3
Resultados e Análises

Neste capítulo serão apresentados os resultados obtidos com a aplicação da metodologia ao banco de dados descrito no capítulo anterior.

As possibilidades de medições para avaliar o desempenho da metodologia proposta são descritas a seguir:

(i) – Percentual de acerto dos não-doentes (PND):

É definido como:

$$PND = \frac{ND_R}{ND} x100 \tag{3.1}$$

Sendo:

ND_R → número de indivíduos que a Rede estima como não-doente quando efetivamente o indivíduo é não-doente;

ND → número total de indivíduos não-doentes.

(ii) – Percentual de acerto dos doentes (PD):

É definido como:

$$PD = \frac{D_R}{D} x100 \tag{3.2}$$

Sendo:

D_R → número de indivíduos que a Rede estima como doente quando efetivamente o indivíduo é doente;

D → número total de indivíduos doentes.

(iii) – Percentual de acerto médio (PM):

É definido como:

$$PM = \frac{PND \cdot ND + PD \cdot D}{ND + D} \tag{3.3}$$

De acordo com os valores de PND, PD e PM, obtidos nos experimentos aqui realizados, pode-se avaliar a qualidade do método proposto, estabelecendo-se comparações dos valores aqui obtidos com valores encontrados em experimentos realizados por outros autores e encontrados na literatura.

3.1
Analisando os Resultados da Rede Neural

Na tabela 2.1, tem-se a ordenação das 13 variáveis presentes no banco e seus respectivos valores de informação mútua (IM) com relação ao desfecho. É importante salientar que o fato de uma variável ter IM nula com relação ao desfecho não significa que conjuntamente com outras variáveis ela não possa vir a ter importância. É por isso que a oitava variável em ordem de importância (FBS) também foi utilizada nas simulações, de forma que se pudesse verificar sua contribuição para a classificação na presença das outras variáveis. As variáveis ordenadas da nona à décima terceira posição não apresentam uma contribuição significativa ao percentual de acerto da rede quando utilizadas nas simulações. Assim, foram utilizadas as 8 primeiras variáveis da tabela 2.1 como entrada da rede neural, da seguinte forma: o número de entradas da rede foi variado de 1 a 8, seguindo a ordem de importância na qual as variáveis foram ordenadas. Desta forma, a Rede Neural foi executada inicialmente com a primeira variável como entrada (THAL), depois com as duas variáveis mais importantes como entrada (THAL e CA), depois com as três variáveis mais importantes como entrada (THAL, CA e CP), e assim sucessivamente, até ter sido executada com as 8 variáveis ordenadas. Os resultados obtidos, para 50 conjuntos de Treinamento/Generalização diferentes, foram os seguintes:

Tabela 3.1 – Percentual de Acerto no Treinamento

Variáveis de entrada da rede	PND	PD	PM
THAL	80,0	72,5	76,5
THAL,CA	79,3	80,1	79,6
THAL,CA,CP	89,5	81,7	85,9
THAL,CA,CP,EXANG	89,8	84,3	87,3
THAL,CA,CP,EXANG,SLOPE	93,3	86,0	90,0
THAL,CA,CP,EXANG,SLOPE,SEX	93,2	89,7	91,6
THAL,CA,CP,EXANG,SLOPE,SEX,RESTECG	95,4	92,0	93,9
THAL,CA,CP,EXANG,SLOPE,SEX,RESTECG,FBS	96,0	92,8	94,5

Tabela 3.2 – Percentual de Acerto na Generalização

Variáveis de entrada da rede	PND	PD	PM
THAL	79,7	70,3	75,4
THAL,CA	76,9	79,2	78,0
THAL,CA,CP	85,9	76,3	81,4
THAL,CA,CP,EXANG	82,8	76,0	79,6
THAL,CA,CP,EXANG,SLOPE	85,1	75,6	80,7
THAL,CA,CP,EXANG,SLOPE,SEX	79,9	78,2	79,1
THAL,CA,CP,EXANG,SLOPE,SEX,RESTECG	80,7	77,8	79,4
THAL,CA,CP,EXANG,SLOPE,SEX,RESTECG,FBS	79,0	79,3	79,2

Como o que se deseja é analisar a qualidade da classificação feita pelo modelo com relação aos indivíduos que não pertenciam ao conjunto de treinamento, deve-se analisar então os resultados do acerto na generalização (fora-da-amostra). O resultado mais representativo encontrado foi então aquele obtido quando se utilizou 3 variáveis de entrada (THAL, CA e CP), obtendo-se um percentual de acerto médio (PM) fora-da-amostra de 81,4 % . É importante que os percentuais de acerto fora-da-amostra para os casos de não-doente (PND = 85,9 %) e doente (PD = 76,3 %), e na amostra (ou seja, no treinamento) para os casos de não-doente e doente (89,5 % e 81,7 %, respectivamente), também apresentem bons resultados, o que foi constatado neste caso.

Apesar de ter sido encontrado um bom resultado, percebeu-se ao longo das simulações que, para os melhores resultados obtidos, classificavam-se

erroneamente quase sempre os mesmos indivíduos. Ou seja, existia um patamar ótimo de percentagem de acerto, do qual não se conseguia ir além, pois existia um grupo de indivíduos que continuamente era classificado erroneamente, mesmo que o número de variáveis de entrada da rede fosse alterado. Por isso, decidiu-se separar e estudar com mais atenção este grupo de indivíduos e fazer simulações sem esses indivíduos para ver se o resultado melhorava e o quanto ele melhorava. Para isso foram excluídos 10 indivíduos não-doentes e 21 indivíduos doentes (31 indivíduos no total), permanecendo agora 150 indivíduos não-doentes e 116 indivíduos doentes para serem utilizados em uma nova simulação (266 indivíduos no total).

Feito isto, uma nova simulação foi realizada somente com os 266 indivíduos que restaram após a exclusão dos 31 que eram sempre classificados erroneamente. Mantendo-se a mesma proporção de amostras treinamento/generalização da simulação anterior, utilizou-se então 2/3 das amostras para treinamento (177 indivíduos) e 1/3 para generalização (89 indivíduos). Os resultados obtidos, para 50 conjuntos de Treinamento/Generalização diferentes, foram os seguintes:

Tabela 3.3 – Percentual de Acerto no Treinamento (sem os 31 indivíduos)

Variáveis de entrada da rede	PND	PD	PM
THAL	82,8	81,8	82,4
THAL,CA	84,3	90,3	86,9
THAL,CA,CP	95,0	93,6	94,4
THAL,CA,CP,EXANG	95,2	95,1	95,2
THAL,CA,CP,EXANG,SLOPE	96,7	96,7	96,7
THAL,CA,CP,EXANG,SLOPE,SEX	97,2	97,3	97,2
THAL,CA,CP,EXANG,SLOPE,SEX,RESTECG	98,1	97,8	98,0
THAL,CA,CP,EXANG,SLOPE,SEX,RESTECG,FBS	98,7	98,4	98,6

Tabela 3.4 – Percentual de Acerto na Generalização (sem os 31 indivíduos)

Variáveis de entrada da rede	PND	PD	PM
THAL	83,9	80,6	82,4
THAL,CA	83,2	88,2	85,3
THAL,CA,CP	94,4	92,7	93,7
THAL,CA,CP,EXANG	93,2	90,5	92,0
THAL,CA,CP,EXANG,SLOPE	91,6	88,7	90,4
THAL,CA,CP,EXANG,SLOPE,SEX	91,0	90,8	90,9
THAL,CA,CP,EXANG,SLOPE,SEX,RESTECG	89,9	89,5	89,7
THAL,CA,CP,EXANG,SLOPE,SEX,RESTECG,FBS	91,0	88,7	90,0

Desta forma, pode-se perceber o quanto estes 31 indivíduos interferem na qualidade da classificação, visto que com eles inseridos no banco, o resultado de PM na generalização e usando as 3 variáveis (THAL, CA e CP) era 81,4 %, passando para 93,7 % quando estes foram excluímos do banco.

Utilizando-se novamente todo o banco de dados (297 indivíduos), foram feitas novas simulações da seguinte forma: com a mesma proporção de 2/3 do banco de dados para treinamento e 1/3 para generalização. Os indivíduos de cada grupo foram escolhidos de forma a se garantir que os 31 que foram classificados como pertencentes ao grupo de **"indivíduos sempre classificados erroneamente", pertençam sempre ao conjunto de treinamento**. Os resultados obtidos, para 50 conjuntos de Treinamento/Generalização diferentes, foram os seguintes:

Tabela 3.5 – Percentual de Acerto no Treinamento
(com os 31 indivíduos no treinamento)

Variáveis de entrada da rede	PND	PD	PM
THAL	78,3	67,6	73,4
THAL,CA	75,8	74,2	75,0
THAL,CA,CP	85,8	73,8	80,3
THAL,CA,CP,EXANG	86,9	77,5	82,6
THAL,CA,CP,EXANG,SLOPE	91,4	82,4	87,3
THAL,CA,CP,EXANG,SLOPE,SEX	91,6	85,3	88,7
THAL,CA,CP,EXANG,SLOPE,SEX,RESTECG	93,9	89,9	92,1
THAL,CA,CP,EXANG,SLOPE,SEX,RESTECG,FBS	94,8	90,9	93,0

Tabela 3.6 – Percentual de Acerto na Generalização
(com os 31 indivíduos no treinamento)

Variáveis de entrada da rede	PND	PD	PM
THAL	82,0	82,7	82,3
THAL,CA	83,9	90,8	87,1
THAL,CA,CP	91,9	89,8	91,0
THAL,CA,CP,EXANG	87,3	87,7	87,5
THAL,CA,CP,EXANG,SLOPE	85,6	83,7	84,7
THAL,CA,CP,EXANG,SLOPE,SEX	83,6	86,4	84,9
THAL,CA,CP,EXANG,SLOPE,SEX,RESTECG	82,7	82,6	82,7
THAL,CA,CP,EXANG,SLOPE,SEX,RESTECG,FBS	80,8	85,7	83,1

Assim, pode-se perceber o quanto o resultado da classificação dos indivíduos melhora quando se compara os resultados da tabela 3.2 com os resultados mostrados na tabela 3.6. Utilizando-se as 3 variáveis (THAL, CA e CP) e todo o banco de dados, quando se escolhe os indivíduos de cada grupo de forma a garantir que os 31 indivíduos pertençam sempre ao conjunto de treinamento, o resultado de PM na generalização passou de 81,4 % (tabela 3.2) para 91,0 % (tabela 3.6).

A idéia é que, se este grupo de 31 indivíduos (ou parte deste grupo) apresentar um padrão de comportamento especial, este padrão será assimilado (aprendido) pela rede neural durante o treinamento, e possíveis indivíduos que

apresentem este padrão de comportamento especial no conjunto de generalização (teste) poderão ser classificados corretamente.

A seguir são apresentados os gráficos de colunas das duas primeiras variáveis na ordem de importância, THAL (figura 3.1) e CA (figura 3.2), para os 266 indivíduos que sobram quando se retira os 31 indivíduos que são sempre classificados erroneamente, e para estes 31 indivíduos, ambos separados para não-doentes e doentes:

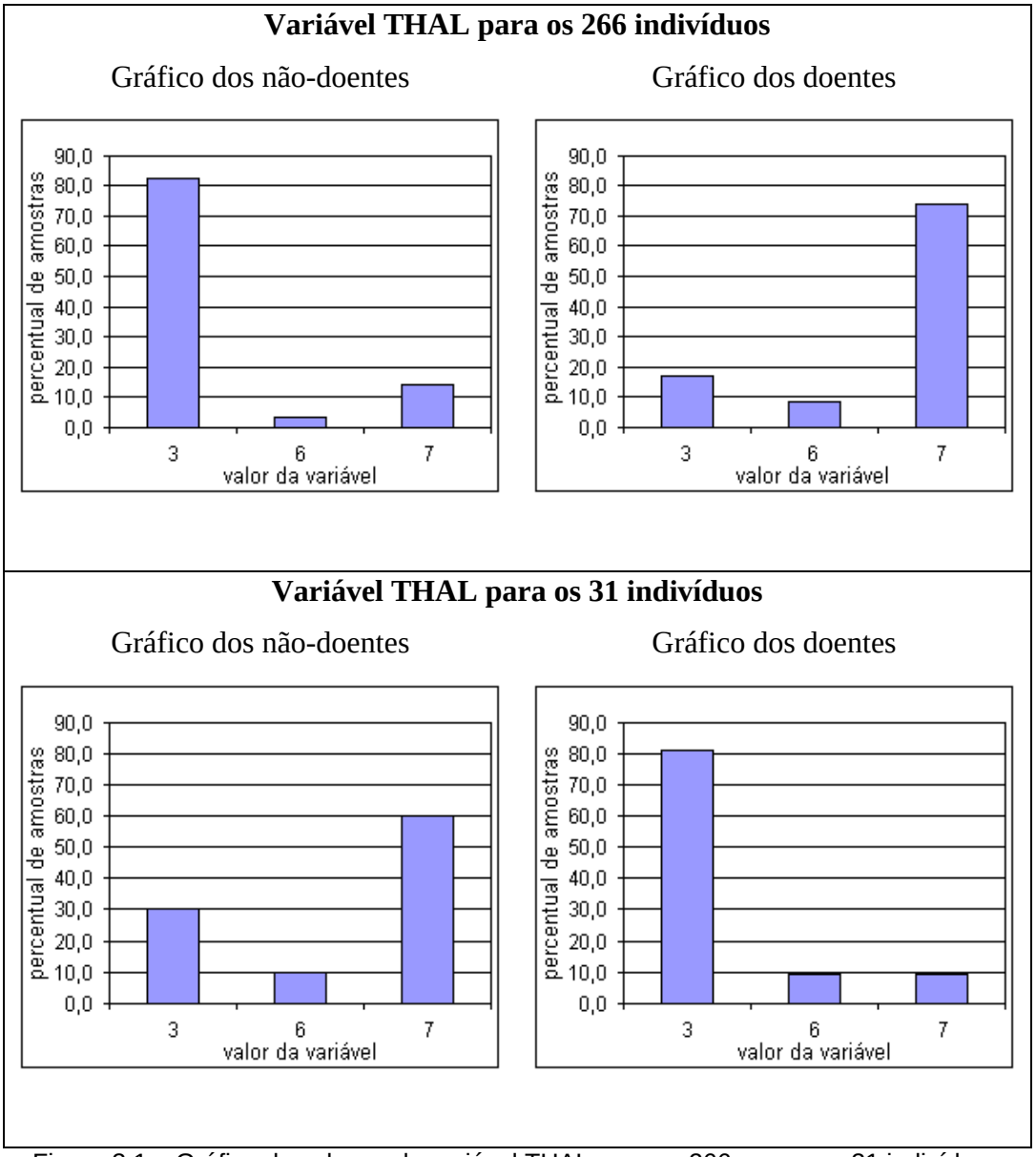

Figura 3.1 – Gráfico de colunas da variável THAL para os 266 e para os 31 indivíduos

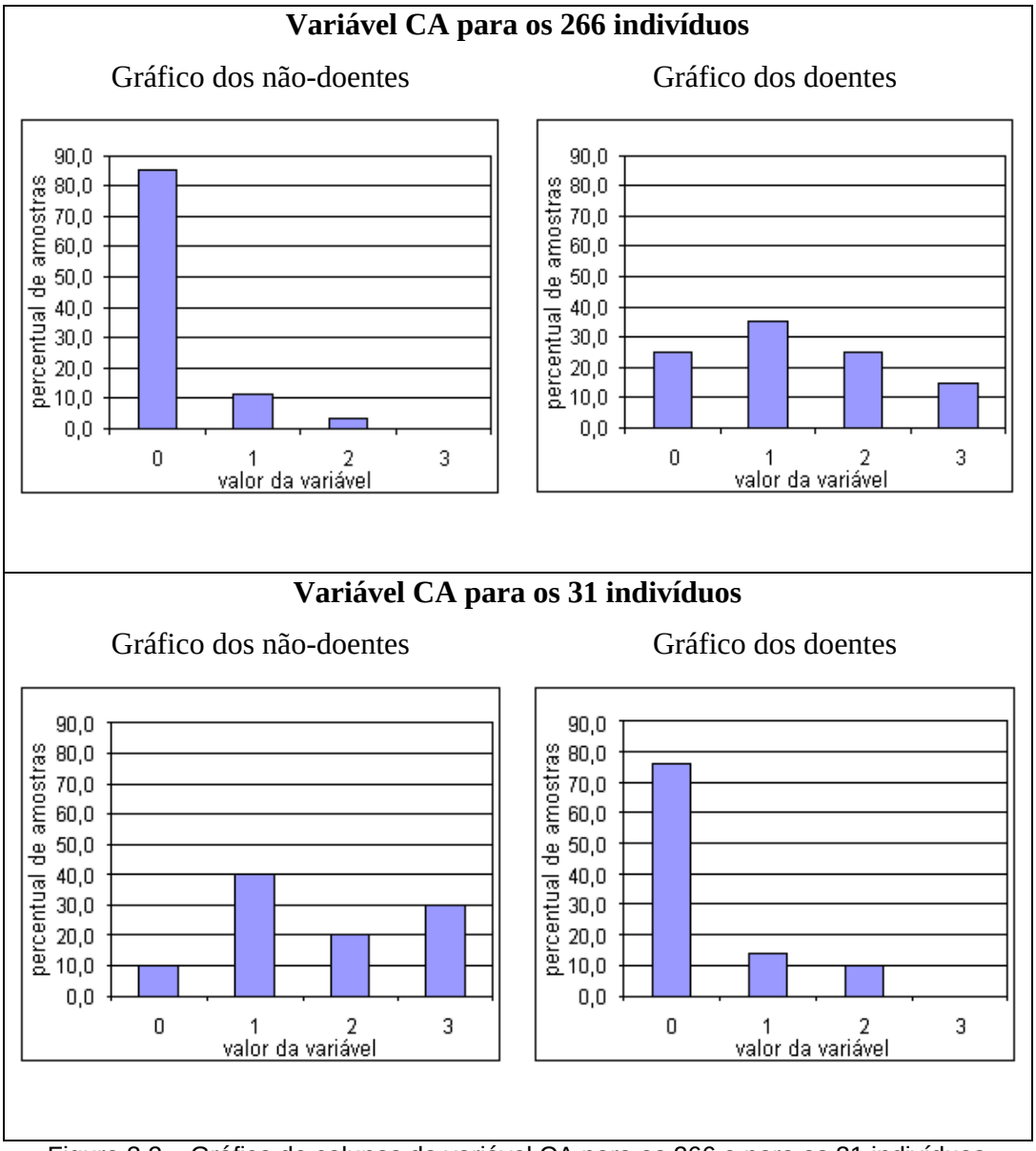

Figura 3.2 – Gráfico de colunas da variável CA para os 266 e para os 31 indivíduos

Como pode ser observado nestes gráficos, o comportamento dos 31 indivíduos é praticamente o oposto ao restante dos indivíduos, visto que o padrão que deveria ser apresentado para os não-doentes, está sendo apresentado para os doentes, e o padrão que deveria ser apresentado para os doentes, está sendo apresentado para os não-doentes. Isto reforça a hipótese de que os 31 indivíduos apresentam um padrão de comportamento especial.

Quando se compara o resultado obtido neste experimento, a partir da metodologia proposta, com resultados obtidos por outros autores encontrados na literatura, percebe-se que o resultado obtido nesta dissertação apresenta um melhor percentual de acerto. Os resultados obtidos por Ho & Chou (2001),

utilizando “Lógica Fuzzy”, e por Hu, Li, Cai & Xu (2004), utilizando “Máquina de Vetor de Suporte”, sendo utilizado em ambos os estudos a mesma proporção aqui escolhida para dividir o banco de dados em conjunto de treinamento e conjunto de generalização, são da ordem de PM = 83,0 % e PM = 83,5 %, respectivamente, apresentados para o conjunto de generalização. O resultado obtido neste experimento, utilizando “Redes Neurais”, com a metodologia aqui proposta, apresentou PM = 91,0 %, para o conjunto de generalização. Portanto, um ganho considerável no percentual médio de acerto, demonstrando a qualidade da classificação segundo a metodologia adotada. Cabe salientar que os outros autores utilizaram todas as 13 variáveis previamente selecionadas, enquanto nesta metodologia bastaram apenas 3 variáveis.

3.2
Comparando os Resultados da Rede Neural aos de outros Métodos de Classificação de Padrões

É interessante também, utilizar-se outras técnicas de classificação de padrões conhecidas na literatura de forma a se estabelecer comparações com os resultados obtidos nesta dissertação, e também para avaliar o impacto sobre estas técnicas de se utilizar a metodologia de divisão dos conjuntos de treinamento/generalização aqui sugerida, ou seja, utilizando-se 2/3 do banco para treinamento e 1/3 para generalização e, da mesma forma, considerando os “indivíduos sempre classificados erroneamente” pertencendo sempre ao conjunto de treinamento.

Foram escolhidas duas técnicas de classificação de padrões encontradas na literatura, que são a **Análise Discriminante** (ver detalhes no Anexo C) e o **Algoritmo C4.5** (ver detalhes no Anexo D). A **Análise Discriminante** é útil para as situações onde se quer construir um modelo preditivo de um grupo de amostras baseado em características observadas de cada caso. O procedimento gera uma função discriminante (ou, para mais de dois grupos, um conjunto de funções discriminantes) baseada em combinações lineares das variáveis preditas que fornecem a melhor discriminação entre as amostras. As funções são geradas a partir de uma parte dos casos para os quais a classe das amostras é conhecida; as funções podem então ser aplicadas aos casos novos, para os quais a classe das amostras é desconhecida. Já o **Algoritmo C4.5** gera um classificador na forma de

uma árvore de decisão, com uma estrutura composta por: 1) uma folha, indicando uma classe; 2) um nó de decisão que especifica um teste a ser realizado no valor de um atributo, com um galho para cada resposta possível do teste, que levará para uma sub-árvore ou uma folha. Em uma árvore de decisão a classificação de um caso se inicia pela raiz da árvore, e esta árvore é percorrida até que se chegue a uma folha. Em cada nó de decisão será feito um teste que irá direcionar o caso para uma sub-árvore. Este processo irá guiar-se para uma folha. A classe do caso pressupõe-se que seja a mesma que está armazenada nesta folha.

Segundo as simulações realizadas, pôde-se comprovar que as 3 primeiras variáveis ordenadas (THAL, CA e CP) são as que permitem uma melhor classificação dos indivíduos com relação à determinação do desfecho. Por isso, de forma a se estabelecer comparações, o número de variáveis de entrada nesta etapa foi variado de 1 a 3, seguindo-se a ordem de importância nas quais as variáveis foram colocadas (primeiro com THAL, depois com THAL e CA, e por fim com THAL, CA e CP). Utilizando-se 2/3 do banco para treinamento e 1/3 para generalização, sem nenhum outro refinamento, e aplicando no algoritmo C4.5 e na análise discriminante, os resultados obtidos, para 50 conjuntos de Treinamento/Generalização diferentes, foram os seguintes:

Tabela 3.7 – Percentual de Acerto no Algoritmo C4.5 e na Análise Discriminante (sem refinamento)

Variáveis de entrada	**C4.5**			**Análise Discriminante**		
	PND	**PD**	**PM**	**PND**	**PD**	**PM**
THAL	79,4	71,1	**75,5**	79,2	72,7	**76,2**
THAL,CA	79,6	71,4	**75,8**	77,4	78,6	**78,0**
THAL,CA,CP	76,9	74,8	**75,9**	84,0	76,8	**80,6**

A tabela 3.8 a seguir, apresenta os resultados para os 3 métodos de classificação de padrões que estão sendo comparados, considerando-se apenas os valores de PM, e utilizando-se o banco sem nenhum refinamento:

Tabela 3.8 – Tabela Comparativa, Valores de PM (sem refinamento)

Variáveis de entrada	C4.5	Análise Discriminante	Rede Neural
THAL	75,5	76,2	75,4
THAL,CA	75,8	78,0	78,0
THAL,CA,CP	75,9	80,6	81,4

Pela análise da tabela 3.8, pode-se perceber que a rede neural, com a configuração aqui utilizada, fornece a melhor classificação entre os 3 métodos avaliados, quando as 3 variáveis (THAL, CA e CP) são utilizadas.

Mantendo-se a mesma proporção de amostras treinamento/generalização da simulação anterior, faz-se agora uma nova simulação, com o algoritmo C4.5 e a análise discriminante, somente com os 266 indivíduos que restam após serem excluídos os 31 indivíduos. Os resultados obtidos, para 50 conjuntos de Treinamento/Generalização diferentes, são os seguintes:

Tabela 3.9 – Percentual de Acerto no Algoritmo C4.5 e na Análise Discriminante (sem os 31 indivíduos)

Variáveis de entrada	C4.5			Análise Discriminante		
	PND	**PD**	**PM**	**PND**	**PD**	**PM**
THAL	82,8	82,3	**82,6**	83,0	82,3	**82,7**
THAL,CA	82,5	80,9	**81,8**	81,4	91,2	**85,7**
THAL,CA,CP	77,2	82,1	**79,3**	90,6	87,1	**89,0**

A tabela 3.10 a seguir, apresenta os resultados para os 3 métodos de classificação de padrões que estão sendo comparados, considerando-se apenas os valores de PM, e utilizando-se o banco sem os 31 indivíduos classificados sempre incorretamente:

Tabela 3.10 – Tabela Comparativa, Valores de PM (sem os 31 indivíduos)

Variáveis de entrada	C4.5	Análise Discriminante	Rede Neural
THAL	82,6	82,7	82,4
THAL,CA	81,8	85,7	85,3
THAL,CA,CP	79,3	89,0	93,7

Pela análise da tabela 3.10, onde os 31 indivíduos não são utilizados no banco, pode-se perceber que a rede neural, com a configuração aqui utilizada, continua fornecendo a melhor classificação entre os 3 métodos avaliados, quando as 3 variáveis (THAL, CA e CP) são utilizadas.

Finalmente, utilizando-se novamente todo o banco de dados, as simulações são feitas garantindo-se que os 31 indivíduos que foram classificados como pertencentes ao grupo de **"indivíduos sempre classificados erroneamente" pertençam sempre ao conjunto de treinamento**. Aplicando-se estes conjuntos no algoritmo C4.5 e na análise discriminante, os resultados obtidos, para 50 conjuntos de Treinamento/Generalização diferentes, são os seguintes:

Tabela 3.11 – Percentual de Acerto no Algoritmo C4.5 e na Análise Discriminante (com os 31 indivíduos no treinamento)

Variáveis de entrada	C4.5			Análise Discriminante		
	PND	**PD**	**PM**	**PND**	**PD**	**PM**
THAL	82,3	83,4	**82,8**	82,6	83,0	**82,8**
THAL,CA	82,8	81,6	**82,2**	82,2	91,7	**86,6**
THAL,CA,CP	76,5	81,7	**78,9**	88,5	89,6	**89,0**

A tabela 3.12 a seguir, apresenta os resultados para os 3 métodos de classificação de padrões que estão sendo comparados, considerando-se apenas os valores de PM:

Tabela 3.12 – Tabela Comparativa, Valores de PM
(com os 31 indivíduos no treinamento)

Variáveis de entrada	C4.5	Análise Discriminante	Rede Neural
THAL	82,8	82,8	82,3
THAL,CA	82,2	86,6	87,1
THAL,CA,CP	78,9	89,0	91,0

Pela análise da tabela 3.12, pode-se perceber que a rede neural ainda continua fornecendo a melhor classificação entre os 3 métodos avaliados, quando as 3 variáveis (THAL, CA e CP) são utilizadas. Uma das conclusões importantes que podem ser tiradas desses resultados, ao serem comparadas as tabelas 3.10 e 3.12, é que os valores de PM praticamente não se alteram ou se alteram pouco, principalmente os resultados do algoritmo C4.5 e da análise discriminante.

Quando se mantém os 31 indivíduos no conjunto treinamento, busca-se uma alternativa de extrair informação útil deste conjunto de indivíduos. A conclusão que se pode tirar das análises realizadas, é que ter o "grupo de indivíduos sempre classificados erroneamente" dentro do conjunto de treinamento, pode auxiliar na classificação correta de indivíduos presentes no conjunto de generalização. Isto provavelmente pode ser estendido para qualquer banco de dados.