável que a determinação dos novos níveis de criticidade siga o mesmo padrão daqueles realizados pelo(s) especialista(s), para que a equipe de administração da via possa continuar agindo com o máximo de equilíbrio no planejamento das intervenções.

Este é um problema típico em que um classificador pode ser empregado, criando-se um novo módulo no *InfraSystem*, com execução automática do treinamento

na finalização da entrada dos dados da vistoria realizada pelo(s) especialista(s) e com predição, também automática, do nível de criticidade de novos locais incluídos posteriormente ou daqueles que tiverem qualquer parâmetro de decisão alterado.

A seguir são apresentados os testes feitos com o CBS, desenvolvido no Capítulo anterior, usando uma base de dados real, para verificar a possibilidade da sua inclusão como o módulo inteligente do *InfraSystem*.

4.2 O CBS na Avaliação de Riscos

A base de dados utilizada foi gerada na vistoria de um trecho de ferrovia com 226 locais cadastrados.

Os parâmetros de decisão utilizam externamente, isto é, nas planilhas de campo e nos relatórios, as escalas descritas abaixo. Por questões operacionais, esses valores são digitados e registrados na base de dados como valores discretos de 1 a 4.

Parâmetros de decisão:

- Registro de Acidentes Anteriores: Sem registros, Com registros de pouca relevância, Pequeno número de registros e Grande número de registros.
- Extensão(L): $L \leq 10$ m., 10 m. $< L \leq 30$ m., 30m. $< L \leq 70$ m. e $L > 70$ m.
- Risco de Interrupção do Tráfego: Não tem, Baixo, Médio e Alto.
- Complexidade da Solução: Sem complexidade, Baixa, Média e Alta.
- Grau de Interferência: Sem interferência, Moradias, Drenagem e Grandes interferências.
- Grau de Deterioração: Não tem, Baixo, Médio e Alto.

Para o nível de criticidade foi usada externamente, a escala: Sem criticidade, Baixa, Média, Alta e Muito Alta e na digitação e registro, a escala discreta de 1 a 5.

Para efeito do uso do CBS, cada local é uma amostra, cada parâmetro de decisão é um atributo e o nível de criticidade é a classe.

Foram realizados dois tipos de testes. No primeiro, usou-se o programa de validação cruzada, que permite uma visão geral do funcionamento do CBS para esta base de dados. No segundo, foram feitos testes comparativos entre o CBS e a rede neural aiNet versão 1.25 (www.ainet-sp.si), para verificar a diferença nos percentuais de acertos de predição de cada classificador.

4.2.1 Validação Cruzada

O resultado completo da validação cruzada realizada foi incluída no ANEXO III. Deste resultado, pode-se destacar:

- A distribuição de amostras por classe na Tabela 4.1 e Figura 4.1 representam a atual situação de risco do trecho da via. Evidentemente, quanto maior a concentração à direita (nível de criticidade = muito alto), pior a situação geral do trecho e, à medida que as correções forem sendo efetuadas, a concentração tenderá para a esquerda.

VALIDAÇÃO CRUZADA						
BASE DE DADOS (BD): InfraSystem						
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE						
Mínimo	**Máximo**					
3	10					
TOTAL DE TESTES REALIZADOS	52					
NÚMERO DE AMOSTRAS POR CLASSE						
Classe	1	2	3	4	5	**Total**
Abs.	28	80	54	40	24	226
Per.	12.39	35.40	23.89	17.70	10.62	100.00

Tabela 4.1. Distribuição de Amostras por Classe.

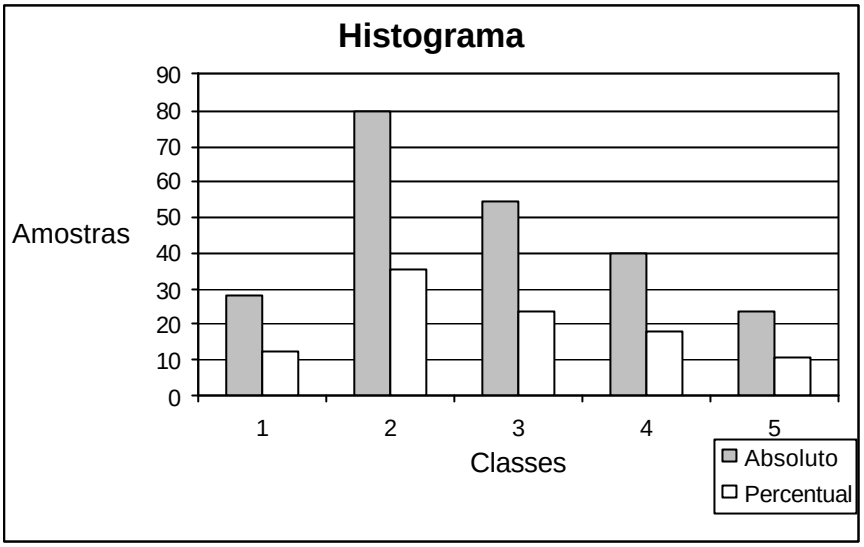

Figura 4.1. Distribuição de Amostras por Classe.

- A Tabela 4.2 mostra o resumo de acertos nos testes realizados. Embora a acurácia de predição de todos os testes tenha sido de 58,46%, observa-se que o valor da acurácia (Acerto Médio Percentual) é crescente, quanto maior for o conjunto de treinamento. Isso é demonstrado claramente no Figura 4.2.

  Outro fator importante nesta análise é que, segundo entrevista com os engenheiros responsáveis pela metodologia, a base de dados em questão é considerada pequena e que a previsão para a grande maioria dos casos, são bases de dados contendo mais de 500 amostras, o que, com certeza, contribui para o aumento da acurácia.

RESUMO - ACERTOS									
Conjs. de Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		74	56	43	36	30	27	23	21	
Mínimo	**Absoluto**	31	28	19	11	9	10	9	10	
	Percentual	41.89	50.00	44.19	30.56	30.00	37.04	39.13	47.62	**30.00**
Máximo	**Absoluto**	35	34	26	29	24	20	17	15	
	Percentual	47.30	60.71	60.47	80.56	80.00	74.07	73.91	71.43	**80.56**
Médio	**Absoluto**	33.33	29.5	23.6	20.5	18.29	16	13.78	13.4	
	Percentual	45.05	52.68	54.88	56.94	60.95	59.26	59.90	63.81	**58.46**

Tabela 4.2. Resumo de Acertos.

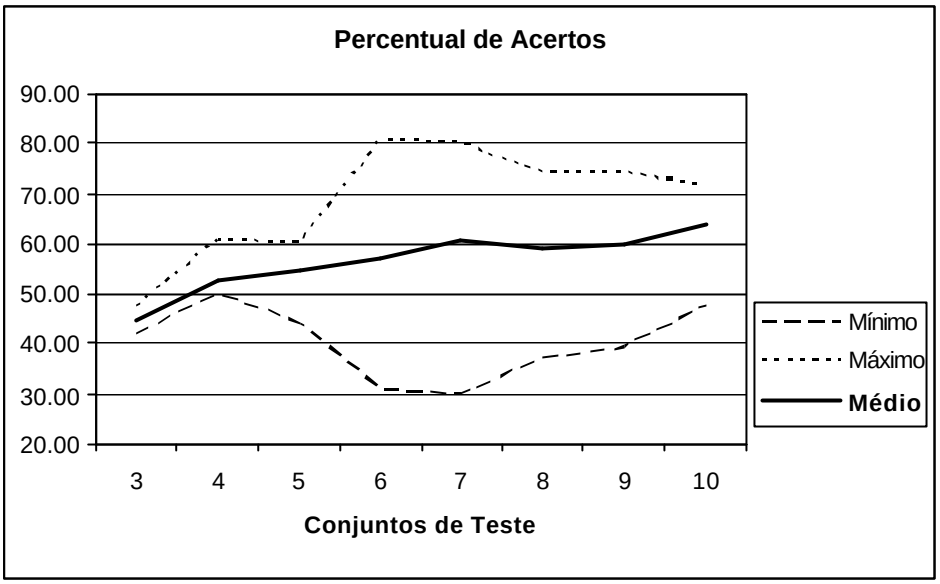

Figura 4.2. Percentual de Acertos.

- O comportamento geral do CBS para esta base de dados é apresentado na matriz de confusão de valores médios percentuais na Tabela 4.3. Nela, pode-se observar que a grande maioria dos erros de predição cometidos ocorrem na vizinhança da diagonal principal, não caracterizando erros de grande gravidade.

ACERTO MÉDIO PERCENTUAL						58.46
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL						
		PREDITO				
		1	2	3	4	5
ESPERADO	1	10.60	1.05	0.00	0.00	0.00
	2	0.79	26.36	8.20	1.14	0.00
	3	0.94	6.40	9.91	6.71	0.00
	4	0.65	3.13	5.94	5.35	2.80
	5	0.16	0.00	0.00	3.63	6.23

Tabela 4.3. Matriz de Confusão – Valores Médios Percentuais.

4.2.2 Testes Comparativos

Foram realizados cinco testes comparativos entre o CBS e o programa aiNet versão 1.25 que utiliza uma Rede Neural de duas camadas. Em todos os testes, a base de dados foi dividida em dois conjuntos: conjunto de treinamento com 166 amostras e conjunto de teste com as 60 amostras restantes. As amostras foram escolhidas aleatoriamente na base de dados e os resultados são apresentados na Tabela 4.4 abaixo.

Percentual de Acertos		
Teste	**CBS**	**aiNet**
1	66.67	68.33
2	65.00	60.00
3	48.67	41.67
4	96.67	100.00
5	78.33	85.00
Média	71.07	71.00

Tabela 4.4. Percentual de Acertos do CBS e do aiNet.

Nota-se que os dois classificadores se comportaram de forma bastante semelhante, sendo o CBS um pouco mais uniforme, isto é, ele foi pior que o aiNet na predição das amostras mais fáceis (testes 1, 4 e 5) e melhor nas mais difíceis (testes 2 e 3).

A Tabela 4.5 abaixo mostra a matriz de confusão de cada teste realizado, onde se observa que os erros cometidos pelo CBS foram, em geral, mais concentrados na vizinhança da diagonal principal, que os erros do aiNet.

MATRIZ DE CONFUSÃO																
TESTE	**CBS**								**aiNet**							
1			**PREDITO**								**PREDITO**					
			1	**2**	**3**	**4**	**5**				**1**	**2**	**3**	**4**	**5**
	ESPERADO	**1**	3	3					**ESPERADO**	**1**	1	5			
		2		21						**2**	1	19		1	
		3		3	9	3				**3**		3	11		1
		4		4	6	4				**4**		3	3	7	1
		5				1	3			**5**			1		3
2			**PREDITO**								**PREDITO**					
			1	**2**	**3**	**4**	**5**				**1**	**2**	**3**	**4**	**5**
	ESPERADO	**1**	13	1					**ESPERADO**	**1**	8	5	1		
		2	2	13	3					**2**		13	3	2	
		3		2	7	5				**3**		2	6	6	
		4		1	4	3				**4**			3	5	
		5				3	3			**5**			1	1	4
3			**PREDITO**								**PREDITO**					
			1	**2**	**3**	**4**	**5**				**1**	**2**	**3**	**4**	**5**
	ESPERADO	**1**	4						**ESPERADO**	**1**	1	2	1		
		2	1	8						**2**		9			
		3		5	10	5				**3**		11	4	5	
		4		2	11	4				**4**		5	3	7	2
		5			1	6	3			**5**			1	5	4
4			**PREDITO**								**PREDITO**					
			1	**2**	**3**	**4**	**5**				**1**	**2**	**3**	**4**	**5**
	ESPERADO	**1**	14						**ESPERADO**	**1**	14				
		2		20						**2**		20			
		3			12					**3**			12		
		4				4	1			**4**				5	
		5				1	8			**5**					9
5			**PREDITO**								**PREDITO**					
			1	**2**	**3**	**4**	**5**				**1**	**2**	**3**	**4**	**5**
	ESPERADO	**1**							**ESPERADO**	**1**					
		2	2	39		1				**2**	4	36	1	1	
		3		2	8	6				**3**		1	14	1	
		4			2					**4**			1	1	
		5								**5**					

Tabela 4.5. Matrizes de Confusão dos Testes Comparativos.

CAPÍTULO 5
CONCLUSÃO

A degradação das malhas viárias terrestres brasileiras, verificada nas últimas três décadas, através de uma brusca mudança nos procedimentos de conservação e manutenção, fazendo com que a engenharia passasse a atuar somente nos moldes corretivos e emergências, levaram dois especialistas em geotecnia a desenvolver uma metodologia que disponibiliza uma visão completa e integrada de todo o trecho da via, possibilitando a elaboração do planejamento estratégico das intervenções necessárias.

Essa metodologia foi apresentada em resumo, destacando que um de seus pontos mais importantes é o relacionamento do risco representado por cada local. Como o valor proposto para medir o risco é discreto e sua determinação é feita a partir de parâmetros coletados em vistoria no local, a solução que se apresenta é o desenvolvimento de uma atividade preditiva de DM para classificar esse valor.

Foi realizado um estudo sobre classificação de dados, detalhando os três métodos considerados mais usuais e escolhido o classificador Bayesiano simples como o algoritmo a ser adotado.

A implementação do classificador CBS foi submetida a diversos experimentos, demonstrando bons resultados tanto na acurácia de predição como no desempenho, que são os critérios mais importantes para o problema em questão.

De acordo com o que foi apresentado, pode-se concluir que o classificador CBS representa uma boa solução para a avaliação de riscos em vias de transporte terrestre, podendo ser incluído como módulo inteligente no sistema gerenciador da base de dados da metodologia proposta.

REFERÊNCIAS BIBLIOGRÁFICAS

CUROTTO, C. L., *Integração de Recursos de Data Mining com Gerenciadores de Bancos de Dados Relacionais*, Tese de Doutorado, COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2003.

FAYYAD, U.M., PIATETSKY-SHAPIRO, G. & SMITH, P. "From Data Mining to Knowledge Discovery: An Overview", In: *Advances in Knowledge Discovery and Data Mining*, FAYYAD, U.M. *et alii* (eds.), AAAI/MIT Press, pp. 1-36, 1995.

FAYYAD, U.M., PIATETSKY-SHAPIRO, G. & SMITH, P. "From Data Mining to Knowledge Discovery in Databases", *Al Magazine*, Vol. 17, No. 3, pp. 37-54, 1996[1].

FAYYAD, U.M., PIATETSKY-SHAPIRO, G. & SMITH, P. "Knowledge Discovery and Data Mining: Towards a Unifying Framework", In: *Proc. Of the KDD'96, 2nd Int'l Conf. on Knowledge Discovery and Data Mining*, Portland, Oregon, USA, pp. 82-88, 1996[2].

HAN, J. & KAMBER, M., *Data Mining: Concepts and Techniques*, 1st ed., San Francisco California, USA, Morgan Kaufmann Publishers, 2001.

HAYKIN S., *Redes Neurais: Princípios e Prática*, 2ª ed., Porto Alegre, RGS, Brasil, Bookman Companhia Editora, 2001.

MITCHELL, T. M., *Machine Learning*, Portland, Oregon, USA, McGraw-Hill Companies, Inc., 1997.

REZENDE, S. O., "Introdução", In: *Sistemas Inteligentes: Fundamentos e Aplicações*, Barueri, SP, Brasil, Rezende, S. O. (coord.), Editora Manole Ltda., Cap. 1, pp. 3-11, 2003.

REZENDE, S. O., PUGLIESI, J. B., MELANDA E. A. & DE PAULA, M. F., "Mineração de Dados", In: *Sistemas Inteligentes: Fundamentos e Aplicações*,

Barueri, SP, Brasil, Rezende, S. O. (coord.), Editora Manole Ltda., Cap. 12, pp. 307-336, 2003.

SILVER, D. L., *Knowledge Discovery and Data Mining*, MBA course notes of Dalhousie University, Nova Scotia, Canada, 1998.

THEODORIDIS, S. & KOUTROUMBAS, K, *Pattern Recognition*, Academic Press, 1999.

ANEXO I

Algoritmos do Classificador CBS

A seguir, são apresentadas as versões detalhadas dos algoritmos de treinamento e de classificação do CBS, nas Listagens I.1 e I.2, respectivamente.

```
início   {TREINA}

    N, M, ND, NC, MAXAD, C, S: int;
    XD: vet [1 . . ND] int;
    XC: vet [1 . . NC] real;
    Si: vet [1 . . M] int;
    PXCD: mat [1 . . M, 1 . . MAXAD, 1 . . ND] real;
    MEDIA, DESVIO, CTE: mat [1 . . M, 1 . . NC] real;
    PCi: vet [1 . . M] real;

    {zera S, Si, PXCD, MEDIA, DESVIO}
    {leitura dos parâmetros}
    leia (N, ND, M, MAXAD);
    NC = N – ND;
    repita
            {leitura dos vetores de atributos de uma amostra de treinamento}
            LERX (ND, NC, XD, XC);
            {leitura da classe}
            leia (C);
            {cálculo de S e Si}
            CALCSSi (M, C, S, Si);
            {preparação do cálculo de P (X | Ci)}
            PREPXCi (C, ND, NC, XD, XC, PXCD, MEDIA, DESVIO);
    até EOF;
    {cálculo de P(Ci)}
    CALCPCi (M, S, Si, PCi);
    {cálculo de P (X | Ci), MEDIA, DESVIO e CTE}
    CALCPXCi (M, MAXAD, ND, NC, Si, PXCD, MEDIA, DESVIO, CTE);
    {gravação de N, ND, M, MAXAD, PCi, PXCD, MEDIA, DESVIO e CTE}
    GRAVAR (N, ND, NC, M, MAXAD, PCi, PXCD, MEDIA, DESVIO, CTE);
fim.  {TREINA}

refinamento de  "ler vetores de atributos de uma amostra de treinamento";
    procedimento LERX (ND, NC, XD, XC);

            int: I;

            para I de 1 até ND faça
                    leia (XD[I]);
            fim – para;
            para I de 1 até NC faça
                    leia (XC[I]);
            fim – para;
    fim – procedimento;        { LERX }
```

```
refinamento de "calcular S e Si";
    procedimento CALCSSi (M, C, S, Si);

            S = S + 1;
            Si[C] = Si[C] + 1;

    fim – procedimento;        { CALCSSi }

refinamento de "preparar cálculo de P (X | Ci)";
    procedimento PREPXCi (C, ND, NC, XD, XC, PXCD, MEDIA, DESVIO);

            int: K;

            para K de 1 até ND faça
                    PXCD[C, XD[K], K] = PXCD[C, XD[K], K] + 1;
            fim – para;
            para K de 1 até NC faça
                    MEDIA [C, K] = MEDIA [C, K] + XC[K];
                    DESVIO [C, K] = DESVIO [C, K] + XC[K]**2;
            fim – para;
    fim – procedimento;        { PREPPXCi }

refinamento de "calcular P (Ci)";
    procedimento CALCPCi (M, S, Si, PCi);

            int: I;

            para I de 1 até M faça
                    PCi[I] = Si[I] / S;
            fim – para;
    fim – procedimento;        { CALCPCi }


refinamento de "calcular P (X | Ci), MEDIA, DESVIO e CTE";
    procedimento CALCPXCi (M, MAXAD, ND, NC, Si, PXCD, MEDIA, DESVIO,
CTE);

            int: I, J, K;

            para I de 1 até M faça
                    para J de 1 até MAXAD faça
                            para K de 1 até ND faça
                                    PXCD [I, J, K] = PXCD [I, J, K] / Si[I];
                            fim – para;
                    fim – para;
                    para K de 1 até NC faça
                            MEDIA [I, K] = MEDIA [I, K] / Si[I];
                            DESVIO [I, K] = raiz ((DESVIO [I, K] - MEDIA [I,K]**2) /
                                                    (Si[I] – 1));
                            CTE [I, K] = 1 / raiz (2 * π * DESVIO [I, K]);
                    fim – para;
            fim – para;
    fim – procedimento;        { CALCPXCi }
```

```
refinamento de "gravar N, ND, M, MAXAD, PCi, PXCD, MEDIA, DESVIO e CTE";
    procedimento GRAVAR (N, ND, NC, M, MAXAD, PCi, PXCD, MEDIA, DESVIO,
                         CTE);

        int: I, J, K;

        grave (N, ND, M, MAXAD);
        para I de 1 até M faça
        grave (PCi [I]);
        fim – para;
        para I de 1 até M faça
            para J de 1 até MAXAD faça
                para K de 1 até ND faça
                    grave (PXCD [I, J, K]);
                fim – para;
            fim – para;
        fim – para;
        para I de 1 até M faça
            para K de 1 até NC faça
                grave (MEDIA [I, K]);
            fim – para;
        fim – para;
        para I de 1 até M faça
            para K de 1 até NC faça
                grave (DESVIO [I, K]);
            fim – para;
        fim – para;
        para I de 1 até M faça
            para K de 1 até NC faça
                grave (CTE [I, K]);
            fim – para;
        fim – para;
    fim – procedimento;        { GRAVAR }
```

Listagem I.1. Algoritmo de Treinamento - TREINA.

```
Início  {CLASS}
    N, M, ND, NC, MAXAD, CLASSE: int;
    XD: vet [1 . . ND] int;
    XC: vet [1 . . NC] real;
    PXCD: mat [1 . . M, 1 . . MAXAD, 1 . . ND] real;
    MEDIA, DESVIO, CTE: mat [1 . . M, 1 . . NC] real;
    PCi, PXCi, PCiX: vet [1 . . M] real;

    {leitura de N, ND, M, MAXAD, PCi, PXCD, MEDIA, DESVIO e CTE do arquivo
    treinado}
    LERT (N, ND, NC, M, MAXAD, PCi, PXCD, MEDIA, DESVIO, CTE);
    repita
            {leitura do vetor característico de teste}
            LERX (ND, NC, XD, XC);
            {cálculo de P (X | Ci)}
            CALCPXCi (ND, NC, M, XD, XC, PXCD, MEDIA, DESVIO, CTE, PXCi);
            {cálculo de P (Ci | X) e CLASSE}
            CLASSIFICA (M, PCi, PXCi, PCiX, CLASSE);
            grave (CLASSE);
    até EOF;
fim.  {CLASS}

refinamento de "ler N, ND, M, MAXAD, PCi, PXCD, MEDIA, DESVIO e CTE do
arquivo treinado";
    procedimento LERT (N, ND, NC, M, MAXAD, PCi, PXCD, MEDIA, DESVIO,
                        CTE);
            int: I, J, K;
            leia (N, ND, M, MAXAD);
            NC = N – ND;
            para I de 1 até M faça
            leia (PCi [I]);
            fim – para;
            para I de 1 até M faça
                    para J de 1 até MAXAD faça
                            para K de 1 até ND faça
                                    leia (PXCD [I, J, K]);
                            fim – para;
                    fim – para;
            fim – para;
            para I de 1 até M faça
                    para K de 1 até NC faça
                            leia (MEDIA [I, K]);
                    fim – para;
            fim – para;
            para I de 1 até M faça
                    para K de 1 até NC faça
                            leia (DESVIO [I, K]);
                    fim – para;
            fim – para;
            para I de 1 até M faça
                    para K de 1 até NC faça
                            leia (CTE [I, K]);
                    fim – para;
            fim – para;
    fim – procedimento;        { LERT }
```

```
refinamento de "calcular P (X | Ci)";
    procedimento CALCPXCi (ND, NC, M, XD, XC, PXCD, MEDIA, DESVIO, CTE,
                                    PXCi);

            int: I, K;

            para I de 1 até M faça
                    PXCi[I] = 1;
                    para K de 1 até ND faça
                            PXCi[I] = PXCi[I] * PXCD[I, XD[K], K];
                    fim – para;
            para K de 1 até NC faça
                            PXCi[I] = PXCi[I] * CTE[I, K] * exp (-(XC[K] – MEDIA[I,
                                    K]**2) / (2 * DESVIO[I, K]**2));
                    fim – para;
            fim – para;
    fim – procedimento;         { CALCPXCi }

refinamento de "calcular P (Ci | X) e CLASSE";
    procedimento CLASSIFICA (M, PCi, PXCi, PCiX, CLASSE);

            int: I;
            real: PMAX;

            PMAX = PCi[1] * PXCi[1];
            CLASSE = 1;
            para I de 1 até M faça
                    PCiX[I] = PCi[I] * PXCi[I];
                    se PCiX[I] > PMAX
                            então
                                    PMAX = PCiX[I];
                                    CLASSE = I;
                    fim-se;
            fim – para;
    fim – procedimento;         { CLASSIFICA }
```

Listagem I.2. Algoritmo de Classificação - CLASS.

ANEXO II

Implementações do Classificador CBS

As Listagens II.1 e II.2 mostram, respectivamente, os programas TREINA.CPP e CLASS.CPP, as Tabelas II.1 a II.4 contêm as organizações dos arquivos de entrada e saída e a Tabela II.5, as formas de chamada para execução dos programas.

```
#include <stdio.h>
#include <stdlib.h>
#include <math.h>

#define quad(x)  (x) * (x)

//*** Definição dos valores máximos considerados - dimensões das matrizes ***
#define TM 11
#define TND 22
#define TNC 60
#define TMAXAD 14

//*** PROGRAMA DE TREINAMENTO – TREINA.CPP ***

//*** Protótipos das Funções ***
int LERX (int ND, int NC, int XD[], double XC[], int* PtC);
void CALCSSi (int C, int* PtS, int Si[]);
void PREPXCi (int C, int ND, int NC, int XD[], double XC[], double PXCD[][TMAXAD][TND],
              double MEDIA[][TNC], double DESVIO[][TNC]);
void CALCPCi (int M, int S, int Si[], double PCi[]);
void CALCPXCi (int M, int MAXAD, int ND, int NC, int Si[],
                 double PXCD[][TMAXAD][TND], double MEDIA[][TNC],
                 double DESVIO[][TNC], double CTE[][TNC]);
void GRAVAR (int N, int ND, int NC, int M, int MAXAD, double PCi[],
                 double PXCD[][TMAXAD][TND], double MEDIA[][TNC],
                 double DESVIO[][TNC], double CTE[][TNC]);

//*** Arquivos de entrada XTREINA e de saída TREINADO ***
FILE *XTREINA, *TREINADO;

void main (int argc,char *argv[])
  {
  int N, M, ND, NC, MAXAD, C, S = 0;
  static int XD[TND], Si[TM];
  static double XC[TNC], PXCD[TM][TMAXAD][TND], MEDIA[TM][TNC],
                DESVIO[TM][TNC], CTE[TM][TNC], PCi[TM];

  if(argc!=3
    {
    printf(" \n Erro: Numero de parâmetros incorreto ");
    exit(1);
    }
```

```
   if((XTREINA = fopen(argv[1], "r")) == NULL)
      {
      printf(" \n Erro: Abertura de arquivo para leitura ");
      exit(1);
      }
   //*** Leitura dos parametros do arquivo de treinamento ***
   fscanf(XTREINA, "%d", &N);
   fscanf(XTREINA, "%d", &ND);
   fscanf(XTREINA, "%d", &M);
   fscanf(XTREINA, "%d", &MAXAD);
   NC = N - ND;
   //*** leitura dos vetores de atributos de uma amostra de treinamento e sua classe ***
   while (LERX(ND, NC, XD, XC, &C))
      {
      //*** cálculo de S e Si ***
      CALCSSi (C, &S, Si);
      //*** preparação do cálculo de P (X | Ci), MEDIA e DESVIO ***
      PREPXCi (C, ND, NC, XD, XC, PXCD, MEDIA, DESVIO);
      }
   fclose(XTREINA);
   //*** cálculo de P(Ci) ***
   CALCPCi (M, S, Si, PCi);
   //*** cálculo de P (X | Ci), MEDIA, DESVIO e CTE ***
   CALCPXCi (M, MAXAD, ND, NC, Si, PXCD, MEDIA, DESVIO, CTE);
   if((TREINADO = fopen(argv[2], "w")) == NULL)
      {
      printf(" \n Erro: abertura de arquivo para gravação ");
      exit(1);
      }
   //*** gravação de N, ND, M, MAXAD, PCi, PXCD, MEDIA, DESVIO e CTE ***
   GRAVAR (N, ND, NC, M, MAXAD, PCi, PXCD, MEDIA, DESVIO, CTE);
   fclose(TREINADO);
   }   //*** Fim main – TREINA.CPP ***


//*** lê vetores de atributos de uma amostra de treinamento e sua  classe ***
int LERX (int ND, int NC, int XD[], double XC[], int* PtC)
   {
   register int I;
   int *PtXD = XD;
   double *PtXC = XC;

   for (I = 0; I < ND; I++)
      {
      if (fscanf(XTREINA, "%d ", PtXD++) == EOF)
         return (0);
      }
   for (I = 0; I < NC; I++)
      {
      if (fscanf(XTREINA, "%lf ", PtXC++) == EOF)
         return (0);
      }
   fscanf(XTREINA, "%d\n", PtC);
   return (1);
   }   //*** Fim LERX ***
```

```
//*** calcula S e Si ***
void CALCSSi ( int C, int* PtS, int Si[])
   {
   int *PtSi = Si;

   *PtS = *PtS + 1;
   PtSi = PtSi + C - 1;
   *PtSi = *PtSi + 1;
   }   //*** Fim CALCSSi ***


//*** prepara cálculo de P (X | Ci), MEDIA e DESVIO ***
void PREPXCi (int C, int ND, int NC, int XD[], double XC[],
                     double PXCD[][TMAXAD][TND], double MEDIA[][TNC],
                     double DESVIO[][TNC])
   {
   register int I;
   int *PtXD;
   double *PtXC, (*PtPXCD)[TMAXAD][TND] = PXCD, (*PtMEDIA)[TNC] = MEDIA,
            (*PtDESVIO)[TNC] = DESVIO;

   PtPXCD = PtPXCD + C - 1;
   for (I = 0; I < ND; I++)
      {
      PtXD = XD + I;
      (*PtPXCD)[*PtXD - 1][I] = (*PtPXCD)[*PtXD - 1][I] + 1;
      }
   PtMEDIA = PtMEDIA + C - 1;
   PtDESVIO = PtDESVIO + C - 1;
   for (I = 0; I < NC; I++)
      {
      PtXC = XC + I;
      (*PtMEDIA)[I] = (*PtMEDIA)[I] + *PtXC;
      (*PtDESVIO)[I] = (*PtDESVIO)[I] + quad((*PtXC));
      }
   }   //*** Fim PREPXCi ***


//*** calcula P(Ci) ***
void CALCPCi (int M, int S, int Si[], double PCi[])
   {
   register int I;
   int *PtSi = Si;
   double *PtPCi = PCi;

   for (I = 0; I < M; I++, PtSi++, PtPCi++)
      *PtPCi = (double) *PtSi / S;
   }   //*** Fim CALCPCi ***
```

```
//*** calcula P (X | Ci), MEDIA, DESVIO e CTE ***
void CALCPXCi (int M, int MAXAD, int ND, int NC, int Si[],
                        double PXCD[][TMAXAD][TND], double MEDIA[][TNC],
                        double DESVIO[][TNC], double CTE[][TNC])
  {
  register int I, J, K;
  int *PtSi = Si;
  double SOMA, (*PtPXCD)[TMAXAD][TND] = PXCD, (*PtMEDIA)[TNC] = MEDIA,
           (*PtDESVIO)[TNC] = DESVIO, (*PtCTE)[TNC] = CTE;

  for (I = 0; I < M; I++, PtPXCD++, PtSi++, PtMEDIA++, PtDESVIO++, PtCTE++)
    {
    for (J = 0; J < MAXAD; J++)
      {
      for (K = 0; K < ND; K++)
        {
        if (*PtSi)
          (*PtPXCD)[J][K] = (*PtPXCD)[J][K] / *PtSi;
        else
          (*PtPXCD)[J][K] = 0.0;
        }
      }
      for (K = 0; K < NC; K++)
        {
        SOMA = (*PtMEDIA)[K];
        if (*PtSi)
          {
          (*PtMEDIA)[K] = SOMA / *PtSi;
          (*PtDESVIO)[K] = sqrt(((*PtDESVIO)[K] - (quad(SOMA) / *PtSi)) / (*PtSi - 1));
          if ((*PtDESVIO)[K])
            (*PtCTE)[K] = (double) 1 / sqrt(2 * M_PI * (*PtDESVIO)[K]);
          else
            (*PtCTE)[K] = 0.0;
          }
        else
          {
          (*PtMEDIA)[K] = 0.0;
          (*PtDESVIO)[K] = 0.0;
          (*PtCTE)[K] = 0.0;
          }
        }
    }
  }   //*** Fim CALCPXCi ***
```

```
//*** grava N, ND, M, MAXAD, PCi, P (X | Ci), MEDIA, DESVIO e CTE ***
void GRAVAR (int N, int ND, int NC, int M, int MAXAD, double PCi[],
                    double PXCD[][TMAXAD][TND], double MEDIA[][TNC],
                    double DESVIO[][TNC], double CTE[][TNC])
  {
  register int I, J, K;
  double *PtPCi = PCi, (*PtPXCD)[TMAXAD][TND] = PXCD, (*PtMEDIA)[TNC] = MEDIA,
         (*PtDESVIO)[TNC] = DESVIO, (*PtCTE)[TNC] = CTE;
  fprintf(TREINADO, "%d\n", N);
  fprintf(TREINADO, "%d\n", ND);
  fprintf(TREINADO, "%d\n", M);
  fprintf(TREINADO, "%d\n", MAXAD);
  for (I = 0; I < M; I++)
    {
    fprintf(TREINADO, "%lf ", *PtPCi++);
    }
  fprintf(TREINADO, "\n", I);
  for (I = 0; I < M; I++, PtPXCD++)
    {
    for (J = 0; J < MAXAD; J++)
      {
      for (K = 0; K < ND; K++)
        {
        fprintf(TREINADO, "%lf ", (*PtPXCD)[J][K]);
        }
      fprintf(TREINADO, "\n", K);
      }
    }
  for (I = 0; I < M; I++, PtMEDIA++)
    {
    for (J = 0; J < NC; J++)
      {
      fprintf(TREINADO, "%lf ", (*PtMEDIA)[J]);
      }
    fprintf(TREINADO, "\n", J);
    }
  for (I = 0; I < M; I++, PtDESVIO++)
    {
    for (J = 0; J < NC; J++)
      {
      fprintf(TREINADO, "%lf ", (*PtDESVIO)[J]);
      }
    fprintf(TREINADO, "\n", J);
    }
  for (I = 0; I < M; I++, PtCTE++)
    {
    for (J = 0; J < NC; J++)
      {
      fprintf(TREINADO, "%lf ", (*PtCTE)[J]);
      }
    fprintf(TREINADO, "\n", J);
    }
  }  //*** Fim GRAVAR  ***
```

Listagem II.1. Programa de Treinamento – TREINA.CPP.

```
#include <stdio.h>
#include <stdlib.h>
#include <math.h>

#define quad(x)  (x) * (x)

//*** Definição dos valores máximos considerados - dimensões das matrizes ***
#define TM 11
#define TND 22
#define TNC 60
#define TMAXAD 14

//*** PROGRAMA DE CLASSIFICAÇÃO – CLASS.CPP ***

//*** Protótipos das Funções ***
void LERPR (int* PtND, int* PtNC, int* PtM, int* PtMAXAD, double PCi[],
                  double PXCD[][TMAXAD][TND], double MEDIA[][TNC],
                  double DESVIO[][TNC], double CTE[][TNC]);
int LERXT (int ND, int NC, int XD[], double XC[]);
void CALCPXCi (int ND, int NC, int M, int XD[], double XC[],
                      double PXCD[][TMAXAD][TND], double MEDIA[][TNC],
                      double DESVIO[][TNC], double CTE[][TNC], double PXCi[]);
void CLASSIFICA (int M, double PCi[], double PXCi[], double PCiX[], int* PtCLASSE);

//*** Arquivos de entrada TREINADO e XTESTE e de saída CLASSIFICADO ***
FILE *TREINADO, *XTESTE, *CLASSIFICADO;

void main (int argc,char *argv[])
  {
  int M, ND, NC, MAXAD, CLASSE;
  static int XD[TND];
  static double XC[TNC], PXCD[TM][TMAXAD][TND], MEDIA[TM][TNC],
                   DESVIO[TM][TNC], CTE[TM][TNC], PCi[TM], PXCi[TM], PCiX[TM];

  if(argc!=4)
    {
    printf(" \n Erro: Numero de parâmetros incorreto ");
    exit(1);
    }
  if((TREINADO = fopen(argv[1], "r")) == NULL)
    {
    printf(" \n Erro: Abertura de arquivo treinado ");
    exit(1);
    }

  //*** leitura do arquivo TREINADO ***
  LERPR (&ND, &NC, &M, &MAXAD, PCi, PXCD, MEDIA, DESVIO, CTE);
  fclose(TREINADO);

  if((XTESTE = fopen(argv[2], "r")) == NULL)
    {
    printf(" \n Erro: Abertura de arquivo de teste ");
    exit(1);
    }
```

```
  if((CLASSIFICADO = fopen(argv[3], "w")) == NULL)
    {
    printf(" \n Erro: Abertura de arquivo de saída ");
    exit(1);
    }

  //*** leitura dos vetores de atributos de uma amostra de teste ***
  while (LERXT(ND, NC, XD, XC))
    {
    //*** cálculo de P (X | Ci) ***
    CALCPXCi (ND, NC, M, XD, XC, PXCD, MEDIA, DESVIO, CTE, PXCi);
    //*** cálculo de P (Ci | X) e CLASSE ***
    CLASSIFICA (M, PCi, PXCi, PCiX, &CLASSE);
    fprintf(CLASSIFICADO, "%d\n", CLASSE);
    }
  fclose(XTESTE);
  fclose(CLASSIFICADO);
  }   //*** Fim main – CLASS.CPP ***

//*** lê arquivo TREINADO ***
void LERPR (int* PtND, int* PtNC, int* PtM, int* PtMAXAD, double PCi[],
                  double PXCD[][TMAXAD][TND], double MEDIA[][TNC],
                  double DESVIO[][TNC], double CTE[][TNC])
  {
  register int I, J, K;
  int N;
  double *PtPCi = PCi, (*PtPXCD)[TMAXAD][TND] = PXCD, (*PtMEDIA)[TNC] = MEDIA,
           (*PtDESVIO)[TNC] = DESVIO, (*PtCTE)[TNC] = CTE;

  fscanf(TREINADO, "%d", &N);
  fscanf(TREINADO, "%d", PtND);
  fscanf(TREINADO, "%d", PtM);
  fscanf(TREINADO, "%d", PtMAXAD);
  *PtNC = N - *PtND;
  for (I = 0; I < *PtM; I++)
    {
    fscanf(TREINADO, "%lf ", PtPCi++);
    }
  for (I = 0; I < *PtM; I++, PtPXCD++)
    {
    for (J = 0; J < *PtMAXAD; J++)
      {
      for (K = 0; K < *PtND; K++)
        {
        fscanf(TREINADO, "%lf ", &(*PtPXCD)[J][K]);
        }
      }
    }
  for (I = 0; I < *PtM; I++, PtMEDIA++)
    {
    for (J = 0; J < *PtNC; J++)
      {
      fscanf(TREINADO, "%lf ", &(*PtMEDIA)[J]);
      }
    }
```

```
   for (I = 0; I < *PtM; I++, PtDESVIO++)
      {
      for (J = 0; J < *PtNC; J++)
         {
         fscanf(TREINADO, "%lf ", &(*PtDESVIO)[J]);
         }
      }
   for (I = 0; I < *PtM; I++, PtCTE++)
      {
      for (J = 0; J < *PtNC; J++)
         {
         fscanf(TREINADO, "%lf ", &(*PtCTE)[J]);
         }
      }
   }   //*** Fim LERPR ***


//*** le vetor característico de teste ***
int LERXT (int ND, int NC, int XD[], double XC[])
   {
   register int I;
   int *PtXD = XD;
   double *PtXC = XC;

   for (I = 0; I < ND; I++)
      {
      if (fscanf(XTESTE, "%d ", PtXD++) == EOF)
         return (0);
      }
   for (I = 0; I < NC; I++)
      {
      if (fscanf(XTESTE, "%lf ", PtXC++) == EOF)
         return (0);
      }
   return (1);
   }   //*** Fim LERX ***


//*** calcula P (X | Ci) ***
void CALCPXCi (int ND, int NC, int M, int XD[], double XC[],
                        double PXCD[][TMAXAD][TND], double MEDIA[][TNC],
                        double DESVIO[][TNC], double CTE[][TNC], double PXCi[])
   {
   register int I, K;
   int *PtXD;
   double *PtXC, (*PtPXCD)[TMAXAD][TND] = PXCD, (*PtMEDIA)[TNC] = MEDIA,
               (*PtDESVIO)[TNC] = DESVIO, (*PtCTE)[TNC] = CTE, *PtPXCi = PXCi;

   for (I = 0; I < M; I++, PtPXCi++, PtPXCD++, PtCTE++, PtMEDIA++, PtDESVIO++)
      {
      *PtPXCi = 1.0;
      for (K = 0, PtXD = XD; K < ND; K++, PtXD++)
         {
         *PtPXCi = *PtPXCi * (*PtPXCD)[*PtXD - 1][K];
         }
```

```
      for (K = 0, PtXC = XC; K < NC; K++, PtXC++)
        {
        if ((*PtDESVIO)[K])
          *PtPXCi = *PtPXCi * (*PtCTE)[K] * exp(-quad(*PtXC - (*PtMEDIA)[K]) /
                    (2 * quad((*PtDESVIO)[K])));
        else
          *PtPXCi = 0.0;
        }
      }
   }   //*** Fim CALCPXCi ***


//*** calcula P (Ci | X) e CLASSE ***
void CLASSIFICA (int M, double PCi[], double PXCi[], double PCiX[], int* PtCLASSE)
   {
   register int I;
   double PMAX, *PtPCi = PCi, *PtPXCi = PXCi, *PtPCiX = PCiX;

   PMAX = *PtPCi * *PtPXCi;
   *PtCLASSE = 1;
   for (I = 0; I < M; I++, PtPCi++, PtPXCi++, PtPCiX++)
      {
      *PtPCiX = *PtPCi * *PtPXCi;
      if (*PtPCiX > PMAX)
         {
         PMAX = *PtPCiX;
         *PtCLASSE = I + 1;
         }
      }
   }   //*** Fim CLASSIFICA ***
```

Listagem II.2. Programa de Classificação – CLASS.CPP.

Nome Interno do Arquivo: XTREINA	
Tipo	Texto
Função	Entrada para o programa TREINA.CPP
Objetivo	Fornece quatro parâmetros que identificam a constituição das amostras de treinamento e, em seguida, todas as amostras de treinamento

Estrutura de Dados	
Identificador	**Descrição**
N	Parâmetro com o número total de atributos
ND	Parâmetro com o número de atributos discretos
M	Parâmetro com o número total de classes
MAXAD	Parâmetro com o número máximo de valores distintos que um atributo discreto qualquer pode assumir
Amostras de treinamento contendo cada uma:	
ND	Vetor de atributos discretos (cada atributo é representado por um valor inteiro a partir de 1)
NC	Vetor de atributos contínuos
C	Classe da amostra

Tabela II.1. Arquivo XTREINA.

Nome Interno do Arquivo: TREINADO	
Tipo	Texto
Função	Saída do programa TREINA.CPP e entrada para o programa CLASS.CPP
Objetivo	Contem o conhecimento extraído das amostras de treinamento

Estrutura de Dados	
Identificador	**Descrição**
N	Parâmetro com o número total de atributos
ND	Parâmetro com o número de atributos discretos
M	Parâmetro com o número total de classes
MAXAD	Parâmetro com o número máximo de valores distintos que um atributo discreto qualquer pode assumir
PCi	Probabilidade *a priori* da classe Ci
PXCD	Probabilidade que ocorra um valor X_k para um atributo discreto A_k, numa amostra X de classe Ci
MEDIA	Média aritmética de todos os valores de cada atributo contínuo A_k pertencentes às amostras de treinamento de classe Ci
DESVIO	Desvio padrão de todos os valores de cada atributo contínuo A_k pertencentes às amostras de treinamento de classe Ci
CTE	Constantes utilizadas na distribuição normal, relativas a cada atributo contínuo A_k pertencentes às amostras de treinamento de classe Ci

Tabela II.2. Arquivo TREINADO.

Nome Interno do Arquivo: XTESTE	
Tipo	Texto
Função	Entrada para o programa CLASS.CPP
Objetivo	Fornece todas as amostras de teste

Estrutura de Dados	
Identificador	**Descrição**
Amostras de teste contendo cada uma:	
ND	Vetor de atributos discretos (cada atributo é representado por um valor inteiro a partir de 1)
NC	Vetor de atributos contínuos

Tabela II.3. Arquivo XTESTE.

Nome Interno do Arquivo: CLASSIFICADO	
Tipo	Texto
Função	Saída do programa CLASS.CPP
Objetivo	Fornece as respectivas classes das amostras de teste

Estrutura de Dados	
Identificador	**Descrição**
CLASSE	Classe predita para uma amostra desconhecida X

Tabela II.4. Arquivo CLASSIFICADO.

Formas de Chamada dos Programas para Execução
treina <nome1>.txt <nome2>.txt
class <nome2>.txt <nome3>.txt <nome4>.txt

Onde:	
nome1	Nome externo do arquivo XTREINA
nome2	Nome externo do arquivo TREINADO
nome3	Nome externo do arquivo XTESTE
nome4	Nome externo do arquivo CLASSIFICADO

Tabela II.5. Formas de Chamada para Execução.

ANEXO III

Experimentos Computacionais

Os resultados específicos obtidos nas validações cruzadas com as bases de dados acadêmicas e com o *InfraSystem* são apresentadas nas Tabelas III.1 a III.22, conforme a seguinte relação:

Base de Dados	**Tabela**
Blood Testing	III.1
Breast Cancer 1	III.2
Breast Cancer 2	III.3
Credit Screening	III.4
Diabetes	III.5
Echocardiogram	III.6
Glass	III.7
Images	III.8
InfraSystem	III.9
Iris	III.10
Mushroom	III.11
Parity 3	III.12
Parity 4	III.13
Parity 5	III.14
Sleepdata1	III.15
Sleepdata2	III.16
Sonar	III.17
Spiral	III.18
Synthetic1	III.19
Vowel	III.20
Wine	III.21
WNBA	III.22

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Blood Testing	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° Amostras / Conjunto**	69
Teste		1	2	3		
Acertos	**Abs.**	62	61	61		
	Per.	89.86	88.41	88.41		

Conjuntos de Teste		4				**N° Amostras / Conjunto**	51
Teste		1	2	3	4		
Acertos	**Abs.**	45	48	43	44		
	Per.	88.24	94.12	84.31	86.27		

Conjuntos de Teste		5					**N° Amostras / Conjunto**	41
Teste		1	2	3	4	5		
Acertos	**Abs.**	37	37	33	36	38		
	Per.	90.24	90.24	80.49	87.80	92.68		

Conjuntos de Teste		6						**N° Amostras / Conjunto**	34
Teste		1	2	3	4	5	6		
Acertos	**Abs.**	31	30	33	27	29	31		
	Per.	91.18	88.24	97.06	79.41	85.29	91.18		

Conjuntos de Teste		7							**N° Amostras / Conjunto**	29
Teste		1	2	3	4	5	6	7		
Acertos	**Abs.**	26	26	28	24	26	24	26		
	Per.	89.66	89.66	96.55	82.76	89.66	82.76	89.66		

Conjuntos de Teste		8								**N° Amostras / Conjunto**	25
Teste		1	2	3	4	5	6	7	8		
Acertos	**Abs.**	22	23	23	24	18	24	21	22		
	Per.	88.00	92.00	92.00	96.00	72.00	96.00	84.00	88.00		

Conjuntos de Teste		9									**N° Amostras / Conjunto**	22
Teste		1	2	3	4	5	6	7	8	9		
Acertos	**Abs.**	19	22	18	22	18	18	21	18	19		
	Per.	86.36	100.00	81.82	100.00	81.82	81.82	95.45	81.82	86.36		

Conjuntos de Teste		10										**N° Amostras / Conjunto**	20
Teste		1	2	3	4	5	6	7	8	9	10		
Acertos	**Abs.**	17	20	16	20	19	15	18	17	18	17		
	Per.	85.00	100.00	80.00	100.00	95.00	75.00	90.00	85.00	90.00	85.00		

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Blood Testing	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		69	51	41	34	29	25	22	20	
Mínimo	**Absoluto**	61	43	33	27	24	18	18	15	
	Percentual	88.41	84.31	80.49	79.41	82.76	72.00	81.82	75.00	**72.00**
Máximo	**Absoluto**	62	48	38	33	28	24	22	20	
	Percentual	89.86	94.12	92.68	97.06	96.55	96.00	100.00	100.00	**100.00**
Médio	**Absoluto**	61.33	45	36.2	30.17	25.71	22.12	19.44	17.7	
	Percentual	88.89	88.24	88.29	88.73	88.67	88.50	88.38	88.50	**88.51**

NÚMERO DE AMOSTRAS POR CLASSE			
Classe	1	2	**Total**
Abs.	75	134	209
Per.	35.89	64.11	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	**72.00**
CONJUNTOS DE TESTE	8
AMOSTRAS / CONJUNTO	25
TESTE	5
ACERTOS - ABS.	18

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	3	6
	2	1	15

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	**100.00**
CONJUNTOS DE TESTE	9
AMOSTRAS / CONJUNTO	22
TESTE	2
ACERTOS - ABS.	22

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	8	0
	2	0	14

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Blood Testing	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL			**88.51**
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL			
		PREDITO	
		1	**2**
ESPERADO	**1**	26.63	8.97
	2	2.52	61.88

Tabela III.1. Validação Cruzada – Base de Dados *Blood Testing*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Breast Cancer 1 - Atributos Contínuos	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° de Amostras / Conjunto**	227
Teste		1	2	3		
Acertos	**Abs.**	211	218	222		
	Per.	92.95	96.04	97.80		

Conjuntos de Teste		4				**N° de Amostras / Conjunto**	170
Teste		1	2	3	4		
Acertos	**Abs.**	160	161	165	165		
	Per.	94.12	94.71	97.06	97.06		

Conjuntos de Teste		5				**N° de Amostras / Conjunto**	135
Teste		1	2	3	4	5	
Acertos	**Abs.**	127	126	129	133	131	
	Per.	94.07	93.33	95.56	98.52	97.04	

Conjuntos de Teste		6					**N° de Amostras / Conjunto**	113
Teste		1	2	3	4	5	6	
Acertos	**Abs.**	106	105	109	108	111	110	
	Per.	93.81	92.92	96.46	95.58	98.23	97.35	

Conjuntos de Teste		7						**N° de Amostras / Conjunto**	97
Teste		1	2	3	4	5	6	7	
Acertos	**Abs.**	91	91	90	93	96	95	94	
	Per.	93.81	93.81	92.78	95.88	98.97	97.94	96.91	

Conjuntos de Teste		8							**N° de Amostras / Conjunto**	84
Teste		1	2	3	4	5	6	7	8	
Acertos	**Abs.**	78	80	77	82	80	83	81	82	
	Per.	92.86	95.24	91.67	97.62	95.24	98.81	96.43	97.62	

Conjuntos de Teste		9								**N° de Amostras / Conjunto**	75
Teste		1	2	3	4	5	6	7	8	9	
Acertos	**Abs.**	69	72	69	71	72	73	74	73	73	
	Per.	92.00	96.00	92.00	94.67	96.00	97.33	98.67	97.33	97.33	

Conjuntos de Teste		10									**N° de Amostras / Conjunto**	67	
Teste		1	2	3	4	5	6	7	8	9	10		
Acertos	**Abs.**	62	65	63	63	65	63	66	66	65	65		
	Per.	92.54	97.01	94.03	94.03	97.01	94.03	98.51	98.51	97.01	97.01		

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Breast Cancer 1 - Atributos Contínuos	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		227	170	135	113	97	84	75	67	
Mínimo	**Absoluto**	211	160	126	105	90	77	69	62	
	Percentual	92.95	94.12	93.33	92.92	92.78	91.67	92.00	92.54	**91.67**
Máximo	**Absoluto**	222	165	133	111	96	83	74	66	
	Percentual	97.80	97.06	98.52	98.23	98.97	98.81	98.67	98.51	**98.97**
Médio	**Absoluto**	217	162.75	129.2	108.17	92.86	80.38	71.78	64.3	
	Percentual	95.59	95.74	95.70	95.72	95.73	95.68	95.70	95.97	**95.75**

NÚMERO DE AMOSTRAS POR CLASSE			
Classe	1	2	**Total**
Abs.	239	444	683
Per.	34.99	65.01	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	91.67
CONJUNTOS DE TESTE	8
AMOSTRAS / CONJUNTO	84
TESTE	3
ACERTOS - ABS.	77

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	29	0
	2	7	48

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	98.97
CONJUNTOS DE TESTE	7
AMOSTRAS / CONJUNTO	97
TESTE	5
ACERTOS - ABS.	96

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	34	0
	2	1	62

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Breast Cancer 1 - Atributos Contínuos

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL				**95.75**
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL				
		PREDITO		
		1	**2**	
ESPERADO	**1**	34.34	0.30	
	2	3.95	61.41	

Tabela III.2. Validação Cruzada – Base de Dados *Breast Cancer 1*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Breast Cancer 2 - Atributos Discretos	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3		**N° de Amostras / Conjunto**	227
Teste		1	2	3	
Acertos	**Abs.**	216	220	222	
	Per.	95.15	96.92	97.80	

Conjuntos de Teste		4			**N° de Amostras / Conjunto**	170
Teste		1	2	3	4	
Acertos	**Abs.**	163	162	167	165	
	Per.	95.88	95.29	98.24	97.06	

Conjuntos de Teste		5				**N° de Amostras / Conjunto**	135
Teste		1	2	3	4	5	
Acertos	**Abs.**	130	128	131	134	130	
	Per.	96.30	94.81	97.04	99.26	96.30	

Conjuntos de Teste		6					**N° de Amostras / Conjunto**	113
Teste		1	2	3	4	5	6	
Acertos	**Abs.**	109	108	109	110	112	109	
	Per.	96.46	95.58	96.46	97.35	99.12	96.46	

Conjuntos de Teste		7						**N° de Amostras / Conjunto**	97
Teste		1	2	3	4	5	6	7	
Acertos	**Abs.**	94	92	91	95	96	96	93	
	Per.	96.91	94.85	93.81	97.94	98.97	98.97	95.88	

Conjuntos de Teste		8							**N° de Amostras / Conjunto**	84
Teste		1	2	3	4	5	6	7	8	
Acertos	**Abs.**	81	80	79	82	82	83	81	82	
	Per.	96.43	95.24	94.05	97.62	97.62	98.81	96.43	97.62	

Conjuntos de Teste		9								**N° de Amostras / Conjunto**	75
Teste		1	2	3	4	5	6	7	8	9	
Acertos	**Abs.**	72	72	71	71	73	74	75	72	73	
	Per.	96.00	96.00	94.67	94.67	97.33	98.67	100.00	96.00	97.33	

Conjuntos de Teste		10									**N° de Amostras / Conjunto**	67	
Teste		1	2	3	4	5	6	7	8	9	10		
Acertos	**Abs.**	64	66	64	63	65	65	66	67	64	65		
	Per.	95.52	98.51	95.52	94.03	97.01	97.01	98.51	100.00	95.52	97.01		

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Breast Cancer 2 - Atributos Discretos	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		227	170	135	113	97	84	75	67	
Mínimo	**Absoluto**	216	162	128	108	91	79	71	63	
	Percentual	95.15	95.29	94.81	95.58	93.81	94.05	94.67	94.03	**93.81**
Máximo	**Absoluto**	222	167	134	112	96	83	75	67	
	Percentual	97.80	98.24	99.26	99.12	98.97	98.81	100.00	100.00	**100.00**
Médio	**Absoluto**	219.33	164.25	130.6	109.5	93.86	81.25	72.56	64.9	
	Percentual	96.62	96.62	96.74	96.90	96.76	96.73	96.74	96.87	**96.77**

NÚMERO DE AMOSTRAS POR CLASSE			
Classe	1	2	**Total**
Abs.	239	444	683
Per.	34.99	65.01	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL		**93.81**
CONJUNTOS DE TESTE	7	
AMOSTRAS / CONJUNTO	97	
TESTE	3	
ACERTOS - ABS.	91	

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	33	1
	2	5	58

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL		**100.00**
CONJUNTOS DE TESTE	9	
AMOSTRAS / CONJUNTO	75	
TESTE	7	
ACERTOS - ABS.	75	

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	26	0
	2	0	49

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Breast Cancer 2 - Atributos Discretos

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL				**96.77**
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL				
		PREDITO		
		1	**2**	
ESPERADO	**1**	34.24	0.40	
	2	2.83	62.53	

Tabela III.3. Validação Cruzada – Base de Dados *Breast Cancer 2*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Credit Screening	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° Amostras / Conjunto**	217
Teste		1	2	3		
Acertos	**Abs.**	178	167	177		
	Per.	82.03	76.96	81.57		

Conjuntos de Teste		4				**N° Amostras / Conjunto**	163
Teste		1	2	3	4		
Acertos	**Abs.**	128	135	126	135		
	Per.	78.53	82.82	77.30	82.82		

Conjuntos de Teste		5				**N° Amostras / Conjunto**	130
Teste		1	2	3	4	5	
Acertos	**Abs.**	98	115	103	95	116	
	Per.	75.38	88.46	79.23	73.08	89.23	

Conjuntos de Teste		6					**N° Amostras / Conjunto**	108
Teste		1	2	3	4	5	6	
Acertos	**Abs.**	86	92	89	84	83	97	
	Per.	79.63	85.19	82.41	77.78	76.85	89.81	

Conjuntos de Teste		7						**N° Amostras / Conjunto**	93
Teste		1	2	3	4	5	6	7	
Acertos	**Abs.**	72	77	76	79	66	76	82	
	Per.	77.42	82.80	81.72	84.95	70.97	81.72	88.17	

Conjuntos de Teste		8							**N° Amostras / Conjunto**	81
Teste		1	2	3	4	5	6	7	8	
Acertos	**Abs.**	61	65	75	64	64	59	67	69	
	Per.	75.31	80.25	92.59	79.01	79.01	72.84	82.72	85.19	

Conjuntos de Teste		9								**N° Amostras / Conjunto**	71
Teste		1	2	3	4	5	6	7	8	9	
Acertos	**Abs.**	54	54	65	58	61	53	52	58	62	
	Per.	76.06	76.06	91.55	81.69	85.92	74.65	73.24	81.69	87.32	

Conjuntos de Teste		10									**N° Amostras / Conjunto**	64	
Teste		1	2	3	4	5	6	7	8	9	10		
Acertos	**Abs.**	49	48	58	57	51	53	49	45	56	55		
	Per.	76.56	75.00	90.62	89.06	79.69	82.81	76.56	70.31	87.50	85.94		

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Credit Screening	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		217	163	130	108	93	81	71	64	
Mínimo	**Absoluto**	167	126	95	83	66	59	52	45	
	Percentual	76.96	77.30	73.08	76.85	70.97	72.84	73.24	70.31	**70.31**
Máximo	**Absoluto**	178	135	116	97	82	75	65	58	
	Percentual	82.03	82.82	89.23	89.81	88.17	92.59	91.55	90.62	**92.59**
Médio	**Absoluto**	174	131	105.4	88.5	75.43	65.5	57.44	52.1	
	Percentual	80.18	80.37	81.08	81.94	81.11	80.86	80.91	81.41	**81.08**

NÚMERO DE AMOSTRAS POR CLASSE			
Classe	1	2	**Total**
Abs.	357	296	653
Per.	54.67	45.33	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	70.31
CONJUNTOS DE TESTE	10
AMOSTRAS / CONJUNTO	64
TESTE	8
ACERTOS - ABS.	45

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	26	9
	2	10	19

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	92.59
CONJUNTOS DE TESTE	8
AMOSTRAS / CONJUNTO	81
TESTE	3
ACERTOS - ABS.	75

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	41	3
	2	3	34

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Credit Screening

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL				**81.08**
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL				
		PREDITO		
		1	**2**	
ESPERADO	**1**	48.01	6.67	
	2	12.26	33.06	

Tabela III.4. Validação Cruzada – Base de Dados *Credit Screening*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Pima Indians Diabetes	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° Amostras / Conjunto**	255
Teste		1	2	3		
Acertos	**Abs.**	187	174	199		
	Per.	73.33	68.24	78.04		

Conjuntos de Teste		4				**N° Amostras / Conjunto**	192
Teste		1	2	3	4		
Acertos	**Abs.**	147	131	145	149		
	Per.	76.56	68.23	75.52	77.60		

Conjuntos de Teste		5					**N° Amostras / Conjunto**	153
Teste		1	2	3	4	5		
Acertos	**Abs.**	114	110	113	119	116		
	Per.	74.51	71.90	73.86	77.78	75.82		

Conjuntos de Teste		6						**N° Amostras / Conjunto**	127
Teste		1	2	3	4	5	6		
Acertos	**Abs.**	99	86	87	93	101	96		
	Per.	77.95	67.72	68.50	73.23	79.53	75.59		

Conjuntos de Teste		7							**N° Amostras / Conjunto**	109
Teste		1	2	3	4	5	6	7		
Acertos	**Abs.**	83	78	75	84	78	89	80		
	Per.	76.15	71.56	68.81	77.06	71.56	81.65	73.39		

Conjuntos de Teste		8								**N° Amostras / Conjunto**	95
Teste		1	2	3	4	5	6	7	8		
Acertos	**Abs.**	70	73	64	68	72	74	76	71		
	Per.	73.68	76.84	67.37	71.58	75.79	77.89	80.00	74.74		

Conjuntos de Teste		9									**N° Amostras / Conjunto**	84
Teste		1	2	3	4	5	6	7	8	9		
Acertos	**Abs.**	62	64	58	59	64	58	67	66	64		
	Per.	73.81	76.19	69.05	70.24	76.19	69.05	79.76	78.57	76.19		

Conjuntos de Teste		10									**N° Amostras / Conjunto**	76	
Teste		1	2	3	4	5	6	7	8	9	10		
Acertos	**Abs.**	56	58	56	52	52	59	58	59	58	59		
	Per.	73.68	76.32	73.68	68.42	68.42	77.63	76.32	77.63	76.32	77.63		

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Pima Indians Diabetes	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		255	192	153	127	109	95	84	76	
Mínimo	**Absoluto**	174	131	110	86	75	64	58	52	
	Percentual	68.24	68.23	71.90	67.72	68.81	67.37	69.05	68.42	**67.37**
Máximo	**Absoluto**	199	149	119	101	89	76	67	59	
	Percentual	78.04	77.60	77.78	79.53	81.65	80.00	79.76	77.63	**81.65**
Médio	**Absoluto**	186.67	143	114.4	93.67	81	71	62.44	56.7	
	Percentual	73.20	74.48	74.77	73.75	74.31	74.74	74.34	74.61	**74.37**

NÚMERO DE AMOSTRAS POR CLASSE			
Classe	1	2	**Total**
Abs.	500	268	768
Per.	65.10	34.90	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	67.37
CONJUNTOS DE TESTE	8
AMOSTRAS / CONJUNTO	95
TESTE	3
ACERTOS - ABS.	64

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	41	21
	2	10	23

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	81.65
CONJUNTOS DE TESTE	7
AMOSTRAS / CONJUNTO	109
TESTE	6
ACERTOS - ABS.	89

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	60	11
	2	9	29

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Pima Indians Diabetes

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL				**74.37**
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL				
		PREDITO		
		1	**2**	
ESPERADO	**1**	50.90	14.48	
	2	11.15	23.47	

Tabela III.5. Validação Cruzada – Base de Dados *Pima Indians Diabetes*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Echocardiogram	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° de Amostras / Conjunto**	20
Teste		1	2	3		
Acertos	**Abs.**	17	14	14		
	Per.	85.00	70.00	70.00		

Conjuntos de Teste		4				**N° de Amostras / Conjunto**	15
Teste		1	2	3	4		
Acertos	**Abs.**	12	12	10	12		
	Per.	80.00	80.00	66.67	80.00		

Conjuntos de Teste		5				**N° de Amostras / Conjunto**	11
Teste		1	2	3	4	5	
Acertos	**Abs.**	8	10	7	6	9	
	Per.	72.73	90.91	63.64	54.55	81.82	

Conjuntos de Teste		6					**N° de Amostras / Conjunto**	10
Teste		1	2	3	4	5	6	
Acertos	**Abs.**	7	10	6	8	6	8	
	Per.	70.00	100.00	60.00	80.00	60.00	80.00	

Conjuntos de Teste		7						**N° de Amostras / Conjunto**	8
Teste		1	2	3	4	5	6	7	
Acertos	**Abs.**	7	6	7	5	5	6	6	
	Per.	87.50	75.00	87.50	62.50	62.50	75.00	75.00	

Conjuntos de Teste		8							**N° de Amostras / Conjunto**	7
Teste		1	2	3	4	5	6	7	8	
Acertos	**Abs.**	6	5	6	5	5	4	4	6	
	Per.	85.71	71.43	85.71	71.43	71.43	57.14	57.14	85.71	

Conjuntos de Teste		9								**N° de Amostras / Conjunto**	6
Teste		1	2	3	4	5	6	7	8	9	
Acertos	**Abs.**	6	3	6	5	3	4	3	3	5	
	Per.	100.00	50.00	100.00	83.33	50.00	66.67	50.00	50.00	83.33	

Conjuntos de Teste		10									**N° de Amostras / Conjunto**	5
Teste		1	2	3	4	5	6	7	8	9	10	
Acertos	**Abs.**	5	3	4	4	3	4	4	3	4	5	
	Per.	100.00	60.00	80.00	80.00	60.00	80.00	80.00	60.00	80.00	100.00	

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Echocardiogram	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		20	15	11	10	8	7	6	5	
Mínimo	**Absoluto**	14	10	6	6	5	4	3	3	
	Percentual	70.00	66.67	54.55	60.00	62.50	57.14	50.00	60.00	**50.00**
Máximo	**Absoluto**	17	12	10	10	7	6	6	5	
	Percentual	85.00	80.00	90.91	100.00	87.50	85.71	100.00	100.00	**100.00**
Médio	**Absoluto**	15	11.5	8	7.5	6	5.12	4.22	3.9	
	Percentual	75.00	76.67	72.73	75.00	75.00	73.21	70.37	78.00	**74.41**

NÚMERO DE AMOSTRAS POR CLASSE			
Classe	1	2	**Total**
Abs.	44	18	62
Per.	70.97	29.03	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL		**50.00**
CONJUNTOS DE TESTE	9	
AMOSTRAS / CONJUNTO	6	
TESTE	2	
ACERTOS - ABS.	3	

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	2	2
	2	1	1

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL		**100.00**
CONJUNTOS DE TESTE	6	
AMOSTRAS / CONJUNTO	10	
TESTE	2	
ACERTOS - ABS.	10	

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	7	0
	2	0	3

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Echocardiogram

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL				**74.41**
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL				
		PREDITO		
		1	**2**	
ESPERADO	**1**	57.38	15.37	
	2	10.22	17.03	

Tabela III.6. Validação Cruzada – Base de Dados *Echocardiogram*.

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Glass

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE

Mínimo	Máximo
3	9

TOTAL DE TESTES REALIZADOS	42

TESTES REALIZADOS

Conjuntos de Teste		3			N° de Amostras / Conjunto	69
Teste		1	2	3		
Acertos	Abs.	26	32	24		
	Per.	37.68	46.38	34.78		

Conjuntos de Teste		4				N° de Amostras / Conjunto	52
Teste		1	2	3	4		
Acertos	Abs.	22	27	27	25		
	Per.	42.31	51.92	51.92	48.08		

Conjuntos de Teste		5				N° de Amostras / Conjunto	40
Teste		1	2	3	4	5	
Acertos	Abs.	20	19	17	15	16	
	Per.	50.00	47.50	42.50	37.50	40.00	

Conjuntos de Teste		6				N° de Amostras / Conjunto	32
Teste		1	2	3	4	5	6
Acertos	Abs.	16	12	25	19	13	13
	Per.	50.00	37.50	78.12	59.38	40.62	40.62

Conjuntos de Teste		7				N° de Amostras / Conjunto	28	
Teste		1	2	3	4	5	6	7
Acertos	Abs.	14	12	18	14	15	13	9
	Per.	50.00	42.86	64.29	50.00	53.57	46.43	32.14

Conjuntos de Teste		8				N° de Amostras / Conjunto	24		
Teste		1	2	3	4	5	6	7	8
Acertos	Abs.	11	9	8	19	10	12	11	10
	Per.	45.83	37.50	33.33	79.17	41.67	50.00	45.83	41.67

Conjuntos de Teste		9				N° de Amostras / Conjunto	21			
Teste		1	2	3	4	5	6	7	8	9
Acertos	Abs.	9	8	7	14	8	11	10	11	10
	Per.	42.86	38.10	33.33	66.67	38.10	52.38	47.62	52.38	47.62

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Glass

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	9

TOTAL DE TESTES REALIZADOS	42

RESUMO - ACERTOS									
Conjs. de Teste		3	4	5	6	7	8	9	**Total**
Amostras / Conjunto		69	52	40	32	28	24	21	
Mínimo	**Absoluto**	24	22	15	12	9	8	7	
	Percentual	34.78	42.31	37.50	37.50	32.14	33.33	33.33	**32.14**
Máximo	**Absoluto**	32	27	20	25	18	19	14	
	Percentual	46.38	51.92	50.00	78.12	64.29	79.17	66.67	**79.17**
Médio	**Absoluto**	27.33	25.25	17.4	16.33	13.57	11.25	9.78	
	Percentual	39.61	48.56	43.5	51.04	48.47	46.88	46.56	**46.91**

NÚMERO DE AMOSTRAS POR CLASSE								
Classe	1	2	3	4	5	6	7	**Total**
Abs.	70	76	17	0	13	9	29	214
Per.	32.71	35.51	7.94	0.00	6.07	4.21	13.55	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	**32.14**
CONJUNTOS DE TESTE	7
AMOSTRAS / CONJUNTO	28
TESTE	7
ACERTOS - ABS.	9

MATRIZ DE CONFUSÃO - ABS.

		PREDITO						
		1	**2**	**3**	**4**	**5**	**6**	**7**
ESPERADO	**1**	4	5	1	0	0	0	0
	2	8	1	0	0	1	0	0
	3	2	0	0	0	0	0	0
	4	0	0	0	0	0	0	0
	5	0	1	0	0	0	0	0
	6	0	0	0	0	0	0	1
	7	0	0	0	0	0	0	4

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Glass

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	9

TOTAL DE TESTES REALIZADOS	42

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	**79.17**
CONJUNTOS DE TESTE	8
AMOSTRAS / CONJUNTO	24
TESTE	4
ACERTOS - ABS.	19

MATRIZ DE CONFUSÃO - ABS.

		PREDITO						
		1	**2**	**3**	**4**	**5**	**6**	**7**
ESPERADO	**1**	8	0	0	0	0	0	0
	2	0	8	0	0	1	0	0
	3	2	0	0	0	0	0	0
	4	0	0	0	0	0	0	0
	5	0	1	0	0	0	0	0
	6	0	0	0	0	0	0	1
	7	0	0	0	0	0	0	3

ACERTO MÉDIO PERCENTUAL	**46.91**

MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL

		PREDITO						
		1	**2**	**3**	**4**	**5**	**6**	**7**
ESPERADO	**1**	22.41	10.19	1.42	0.00	0.00	0.00	0.00
	2	18.02	13.45	0.08	0.00	5.22	0.00	0.37
	3	5.62	0.93	0.29	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.00	2.48	0.00	0.00	1.67	0.00	0.71
	6	0.00	1.10	0.00	0.00	0.37	0.00	2.35
	7	0.51	1.53	0.00	0.00	2.18	0.00	9.09

Tabela III.7. Validação Cruzada – Base de Dados *Glass*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Images	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° de Amostras / Conjunto**	70
Teste		1	2	3		
Acertos	**Abs.**	16	20	20		
	Per.	22.86	28.57	28.57		

Conjuntos de Teste		4				**N° de Amostras / Conjunto**	49
Teste		1	2	3	4		
Acertos	**Abs.**	12	17	16	14		
	Per.	24.49	34.69	32.65	28.57		

Conjuntos de Teste		5					**N° de Amostras / Conjunto**	42
Teste		1	2	3	4	5		
Acertos	**Abs.**	10	15	12	15	12		
	Per.	23.81	35.71	28.57	35.71	28.57		

Conjuntos de Teste		6						**N° de Amostras / Conjunto**	35
Teste		1	2	3	4	5	6		
Acertos	**Abs.**	8	14	12	10	15	10		
	Per.	22.86	40.00	34.29	28.57	42.86	28.57		

Conjuntos de Teste		7							**N° de Amostras / Conjunto**	28
Teste		1	2	3	4	5	6	7		
Acertos	**Abs.**	6	11	10	9	9	12	8		
	Per.	21.43	39.29	35.71	32.14	32.14	42.86	28.57		

Conjuntos de Teste		8								**N° de Amostras / Conjunto**	21
Teste		1	2	3	4	5	6	7	8		
Acertos	**Abs.**	5	8	8	8	7	7	7	9		
	Per.	23.81	38.10	38.10	38.10	33.33	33.33	33.33	42.86		

Conjuntos de Teste		9									**N° de Amostras / Conjunto**	21
Teste		1	2	3	4	5	6	7	8	9		
Acertos	**Abs.**	5	8	8	8	7	7	7	9	6		
	Per.	23.81	38.10	38.10	38.10	33.33	33.33	33.33	42.86	28.57		

Conjuntos de Teste		10									**N° de Amostras / Conjunto**	21	
Teste		1	2	3	4	5	6	7	8	9	10		
Acertos	**Abs.**	5	8	8	8	7	7	7	9	6	8		
	Per.	23.81	38.10	38.10	38.10	33.33	33.33	33.33	42.86	28.57	38.10		

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Images

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. de Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		70	49	42	35	28	21	21	21	
Mínimo	**Absoluto**	16	12	10	8	6	5	5	5	
	Percentual	22.86	24.49	23.81	22.86	21.43	23.81	23.81	23.81	**21.43**
Máximo	**Absoluto**	20	17	15	15	12	9	9	9	
	Percentual	28.57	34.69	35.71	42.86	42.86	42.86	42.86	42.86	**42.86**
Médio	**Absoluto**	18.67	14.75	12.8	11.5	9.29	7.38	7.22	7.3	
	Percentual	26.67	30.10	30.48	32.86	33.16	35.12	34.39	34.76	**33.08**

NÚMERO DE AMOSTRAS POR CLASSE								
Classe	1	2	3	4	5	6	7	**Total**
Abs.	30	30	30	30	30	30	30	210
Per.	14.29	14.29	14.29	14.29	14.29	14.29	14.29	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	**21.43**
CONJUNTOS DE TESTE	7
AMOSTRAS / CONJUNTO	28
TESTE	1
ACERTOS - ABS.	6

MATRIZ DE CONFUSÃO - ABS.								
		PREDITO						
		1	**2**	**3**	**4**	**5**	**6**	**7**
ESPERADO	**1**	0	0	0	4	0	0	0
	2	0	0	0	4	0	0	0
	3	0	0	0	4	0	0	0
	4	0	0	0	4	0	0	0
	5	0	0	0	4	0	0	0
	6	0	0	0	2	0	2	0
	7	3	0	0	1	0	0	0

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Images

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	42.86
CONJUNTOS DE TESTE	6
AMOSTRAS / CONJUNTO	35
TESTE	5
ACERTOS - ABS.	15

MATRIZ DE CONFUSÃO - ABS.

		PREDITO						
		1	**2**	**3**	**4**	**5**	**6**	**7**
ESPERADO	**1**	0	0	3	2	0	0	0
	2	0	0	0	5	0	0	0
	3	0	0	5	0	0	0	0
	4	0	0	0	5	0	0	0
	5	0	0	2	3	0	0	0
	6	0	0	0	0	0	5	0
	7	0	0	5	0	0	0	0

ACERTO MÉDIO PERCENTUAL	33.08

MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL

		PREDITO						
		1	**2**	**3**	**4**	**5**	**6**	**7**
ESPERADO	**1**	0.00	0.00	8.17	6.12	0.00	0.00	0.00
	2	0.00	0.00	2.14	12.15	0.00	0.00	0.00
	3	0.07	0.00	11.75	2.47	0.00	0.00	0.00
	4	0.00	0.00	3.69	10.10	0.00	0.50	0.00
	5	0.00	0.00	10.46	3.83	0.00	0.00	0.00
	6	0.00	0.00	0.51	2.54	0.00	11.23	0.00
	7	1.28	0.00	11.81	1.19	0.00	0.00	0.00

Tabela III.8. Validação Cruzada – Base de Dados *Images*.

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): InfraSystem

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE

Mínimo	Máximo
3	10

TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3		N° de Amostras / Conjunto	74
Teste		1	2	3	
Acertos	Abs.	34	35	31	
	Per.	45.95	47.30	41.89	

Conjuntos de Teste		4		N° de Amostras / Conjunto	56
Teste		1	2	3	4
Acertos	Abs.	28	34	28	28
	Per.	50.00	60.71	50.00	50.00

Conjuntos de Teste		5		N° de Amostras / Conjunto	43
Teste		1	2	3	4	5
Acertos	Abs.	26	26	24	23	19
	Per.	60.47	60.47	55.81	53.49	44.19

Conjuntos de Teste		6		N° de Amostras / Conjunto	36
Teste		1	2	3	4	5	6
Acertos	Abs.	25	20	19	19	11	29
	Per.	69.44	55.56	52.78	52.78	30.56	80.56

Conjuntos de Teste		7		N° de Amostras / Conjunto	30
Teste		1	2	3	4	5	6	7
Acertos	Abs.	22	16	21	19	17	9	24
	Per.	73.33	53.33	70.00	63.33	56.67	30.00	80.00

Conjuntos de Teste		8		N° de Amostras / Conjunto	27
Teste		1	2	3	4	5	6	7	8
Acertos	Abs.	20	15	18	13	17	10	17	18
	Per.	74.07	55.56	66.67	48.15	62.96	37.04	62.96	66.67

Conjuntos de Teste		9		N° de Amostras / Conjunto	23
Teste		1	2	3	4	5	6	7	8	9
Acertos	Abs.	16	16	17	15	11	14	9	10	16
	Per.	69.57	69.57	73.91	65.22	47.83	60.87	39.13	43.48	69.57

Conjuntos de Teste		10		N° de Amostras / Conjunto	21	
Teste		1	2	3	4	5	6	7	8	9	10
Acertos	Abs.	15	15	13	14	14	12	10	13	15	13
	Per.	71.43	71.43	61.90	66.67	66.67	57.14	47.62	61.90	71.43	61.90

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): InfraSystem

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE

Mínimo	Máximo
3	10

TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS

Conjs. de Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		74	56	43	36	30	27	23	21	
Mínimo	**Absoluto**	31	28	19	11	9	10	9	10	
	Percentual	41.89	50.00	44.19	30.56	30.00	37.04	39.13	47.62	**30.00**
Máximo	**Absoluto**	35	34	26	29	24	20	17	15	
	Percentual	47.30	60.71	60.47	80.56	80.00	74.07	73.91	71.43	**80.56**
Médio	**Absoluto**	33.33	29.5	23.6	20.5	18.29	16	13.78	13.4	
	Percentual	45.05	52.68	54.88	56.94	60.95	59.26	59.90	63.81	**58.46**

NÚMERO DE AMOSTRAS POR CLASSE

Classe	1	2	3	4	5	**Total**
Abs.	28	80	54	40	24	226
Per.	12.39	35.40	23.89	17.70	10.62	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	**30.00**
CONJUNTOS DE TESTE	7
AMOSTRAS / CONJUNTO	30
TESTE	6
ACERTOS - ABS.	9

MATRIZ DE CONFUSÃO - ABS.

		PREDITO				
		1	**2**	**3**	**4**	**5**
ESPERADO	**1**	4	0	0	0	0
	2	1	1	6	3	0
	3	1	0	3	3	0
	4	0	0	4	0	1
	5	0	0	0	2	1

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): InfraSystem

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	80.56
CONJUNTOS DE TESTE	6
AMOSTRAS / CONJUNTO	36
TESTE	6
ACERTOS - ABS.	29

MATRIZ DE CONFUSÃO - ABS.						
		PREDITO				
		1	**2**	**3**	**4**	**5**
ESPERADO	**1**	4	0	0	0	0
	2	1	12	0	0	0
	3	0	2	5	2	0
	4	0	2	0	4	0
	5	0	0	0	0	4

ACERTO MÉDIO PERCENTUAL	58.46

MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL						
		PREDITO				
		1	**2**	**3**	**4**	**5**
ESPERADO	**1**	10.60	1.05	0.00	0.00	0.00
	2	0.79	26.36	8.20	1.14	0.00
	3	0.94	6.40	9.91	6.71	0.00
	4	0.65	3.13	5.94	5.35	2.80
	5	0.16	0.00	0.00	3.63	6.23

Tabela III.9. Validação Cruzada – Base de Dados *InfraSystem*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Iris	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° de Amostras / Conjunto**	48
Teste		1	2	3		
Acertos	**Abs.**	46	46	42		
	Per.	95.83	95.83	87.50		

Conjuntos de Teste		4				**N° de Amostras / Conjunto**	36
Teste		1	2	3	4		
Acertos	**Abs.**	34	34	34	33		
	Per.	94.44	94.44	94.44	91.67		

Conjuntos de Teste		5					**N° de Amostras / Conjunto**	30
Teste		1	2	3	4	5		
Acertos	**Abs.**	29	27	30	26	29		
	Per.	96.67	90.00	100.00	86.67	96.67		

Conjuntos de Teste		6						**N° de Amostras / Conjunto**	24
Teste		1	2	3	4	5	6		
Acertos	**Abs.**	23	23	22	24	20	23		
	Per.	95.83	95.83	91.67	100.00	83.33	95.83		

Conjuntos de Teste		7							**N° de Amostras / Conjunto**	21
Teste		1	2	3	4	5	6	7		
Acertos	**Abs.**	20	20	19	21	19	19	20		
	Per.	95.24	95.24	90.48	100.00	90.48	90.48	95.24		

Conjuntos de Teste		8								**N° de Amostras / Conjunto**	18
Teste		1	2	3	4	5	6	7	8		
Acertos	**Abs.**	17	17	17	17	18	16	16	17		
	Per.	94.44	94.44	94.44	94.44	100.00	88.89	88.89	94.44		

Conjuntos de Teste		9									**N° de Amostras / Conjunto**	15
Teste		1	2	3	4	5	6	7	8	9		
Acertos	**Abs.**	14	15	14	13	15	15	13	13	14		
	Per.	93.33	100.00	93.33	86.67	100.00	100.00	86.67	86.67	93.33		

Conjuntos de Teste		10										**N° de Amostras / Conjunto**	15
Teste		1	2	3	4	5	6	7	8	9	10		
Acertos	**Abs.**	14	15	14	13	15	15	13	13	14	15		
	Per.	93.33	100.00	93.33	86.67	100.00	100.00	86.67	86.67	93.33	100.00		

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Iris	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. de Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		48	36	30	24	21	18	15	15	
Mínimo	**Absoluto**	42	33	26	20	19	16	13	13	
	Percentual	87.50	91.67	86.67	83.33	90.48	88.89	86.67	86.67	**83.33**
Máximo	**Absoluto**	46	34	30	24	21	18	15	15	
	Percentual	95.83	94.44	100.00	100.00	100.00	100.00	100.00	100.00	**100.00**
Médio	**Absoluto**	44.67	33.75	28.2	22.5	19.71	16.88	14	14.1	
	Percentual	93.06	93.75	94.00	93.75	93.88	93.75	93.33	94.00	**93.73**

NÚMERO DE AMOSTRAS POR CLASSE				
Classe	1	2	3	**Total**
Abs.	50	50	50	150
Per.	33.33	33.33	33.33	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	**83.33**
CONJUNTOS DE TESTE	6
AMOSTRAS / CONJUNTO	24
TESTE	5
ACERTOS - ABS.	20

MATRIZ DE CONFUSÃO - ABS.

		PREDITO		
		1	**2**	**3**
ESPERADO	**1**	8	0	0
	2	0	6	2
	3	0	2	6

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	**100.00**
CONJUNTOS DE TESTE	5
AMOSTRAS / CONJUNTO	30
TESTE	3
ACERTOS - ABS.	30

MATRIZ DE CONFUSÃO - ABS.

		PREDITO		
		1	**2**	**3**
ESPERADO	**1**	10	0	0
	2	0	10	0
	3	0	0	10

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Iris

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL					93.73
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL					
		PREDITO			
		1	2	3	
ESPERADO	1	33.33	0.00	0.00	
	2	0.00	30.52	2.81	
	3	0.00	3.46	29.87	

Tabela III.10. Validação Cruzada – Base de Dados *Iris*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Mushroom	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3		**N° de Amostras / Conjunto**	1880
Teste		1	2	3	
Acertos	**Abs.**	1624	1880	1806	
	Per.	86.38	100.00	96.06	

Conjuntos de Teste		4			**N° de Amostras / Conjunto**	1411
Teste		1	2	3	4	
Acertos	**Abs.**	1155	1410	1411	1313	
	Per.	81.86	99.93	100.00	93.05	

Conjuntos de Teste		5				**N° de Amostras / Conjunto**	1128
Teste		1	2	3	4	5	
Acertos	**Abs.**	872	1128	1128	1128	1092	
	Per.	77.30	100.00	100.00	100.00	96.81	

Conjuntos de Teste		6					**N° de Amostras / Conjunto**	940
Teste		1	2	3	4	5	6	
Acertos	**Abs.**	684	940	940	940	940	904	
	Per.	72.77	100.00	100.00	100.00	100.00	96.17	

Conjuntos de Teste		7						**N° de Amostras / Conjunto**	806
Teste		1	2	3	4	5	6	7	
Acertos	**Abs.**	550	806	806	806	806	806	754	
	Per.	68.24	100.00	100.00	100.00	100.00	100.00	93.55	

Conjuntos de Teste		8							**N° de Amostras / Conjunto**	705
Teste		1	2	3	4	5	6	7	8	
Acertos	**Abs.**	449	705	705	705	705	705	705	665	
	Per.	63.69	100.00	100.00	100.00	100.00	100.00	100.00	94.33	

Conjuntos de Teste		9								**N° de Amostras / Conjunto**	626
Teste		1	2	3	4	5	6	7	8	9	
Acertos	**Abs.**	626	626	626	626	626	626	626	626	602	
	Per.	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	96.17	

Conjuntos de Teste		10									**N° de Amostras / Conjunto**	563	
Teste		1	2	3	4	5	6	7	8	9	10		
Acertos	**Abs.**	563	563	563	563	563	563	563	563	563	547		
	Per.	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	97.16		

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Mushroom	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS

Conjs. Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		1880	1411	1128	940	806	705	626	563	
Mínimo	**Absoluto**	1624	1155	872	684	550	449	602	547	
	Percentual	86.38	81.86	77.30	72.77	68.24	63.69	96.17	97.16	**63.69**
Máximo	**Absoluto**	1880	1411	1128	940	806	705	626	563	
	Percentual	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	**100.00**
Médio	**Absoluto**	1770	1322.3	1069.6	891.33	762	668	623.33	561.4	
	Percentual	94.15	93.71	94.82	94.82	94.54	94.75	99.57	99.72	**96.41**

NÚMERO DE AMOSTRAS POR CLASSE

Classe	1	2	**Total**
Abs.	3488	2156	5644
Per.	61.80	38.20	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	**63.69**
CONJUNTOS DE TESTE	8
AMOSTRAS / CONJUNTO	705
TESTE	1
ACERTOS - ABS.	449

MATRIZ DE CONFUSÃO - ABS.

		PREDITO	
		1	**2**
ESPERADO	**1**	436	0
	2	256	13

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	**100.00**
CONJUNTOS DE TESTE	3
AMOSTRAS / CONJUNTO	1880
TESTE	2
ACERTOS - ABS.	1880

MATRIZ DE CONFUSÃO - ABS.

		PREDITO	
		1	**2**
ESPERADO	**1**	1162	0
	2	0	718

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Mushroom

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL				**96.41**
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL				
		PREDITO		
		1	**2**	
ESPERADO	**1**	61.52	0.29	
	2	3.29	34.89	

Tabela III.11. Validação Cruzada – Base de Dados *Mushroom*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Odd Parity (3-bit Parity)	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° Amostras / Conjunto**	26
Teste		1	2	3		
Acertos	**Abs.**	11	16	15		
	Per.	42.31	61.54	57.69		

Conjuntos de Teste		4				**N° Amostras / Conjunto**	20
Teste		1	2	3	4		
Acertos	**Abs.**	7	7	1	6		
	Per.	35.00	35.00	5.00	30.00		

Conjuntos de Teste		5				**N° Amostras / Conjunto**	16
Teste		1	2	3	4	5	
Acertos	**Abs.**	4	6	4	5	5	
	Per.	25.00	37.50	25.00	31.25	31.25	

Conjuntos de Teste		6				**N° Amostras / Conjunto**	12
Teste		1	2	3	4	5	6
Acertos	**Abs.**	5	5	4	4	2	2
	Per.	41.67	41.67	33.33	33.33	16.67	16.67

Conjuntos de Teste		7				**N° Amostras / Conjunto**	10	
Teste		1	2	3	4	5	6	7
Acertos	**Abs.**	6	2	4	2	5	2	2
	Per.	60.00	20.00	40.00	20.00	50.00	20.00	20.00

Conjuntos de Teste		8				**N° Amostras / Conjunto**	10		
Teste		1	2	3	4	5	6	7	8
Acertos	**Abs.**	6	2	4	2	5	2	2	6
	Per.	60.00	20.00	40.00	20.00	50.00	20.00	20.00	60.00

Conjuntos de Teste		9				**N° Amostras / Conjunto**	8			
Teste		1	2	3	4	5	6	7	8	9
Acertos	**Abs.**	3	1	2	3	2	4	2	3	1
	Per.	37.50	12.50	25.00	37.50	25.00	50.00	25.00	37.50	12.50

Conjuntos de Teste		10				**N° Amostras / Conjunto**	8				
Teste		1	2	3	4	5	6	7	8	9	10
Acertos	**Abs.**	3	1	2	3	2	4	2	3	1	4
	Per.	37.50	12.50	25.00	37.50	25.00	50.00	25.00	37.50	12.50	50.00

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Odd Parity (3-bit Parity)	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		26	20	16	12	10	10	8	8	
Mínimo	**Absoluto**	11	1	4	2	2	2	1	1	
	Percentual	42.31	5.00	25.00	16.67	20.00	20.00	12.50	12.50	**5.00**
Máximo	**Absoluto**	16	7	6	5	6	6	4	4	
	Percentual	61.54	35.00	37.50	41.67	60.00	60.00	50.00	50.00	**61.54**
Médio	**Absoluto**	14	5.25	4.8	3.67	3.29	3.62	2.33	2.5	
	Percentual	53.85	26.25	30.00	30.56	32.86	36.25	29.17	31.25	**32.59**

NÚMERO DE AMOSTRAS POR CLASSE			
Classe	1	2	**Total**
Abs.	40	40	80
Per.	50.00	50.00	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL		**5.00**
CONJUNTOS DE TESTE	4	
AMOSTRAS / CONJUNTO	20	
TESTE	3	
ACERTOS - ABS.	1	

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	1	9
	2	10	0

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL		**61.54**
CONJUNTOS DE TESTE	3	
AMOSTRAS / CONJUNTO	26	
TESTE	2	
ACERTOS - ABS.	16	

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	8	5
	2	5	8

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Odd Parity (3-bit Parity)

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL				**32.59**
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL				
		PREDITO		
		1	**2**	
ESPERADO	**1**	12.81	37.19	
	2	30.22	19.78	

Tabela III.12. Validação Cruzada – Base de Dados *Odd Parity (3-bit Parity)*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Odd Parity (4-bit Parity)	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° Amostras / Conjunto**	52
Teste		1	2	3		
Acertos	**Abs.**	24	17	19		
	Per.	46.15	32.69	36.54		

Conjuntos de Teste		4				**N° Amostras / Conjunto**	40
Teste		1	2	3	4		
Acertos	**Abs.**	13	12	18	14		
	Per.	32.50	30.00	45.00	35.00		

Conjuntos de Teste		5				**N° Amostras / Conjunto**	32
Teste		1	2	3	4	5	
Acertos	**Abs.**	12	12	14	15	11	
	Per.	37.50	37.50	43.75	46.88	34.38	

Conjuntos de Teste		6					**N° Amostras / Conjunto**	26
Teste		1	2	3	4	5	6	
Acertos	**Abs.**	8	9	10	10	7	7	
	Per.	30.77	34.62	38.46	38.46	26.92	26.92	

Conjuntos de Teste		7						**N° Amostras / Conjunto**	22
Teste		1	2	3	4	5	6	7	
Acertos	**Abs.**	5	8	7	7	8	6	6	
	Per.	22.73	36.36	31.82	31.82	36.36	27.27	27.27	

Conjuntos de Teste		8							**N° Amostras / Conjunto**	20
Teste		1	2	3	4	5	6	7	8	
Acertos	**Abs.**	5	10	9	9	8	8	5	6	
	Per.	25.00	50.00	45.00	45.00	40.00	40.00	25.00	30.00	

Conjuntos de Teste		9								**N° Amostras / Conjunto**	16
Teste		1	2	3	4	5	6	7	8	9	
Acertos	**Abs.**	5	4	6	2	8	7	7	7	2	
	Per.	31.25	25.00	37.50	12.50	50.00	43.75	43.75	43.75	12.50	

Conjuntos de Teste		10									**N° Amostras / Conjunto**	16	
Teste		1	2	3	4	5	6	7	8	9	10		
Acertos	**Abs.**	5	4	6	2	8	7	7	7	2	6		
	Per.	31.25	25.00	37.50	12.50	50.00	43.75	43.75	43.75	12.50	37.50		

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Odd Parity (4-bit Parity)	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		52	40	32	26	22	20	16	16	
Mínimo	**Absoluto**	17	12	11	7	5	5	2	2	
	Percentual	32.69	30.00	34.38	26.92	22.73	25.00	12.50	12.50	**12.50**
Máximo	**Absoluto**	24	18	15	10	8	10	8	8	
	Percentual	46.15	45.00	46.88	38.46	36.36	50.00	50.00	50.00	**50.00**
Médio	**Absoluto**	20	14.25	12.8	8.5	6.71	7.5	5.33	5.4	
	Percentual	38.46	35.62	40.00	32.69	30.52	37.50	33.33	33.75	**34.71**

NÚMERO DE AMOSTRAS POR CLASSE			
Classe	1	2	**Total**
Abs.	80	80	160
Per.	50.00	50.00	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL		**12.50**
CONJUNTOS DE TESTE	9	
AMOSTRAS / CONJUNTO	16	
TESTE	4	
ACERTOS - ABS.	2	

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	1	7
	2	7	1

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL		**50.00**
CONJUNTOS DE TESTE	8	
AMOSTRAS / CONJUNTO	20	
TESTE	2	
ACERTOS - ABS.	10	

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	2	8
	2	2	8

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Odd Parity (4-bit Parity)

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL				**34.71**
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL				
		PREDITO		
		1	**2**	
ESPERADO	**1**	17.63	32.37	
	2	32.91	17.09	

Tabela III.13. Validação Cruzada – Base de Dados *Odd Parity (4-bit Parity)*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Odd Parity (5-bit Parity)	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° Amostras / Conjunto**	106
Teste		1	2	3		
Acertos	**Abs.**	40	44	44		
	Per.	37.74	41.51	41.51		

Conjuntos de Teste		4				**N° Amostras / Conjunto**	80
Teste		1	2	3	4		
Acertos	**Abs.**	29	33	16	21		
	Per.	36.25	41.25	20.00	26.25		

Conjuntos de Teste		5					**N° Amostras / Conjunto**	64
Teste		1	2	3	4	5		
Acertos	**Abs.**	23	30	30	26	21		
	Per.	35.94	46.88	46.88	40.62	32.81		

Conjuntos de Teste		6						**N° Amostras / Conjunto**	52
Teste		1	2	3	4	5	6		
Acertos	**Abs.**	23	16	20	23	24	19		
	Per.	44.23	30.77	38.46	44.23	46.15	36.54		

Conjuntos de Teste		7							**N° Amostras / Conjunto**	44
Teste		1	2	3	4	5	6	7		
Acertos	**Abs.**	11	12	22	16	17	18	20		
	Per.	25.00	27.27	50.00	36.36	38.64	40.91	45.45		

Conjuntos de Teste		8								**N° Amostras / Conjunto**	40
Teste		1	2	3	4	5	6	7	8		
Acertos	**Abs.**	14	13	17	15	11	15	13	22		
	Per.	35.00	32.50	42.50	37.50	27.50	37.50	32.50	55.00		

Conjuntos de Teste		9									**N° Amostras / Conjunto**	34
Teste		1	2	3	4	5	6	7	8	9		
Acertos	**Abs.**	11	9	9	15	13	8	13	12	10		
	Per.	32.35	26.47	26.47	44.12	38.24	23.53	38.24	35.29	29.41		

Conjuntos de Teste		10										**N° Amostras / Conjunto**	32	
Teste		1	2	3	4	5	6	7	8	9	10			
Acertos	**Abs.**	13	8	7	12	10	10	12	14	6	14			
	Per.	40.62	25.00	21.88	37.50	31.25	31.25	37.50	43.75	18.75	43.75			

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Odd Parity (5-bit Parity)	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		106	80	64	52	44	40	34	32	
Mínimo	**Absoluto**	40	16	21	16	11	11	8	6	
	Percentual	37.74	20.00	32.81	30.77	25.00	27.50	23.53	18.75	**18.75**
Máximo	**Absoluto**	44	33	30	24	22	22	15	14	
	Percentual	41.51	41.25	46.88	46.15	50.00	55.00	44.12	43.75	**55.00**
Médio	**Absoluto**	42.67	24.75	26	20.83	16.57	15	11.11	10.6	
	Percentual	40.25	30.94	40.62	40.06	37.66	37.50	32.68	33.12	**36.10**

NÚMERO DE AMOSTRAS POR CLASSE			
Classe	1	2	**Total**
Abs.	160	160	320
Per.	50.00	50.00	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL		**18.75**
CONJUNTOS DE TESTE	10	
AMOSTRAS / CONJUNTO	32	
TESTE	9	
ACERTOS - ABS.	6	

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	2	14
	2	12	4

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL		**55.00**
CONJUNTOS DE TESTE	8	
AMOSTRAS / CONJUNTO	40	
TESTE	8	
ACERTOS - ABS.	22	

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	14	6
	2	12	8

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Odd Parity (5-bit Parity)

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL				**36.10**
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL				
		PREDITO		
		1	**2**	
ESPERADO	**1**	16.29	33.71	
	2	30.19	19.81	

Tabela III.14. Validação Cruzada – Base de Dados *Odd Parity (5-bit Parity)*.

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Sleepdata1

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	7

TOTAL DE TESTES REALIZADOS	25

TESTES REALIZADOS

Conjuntos de Teste		3			**N° Amostras / Conjunto**	154
Teste		1	2	3		
Acertos	**Abs.**	94	87	128		
	Per.	61.04	56.49	83.12		

Conjuntos de Teste		4				**N° Amostras / Conjunto**	115
Teste		1	2	3	4		
Acertos	**Abs.**	66	102	73	92		
	Per.	57.39	88.70	63.48	80.00		

Conjuntos de Teste		5				**N° Amostras / Conjunto**	92
Teste		1	2	3	4	5	
Acertos	**Abs.**	56	74	75	59	67	
	Per.	60.87	80.43	81.52	64.13	72.83	

Conjuntos de Teste		6				**N° Amostras / Conjunto**	75
Teste		1	2	3	4	5	6
Acertos	**Abs.**	44	60	64	44	67	62
	Per.	58.67	80.00	85.33	58.67	89.33	82.67

Conjuntos de Teste		7				**N° Amostras / Conjunto**	65
Teste		1	2	3	4	5	6	7
Acertos	**Abs.**	35	50	61	50	42	55	51
	Per.	53.85	76.92	93.85	76.92	64.62	84.62	78.46

RESUMO - ACERTOS							
Conjs. de Teste		3	4	5	6	7	**Total**
Amostras / Conjunto		154	115	92	75	65	
Mínimo	**Absoluto**	87	66	56	44	35	
	Percentual	56.49	57.39	60.87	58.67	53.85	**53.85**
Máximo	**Absoluto**	128	102	75	67	61	
	Percentual	83.12	88.70	81.52	89.33	93.85	**93.85**
Médio	**Absoluto**	103	83.25	66.2	56.83	49.14	
	Percentual	66.88	72.39	71.96	75.78	75.60	**73.36**

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Sleepdata1

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	7

TOTAL DE TESTES REALIZADOS	25

NÚMERO DE AMOSTRAS POR CLASSE							
Classe	1	2	3	4	5	6	**Total**
Abs.	138	65	16	221	21	7	468
Per.	29.49	13.89	3.42	47.22	4.49	1.50	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	53.85
CONJUNTOS DE TESTE	7
AMOSTRAS / CONJUNTO	65
TESTE	1
ACERTOS - ABS.	35

MATRIZ DE CONFUSÃO - ABS.

		PREDITO					
		1	**2**	**3**	**4**	**5**	**6**
ESPERADO	**1**	18	0	0	1	0	0
	2	0	7	2	0	0	0
	3	0	1	1	0	0	0
	4	0	4	12	6	9	0
	5	0	0	0	0	3	0
	6	0	0	0	0	1	0

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	93.85
CONJUNTOS DE TESTE	7
AMOSTRAS / CONJUNTO	65
TESTE	3
ACERTOS - ABS.	61

MATRIZ DE CONFUSÃO - ABS.

		PREDITO					
		1	**2**	**3**	**4**	**5**	**6**
ESPERADO	**1**	19	0	0	0	0	0
	2	0	9	0	0	0	0
	3	0	1	0	0	1	0
	4	0	1	0	30	0	0
	5	0	0	0	0	3	0
	6	0	0	0	0	1	0

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Sleepdata1

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	7

TOTAL DE TESTES REALIZADOS	25

ACERTO MÉDIO PERCENTUAL							**73.36**
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL							
		PREDITO					
		1	**2**	**3**	**4**	**5**	**6**
ESPERADO	**1**	21.52	0.00	0.00	5.97	2.23	0.00
	2	0.00	8.75	2.64	1.77	0.61	0.00
	3	0.00	1.05	1.19	0.53	0.32	0.00
	4	0.89	1.55	2.71	39.68	2.95	0.00
	5	0.27	0.18	0.48	1.22	2.21	0.00
	6	0.25	0.00	0.00	0.00	1.02	0.00

Tabela III.15. Validação Cruzada – Base de Dados *Sleepdata1*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Sleepdata2	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	8
TOTAL DE TESTES REALIZADOS	33

TESTES REALIZADOS

Conjuntos de Teste		3			**N° de Amostras / Conjunto**	129
Teste		1	2	3		
Acertos	**Abs.**	19	86	83		
	Per.	14.73	66.67	64.34		

Conjuntos de Teste		4				**N° de Amostras / Conjunto**	97
Teste		1	2	3	4		
Acertos	**Abs.**	40	52	64	61		
	Per.	41.24	53.61	65.98	62.89		

Conjuntos de Teste		5					**N° de Amostras / Conjunto**	77
Teste		1	2	3	4	5		
Acertos	**Abs.**	25	43	50	52	49		
	Per.	32.47	55.84	64.94	67.53	63.64		

Conjuntos de Teste		6						**N° de Amostras / Conjunto**	63
Teste		1	2	3	4	5	6		
Acertos	**Abs.**	38	38	38	41	40	42		
	Per.	60.32	60.32	60.32	65.08	63.49	66.67		

Conjuntos de Teste		7							**N° de Amostras / Conjunto**	54
Teste		1	2	3	4	5	6	7		
Acertos	**Abs.**	35	33	32	34	38	33	36		
	Per.	64.81	61.11	59.26	62.96	70.37	61.11	66.67		

Conjuntos de Teste		8								**N° de Amostras / Conjunto**	48
Teste		1	2	3	4	5	6	7	8		
Acertos	**Abs.**	31	32	23	34	31	33	30	31		
	Per.	64.58	66.67	47.92	70.83	64.58	68.75	62.50	64.58		

RESUMO - ACERTOS

Conjs. de Teste		3	4	5	6	7	8	**Total**
Amostras / Conjunto		129	97	77	63	54	48	
Mínimo	**Absoluto**	19	40	25	38	32	23	
	Percentual	14.73	41.24	32.47	60.32	59.26	47.92	**14.73**
Máximo	**Absoluto**	86	64	52	42	38	34	
	Percentual	66.67	65.98	67.53	66.67	70.37	70.83	**70.83**
Médio	**Absoluto**	62.67	54.25	43.8	39.5	34.43	30.62	
	Percentual	48.58	55.93	56.88	62.70	63.76	63.80	**60.21**

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Sleepdata2

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	8

TOTAL DE TESTES REALIZADOS	33

NÚMERO DE AMOSTRAS POR CLASSE								
Classe	1	2	3	4	5	6	**Total**	
Abs.	58	61	11	209	8	50	397	
Per.	14.61	15.37	2.77	52.64	2.02	12.59	100.00	

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL		**14.73**
CONJUNTOS DE TESTE	3	
AMOSTRAS / CONJUNTO	129	
TESTE	1	
ACERTOS - ABS.	19	

MATRIZ DE CONFUSÃO - ABS.

		PREDITO					
		1	**2**	**3**	**4**	**5**	**6**
ESPERADO	**1**	19	0	0	0	0	0
	2	20	0	0	0	0	0
	3	3	0	0	0	0	0
	4	69	0	0	0	0	0
	5	2	0	0	0	0	0
	6	16	0	0	0	0	0

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Sleepdata2

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	8

TOTAL DE TESTES REALIZADOS	33

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	**70.83**
CONJUNTOS DE TESTE	8
AMOSTRAS / CONJUNTO	48
TESTE	4
ACERTOS - ABS.	34

MATRIZ DE CONFUSÃO - ABS.

		PREDITO					
		1	**2**	**3**	**4**	**5**	**6**
ESPERADO	**1**	7	0	0	0	0	0
	2	0	0	7	0	0	0
	3	0	0	1	0	0	0
	4	0	0	0	26	0	0
	5	0	0	0	1	0	0
	6	1	0	0	5	0	0

ACERTO MÉDIO PERCENTUAL	**60.21**

MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL

		PREDITO					
		1	**2**	**3**	**4**	**5**	**6**
ESPERADO	**1**	12.45	0.00	0.26	1.82	0.00	0.00
	2	0.75	0.00	11.48	2.97	0.00	0.00
	3	0.07	0.00	1.22	0.75	0.00	0.00
	4	3.55	0.00	3.67	46.54	0.00	0.00
	5	0.41	0.00	0.00	1.36	0.00	0.00
	6	4.38	0.00	0.00	8.30	0.00	0.00

Tabela III.16. Validação Cruzada – Base de Dados *Sleepdata2*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Sonar Return	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° Amostras / Conjunto**	69
Teste		1	2	3		
Acertos	**Abs.**	50	52	43		
	Per.	72.46	75.36	62.32		

Conjuntos de Teste		4				**N° Amostras / Conjunto**	51
Teste		1	2	3	4		
Acertos	**Abs.**	35	33	42	33		
	Per.	68.63	64.71	82.35	64.71		

Conjuntos de Teste		5				**N° Amostras / Conjunto**	41
Teste		1	2	3	4	5	
Acertos	**Abs.**	29	27	31	32	24	
	Per.	70.73	65.85	75.61	78.05	58.54	

Conjuntos de Teste		6					**N° Amostras / Conjunto**	34
Teste		1	2	3	4	5	6	
Acertos	**Abs.**	23	23	22	29	25	20	
	Per.	67.65	67.65	64.71	85.29	73.53	58.82	

Conjuntos de Teste		7						**N° Amostras / Conjunto**	28
Teste		1	2	3	4	5	6	7	
Acertos	**Abs.**	18	21	16	20	26	19	18	
	Per.	64.29	75.00	57.14	71.43	92.86	67.86	64.29	

Conjuntos de Teste		8							**N° Amostras / Conjunto**	25
Teste		1	2	3	4	5	6	7	8	
Acertos	**Abs.**	15	19	14	18	17	24	18	12	
	Per.	60.00	76.00	56.00	72.00	68.00	96.00	72.00	48.00	

Conjuntos de Teste		9								**N° Amostras / Conjunto**	22
Teste		1	2	3	4	5	6	7	8	9	
Acertos	**Abs.**	12	17	13	14	16	19	17	18	13	
	Per.	54.55	77.27	59.09	63.64	72.73	86.36	77.27	81.82	59.09	

Conjuntos de Teste		10									**N° Amostras / Conjunto**	20	
Teste		1	2	3	4	5	6	7	8	9	10		
Acertos	**Abs.**	10	15	15	10	16	14	19	13	17	12		
	Per.	50.00	75.00	75.00	50.00	80.00	70.00	95.00	65.00	85.00	60.00		

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Sonar Return	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		69	51	41	34	28	25	22	20	
Mínimo	**Absoluto**	43	33	24	20	16	12	12	10	
	Percentual	62.32	64.71	58.54	58.82	57.14	48.00	54.55	50.00	**48.00**
Máximo	**Absoluto**	52	42	32	29	26	24	19	19	
	Percentual	75.36	82.35	78.05	85.29	92.86	96.00	86.36	95.00	**96.00**
Médio	**Absoluto**	48.33	35.75	28.6	23.67	19.71	17.12	15.44	14.1	
	Percentual	70.05	70.10	69.76	69.61	70.41	68.50	70.20	70.50	**69.90**

NÚMERO DE AMOSTRAS POR CLASSE			
Classe	1	2	**Total**
Abs.	111	97	208
Per.	53.37	46.63	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL		**48.00**
CONJUNTOS DE TESTE	8	
AMOSTRAS / CONJUNTO	25	
TESTE	8	
ACERTOS - ABS.	12	

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	6	7
	2	6	6

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL		**96.00**
CONJUNTOS DE TESTE	8	
AMOSTRAS / CONJUNTO	25	
TESTE	6	
ACERTOS - ABS.	24	

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	12	1
	2	0	12

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Sonar Return

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL			**69.90**
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL			
		PREDITO	
		1	**2**
ESPERADO	**1**	36.33	17.34
	2	12.77	33.57

Tabela III.17. Validação Cruzada – Base de Dados *Sonar Return*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Spiral	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° Amostras / Conjunto**	30
Teste		1	2	3		
Acertos	**Abs.**	14	16	16		
	Per.	46.67	53.33	53.33		

Conjuntos de Teste		4				**N° Amostras / Conjunto**	22
Teste		1	2	3	4		
Acertos	**Abs.**	6	8	6	12		
	Per.	27.27	36.36	27.27	54.55		

Conjuntos de Teste		5					**N° Amostras / Conjunto**	18
Teste		1	2	3	4	5		
Acertos	**Abs.**	2	2	6	6	6		
	Per.	11.11	11.11	33.33	33.33	33.33		

Conjuntos de Teste		6						**N° Amostras / Conjunto**	14
Teste		1	2	3	4	5	6		
Acertos	**Abs.**	0	0	4	2	4	8		
	Per.	0.00	0.00	28.57	14.29	28.57	57.14		

Conjuntos de Teste		7							**N° Amostras / Conjunto**	12
Teste		1	2	3	4	5	6	7		
Acertos	**Abs.**	0	0	0	2	0	4	6		
	Per.	0.00	0.00	0.00	16.67	0.00	33.33	50.00		

Conjuntos de Teste		8								**N° Amostras / Conjunto**	10
Teste		1	2	3	4	5	6	7	8		
Acertos	**Abs.**	0	0	0	2	2	0	2	8		
	Per.	0.00	0.00	0.00	20.00	20.00	0.00	20.00	80.00		

Conjuntos de Teste		9									**N° Amostras / Conjunto**	10
Teste		1	2	3	4	5	6	7	8	9		
Acertos	**Abs.**	0	0	0	2	2	0	2	8	0		
	Per.	0.00	0.00	0.00	20.00	20.00	0.00	20.00	80.00	0.00		

Conjuntos de Teste		10									**N° Amostras / Conjunto**	8	
Teste		1	2	3	4	5	6	7	8	9	10		
Acertos	**Abs.**	0	0	0	0	2	4	0	0	4	6		
	Per.	0.00	0.00	0.00	0.00	25.00	50.00	0.00	0.00	50.00	75.00		

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Spiral	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		30	22	18	14	12	10	10	8	
Mínimo	**Absoluto**	14	6	2	0	0	0	0	0	
	Percentual	46.67	27.27	11.11	0.00	0.00	0.00	0.00	0.00	**0.00**
Máximo	**Absoluto**	16	12	6	8	6	8	8	6	
	Percentual	53.33	54.55	33.33	57.14	50.00	80.00	80.00	75.00	**80.00**
Médio	**Absoluto**	15.33	8	4.4	3	1.71	1.75	1.56	1.6	
	Percentual	51.11	36.36	24.44	21.43	14.29	17.50	15.56	20.00	**21.72**

NÚMERO DE AMOSTRAS POR CLASSE			
Classe	1	2	**Total**
Abs.	46	46	92
Per.	50.00	50.00	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	0.00
CONJUNTOS DE TESTE	6
AMOSTRAS / CONJUNTO	14
TESTE	1
ACERTOS - ABS.	0

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	0	7
	2	7	0

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	80.00
CONJUNTOS DE TESTE	8
AMOSTRAS / CONJUNTO	10
TESTE	8
ACERTOS - ABS.	8

MATRIZ DE CONFUSÃO - ABS.			
		PREDITO	
		1	**2**
ESPERADO	**1**	4	1
	2	1	4

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Spiral

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL				21.72
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL				
		PREDITO		
		1	2	
ESPERADO	1	10.86	39.14	
	2	39.14	10.86	

Tabela III.18. Validação Cruzada – Base de Dados *Spiral*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Synthetic	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° de Amostras / Conjunto**	416
Teste		1	2	3		
Acertos	**Abs.**	359	375	366		
	Per.	86.30	90.14	87.98		

Conjuntos de Teste		4				**N° de Amostras / Conjunto**	312
Teste		1	2	3	4		
Acertos	**Abs.**	262	287	276	282		
	Per.	83.97	91.99	88.46	90.38		

Conjuntos de Teste		5				**N° de Amostras / Conjunto**	250
Teste		1	2	3	4	5	
Acertos	**Abs.**	211	226	228	222	225	
	Per.	84.40	90.40	91.20	88.80	90.00	

Conjuntos de Teste		6					**N° de Amostras / Conjunto**	208
Teste		1	2	3	4	5	6	
Acertos	**Abs.**	175	186	188	186	182	188	
	Per.	84.13	89.42	90.38	89.42	87.50	90.38	

Conjuntos de Teste		7						**N° de Amostras / Conjunto**	178
Teste		1	2	3	4	5	6	7	
Acertos	**Abs.**	150	153	160	164	159	155	161	
	Per.	84.27	85.96	89.89	92.13	89.33	87.08	90.45	

Conjuntos de Teste		8							**N° de Amostras / Conjunto**	156
Teste		1	2	3	4	5	6	7	8	
Acertos	**Abs.**	132	130	144	142	140	136	142	140	
	Per.	84.62	83.33	92.31	91.03	89.74	87.18	91.03	89.74	

Conjuntos de Teste		9								**N° de Amostras / Conjunto**	138
Teste		1	2	3	4	5	6	7	8	9	
Acertos	**Abs.**	116	114	128	122	131	120	123	125	123	
	Per.	84.06	82.61	92.75	88.41	94.93	86.96	89.13	90.58	89.13	

Conjuntos de Teste		10								**N° de Amostras / Conjunto**	124	
Teste		1	2	3	4	5	6	7	8	9	10	
Acertos	**Abs.**	103	105	111	111	114	113	111	110	113	109	
	Per.	83.06	84.68	89.52	89.52	91.94	91.13	89.52	88.71	91.13	87.90	

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Synthetic	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		416	312	250	208	178	156	138	124	
Mínimo	**Absoluto**	359	262	211	175	150	130	114	103	
	Percentual	86.30	83.97	84.40	84.13	84.27	83.33	82.61	83.06	**82.61**
Máximo	**Absoluto**	375	287	228	188	164	144	131	114	
	Percentual	90.14	91.99	91.20	90.38	92.13	92.31	94.93	91.94	**94.93**
Médio	**Absoluto**	366.67	276.75	222.4	184.17	157.43	138.25	122.44	110	
	Percentual	88.14	88.70	88.96	88.54	88.44	88.62	88.73	88.71	**88.63**

NÚMERO DE AMOSTRAS POR CLASSE			
Classe	1	2	**Total**
Abs.	625	625	1250
Per.	50.00	50.00	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	82.61
CONJUNTOS DE TESTE	9
AMOSTRAS / CONJUNTO	138
TESTE	2
ACERTOS - ABS.	114

MATRIZ DE CONFUSÃO - ABS.

		PREDITO	
		1	**2**
ESPERADO	**1**	54	15
	2	9	60

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	94.93
CONJUNTOS DE TESTE	9
AMOSTRAS / CONJUNTO	138
TESTE	5
ACERTOS - ABS.	131

MATRIZ DE CONFUSÃO - ABS.

		PREDITO	
		1	**2**
ESPERADO	**1**	64	5
	2	2	67

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Synthetic

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL				**88.63**
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL				
		PREDITO		
		1	**2**	
ESPERADO	**1**	43.64	6.36	
	2	5.01	44.99	

Tabela III.19. Validação Cruzada – Base de Dados *Synthetic*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Vowel	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° Amostras / Conjunto**	330
Teste		1	2	3		
Acertos	**Abs.**	153	196	148		
	Per.	46.36	59.39	44.85		

Conjuntos de Teste		4				**N° Amostras / Conjunto**	242
Teste		1	2	3	4		
Acertos	**Abs.**	119	121	132	118		
	Per.	49.17	50.00	54.55	48.76		

Conjuntos de Teste		5					**N° Amostras / Conjunto**	198
Teste		1	2	3	4	5		
Acertos	**Abs.**	90	107	104	120	79		
	Per.	45.45	54.04	52.53	60.61	39.90		

Conjuntos de Teste		6						**N° Amostras / Conjunto**	165
Teste		1	2	3	4	5	6		
Acertos	**Abs.**	81	83	99	91	104	63		
	Per.	49.09	50.30	60.00	55.15	63.03	38.18		

Conjuntos de Teste		7							**N° Amostras / Conjunto**	132
Teste		1	2	3	4	5	6	7		
Acertos	**Abs.**	75	60	75	63	86	84	55		
	Per.	56.82	45.45	56.82	47.73	65.15	63.64	41.67		

Conjuntos de Teste		8								**N° Amostras / Conjunto**	121
Teste		1	2	3	4	5	6	7	8		
Acertos	**Abs.**	66	65	70	80	55	85	80	53		
	Per.	54.55	53.72	57.85	66.12	45.45	70.25	66.12	43.80		

Conjuntos de Teste		9								**N° Amostras / Conjunto**	110
Teste		1	2	3	4	5	6	7	8	9	
Acertos	**Abs.**	62	64	55	84	46	73	71	67	47	
	Per.	56.36	58.18	50.00	76.36	41.82	66.36	64.55	60.91	42.73	

Conjuntos de Teste		10								**N° Amostras / Conjunto**	99	
Teste		1	2	3	4	5	6	7	8	9	10	
Acertos	**Abs.**	53	53	54	61	61	41	76	66	53	47	
	Per.	53.54	53.54	54.55	61.62	61.62	41.41	76.77	66.67	53.54	47.47	

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Vowel

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. de Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		330	242	198	165	132	121	110	99	
Mínimo	**Absoluto**	148	118	79	63	55	53	46	41	
	Percentual	44.85	48.76	39.90	38.18	41.67	43.80	41.82	41.41	**38.18**
Máximo	**Absoluto**	196	132	120	104	86	85	84	76	
	Percentual	59.39	54.55	60.61	63.03	65.15	70.25	76.36	76.77	**76.77**
Médio	**Absoluto**	165.7	122.5	100	86.83	71.14	69.25	63.22	56.5	
	Percentual	50.20	50.62	50.51	52.63	53.90	57.23	57.47	57.07	**54.70**

NÚMERO DE AMOSTRAS POR CLASSE												
Classe	1	2	3	4	5	6	7	8	9	10	11	**Total**
Abs.	90	90	90	90	90	90	90	90	90	90	90	990
Per.	9.09	9.09	9.09	9.09	9.09	9.09	9.09	9.09	9.09	9.09	9.09	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	**38.18**
CONJUNTOS DE TESTE	6
AMOSTRAS / CONJUNTO	165
TESTE	6
ACERTOS - ABS.	63

MATRIZ DE CONFUSÃO - ABS.

		PREDITO										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
ESPERADO	**1**	8	1	0	0	0	0	0	0	0	6	0
	2	5	2	1	0	0	0	0	0	0	7	0
	3	0	8	2	1	0	0	4	0	0	0	0
	4	0	0	0	7	0	8	0	0	0	0	0
	5	0	0	0	0	4	0	9	0	2	0	0
	6	0	0	0	0	0	7	2	0	0	0	6
	7	0	0	0	0	1	0	6	0	8	0	0
	8	0	0	0	0	0	0	3	0	12	0	0
	9	0	0	0	0	0	0	0	1	8	6	0
	10	1	0	0	0	0	0	0	0	2	12	0
	11	0	0	0	0	0	0	2	0	0	6	7

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Vowel	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	**76.77**
CONJUNTOS DE TESTE	10
AMOSTRAS / CONJUNTO	99
TESTE	7
ACERTOS - ABS.	76

MATRIZ DE CONFUSÃO - ABS.

		PREDITO										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
ESPERADO	**1**	9	0	0	0	0	0	0	0	0	0	0
	2	1	7	0	0	0	0	0	0	0	1	0
	3	0	0	9	0	0	0	0	0	0	0	0
	4	0	0	0	9	0	0	0	0	0	0	0
	5	0	0	0	0	9	0	0	0	0	0	0
	6	0	0	0	0	8	1	0	0	0	0	0
	7	0	0	0	0	0	0	9	0	0	0	0
	8	0	0	0	0	0	0	0	9	0	0	0
	9	0	0	0	0	0	0	0	3	2	4	0
	10	0	0	0	0	0	0	0	0	0	9	0
	11	0	0	0	0	0	0	0	0	6	0	3

ACERTO MÉDIO PERCENTUAL	**54.70**

		PREDITO										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
ESPERADO	**1**	7.37	0.51	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.21	0.00
	2	3.30	4.30	0.22	0.00	0.00	0.00	0.00	0.00	0.02	1.25	0.00
	3	0.71	4.09	3.56	0.21	0.00	0.01	0.22	0.00	0.00	0.16	0.11
	4	0.01	0.07	1.44	4.94	0.46	1.47	0.13	0.00	0.01	0.00	0.55
	5	0.02	0.04	0.08	0.00	4.52	0.56	3.46	0.10	0.23	0.00	0.08
	6	0.00	0.25	0.32	0.34	2.07	3.97	0.74	0.00	0.33	0.04	1.03
	7	0.50	0.24	0.01	0.00	0.14	0.00	6.41	0.96	0.76	0.07	0.00
	8	0.09	0.00	0.00	0.00	0.00	0.00	1.00	6.52	1.11	0.37	0.00
	9	0.03	0.18	0.00	0.00	0.00	0.00	1.54	2.10	3.28	1.97	0.00
	10	0.97	0.79	0.00	0.00	0.00	0.00	0.06	0.13	0.72	6.32	0.09
	11	0.01	0.56	0.02	0.00	0.25	1.44	1.48	0.00	0.76	1.07	3.51

Tabela III.20. Validação Cruzada – Base de Dados *Vowel*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Wine	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3			**N° de Amostras / Conjunto**	59
Teste		1	2	3		
Acertos	**Abs.**	57	58	58		
	Per.	96.61	98.31	98.31		

Conjuntos de Teste		4				**N° de Amostras / Conjunto**	43
Teste		1	2	3	4		
Acertos	**Abs.**	42	41	42	43		
	Per.	97.67	95.35	97.67	100.00		

Conjuntos de Teste		5				**N° de Amostras / Conjunto**	34
Teste		1	2	3	4	5	
Acertos	**Abs.**	33	32	33	34	34	
	Per.	97.06	94.12	97.06	100.00	100.00	

Conjuntos de Teste		6					**N° de Amostras / Conjunto**	29
Teste		1	2	3	4	5	6	
Acertos	**Abs.**	28	28	28	28	29	29	
	Per.	96.55	96.55	96.55	96.55	100.00	100.00	

Conjuntos de Teste		7						**N° de Amostras / Conjunto**	24
Teste		1	2	3	4	5	6	7	
Acertos	**Abs.**	23	24	22	24	24	23	24	
	Per.	95.83	100.00	91.67	100.00	100.00	95.83	100.00	

Conjuntos de Teste		8							**N° de Amostras / Conjunto**	21
Teste		1	2	3	4	5	6	7	8	
Acertos	**Abs.**	20	21	18	21	20	21	21	21	
	Per.	95.24	100.00	85.71	100.00	95.24	100.00	100.00	100.00	

Conjuntos de Teste		9								**N° de Amostras / Conjunto**	18
Teste		1	2	3	4	5	6	7	8	9	
Acertos	**Abs.**	17	18	18	15	18	17	18	18	18	
	Per.	94.44	100.00	100.00	83.33	100.00	94.44	100.00	100.00	100.00	

Conjuntos de Teste		10									**N° de Amostras / Conjunto**	16	
Teste		1	2	3	4	5	6	7	8	9	10		
Acertos	**Abs.**	15	16	16	14	16	16	16	16	16	16		
	Per.	93.75	100.00	100.00	87.50	100.00	100.00	100.00	100.00	100.00	100.00		

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): Wine	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. de Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		59	43	34	29	24	21	18	16	
Mínimo	**Absoluto**	57	41	32	28	22	18	15	14	
	Percentual	96.61	95.35	94.12	96.55	91.67	85.71	83.33	87.50	**83.33**
Máximo	**Absoluto**	58	43	34	29	24	21	18	16	
	Percentual	98.31	100.00	100.00	100.00	100.00	100.00	100.00	100.00	**100.00**
Médio	**Absoluto**	57.67	42	33.2	28.33	23.43	20.38	17.44	15.7	
	Percentual	97.74	97.67	97.65	97.70	97.62	97.02	96.91	98.12	**97.53**

NÚMERO DE AMOSTRAS POR CLASSE				
Classe	1	2	3	**Total**
Abs.	48	51	79	178
Per.	26.97	28.65	44.38	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	**83.33**
CONJUNTOS DE TESTE	9
AMOSTRAS / CONJUNTO	18
TESTE	4
ACERTOS - ABS.	15

MATRIZ DE CONFUSÃO - ABS.

		PREDITO		
		1	**2**	**3**
ESPERADO	**1**	5	0	0
	2	1	3	1
	3	0	1	7

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	**100.00**
CONJUNTOS DE TESTE	4
AMOSTRAS / CONJUNTO	43
TESTE	4
ACERTOS - ABS.	43

MATRIZ DE CONFUSÃO - ABS.

		PREDITO		
		1	**2**	**3**
ESPERADO	**1**	12	0	0
	2	0	12	0
	3	0	0	19

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): Wine

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL					**97.53**
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL					
		PREDITO			
		1	**2**	**3**	
ESPERADO	**1**	26.82	0.00	0.00	
	2	0.83	27.56	0.57	
	3	0.00	1.08	43.15	

Tabela III.21. Validação Cruzada – Base de Dados *Wine*.

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): WNBA	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10
TOTAL DE TESTES REALIZADOS	52

TESTES REALIZADOS

Conjuntos de Teste		3		**N° Amostras / Conjunto**	38
Teste		1	2	3	
Acertos	**Abs.**	30	28	31	
	Per.	78.95	73.68	81.58	

Conjuntos de Teste		4		**N° Amostras / Conjunto**	29
Teste		1	2	3	4
Acertos	**Abs.**	21	24	23	25
	Per.	72.41	82.76	79.31	86.21

Conjuntos de Teste		5		**N° Amostras / Conjunto**	23
Teste		1	2	3	4	5
Acertos	**Abs.**	16	19	19	19	19
	Per.	69.57	82.61	82.61	82.61	82.61

Conjuntos de Teste		6		**N° Amostras / Conjunto**	18
Teste		1	2	3	4	5	6
Acertos	**Abs.**	14	13	14	14	15	15
	Per.	77.78	72.22	77.78	77.78	83.33	83.33

Conjuntos de Teste		7		**N° Amostras / Conjunto**	15
Teste		1	2	3	4	5	6	7
Acertos	**Abs.**	12	10	11	13	11	14	11
	Per.	80.00	66.67	73.33	86.67	73.33	93.33	73.33

Conjuntos de Teste		8		**N° Amostras / Conjunto**	13
Teste		1	2	3	4	5	6	7	8
Acertos	**Abs.**	11	8	12	9	10	11	11	11
	Per.	84.62	61.54	92.31	69.23	76.92	84.62	84.62	84.62

Conjuntos de Teste		9		**N° Amostras / Conjunto**	12
Teste		1	2	3	4	5	6	7	8	9
Acertos	**Abs.**	10	7	11	9	8	11	7	11	8
	Per.	83.33	58.33	91.67	75.00	66.67	91.67	58.33	91.67	66.67

Conjuntos de Teste		10		**N° Amostras / Conjunto**	11	
Teste		1	2	3	4	5	6	7	8	9	10
Acertos	**Abs.**	9	6	11	8	8	10	8	9	10	10
	Per.	81.82	54.55	100.00	72.73	72.73	90.91	72.73	81.82	90.91	90.91

VALIDAÇÃO CRUZADA	
BASE DE DADOS (BD): WNBA	
VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

RESUMO - ACERTOS										
Conjs. de Teste		3	4	5	6	7	8	9	10	**Total**
Amostras / Conjunto		38	29	23	18	15	13	12	11	
Mínimo	**Absoluto**	28	21	16	13	10	8	7	6	
	Percentual	73.68	72.41	69.57	72.22	66.67	61.54	58.33	54.55	**54.55**
Máximo	**Absoluto**	31	25	19	15	14	12	11	11	
	Percentual	81.58	86.21	82.61	83.33	93.33	92.31	91.67	100.00	**100.00**
Médio	**Absoluto**	29.67	23.25	18.4	14.17	11.71	10.38	9.11	8.9	
	Percentual	78.07	80.17	80.00	78.70	78.10	79.81	75.93	80.91	**78.94**

NÚMERO DE AMOSTRAS POR CLASSE				
Classe	1	2	3	**Total**
Abs.	53	47	20	120
Per.	44.17	39.17	16.67	100.00

PRIMEIRA OCORRÊNCIA DE ACERTO MÍNIMO PERCENTUAL	**54.55**
CONJUNTOS DE TESTE	10
AMOSTRAS / CONJUNTO	11
TESTE	2
ACERTOS - ABS.	6

MATRIZ DE CONFUSÃO - ABS.				
		PREDITO		
		1	**2**	**3**
ESPERADO	**1**	4	1	0
	2	1	2	1
	3	0	2	0

PRIMEIRA OCORRÊNCIA DE ACERTO MÁXIMO PERCENTUAL	**100.00**
CONJUNTOS DE TESTE	10
AMOSTRAS / CONJUNTO	11
TESTE	3
ACERTOS - ABS.	11

MATRIZ DE CONFUSÃO - ABS.				
		PREDITO		
		1	**2**	**3**
ESPERADO	**1**	5	0	0
	2	0	4	0
	3	0	0	2

VALIDAÇÃO CRUZADA

BASE DE DADOS (BD): WNBA

VARIAÇÃO DA DIVISÃO DA BD EM CONJUNTOS DE TESTE	
Mínimo	**Máximo**
3	10

TOTAL DE TESTES REALIZADOS	52

ACERTO MÉDIO PERCENTUAL					78.94
MATRIZ DE CONFUSÃO - VALORES MÉDIOS - PERCENTUAL					
		PREDITO			
		1	2	3	
ESPERADO	1	40.13	4.10	0.45	
	2	5.27	28.86	4.82	
	3	0.00	6.43	9.95	

Tabela III.22. Validação Cruzada – Base de Dados *WNBA*.