Agradecimentos

Aos meus pais, por tudo o que me ensinaram, e à minha irmã pela força.

Um agradecimento muito especial ao professor e orientador Doutor Paulo Cortez, pelos ensinamentos, amizade, paciência, confiança, incentivos e solidariedade demonstrada nos momentos mais difíceis da realização deste trabalho.

Gostaria de deixar um reconhecimento ao colega e amigo José Novais, que partilhou comigo muitas noites de longas conversas, ensinamentos e desabafos.

A toda a equipa de trabalho laboratorial, responsáveis pela recolha dos dados, Eng. Sandra Rodrigues, Eng. Vasco Cadavez e Prof. Dr. Alfredo Teixeira, aos quais agradeço todos os esclarecimentos e disponibilidade demonstrada.

Não quero deixar de agradecer a estimada contribuição dos colegas do CiESA, Luís Pires, Nuno Carvalho, Pedro Bastos e Sérgio Deusdado, por todo o apoio e incentivo demonstrado.

Por último, quero deixar um agradecimento a todos aqueles que, directa ou indirectamente, colaboraram para este trabalho.

Redes Neuronais Artificiais para a Previsão da Qualidade em Carnes

Resumo

A recente criação de produtos cárnicos de origem ovina, com Denominação de Origem ou Identificação Geográfica Protegida, é um incentivo ao desenvolvimento de produtos de qualidade, cujas características devem corresponder às expectativas dos consumidores. De facto, estes dão cada vez mais prioridade à qualidade dos produtos em detrimento da quantidade, estando mesmo dispostos a pagar preços extra por artigos superiores. Assim, no presente mercado global e competitivo, a qualidade torna-se um factor económico deveras relevante para a indústria da carne. De entre os diversos factores que condicionam o paladar, a *tenrura* é considerada a característica mais importante na influência da percepção alimentar.

Por outro lado, nas últimas décadas tem sido dedicada uma intensa investigação à extracção de conhecimento de alto nível a partir de dados em bruto, num processo iterativo designado por *Descoberta de Conhecimento em Bases de Dados* e que envolve todo um conjunto de etapas. O *Data Mining* é uma fase crucial deste processo, passando pela aplicação de algoritmos de aprendizagem (*e.g. Redes Neuronais Artificiais*) com vista à procura de padrões úteis.

Neste trabalho é proposto um *Conjunto de Redes Neuronais*, baseado na selecção de atributos via um procedimento de *Análise de Sensibilidade*, para predição da *tenrura* em carne de cordeiros oriundos da região de Trás-os-Montes. Este problema foi modelado através de duas tarefas de regressão, usando *medições instrumentais* e um *painel sensorial*. Em ambos os casos, a solução proposta apresentou melhores resultados que outras configurações de *Redes Neuronais*, bem como o método clássico da *Regressão Múltipla*. Além disso, o modelo neuronal é mais rápido e menos custoso do que os métodos convencionais para a aferição da *tenrura*, abrindo assim caminho para o desenvolvimento de ferramentas automáticas de apoio à tomada de decisão.

Artificial Neural Networks for Meat Quality Prediction

Abstract

The recent creation of lamb meat products with Protected Designation of Origin or Geographic Identification is a stimulus for the development of quality products, whose features should correspond to the consumers' expectations. Indeed, nowadays consumers are giving more priority to quality rather than quantity. They even are willing to pay premium prices for top products. In an increasingly global and competitive market, quality is becoming a crucial economic factor for the meat industry. Among several characteristics, *tenderness* is considered the most important feature that affects the meat's taste.

On the other hand, in the last decades there has been an intense investigation related to the extraction of high level knowledge from raw data, in an iterative process known as *Knowledge Discovery from Databases*, which involves several stages. In particular, the *Data Mining* is an important step of this process, where learning algorithms (e.g. *Artificial Neural Networks*) are applied in search of useful patterns.

In this work, a *Neural Network Ensemble*, based on a feature selection by a *Sensitivity Analysis* procedure, is proposed to predict the lamb meat *tenderness*, defined in terms of two regression tasks: a *instrumental measurement* and a *sensorial panel*. This strategy will be tested on animal data from the Trás-os-Montes region of Portugal. In both tasks, the proposed approach presented better results than other *Neural Networks*, as well as the classical *Multiple Regression*. Furthermore, the neural model is faster and less expensive than the conventional animal science methods, opening room for the development of automatic tools to support decision making.

Conteúdo

Índice de Figuras

Índice de Tabelas

Capítulo 1

Introdução

1.1 – Motivação

Nas últimas décadas, diversas organizações têm investido na construção de bases de dados, sendo que a maioria delas são utilizadas exclusivamente para o armazenamento, actualização e consulta de dados. Porém, todos estes dados podem trazer benefícios e consequentes vantagens competitivas se forem sabiamente explorados. Actualmente, já é possível efectuar análises de dados de forma a auxiliar o processo de tomada de decisão das organizações, tornando-as mais eficientes, produtivas e competitivas. O processo de transformação dos dados em bruto, em conhecimento de alto nível, é designado por *Descoberta de Conhecimento em Bases de Dados (DCBD)*[1] [Fayyad e Piattetsky-Shapiro, 1996]. Trata-se de um processo iterativo que compreende todo um conjunto de fases, sendo o *Data Mining (DM)* uma das etapas cruciais, onde se aplicam algoritmos de aprendizagem com vista à procura de padrões úteis.

Diversos modelos de extracção de conhecimento, ou de aprendizagem, têm sido desenvolvidos, utilizando paradigmas oriundos de diferentes áreas do saber, incluindo, entre outros, a *Estatística*, a *Inteligência Artificial* e a *Aprendizagem Automática*[2] [Mitchell, 1997]. Assim, existem diversos algoritmos de aprendizagem e cada um deles possui vantagens e desvantagens, podendo uns apresentar melhor desempenho do que outros numa determinada amostra de dados. Não existe uma análise matemática que possa determinar se um dado algoritmo terá um desempenho melhor que outro. Para tal, são necessários estudos experimentais.

Por outro lado, nas últimas décadas, o mercado da carne, a exemplo da economia mundial, tem sofrido diversas transformações, alternando os níveis de procura por parte dos seus consumidores. A ênfase dada à selecção dos melhores produtos inclui, até ao

[1] Do inglês *Knowledge Discovery from Databases (KDD)*.

[2] Tradução adoptada para o termo *Machine Learning*. Outros autores preferem o termo *Aprendizagem Máquina*.

momento, basicamente avaliações voltadas apenas para algumas características. A intensificação dos sistemas de produção, a demanda por eficiência, a tendência dos mercados, o desejo por oferecer produtos de maior qualidade e o lucro, tornaram-se objectivos primordiais para a sobrevivência da actividade, daí a procura de métodos que permitam optimizar todo este processo.

Em especial, a recente criação de produtos cárnicos de origem ovina, com *Denominação de Origem* ou *Identificação Geográfica Protegida*, é um incentivo à produção de produtos de qualidade cujas características devem corresponder às expectativas dos consumidores. Estes, por sua vez, são cada vez mais exigentes no que diz respeito à qualidade. Daí a importância tecnológica e económica da avaliação da qualidade da carne, visto que os resultados obtidos podem condicionar o êxito ou o fracasso dos avanços e inovações que se produzem na tecnologia dos alimentos.

A qualidade da carne é condicionada por diversos factores, tais como a *tenrura*, *suculência*, a *aparência*, o *flavour* ou o *aroma*. De entre estes, a *tenrura* é considerado o atributo mais importante na influência da percepção alimentar e os consumidores estão mesmo dispostos a pagar preços de topo por carne mais tenra [Huffman *et al.*, 1997]. Por outro lado, o método ideal para medir a *tenrura* deve ser preciso, rápido, automático e não invasivo. No passado, foram propostas duas grandes abordagens [Arvanitoyannis e Houwelingen-Koukaliaroglou, 2003]: análises instrumentais e sensoriais. As primeiras são baseadas em testes objectivos, sendo o *instron*, equipado com uma célula *Warner-Bratzler (WBS)*, o instrumento mais comummente usado. Por sua vez, os métodos sensoriais são baseados em informação subjectiva, usualmente dados por um painel de provadores. Ambas estas soluções requerem trabalho de laboratório, sendo caras, exigentes em termos de tempo e invasivas.

De entre os critérios de modelação automática utilizados para a previsão da qualidade da carne, poder-se-ão citar alguns métodos estatísticos. Incluem-se aqui, os modelos de regressão linear múltipla, que irão falhar claramente se existir uma relação não linear nos dados. Por conseguinte, uma alternativa a considerar passa pelo uso de *Redes Neuronais Artificiais*. Estes são modelos conexionistas, inspirados no comportamento do cérebro humano, conseguindo adquirir conhecimento a partir de exemplos e reunindo um conjunto de vantagens, tais como: uma aprendizagem não linear e tolerância ao ruído, mesmo quando são utilizados um elevado número de atributos [Haykin, 1999]. Convém referir ainda que a construção de Sistemas de Apoio à Decisão baseados nesta

tecnologia não tem como missão a substituição dos especialistas, mas sim facilitar o seu trabalho, realizando-o de uma forma mais eficiente e eficaz.

As *Redes Neuronais* estão a adquirir uma atenção crescente por parte das comunidades do *Data Mining* e da *Aprendizagem Automática*, devido ao seu potencial em termos de conhecimento preditivo [Mitra *et al.*, 2002]. Apesar de serem bem conhecidas em diversos ramos de actividade (*e.g.* economia ou engenharia), somente nesta última década é que as *Redes Neuronais Artificiais* passaram a ser utilizadas com maior frequência no domínio da *Ciência Animal*. No passado, diversos estudos revelaram resultados promissores na avaliação da qualidade da carne (*e.g.* porco, aves, gado e salchichas) [Arvanitoyannis e Houwelingen-Koukaliaroglou, 2003]. Contudo, a literatura relacionada com a previsão da *tenrura* da carne é escassa, sendo principalmente orientada para a carne de bovino. Este é um factor de forte motivação para o desenvolvimento deste trabalho, uma vez que, dentro do nosso conhecimento, este é o primeiro estudo onde são aplicadas *Redes Neuronais* para a previsão da *tenrura* da carne de cordeiro.

A base de dados, para a qual se irá efectuar o processo de *DCBD*, foi elaborada tendo por base amostras recolhidas pelos investigadores no âmbito do projecto Agro 246 da Escola Superior Agrária do Instituto Politécnico de Bragança. Esta contém informações sobre a *tenrura* da carne de cordeiro. Os dados disponíveis incluem informações da carcaça (*e.g.* sexo, peso e estado de engorda) e da carne (*e.g.* pH e cor), que serão utilizadas como variáveis de entrada dos modelos de aprendizagem. Dado que os custos associados à predição da *tenrura* da carne pelos métodos tradicionais são elevados, foi proposta a utilização das *Redes Neuronais*, com vista a obter-se não só uma minimização de custos, mas também resultados num curto espaço de tempo, com um nível de confiança bastante aceitável.

1.2 – Objectivos

Este trabalho tem por finalidade apresentar uma abordagem da *DCBD*, aplicando *Redes Neuronais Artificiais* para a previsão da *tenrura* da carne de cordeiro. Como metodologia de apoio na condução deste estudo, será utilizado o *CRISP-DM*. Trata-se de uma metodologia de referência internacional, desenvolvida por um consórcio de investigadores e empresas de consultadoria em *Data Mining*.

A relevância deste trabalho de investigação está na oportunidade de se desenvolverem novos métodos de aferição de qualidade da carne com base em modelos gerados a partir de *Redes Neuronais Artificiais*. Assim, pretende-se obter um conhecimento sobre quais as reais capacidades e qual a melhor forma de utilizar as *Redes Neuronais* nesta aplicação. O objectivo final, caso se obtenham bons resultados, é que este estudo possa permitir o desenvolvimento de um *Sistema de Apoio à Decisão* para operar em ambiente real, permitindo uniformizar critérios, dar respostas num curto espaço de tempo e com custos reduzidos.

Assim, os objectivos principais desta dissertação visam o desenvolvimento de modelos de *Redes Neuronais* para:

- Estudar as relações entre as variáveis da base de dados, de forma a ser possível analisar a *tenrura* da carne, com base nas suas características comercialmente mais interessantes;
- Prever a *tenrura* da carne, para predição da sua qualidade, baseada em duas abordagens: medidas instrumentais e análise sensorial;
- Analisar os modelos de previsão obtidos e compará-los com as tradicionais técnicas de medição utilizadas na área.

1.3 – Organização

Neste capítulo foi apresentado o contexto geral do trabalho, bem como as motivações que levaram ao desenvolvimento da investigação e os objectivos que se pretendem atingir. O restante da dissertação encontra-se dividido em duas partes fundamentais. Na primeira (Capítulos 2, 3 e 4) é apresentada uma visão geral de toda a fundamentação teórica que serviu de apoio à realização do trabalho prático. Na segunda parte (Capítulos 4 e 5), é apresentado o trabalho prático que será utilizado para aplicação dos objectivos inicialmente propostos, sendo também analisados e discutidos os resultados obtidos.

À luz destas considerações, esta dissertação possui seis capítulos organizados da seguinte forma (excluindo o capítulo introdutório):

Capítulo 2 – Descoberta de Conhecimento em Bases de Dados e *Data Mining*

É apresentada uma revisão dos conceitos principais sobre o processo de *Descoberta de Conhecimento em Bases de Dados* e o *Data Mining*, explicando as várias etapas desse

processo, bem como, a metodologia utilizada e todos os procedimentos necessários para a sua implementação.

Capítulo 3 – Redes Neuronais Artificiais

Este capítulo é dedicado à fundamentação teórica das *Redes Neuronais Artificiais*, comentando os processos de modelação, topologias e algoritmos de aprendizagem, dando-se uma maior ênfase às redes do tipo *Perceptrão Multicamada* e ao algoritmo de *Retropropagação*.

Capítulo 4 – Qualidade da Carne de Cordeiro

Apresentação do problema objecto de estudo. São analisadas e comentadas as características essenciais para a qualidade da carcaça e da carne dos animais, assim como, as principais características da análise instrumental e sensorial da qualidade da carne. Finalmente, são apresentadas algumas abordagens clássicas e casos de estudo em que foram utilizadas *Redes Neuronais Artificiais*.

Capítulo 5 – Previsão da Tenrura da Carne de Cordeiro

Fornece-se uma descrição do trabalho prático realizado. É apresentada a ferramenta utilizada, sendo também descrito todo o processo de *Descoberta de Conhecimento em Bases de Dados*, que foi conduzido com base na metodologia *CRISP-DM*, compreendendo as fases: estudo do negócio, estudo dos dados, preparação dos dados, modelação e avaliação. Por último, são apresentados e analisados os resultados obtidos.

Capítulo 6 – Conclusão

É feita uma síntese do trabalho realizado, são discutidas as conclusões mais importantes relativas ao trabalho desenvolvido, assim como são apresentadas as contribuições e recomendações para trabalhos futuros.

Capítulo 2

Descoberta de Conhecimento em Bases de Dados e *Data Mining*

É apresentada uma revisão dos conceitos principais sobre o processo de Descoberta de Conhecimento em Bases de Dados e o Data Mining, explicando as várias etapas desse processo, bem como, a metodologia utilizada e todos os procedimentos necessários para a sua implementação.

2.1 – Introdução

A facilidade de se armazenar a informação em grandes volumes de dados está a induzir em muitos investigadores a seguinte questão: "*Agora que temos muitos dados, o que podemos fazer com eles?*". Novas metodologias para extracção de conhecimento estão a ser exploradas, de forma a responder e criar soluções para essa mesma questão [Frawley *et al.*, 1991]. Segundo Cabena e seus colaboradores [1997], a *Descoberta de Conhecimento em Bases de Dados (DCBD)* significa extrair de grandes bases de dados, sem nenhuma formulação prévia de hipóteses, informações genéricas, relevantes e previamente desconhecidas, que podem ser utilizadas para a tomada de decisões. Já Fayyad e seus colaboradores [1996] dizem que a *DCBD* pode ser definida como o processo não trivial de identificação de padrões úteis a partir de dados em bruto.

Desenvolveu-se assim um novo conceito com o objectivo de examinar grandes quantidades de dados, procurando encontrar relações em dados não explícitos, que possam ser usadas em modelos do mundo com capacidade de previsão e descrição. Todo este processo é iterativo, pelo que muitos passos necessitam de ser repetidos para que todo o processo possa ser refinado, na tentativa de obter uma solução apropriada para a análise dos dados do problema.

É importante reforçar a ideia que a *DCBD* é a actividade de extrair informações relevantes, que pode ser vista como um processo de descoberta de novas correlações,

padrões e tendências significativas por meio de uma análise minuciosa de um grande conjunto de dados, sendo vista como uma disciplina mais ampla e não devendo ser confundida com uma das suas fases, o *Data Mining*. Esta é uma das etapas em que o sistema pesquisa os dados e determina os padrões (esta fase será analisada com mais pormenor na Secção 2.4.4).

Para tornar este processo efectivo, é necessário examinar que tipo de características se espera que tenha um sistema de *DCBD* e que tipo de desafios se enfrenta no seu desenvolvimento [Chen *et al.*, 1996]:

1. **Manipulações de diferentes tipos de dados** – Sistemas de *DCBD* devem ser construídos para extrair conhecimentos sobre tipos específicos de dados, tais como, sistemas dedicados para extrair conhecimentos em bases de dados relacionais, bases de dados de transacção, bases de dados espaciais e bases de dados multimédia, etc;

2. **Eficiência e escalonamento dos algoritmos** – Os algoritmos de descoberta de conhecimento devem ser eficientes e possíveis de serem escalonados para grandes bases de dados;

3. **Utilidade, certeza e expressividade** – O conhecimento descoberto deve retratar precisamente os conteúdos da base de dados e ser usual para certas aplicações;

4. **Expressão de vários tipos de requisições e resultados** – Diferentes tipos de conhecimento podem ser descobertos a partir de uma grande quantidade de dados. Também se pode querer examinar os conhecimentos descobertos em diferentes visões e apresentá-los em diferentes formas, para facilitar a especificação de tarefas por pessoas não especializadas e o entendimento do conhecimento descoberto;

5. **Extracção interactiva de conhecimento** – A descoberta interactiva deve ser encorajada, permitindo que um utilizador possa visualizar de forma flexível os dados e resultados a níveis múltiplos de abstracção;

6. **Extracção de informação de diferentes fontes de dados** – Extrair conhecimento de diferentes fontes de dados formatados e não formatados com diversas semânticas propõe novos desafios;

7. **Protecção de privacidade e segurança dos dados** – É importante estudar quando a *DCBD* pode levar a uma invasão de privacidade e que medidas de segurança podem ser desenvolvidas para prevenir a divulgação de informação.

Poderá assim concluir-se que as bases de dados fornecem todo o apoio necessário ao armazenamento e utilização de grandes quantidades de informação. Por outro lado, a sua correcta compreensão e análise, obriga à utilização de teorias, métodos e algoritmos provenientes de diferentes áreas[3] e de ferramentas específicas, que facilitem e executem o processo de análise de dados e descoberta do conhecimento de uma forma automática [Fayyad *et al.*, 1996].

2.2 – Relacionamento da *DCBD* com Outros Campos

Actualmente, os analistas de negócios precisam de fazer uso de ferramentas capazes de responder a perguntas mais complexas como: "*Qual o produto que venderia mais, analisando os dados dos últimos 10 anos?*". Este tipo de informação pode ser crucial para a sobrevivência de uma empresa nos tempos actuais. Desta forma, novas ferramentas de extracção de conhecimento devem ser, cada vez mais, usadas no processo da tomada de decisão.

A implementação do *Data Warehousing* [Gardner, 1998] é considerada como um dos primeiros passos para se tornar factível a análise em grandes quantidades de dados no apoio ao processo de decisão. Essa abordagem tem como objectivo proceder ao armazenamento dos dados, o *Data Warehouse*, permitindo que contenham dados limpos, agregados e consolidados que possam ser analisados por ferramentas *OLAP*[4] [StatSoft, 2005].

As ferramentas utilizadas para a análise de *Data Warehouse*, normalmente, são orientadas às consultas, ou seja, são dirigidas pelos utilizadores, os quais possuem hipóteses que gostariam de comprovar, ou limitam-se a consultas aleatórias. Essas consultas podem impedir a descoberta de padrões escondidos nos dados, uma vez que o utilizador não terá condições de imaginar todas as possíveis relações e associações existentes num grande volume de dados. Devido a estes condicionalismos, é necessário recorrer à utilização de técnicas de análise apoiadas por ferramentas específicas, que permitam a extracção automática (ou semi-automática) de novos conhecimentos de uma

[3] Entre outras, essas áreas poderão incluir a *Inteligência Artificial*, a *Aprendizagem Automática*, a *Estatística* e as *Bases de Dados*.

[4] *On-line Analytical Processing* – Armazenamento multidimensional de dados, em formato de cubo, que permite o rápido agregamento de dados e detalhamento de análises.

base de dados [Bradley *et al.*, 1998]. Assim, as técnicas utilizadas na *DCBD* não devem ser vistas como substitutas de outras formas de análise (*e.g*. *OLAP*), mas sim, como práticas para melhorar os resultados das explorações feitas com as ferramentas actualmente usadas [Fayyad *et al.*, 1996].

2.3 – Elementos de apoio à Análise de Dados

De seguida serão descritos os principais elementos/técnicas de apoio ao processo de *DCBD*, a saber: *Bases de Dados*, *Data Warehouse*, *Ferramentas de Visualização*, *Técnicas Estatísticas* e *Aprendizagem Automática*.

- **Bases de Dados** – Os dados a serem analisados deverão estar armazenados num *Sistema de Gestão de Bases de Dados (SGBD)*[5], sendo esse sistema a fonte dos dados. Porém, a importância do *SGBD* vai muito além de um sistema com dados, auxiliando na tarefa de selecção e preparação dos dados, uma vez que nessa tarefa são seleccionadas e criadas visões da base de dados em função da base operacional, determina-se o tamanho da amostra a ser utilizada e selecciona-se um conjunto de dados para ser utilizado pelo sistema;

- ***Data Warehouse*** – Segundo Inmon [1996] para se realizar o processo de *DCBD* não é necessário que se tenha implementado um *Data Warehouse*, porém, se uma organização qualquer resolver realizar um processo de extracção de conhecimento num determinado domínio da aplicação, deve ser destacado que esta organização deverá possuir um *Data Warehouse* com os seus dados, o que reduziria o esforço gasto no processamento dos mesmos [Han e Kamber, 2001];

- **Ferramentas de Visualização** – As técnicas e ferramentas para visualização de dados são "instrumentos" indispensáveis ao processo de *DCBD*. Estas podem ser usadas durante a execução das várias etapas do processo (Secção 2.4), melhorando a compreensão dos resultados obtidos e até a comunicação entre os utilizadores envolvidos no processo [Hand *et al.*, 2001];

[5] Sistema que armazena e manipula grandes volumes de informação, suporta acesso simultâneo por vários utilizadores, devendo esse acesso ser seguro.

- **Técnicas Estatísticas** – A estatística é a área da matemática que estuda a colecção, organização e interpretação de dados numéricos, especialmente a análise de características da população por inferências a partir de amostras. As técnicas estatísticas possuem uma elevada importância dentro do processo de *DCBD* e muitos dos métodos utilizados neste processo tiveram as suas origens dentro da estatística [Fayyad *et al.*, 1996; Glymour *et al.*, 1997].

- ***Aprendizagem Automática*** – Trata-se de uma das partes centrais da etapa de *Data Mining*, sendo muito utilizada na tarefa de extracção de padrões. Pode ser definida como uma sub-área da *Inteligência Artificial* [Rich e Knight, 1993], que pesquisa métodos computacionais relacionados com a aquisição de novos conhecimentos, novas habilidades e novas formas de organizar o conhecimento já existente. Esses métodos podem ser entendidos como sistemas de aprendizagem que tomam decisões baseadas em experiências acumuladas [Mitchell, 1997].

2.4 – Etapas do Processo de *DCBD*

No processo de *DCBD*, estão envolvidas várias etapas que vão desde a selecção das bases de dados sobre as quais será realizado o processamento, até à disponibilização do conhecimento descoberto (caso se verifique a descoberta de conhecimento). Num elevado nível de abstracção, poder-se-á afirmar que essas etapas fazem parte de três grandes grupos: *pré-processamento*, aplicação de um algoritmo de *Data Mining* e finalmente *pós-processamento* [Michalski e Kaufman, 1998].

Como já foi referido anteriormente, o processo de *DCBD* é iterativo (uma vez que pode existir retrocesso às etapas anteriores), cada ciclo de iteração envolve várias etapas sequenciais [Fayyad *et al.*, 1996], mas também é um processo interactivo, já que requer a participação do utilizador sempre que é necessária a tomada de decisão [Santos, 2001]. A Figura 2.1 mostra a sequência de etapas de um ciclo de processo de *DCBD*:

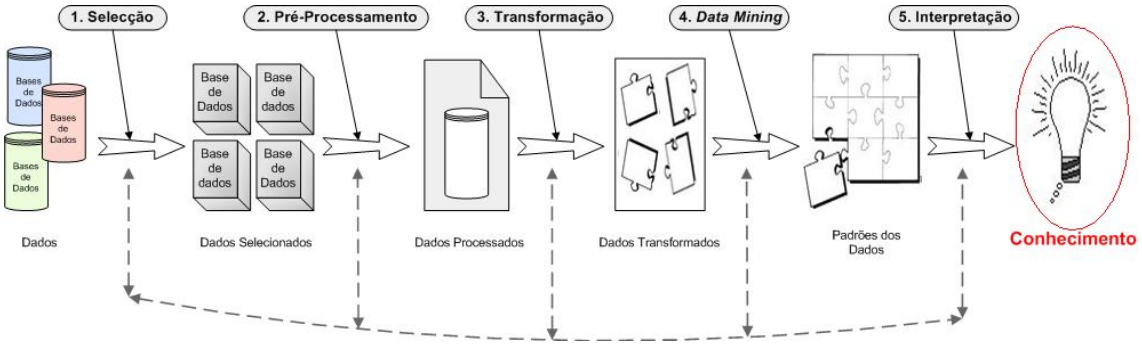


Figura 2.1 – Etapas do processo de *DCBD* (Adaptado de: [Fayyad *et al.*, 1996]).

Assim sendo, poderemos considerar os seguintes passos de forma a efectuar a extracção de conhecimento:

1. Selecção;
2. Pré-Processamento dos Dados;
3. Transformação dos Dados;
4. *Data Mining*;
5. Interpretação de Resultados.

Convém referir que as três primeiras etapas que antecedem o *Data Mining*, compõem-se o que se chama de preparação dos dados. A quarta etapa é o coração do processo de *DCBD*, a última etapa tem como finalidade a análise e assimilação dos resultados produzidos pela etapa que a antecede.

As etapas que antecedem o *Data Mining* requerem um grande dispêndio de tempo. Por exemplo Manilla [1994], entre outros, revela que estas etapas chegam a demorar até 80% de tempo total gasto no processamento de *DCBD*, devido em grande parte às já conhecidas dificuldades de integração em bases de dados heterogéneas. Esta quantidade de tempo gasta é tida como uma ingrata surpresa para os analistas, sendo uma fonte de frustração. Contudo, este esforço poderia ser reduzido, de uma forma substancial, através da construção de um *Data Warehouse* [Rocha, 1999]. Quanto às restantes etapas, o *Data Mining* e a *Interpretação de Resultados* requerem ambas cerca de 10% do tempo total dispendido.

É importante referir que antes de executar cada uma das etapas, deverá ser **Definido e Compreendido o Domínio da Aplicação**, considerando aspectos tais como, os

objectivos da aplicação, as fontes de dados e os requisitos necessários para que seja possível extrair e incorporar o conhecimento conseguido [Mannila, 1996].

Outro aspecto igualmente importante reside na **Consolidação do Conhecimento**. Isto é realizado incorporando-se o conhecimento obtido num sistema de aplicação ou, simplesmente, documentando-o e relatando às partes interessadas. Tal permite que se possam resolver os conflitos potenciais com os conhecimentos anteriores e previamente extraídos ou acreditados.

Existem diversos autores, tais como Rocha [1999], que consideram a **Definição e Compreensão do Domínio da Aplicação** e a **Consolidação do Conhecimento** como duas etapas integrantes do processo de *DCBD*.

Por se tratar de um processo interactivo, as pessoas envolvidas na *DCBD* devem possuir um canal de comunicação que viabilize uma boa troca de informações em qualquer altura do processo. Devido a esta necessidade de informação os utilizadores do processo podem dividir-se em três grupos de diferentes especialidades, necessidades e tipo de participação dentro de todo o processo [Brachman e Arnand, 1996]:

- **Especialistas do Domínio** – Necessitam de visões de alto nível sobre a informação e passam muito menos tempo a utilizar computadores do que os outros grupos. Eventualmente, podem precisar de informação que não está presente nos seus *EIS*[6];

- **Analistas** – Sabem interpretar a informação e usam computadores. Devem possuir amplo conhecimento das etapas que compõem esse processo. Normalmente percebem um pouco de estatística e *SQL*[7];

- **Utilizadores Finais** – Sabem usar computadores, mas não programam. É este grupo de utilizadores que normalmente usa o conhecimento extraído do processo de *DCBD*.

[6] *Executive Information Systems* – Sistemas de Informação de Executivos.

[7] *Structured Query Language*.

2.4.1 – Selecção

Nesta etapa, e consoante os objectivos do processo, são identificadas as bases de dados relevantes para análise do processo, normalmente usadas pelos sistemas de informação das empresas. A selecção dos dados tem como principal objectivo limitar o espaço de pesquisa, eliminando atributos que não têm qualquer interesse no processo de descoberta de conhecimento [Santos, 2001]. Estas bases de dados nem sempre estão de acordo com as exigências definidas pelo domínio apresentado.

Juntar todas as informações numa base de dados centralizada nem sempre é uma tarefa fácil, uma vez que os dados podem ser oriundos de diversos locais (como fontes internas ou externas), o que pode ocasionar uma variação na qualidade dos dados. Algumas bases de dados são actualizadas diariamente, enquanto outras contêm informações datadas ao longo de vários anos. Diferentes bases de dados podem usar distintas técnicas para identificar um certo atributo (*e.g.* uma através de *string* e outra por números), o que deixa bem claro que a selecção de dados não é uma tarefa trivial.

2.4.2 – Pré-Processamento dos Dados

Em geral as bases de dados actuais não estão preparadas para a obtenção de padrões. Assim, deve-se remover as inconsistências e integrá-los, visando adequá-los aos algoritmos de todo o processo de *DCBD*, sendo necessário para isso resolver os diversos tipos de conflitos típicos da integração de dados. As operações mais usuais para a resolução deste tipo de conflitos são a integração de dados heterogéneos, tratamento de ausência de dados, eliminação de dados incompletos, repetição de registos, tratamentos de ruídos, etc. O resultado desta etapa é, em geral, um arquivo de dados distinto das bases de dados originais.

Um outro problema que deve ser resolvido está relacionado com o tamanho do conjunto de dados, pois uma grande quantidade de dados reunida, às vezes pode impossibilitar a realização do processo de *DCBD*, uma vez que algoritmos usados no *Data Mining*, apenas permitem o uso de um número limitado de registos. Por isso, devem utilizar-se algumas técnicas de amostragem de forma a reduzir o tamanho do conjunto de dados, obtendo-se um subconjunto que seja relevante e representativo.

2.4.3 – Transformação dos Dados

Esta etapa reúne um conjunto diverso de transformações. Dependendo do objectivo de *DM*, pode efectuar-se uma procura das entradas mais relevantes e consequentemente reduzir-se o número total de atributos a considerar. Para este efeito podem ser considerados métodos de extracção de características (*e.g.* Análise de Sensibilidade) ou métodos de compressão de características (*e.g.* Análise de Factores Principais) [Fayyad *et al.*, 1996]. Por outro lado, pode pretender-se alterar o modo de representação de um dado atributo, de forma a facilitar o processo de aprendizagem [Pyle, 1999]. Por exemplo, um atributo numérico contendo o número total de filhos pode ser transformado num atributo discreto, com os níveis "nenhum", "um ou dois" e "mais do que dois filhos".

O objectivo fundamental desta etapa é transformar os dados pré-processados na etapa anterior, de modo a torná-los compatíveis com as entradas dos diversos algoritmos da etapa seguinte, o *Data Mining*.

2.4.4 – *Data Mining*

Começa-se por ressalvar que vários autores usam o termo *Data Mining* (*DM*) para referenciar todo o processo de *DCBD*, porém, nesta dissertação os termos *DCBD* e *DM* são usados com significados distintos. *DCBD* refere-se ao processo como um todo, enquanto que o *DM* (como será exposto nesta secção) é tratado como um componente dentro deste processo.

O conceito de *DM* (coração da *DCBD*), explanado na Figura 2.2, consiste na extracção de informação, previamente desconhecida e de máxima abrangência, a partir de uma bases de dados, para usá-la na tomada de decisão [Cabena *et al.*, 1997].

Figura 2.2 – Esquema do funcionamento do *DM*.

São chamadas de *DM* todas a técnicas que permitem extrair conhecimento de um conjunto de dados que, de outra maneira, permaneceria escondida. Basicamente a etapa

de *DM* tem por função a análise de um conjunto de dados e uso de técnicas na procura de padrões, identificando as regras subjacentes nos dados. A ideia consiste em descobrir "ouro" em lugares inesperados, pois os algoritmos de *DM* permitem encontrar padrões por vezes nada óbvios. Embora os algoritmos sejam capazes de descobrir esses padrões ainda não existe uma solução eficaz para determinar padrões valiosos. Por esta razão ainda é necessário uma interacção muito forte com analistas humanos [Fayyad *et al.*, 1996], que são em última instância os principais responsáveis pela determinação do valor dos padrões encontrados.

2.4.4.1 – Tarefas do *DM*

Os algoritmos de *DM* podem ser executados com vista a realizar uma ou mais das seguintes tarefas [Fayyad *et al.*, 1996]:

1. **Classificação** – A classificação é uma tarefa de previsão. Analisa o conjunto de dados de treino, cada um na forma de um vector de entradas e uma saída discreta (a classe pretendida). Um exemplo é o diagnóstico de doenças baseado nos sintomas de pacientes. Existem vários métodos de classificação desenvolvidos nos campos da *Aprendizagem Automática* e *Estatística*, entre outros. A classificação (Figura 2.3) pode ser usada nos segmentos de mercado, modelação de negócios e análises de crédito [Braga e Carvalho, 2000];

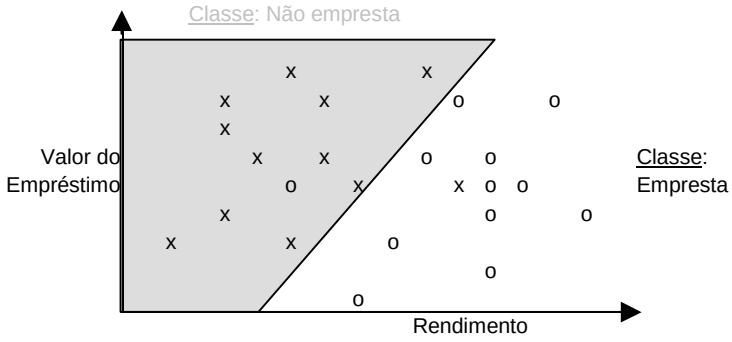


Figura 2.3 – Exemplo da tarefa de classificação [Fayyad *et al.*, 1996].

2. **Regressão** – Tarefa de previsão que pretende encontrar uma função de mapeamento para um registo numérico, de modo a possibilitar a previsão dos valores do mesmo. Existem diversas aplicações de regressão (Figura 2.4), como por exemplo a previsão de um peso fetal com base em medições efectuadas a partir de uma ecografia ou mesmo o problema objecto de estudo nesta tese;

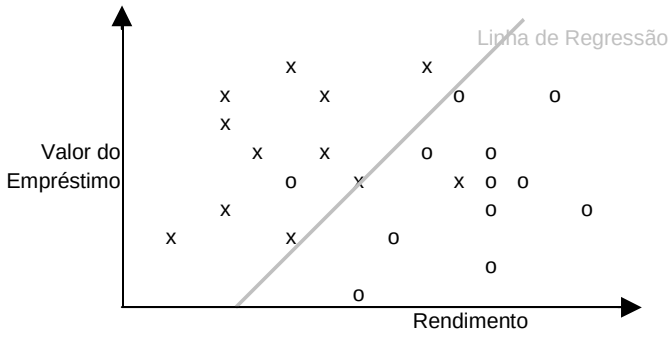


Figura 2.4 – Exemplo de regressão linear [Fayyad *et al.*, 1996]

3. Segmentação – Tarefa de descrição cujo objectivo é procurar identificar um conjunto de categorias finitas ou agrupamentos naturais que possam descrever um certo comportamento nos dados. Esta tarefa não é supervisionada, uma vez que não existe qualquer influência no conjunto de dados por parte do utilizador. Exemplos de aplicações da segmentação (Figura 2.5) no contexto da descoberta do conhecimento, são a descoberta de grupos homogéneos de consumidores de uma base de dados de um supermercado;

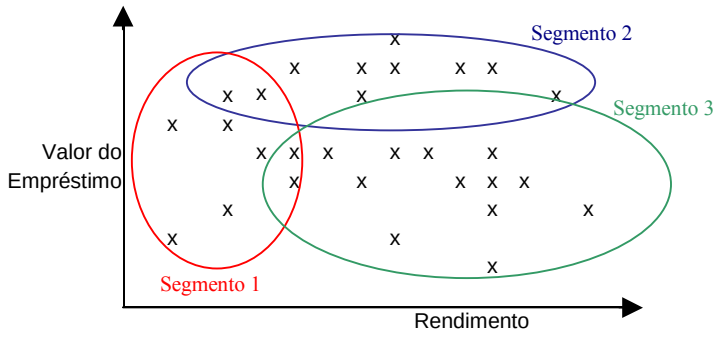


Figura 2.5 – Exemplo da tarefa de segmentação [Fayyad *et al.*, 1996].

4. Sumariação – Tarefa que visa obter uma descrição completa de um conjunto de dados que os distinga de outros, de forma a conseguir fazer análises exploratórias de dados e geração automática de relatórios. A tarefa de sumariação pode ser usada para comparar as vendas de uma empresa sedeada em países diferentes;

5. Associação – Responsável pela descoberta de relações ou correlações entre os atributos de um conjunto de dados [Chen *et al.*, 1996]. São muitas vezes expressas

sobre a forma de regras. Por exemplo, "64% dos compradores de embalagens de leite também compram pão".

2.4.4.2 – Algoritmos Usados no *DM*

A escolha do algoritmo de *DM* não é uma tarefa trivial. De facto, o analista deve ter em atenção as características gerais dos algoritmos de aprendizagem. Cada algoritmo apresenta as suas vantagens e desvantagens, distinguindo-se essencialmente pela sua **estrutura de representação** e **método de optimização** [Mitchell, 1997]. O primeiro factor determina a forma do modelo considerado, enquanto que o segundo define como se procura o melhor modelo de entre o espaço de procura de modelos possíveis.

Dada a importância do *DM*, nas últimas décadas foram propostos inúmeros algoritmos de aprendizagem (*e.g*. *Modelos Bayesianos*, *K-Vizinhos mais próximos* ou *Sistemas de Classificação*), cuja descrição se encontra fora do âmbito desta dissertação. Os métodos mais populares para tarefas de classificação incluem as *Regras de Classificação* e as *Árvores de Decisão* [Hand *et al.*, 2001]. No caso dos problemas de regressão, os modelos baseiam-se tipicamente em expressões matemáticas (*e.g. Regressão Múltipla*) [Mitchell, 1997]. Por sua vez, o uso de *Redes Neuronais Artificiais* têm vindo a generalizar-se, não só em tarefas de regressão mas também de classificação. Tal deve-se às suas capacidades de aprendizagem não linear, permitindo-lhes obter bons resultados onde outros métodos falham.

2.4.5 – Interpretação de Resultados

A descoberta de padrões nos dados não significa que o processo tenha sido finalizado, é necessário que se interpretem os resultados e se possa julgar a veracidade do conhecimento descoberto. Extraídos os padrões, torna-se necessária uma avaliação do conhecimento obtido e, para isso, são utilizados, entre outros, os critérios de precisão, compreensão e interesse. Estes critérios auxiliam na análise dos padrões encontrados, podendo ajudar também na filtragem do que foi apreendido e, remoção dos padrões redundantes e irrelevantes [Fayyad *et al.*, 1996; StatSoft, 2005].

Deverá ser analisada a validade de todo o processo, ou seja, colocar-se-á a seguinte questão: "*O conhecimento gerado é relevante e aplicável?*", se a resposta não for satisfatória, então, provavelmente, deve-se retomar a qualquer uma das etapas anteriores

e tentar refazê-las ou melhorá-las [Fayyad *et al.*, 1996]. Esta iteração pode ocorrer até que se obtenham resultados aceitáveis ou concluir-se que não é possível extrair conhecimento a partir dos dados.

2.5 – Metodologias de Desenvolvimento de Soluções de *DM*

É de consenso comum que é relativamente simples iniciar um projecto de *DM*, a dificuldade está em finalizá-lo de acordo com as expectativas. Em especial, as promessas geradas no início de um projecto podem ser mal interpretadas, dando origem a falsas expectativas que mais tarde se transformam em desilusões.

De facto, a execução de projectos de *DM* é uma actividade relativamente recente, pois a primeira conferência do *ACM Special Interested Group on Knowledge Discovery in Data and Data Mining* ocorreu em 1995. As entidades pioneiras nesta área foram seguindo o seu próprio caminho, definindo as suas próprias estratégias e métodos. Dado este estado caótico, ao fim de algum tempo surgiu a necessidade de uma metodologia de referência para o desenvolvimento de projectos de *DCBD/DM*. Assim, diversas metodologias foram propostas, das quais somente duas estão hoje em dia disseminadas em grande escala: a *SEMMA – Sample, Explorer, Modify, Assessement*, e o *CRISP-DM*[8]*– CRoss-Industry Standard Process for Data Mining*. Esta última metodologia, que é mais completa, foi utilizada na execução deste trabalho, sendo por conseguinte objecto de análise.

O *CRISP-DM* é uma metodologia desenvolvida por um consórcio de pesquisadores e empresas de consultadoria de *DM*, a partir de experiências de quatro empresas pioneiras no sector: a *DaimlerChrysler*[9], que aplica análises de *DM* nos seus negócios desde 1996; a *NCR*[10], que proporciona soluções de *Data Warehouse*; a *SPSS*[11], que disponibiliza soluções baseadas no processo de *DM* desde 1990; e *OHRA*[12], um grupo

[8] http://www.crisp-dm.org.

[9] http://www.daimlerchrysler.com/dccom.

[10] http://www.ncr.com.

[11] http://www.spss.com.

[12] http://www.ohra.nl/nl.

bancário Holandês [Chapman *et al.*, 2000]. Esta metodologia surgiu em 1996 e até 1999 foram realizados muitos progressos, chegamdose ao *CRISP-DM 1.0*. Esta versão foi desenvolvida com base na experiência de aplicações de *DM* e como tal é deveras orientada para os aspectos práticos.

Para a execução de projectos de *DM* e ao correcto ajustamento de recursos para os mesmos, foi desenvolvida esta metodologia padrão não proprietária, que visa, identificar as diferentes fases na implementação de um projecto. Na realidade, o *CRISP-DM* consiste num conjunto de fases e processos padrões para desenvolver projectos de *DCBD/DM*, independentemente da área de negócio e das ferramentas utilizadas, de uma forma estruturada e metódica [CRISP-DM, 1999]. Como é de fácil dedução, esta metodologia tem como objectivo fazer com que grandes projectos de *DM* se tornem mais rápidos, baratos e simples de gerir. Contudo, até projectos de pequena envergadura podem beneficiar com a aplicação desta metodologia [Han e Kamber, 2001].

A sua utilização consiste num modelo hierárquico de processos, representados por um conjunto de tarefas com quatro níveis de abstracção: **Fases**, **Tarefas Genéricas**, **Tarefas Especializadas** e **Instâncias de Processos** [Chapman *et al.*, 2000]. As **Tarefas Genéricas** são apresentadas de forma suficientemente geral para cobrir todas as situações de *DM*. Por sua vez, as **Tarefas Especializadas**, descrevem as acções do nível genérico que devem ser executadas em certas situações específicas. Finalmente, as **Instâncias de Processos** registam as acções, decisões e resultados de um determinado projecto de *DM*.

A implementação de um processo de *DM* consiste no desenvolvimento de várias **Fases** ou etapas, cuja sequência não é rígida, sempre que necessário podem acontecer retornos e avanços entre as diversas fases. Esta interactividade está fortemente dependente dos resultados disponibilizados pela fase anterior. Assim, esta metodologia é descrita como um modelo hierárquico, iterativo e interactivo, com um ciclo de vida que se desenvolve em seis fases principais: **Estudo do Negócio**, **Estudo dos Dados**, **Preparação dos Dados**, **Modelação**, **Avaliação** e **Implementação** (Figura 2.6). De seguida, será retratada cada uma destas fases.

- **Estudo do Negócio** – Visa o entendimento dos objectivos e requisitos do projecto, do ponto de vista do negócio. Baseado no conhecimento adquirido, o

problema de *DM* é definido e um plano preliminar é projectado para alcançar os objectivos;

- **Estudo dos Dados** – Esta fase inicia-se com um conjunto de dados, seguidamente são processadas actividades que visam: buscar familiaridade nos dados, identificar problemas de qualidade nos dados, descobrir discernimentos nos dados ou detectar subconjuntos interessantes para formar hipóteses da informação escondida;

- **Preparação dos Dados** – Engloba as actividades de construção de um conjunto de dados para posterior aplicação do modelo. Geralmente, esta etapa é refeita múltiplas vezes e sem nenhuma ordem predefinida. São realizadas tarefas de selecção, transformação e limpeza no conjunto de dados para uso dos algoritmos de *DM*;

- **Modelação** – Selecção dos algoritmos a serem usados e efectivo processamento do modelo. Alguns algoritmos necessitam de dados em formatos específicos, o que provoca vários retornos à fase anterior;

- **Avaliação** – Antes de avançar para a fase seguinte, é importante rever os passos dados para a construção do modelo, no sentido de garantir que ele cumpre os objectivos do negócio. Um objectivo chave é determinar se existe alguma informação importante de negócio que não foi considerada. Após esta análise, é decidido se se implementam os resultados do processo de *DM*;

- **Implementação** – Consiste na utilização dos modelos gerados, ou seja, um conjunto de acções que conduzam à utilização dos resultados do processo de *DM* no negócio, tendo em vista as avaliações dos resultados, gerando uma estratégia de divulgação. O resultado desta fase é um relatório final que procura explicar os resultados obtidos e as experiências.

Como já foi referido, a sequência destas fases não é rigorosa, podendo existir avanços e retornos. As setas da Figura 2.6 indicam as dependências mais importantes e frequentes entre as fases. O círculo externo simboliza a natureza cíclica do *DM*. Os processos do *DM* subsequentes beneficiarão das experiências anteriores. Resumindo, um projecto de *DM* requer a necessidade de conhecimento do negócio, o entendimento dos dados e da ferramenta escolhida, bem como do algoritmo implementado pela ferramenta.

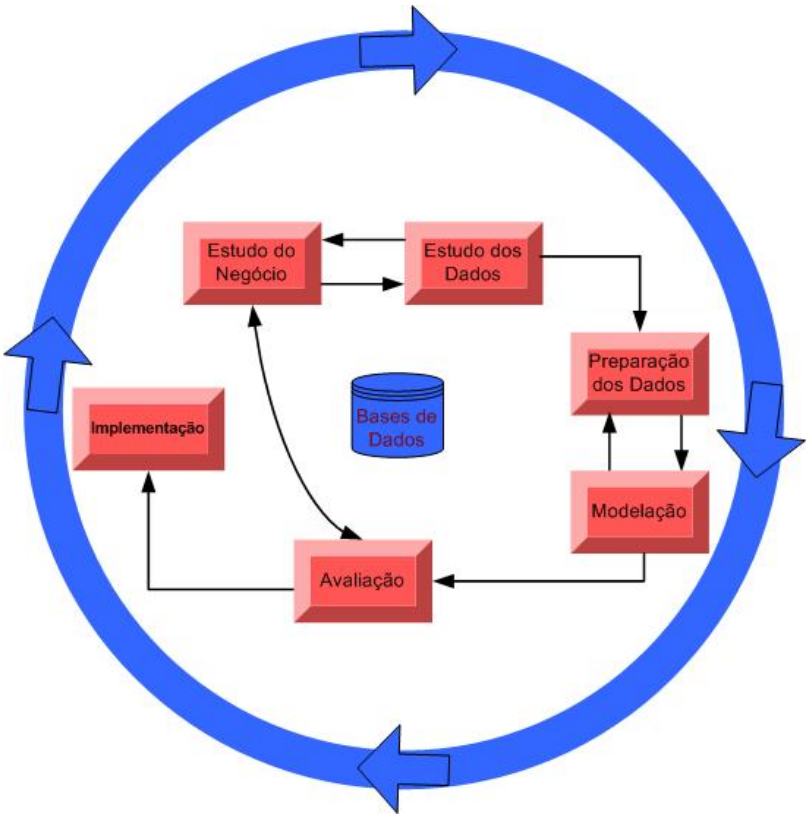


Figura 2.6 – Ciclo de vida da metodologia *CRISP-DM* [Chapman *et al.*, 2000].

2.6– Discussão

A literatura sobre *DCBD* e *DM*, propõe diferentes abordagens na definição destes termos. A primeira pode-se definir como todo o processo, enquanto a segunda é apenas uma das suas etapas, talvez a mais importante, desse processo. Assim, a *DCBD* é uma área da ciência da computação que visa extrair padrões e modelos de dados. Os padrões podem ser extraídos realizando-se alguns passos como Selecção, Pré-Processamento e Transformação dos dados. Após este processamento dos dados, são aplicados os algoritmos de *DM*, sendo estes escolhidos com base nos objectivos do processo de *DCBD*. Finalmente, procede-se a um estudo e avaliação dos resultados obtidos por todo o processo, isto é, o processo de Interpretação de Resultados.

Convém referir que o processo de *DCBD* é iterativo, pelo que a ocorrência de mudanças em qualquer uma das etapas afectará o sucesso de todo o processo. Desta forma, os resultados de uma determina etapa podem acarretar o recomeço de todo o processo [Fayyad *et al.*, 1996]. E por se tratar de um processo interactivo, as pessoas envolvidas

na sua realização devem possuir um canal de comunicação que viabilize uma boa troca de informações.

Por outro lado, uma vez que a execução de um processo de *DCBD* é deveras complexa, surge a necessidade de recorrer a metodologias *standard* para o desenvolvimento de projectos de *DM*, entre as quais se encontra o *CRISP-DM*. Esta metodologia consiste num conjunto de fases e processos padrões que auxiliam os seus utilizadores no desenvolvimento de projectos de *DCBD/DM*. As principais vantagens apontadas para o uso desta metodologia são a sua rapidez e baixo custo, que se torna fiável e de fácil controlo. Por conseguinte, o *CRISP-DM* foi escolhido para auxiliar a organização e a estruturação do trabalho descrito nesta dissertação.

Convém referir que a *DCBD* está em crescimento, pretendendo-se que consolide nos próximos anos como uma sólida tecnologia no campo das bases de dados. Devido à necessidade da existência de ferramentas automáticas para a extracção de conhecimento, inúmeras ferramentas têm vindo a ser disponibilizadas no mercado. De facto, existe um consenso sobre o facto do conhecimento ser de vital importância para o mundo dos negócios. As empresas que detêm e/ou fornecem o conhecimento têm grandes possibilidades de permanecer de forma competitiva no mercado. É neste contexto que o processo da *DCBD* tem um papel importante.

Capítulo 3

Redes Neuronais Artificiais

Este capítulo é dedicado à fundamentação teórica das Redes Neuronais Artificiais, comentando os processos de modelação, topologias e algoritmos de aprendizagem, dando-se uma maior ênfase às redes do tipo Perceptrão Multicamada e ao algoritmo de Retropropagação.

3.1 – Introdução

Uma *Rede Neuronal Artificial (RNA)*, terminologia genérica que abrange uma grande quantidade de arquitecturas e paradigmas, tem como objectivo compreender o funcionamento do cérebro humano e, de alguma forma, procurar reproduzi-lo. As *RNAs* são sistemas paralelos de processamento, compostos por unidades de processamento (neurónios), que calculam determinadas funções matemáticas (normalmente não lineares). Estas unidades, geralmente, estão interligadas por canais de transmissão com um determinado peso. As unidades fazem operações apenas sobre os dados locais, que são entradas recebidas pelas suas conexões. O comportamento "inteligente" de uma *RNA* advém das interacções entre as várias unidades de processamento da rede.

Uma definição bastante precisa para *RNAs*, é a encontrada em Haykin [2001]: "*uma rede neuronal é um processador paralelamente distribuído, constituído por unidades de processamento simples, que têm a propensão natural de armazenar conhecimento experimental e torná-lo disponível para uso*". Isto significa que uma *RNA* é um sistema computacional que tem a capacidade de aprender com o seu próprio uso, ou seja, de produzir saídas adequadas para entradas que não estejam presentes durante a sua aprendizagem, de forma a atingir um objectivo específico. Esta capacidade de generalização, possibilita a resolução de problemas computacionais complexos.

Por conseguinte, podem utilizar-se *RNAs* numa vasta gama de problemas, encontradas nas mais diversas áreas de aplicação: classificação, diagnóstico, análises de sinais e de

imagens, optimização e controlo. Do ponto de vista prático, as *RNAs*, têm como vantagem o facto de não necessitarem de especialistas para a tomada de decisões, dado que se baseiam unicamente nos exemplos que lhes são fornecidos.

Como conseguem reproduzir o comportamento de funções matematicamente desconhecidas à priori, a modelagem através de *RNAs* aparece como um potencial substituto dos modelos estatísticos convencionais, devido à fácil interpretação dos interfaces dos programas por parte do utilizador e a não necessidade de conhecimento prévio da relação entre as variáveis envolvidas. De facto, são diversas as aplicações onde as *RNAs* têm apresentado um desempenho superior aos métodos estatísticos [Subramanian *et al.*, 1993; Falas, 1995].

3.2 – Fundamentos Biológicos

Uma *RNA* é um modelo matemático cuja grande inspiração é o cérebro humano, uma estrutura de processamento altamente complexa, não linear e paralela. Ao contrário da arquitectura tradicional de computadores, o cérebro humano possui uma grande quantidade de processadores, conhecidos por unidades de processamento ou neurónios, que executam funções mais simples. Entre outras funções, o neurónio apresenta a capacidade de transmissão de impulsos nervosos a outros neurónios e células musculares.

O sistema de processamento de informação do cérebro humano possui a capacidade de organizar os neurónios de tal forma que o seu desempenho na execução de certas tarefas (*e.g.* reconhecimento de exemplos, percepção e controlo motor), seja realizado de uma forma muito mais rápida do que o mais potente computador, apesar da sua velocidade de processamento ser relativamente mais baixa. Esta rapidez na execução de certas tarefas deve-se à quantidade de neurónios existentes no cérebro humano, uma vez que, se prevê que a quantidade de neurónios existentes no cérebro esteja na casa dos biliões [Cottrell, 1985].

Como foi referido anteriormente, no corpo humano existe uma grande variedade de tipo de neurónios cujas funções não são totalmente conhecidas. Contudo, um neurónio natural é basicamente constituído por dendrites, sinapses, axónio e corpo celular (soma), como demonstrado na Figura 3.1.

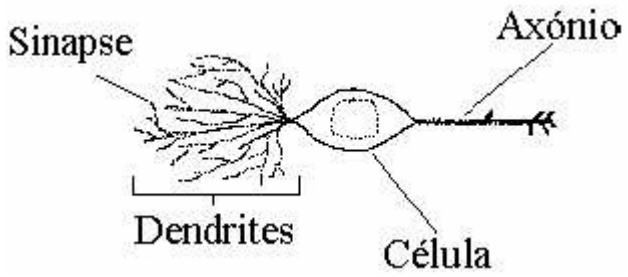


Figura 3.1 – Estrutura de um neurónio natural [Cortez, 2002].

À medida que novos exemplos são apresentados, determinadas ligações entre os neurónios são reforçadas, enquanto outras enfraquecidas. Este ajustamento, que se processa nas ligações entre os neurónios durante o processo de aprendizagem, é uma das características mais importantes das redes neuronais biológicas [Hopfield, 1982]. Contudo, e apesar da investigação contínua sobre o assunto, o conhecimento sobre o funcionamento das redes biológicas ainda não está totalmente adquirido.

Uma vez que, as redes neuronais biológicas têm a sua base de fundamentação na neurotransmissão ocorrida nas células nervosas, é importante que se conheça o funcionamento básico de um neurónio biológico. O **neurónio** tem uma estrutura simples, que permite três funções básicas: entrada, processamento e saída de sinais. As **dendrites** são conexões através das quais os sinais de entrada chegam aos neurónios. A **célula** ou corpo celular é o processador do neurónio, recebe sinais de entrada através das dendrites e soma esses sinais, se o valor resultante estiver acima de um certo limite, o neurónio excita-se e tende a propagar o estímulo, caso contrário, ele fica inibido. O **axónio** serve como canal de saída do neurónio, estando ligado às dendrites de outros neurónios através das **sinapses**. As sinapses não são ligações físicas, mas sim químicas e temporárias [Barreto, 2002].

3.3 – Factos Históricos

As primeiras pesquisas na área das *RNAs* surgiram com o trabalho de McCulloch & Pitts (1943), Hebb (1949) e Rosenblatt (1958). Estas publicações introduziram o primeiro modelo de *RNAs* simulando máquinas, o modelo básico de rede de auto-organização, e o modelo *perceptrão* de aprendizagem supervisionada [Costa, 2003].

Ainda nesta época, Shannon e McCarty [1956] publicaram um livro onde surgiram dois paradigmas da *Inteligência Artificial (IA)*, as vertentes simbólica[13] e conexionista[14].

Porém, em 1969 um estudo elaborado por Minsky e Papert no seu livro intitulado "*Perceptrons*", provou formalmente que uma rede formada por uma única camada, independente do algoritmo de aprendizagem, é capaz de resolver o problema de associação de padrões apenas quando os conjuntos são linearmente separáveis (*e.g.* a função lógica OU) [Minsky e Papert, 1969]. Estes resultados e observações foram devastadores para esta área, e consequência disto, a abordagem conexionista ficou relegada para um plano secundário durante toda a década de setenta e início da década de oitenta.

Somente com o ressurgimento do algoritmo de *Retropropagação*[15] para *Redes Multicamada*[16], proposto por Rumelhart e seus colaboradores, a comunidade científica voltou a interessar-se pelas *RNAs* como ferramenta para reconhecimento de padrões [Rumelhart *et al.*, 1986]. Dois outros factores também foram responsáveis pela retomada do interessa em *RNAs*: o avanço da tecnologia e o facto da vertente simbólica não ter conseguido avanços significativos na resolução de alguns problemas simples para o ser humano.

3.4 – Vantagens e Desvantagens do uso de *RNAs*

As *RNAs* são capazes de aprender pela experiência, generalizar os casos anteriores e encontrar características essenciais a partir de entradas, etc. Isto faz com que ofereçam numerosas vantagens e em consequência disso são inúmeras as áreas onde estão a ser aplicadas [Subramanian *et al.*, 1993]. Assim, destacam-se as seguintes vantagens:

- **Aprendizagem Adaptativa** – Capacidade de aprender a realizar tarefas baseadas num treino ou numa experiência inicial. Esta é uma das características mais

[13] Inteligência Artificial Simbólica: o comportamento global é simulado por um conjunto de regras explícitas.

[14] Inteligência Artificial Conexionista: acredita-se que construindo uma máquina que imite o cérebro humano, ela apresentará conhecimento.

[15] Tradução do inglês *Backpropagation*.

[16] Conhecidas em inglês por *Multilayer Perceptrons (MLP)*.

atractivas das *RNAs*, isto é, aprendem a levar a cabo certas tarefas mediante um treino com exemplos ilustrativos;

- **Não Linearidade** – Capacidade de modelar funções não lineares. Dado que muitos problemas são de natureza não linear, esta é uma das razões principais para que as *RNAs* apresentem desempenhos superiores a outras técnicas, em termos de conhecimento predictivo;
- **Tolerância a Falhas**[17] – Enquanto os meios computacionais tradicionais perdem a sua funcionalidade quando sofrem um pequeno erro de memória, nas *RNAs*, se existir uma falha num número não muito elevado de neurónios, o comportamento da rede não se torna inoperante;
- **Operação em Tempo Real** – Uma das prioridades das várias áreas de aplicação, é a necessidade de realizar processos com dados de forma muito rápida. As *RNAs* são constituídas por um grande número de unidades de processamento trabalhando em paralelo, pelo que podem trabalhar em velocidades consideráveis em relação aos métodos tradicionais comuns;
- **Adaptabilidade às Tecnologias Existentes** – Uma rede pode ser treinada para desenvolver uma ou mais tarefas bem definidas. Com as ferramentas computacionais existentes, a rede pode ser facilmente treinada e verificada. Assim não se apresentam dificuldades para a inserção de *RNAs* em áreas específicas.

Existem, no entanto, situações em que as características do problema tornam desaconselhável a utilização de *RNAs*. Além disso, o campo de estudos onde se inserem é muito recente, e alguns problemas continuam ainda por resolver [Subramanian *et al.*, 1993; Ambrósio, 2002]. Assim, são apontadas algumas desvantagens no seu uso, tais como:

- **Treino Demorado** – O treino de uma rede, dependendo da aplicação e da quantidade de dados, pode ser demorado;
- **Resultados Desconcertantes** – As redes podem chegar a conclusões que contrariem as regras e teorias estabelecidas, bem como considerar dados

[17] Existem dois aspectos distintos a respeito da tolerância a falhas: *i)* as redes podem aprender a reconhecer padrões com ruído, distorcidos ou incompletos. Esta é uma tolerância a falhas de dados; *ii)* as redes podem continuar a realizar as suas funções (com algum grau de degradação), ainda que se destrua um dos seus nós. Esta é uma tolerância a falhas da *RNA*.

irrelevantes como básicos; esta situação só é resolúvel com a participação do profissional da área que está a ser objecto de estudo;

- **Obscuridade** – O conhecimento está armazenado de modo implícito, pelo que é difícil compreender o que realmente uma *RNA* aprendeu;
- **Exigem um Elevado Número de Exemplos** – A carência de dados poderá impossibilitar a obtenção de um bom desempenho pela *RNA*;
- **Preparação dos Dados** – Os dados de entrada necessitam de um tratamento prévio, devem ser normalizados e cuidadosamente seleccionados para que a rede aprenda correctamente.

Contudo, convém referir que muitas destas vantagens podem ser colmatadas. Por exemplo, existem algoritmos de treino mais rápidos que o tradicional de *retropropagação*. Também é possível eliminar dados irrelevantes pela análise de *outliers*. Para além disso, existe uma intensa investigação na extracção de conhecimento, na forma de regras, a partir de redes treinadas [Setiono, 2003].

Em diversas circunstâncias, incluindo problemas com amostras pequenas e funções mais complexas, as *RNAs* apresentam soluções bastantes satisfatórias. De facto, o desempenho das *RNAs* tem-se mostrado bastante superior aos métodos estatísticos usados para os mesmos fins [Falas, 1995; Kwon *et al.*, 1995].

3.5 – Aplicação de *RNAs* no *Data Mining*

As *RNAs* têm sido empregues em diversas tarefas de *DM*, tais como classificação, regressão e segmentação (Secção 2.4.4.1) [Berry e Linoff, 2000; Haykin, 2001]. Para mostra um exemplo de uma aplicação de *RNAs* em *DM*, será apresentada uma analogia com o processo de aprendizagem humano. Imagine-se que o objectivo é a análise de vinhos portugueses através da degustação (*e.g.* oriundos das regiões do Minho, Trás-os-Montes e Alentejo). Na primeira etapa, de aprendizagem, provam-se os vinhos um a um e tentam-se identifica-los através de um conjunto de indicadores (*e.g.* seu paladar ou perfume). De seguida, analisa-se a etiqueta para verificar se a identificação estava correcta. Se a identificação não foi correcta, deve tomar-se isso em linha de conta de forma a aprender alguma coisa com o erro. Volta-se a fazer a prova a fim de tentar assimilar as suas características, através de um processo iterativo, experimentam-se os

vinhos até que seja possível a identificação de todas as garrafas seleccionadas. Conseguida esta identificação está concluída a etapa de aprendizagem. O que acontece ao cérebro humano? A rede de neurónios do cérebro armazenou o conhecimento que permitiu distinguir o tipo de vinho. As *RNAs* aprendem exactamente da mesma forma, através do fornecimento de exemplos sobre o que se pretende que ela aprenda.

Terminada a etapa de aprendizagem passa-se à etapa de utilização do conhecimento adquirido. No exemplo que se está a retratar, deve-se ser capaz de distinguir os vinhos, mesmo que não sejam os mesmos produtores ou os mesmos anos. Pelo menos isto é o que se espera da *RNAs*, uma vez que passou pela fase de aprendizagem. A grande diferença entre o ser humano e a rede, é que o ser humano rapidamente atinge o seu limite, fadiga ou embriaguez. Mas uma *RNA* não se cansa e pode ler tantos exemplos, quantos a base de dados for capaz de fornecer.

Com o objectivo de ilustrar a diversidade de aplicações nas quais as *RNAs* têm sido utilizadas para a *DCBD*, apresentam-se alguns exemplos de áreas comerciais onde têm sido eficazmente utilizadas [Wong *et al.*, 2000; Haykin, 2001; StatSoft, 2005]:

- **Biologia** – Estudos com o objectivo de aprender mais sobre o cérebro e outros sistemas;
- **Indústria** – Análise concorrencial, identificação de trabalhadores para áreas específicas;
- **Meio Ambiente** – Análise de tendências e padrões, previsão do estado do tempo;
- **Economia** – Previsões de evolução dos preços, créditos, falsificações, análises financeiras e taxas de juros;
- **Geologia** – Caracterização de rochas e prospecção mineral;
- **Transportes** – Escalonamento de horários e rotas;
- **Manufactura** – Análise de qualidade, controlo de produção e sistemas robotizados;
- **Telecomunicações** – Compressão de dados e reconhecimento de voz;
- **Desporto** – Previsão de resultados;
- **Medicina** – Monitorização de cirurgias e diagnóstico de doenças;

- **Militares** – Classificação de sinais de radares, armas inteligentes, optimização de recursos escassos e detecção de alvos.

As redes supervisionadas e não supervisionadas (3.9.1) são dois tipos de redes utilizadas nas aplicações de *DM*. Para as não supervisionadas, as redes de Kohenen, são as mais conhecidas, sendo muito aplicadas em problemas que envolvam algum tipo de segmentação ou associação. Na aprendizagem supervisionada, as redes de *retropropagação* são as mais usadas em problemas de classificação e regressão [Bigus, 1991].

As *RNAs* são cada vez mais usadas em aplicações de *DM*, embora, existam técnicas mais simples. Este facto deve-se à eficiência obtida em termos predictivos, quando comparada com o desempenho de outros métodos (*e.g.* as árvores de decisão, devido essencialmente à sua capacidade de aprendizagem não linear [Setiono, 2003]. Por outro lado, existe hoje em dia, cada vez mais, uma panóplia de software, gratuito e comercial, que facilita o uso de *RNAs* em tarefas de *DM*. Assim, o sucesso do *DM* depende muito da experiência e sensibilidade do pesquisador, o qual terá que identificar qual a melhor técnica a ser utilizada, de acordo com o tipo de resposta procurada e com o modo em que se encontram os dados.

3.6 – Principais Componentes das *RNAs*

De acordo com Rumelhart e Clelland [1986] e Costa [2003], um modelo de uma *RNA* pode ser composto por três componentes. Seguidamente relatam-se de uma forma resumida, algumas das principais características desses componentes.

3.6.1 – Unidade de Processamento

A unidade de processamento ou neurónio artificial é o principal elemento das *RNAs*. Na Figura 3.2 é apresentado um neurónio como unidade que representa um limite a ser ultrapassado. Cada entrada possui um sinal x que é adicionado ($\sum$). Depois da soma, este sinal é processado através da função que representa o limite ou a função de activação $f()$, a qual produz um sinal de saída.

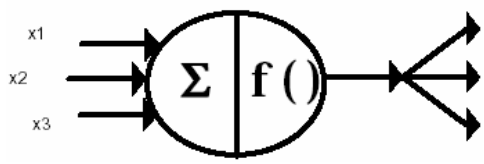


Figura 3.2 – Neurónio como unidade limite.

As unidades de processamento da rede podem ser identificadas pela letra ***x***, seguidas de um índice ***i*** que indica a posição que o neurónio ocupa na rede. Cada neurónio x_i da rede calcula um estado de activação, que é um valor numérico líquido de saída. O cálculo desta activação é realizado a partir dos sinais de saída dos demais neurónios conectados directamente a este neurónio, dos correspondentes pesos dessas conexões, assim como da função ***f0***.

3.6.2 – Função de Activação

A *Função de Activação* calcula o valor de saída do neurónio a partir do seu valor de activação ou integração (Σ). As funções de activação mais utilizadas são: linear (A), rampa (B), degrau (C) e *logística* (D) [18], que estão representadas na Figura 3.3.

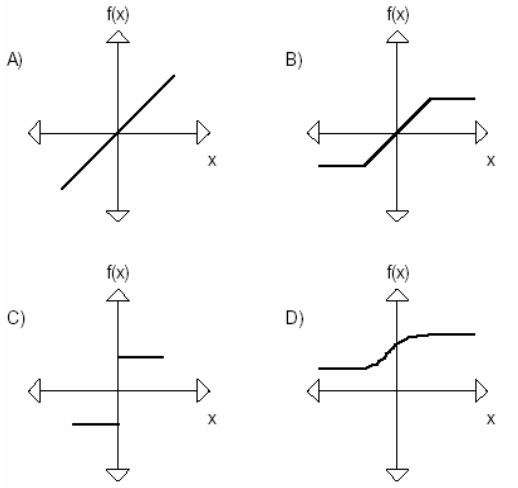


Figura 3.3 – Funções de activação.

De entre estas, a última função é a mais utilizada, visto que permite uma aprendizagem não linear, sendo definida pela Equação 1.

[18] Também conhecida por *Sigmoid*.

$$y = \frac{1}{1 + \exp(-x)} \quad \textbf{(1)}$$

3.6.3 – Ligações entre as Unidades de Processamento

As ligações entre os neurónios podem representar-se por uma matriz de pesos $\boldsymbol{w}$, onde um elemento $\boldsymbol{w_{ij}}$ corresponde à influência do neurónio $\boldsymbol{u_i}$ sobre o neurónio $\boldsymbol{u_j}$. Conexões com pesos positivos, chamadas de excitadoras, indicam o reforço na activação do neurónio $\boldsymbol{u_j}$. Conexões com pesos negativos, chamadas de inibitórias, indicam inibição na activação do neurónio $\boldsymbol{u_j}$. O objectivo principal da aprendizagem, ou treino, consiste na determinação do melhor conjunto de pesos capazes de responder se um modo adequado a um dado problema.

3.7 – Preparação dos Dados

Sobre um ponto de vista simplista, as *RNAs* podem ser encaradas como modelos que recebem estímulos de entrada e geram respostas, de acordo com o conhecimento adquirido durante um processo de aprendizagem.

As fases do processo de aprendizagem das *RNAs* são ilustradas na Figura 3.4. O ponto de partida é a obtenção dos dados que representam um determinado domínio. No pré-processamento os atributos numéricos devem ser normalizados dentro de uma escala de valores. Os atributos discretos precisam de ser codificados em valores numéricos. Os valores desconhecidos precisam de ser preenchidos utilizando-se métodos, como por exemplo, médias dos valores do atributo ou definição de novos valores.

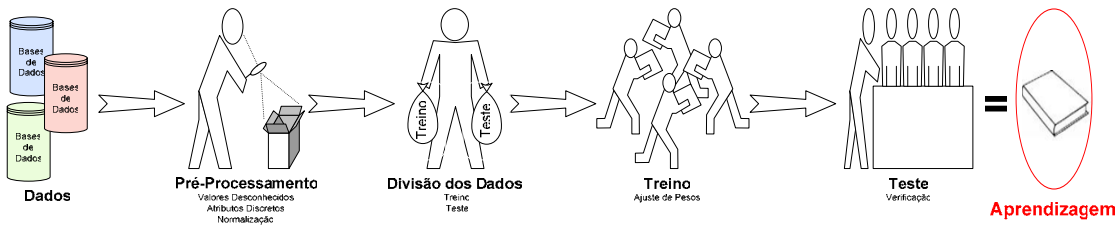


Figura 3.4 – Fases do processo de aprendizagem de uma *RNA*.

O conjunto de dados devidamente pré-processados deve ser então dividido em subconjuntos distintos de treino e teste. Seguidamente, a rede é treinada com os dados seleccionados para treino, de forma a ajustar os seus pesos que lhe permitirão apresentar

conhecimento. A fase de testes é a última etapa do processo e é necessária para verificar se a aprendizagem foi bem sucedida, através da utilização do conjunto de dados de teste, que é constituído por exemplos nunca vistos durante o treino.

Assim, tudo que foi dito na Secção 2.4 é aqui relevante. No entanto, existem algumas recomendações especiais para *RNAs*, essencialmente na fase de pré-processamento dos dados. Esse processamento pode ser simples ou complexo, dependendo dos dados a serem trabalhados e do algoritmo a ser utilizado. Em seguida é apresentado um conjunto de regras que deve ser seguido [Pyle, 1999]:

- **Valores Desconhecidos** – Devem ser tratados de alguma forma. Existem diversas possibilidades, tais como a substituição por um valor médio ou a eliminação de registos, caso sejam em número diminuto.
  - Média – Valores desconhecidos de um atributo podem ser substituídos pela média de todos os seus valores. Em atributos que representam valores ordenados, a média de *n* valores anteriores e posteriores pode ser utilizado para a definição do valor desconhecido;
  - Definição de Novos Valores – De acordo com cada problema, valores arbitrários podem ser definidos para o preenchimento de campos desconhecidos;
  - Eliminação dos Valores em Falta – Caso se trate de valores que sejam de difícil previsão.
- **Variáveis Discretas** – Os valores não numéricos devem ser codificados em valores numéricos, de acordo com o seguinte:
  - Atributos Binários – Devem ser codificados como os valores limite definidos pela função de activação (*e.g.* não → -1.0 e sim → 1.0);
  - Atributos Nominais – os atributos com *C* valores possíveis devem ser transformados com uma codificação *1-of-C*, que gera uma variável binária por cada classe (*e.g.* verde → -1.0 1.0 1.0, azul → -1.0 1.0 -1.0 e vermelho → -1.0 -1.0 1.0).
- **Normalização** – Os atributos de entrada devem ser normalizados, dado que os algoritmos de treino são sensíveis a intervalos e distribuições diferentes nas suas

entradas. Assim, os valores de cada entrada devem ser escalonados de forma a que a média dos seus valores no conjunto de casos de treino seja nula.

3.8 – Topologias de *RNAs*

Em geral, as topologias de *RNAs* podem ser agrupadas em duas classes: Não-**Recorrentes** e **Recorrentes**.

3.8.1 – *RNAs* Não-Recorrentes

As *RNAs* não-recorrentes são aquelas que as suas saídas não possuem realimentação para as suas entradas. A estrutura destas redes é em camadas, podendo ser formadas por uma (*RNA* de uma só camada) ou mais camadas (*RNA multicamadas*).

As *RNAs* de uma só camada (Secção 3.10.1), ilustrada pela Figura 3.6, são sempre unidireccionais (convergentes ou divergentes). A rede é composta por um conjunto de neurónios de entrada (esta não é considerada como camada, uma vez que aí não são efectuados cálculos, apenas contém os valores de entrada da rede [Wasserman, 1989]) e uma camada de saída.

As *RNAs multicamadas* (Secção 3.10.2), ilustrada pela Figura 3.7, são compostas por um conjunto de neurónios de entrada, de saída e neurónios que não pertencem a nenhuma destas camadas, sendo organizados numa camada escondida ou intermédia, podendo existir mais do que uma.

3.8.2 – *RNAs* Recorrentes

São redes em que a saída de um neurónio influência de alguma forma a entrada desse mesmo neurónio, criando-se um circuito fechado (Figura 3.5). Isto implica a existência de conexões cíclicas na rede. Além disso, a sua estrutura não é obrigatoriamente organizada em camadas e, quando o é, podem possuir interligações entre neurónios da mesma camada e entre camadas não consecutivas.

Cohen e Grossberg apresentaram um teorema que determina que para as *RNAs* recorrentes alcançarem um estado estável é necessário possuírem conexões simétricas [Wasserman, 1989].

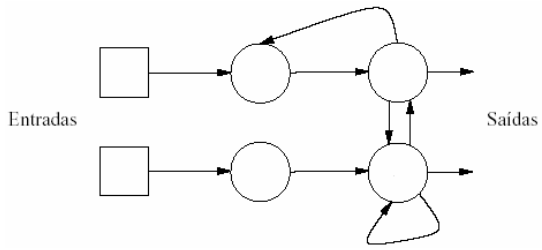


Figura 3.5 – *RNA* recorrente [Cortez, 2002].

Contribuições também importantes foram dadas por Hopfield [1984], sendo que algumas das suas configurações passaram a ser chamadas de redes Hopfield. Hinton e Sejnowski [1986] também tiveram mérito neste campo, introduzindo regras para treino de redes, denominadas por eles de *Máquinas de Boltzmann*, devido ao cálculo das saídas dos neurónios artificiais ser efectuado através da probabilidade segundo a distribuição de Boltzmann.

3.9 – Aprendizagem das *RNAs*

Actualmente, existe uma grande diversidade de arquitecturas de *RNAs* e de correspondentes algoritmos de aprendizagem, cada um com diferentes possibilidades e finalidades, vantagens e inconvenientes. Como já foi mencionado, uma das principais vantagens e importâncias das *RNAs* é sem dúvida a sua capacidade de aprender e de melhorar sempre o seu desempenho baseado nessa aprendizagem.

De seguida**,** são apresentadas duas vertentes importantes das *RNAs*: *i)* **os paradigmas** de aprendizagem, que implementam os procedimentos de treino que levam as redes a aprender determinadas tarefas e; *ii)* **tipo de regras**, que consiste em encontrar os pesos das conexões sinápticas, de forma a resolver determinado problema.

3.9.1 – Paradigmas de Aprendizagem

Tal como no tipo de aprendizagem para o *DM* (Secção 2.4.4.2), e particularizando para as *RNAs*, as redes podem agrupar-se em duas formas de classificação principais: Supervisionada ou Não-Supervisionada. Outro paradigma bastante conhecido é o da aprendizagem por Reforço, que pode ser considerado um caso particular da aprendizagem supervisionada, assim como a aprendizagem por Competição, um caso particular da aprendizagem não-supervisionada [Costa, 2003].

- **Aprendizagem Supervisionada** – São sucessivamente apresentados à rede, conjuntos de padrões de entrada e os seus correspondentes padrões de saída. Durante este processo, a rede realiza o ajustamento dos pesos das conexões entre os elementos de processamento, até que o erro entre os padrões da saída gerados pela rede e os padrões de saída apresentados, alcance um valor mínimo desejado. Os exemplos mais conhecidos de algoritmos para aprendizagem supervisionada são a Regra Delta Generalizada, utilizada no algoritmo de *retropropagação* [Rumelhart *et al.*, 1986];

- **Aprendizagem por Reforço** – Neste tipo de aprendizagem, a rede apenas recebe o valor que indica se o valor de saída está ou não correcto. Consiste na aprendizagem através do método de tentativa de erro, de modo a optimizar um índice de performance chamado sinal de reforço;

- **Aprendizagem Não-Supervisionada** – A rede analisa o conjunto de dados recebidos, determina algumas propriedades desses conjuntos de dados e "aprende" a reflectir essas propriedades. A rede utiliza padrões, regularidades e correlações para agrupar o conjunto de dados em classes. As propriedades adquiridas pela rede sobre os dados pode variar em função da arquitectura e da regra de aprendizagem utilizada. Este tipo de aprendizagem só é possível quando existe uma quantidade considerável de dados de entrada, caso contrário, torna-se difícil, ou mesmo impraticável, apurar quaisquer padrões ou características nos dados de entrada;

- **Aprendizagem por Competição** – O objectivo deste tipo de aprendizagem é segmentar os dados que se introduzem na rede. Desta forma, as informações similares são classificadas formando parte da mesma categoria, e por conseguinte, devem activar a mesma saída, existindo por isso uma competição entre as unidades de saída, para decidir qual delas será a vencedora, e consequentemente, terá a saída activada e os seus pesos actualizados no treino. As categorias devem ser criadas pela própria rede, já que se trata de uma aprendizagem não-supervisionada através das correlações dos dados. A unidade mais forte fica com um maior peso, e o seu efeito inibidor sobre as outras unidades de saída torna-se dominante. Com o tempo, todas as outras unidades de saída ficarão completamente inactivas, excepto a vencedora.

3.9.2 – Regras de Aprendizagem

De acordo com Haykin [1999], as cinco principais regras básicas de aprendizagem são: *i)* aprendizagem por correcção de erro (**gradiente descendente**), fundamentada na filtragem óptima; *ii)* **aprendizagem baseada na memória**, que trabalha na memorização directa dos dados de treino; *iii)* **aprendizagem *hebbiana***, baseada na teoria original de Hebb; *iv)* **aprendizagem competitiva**, também inspirada na neurobiologia; e *v)* **aprendizagem de Boltzmann**, baseada em ideias da mecânica estatística. A regra a aplicar está intimamente relacionada com a arquitectura da rede. De seguida é apresentada uma breve descrição de cada uma destas regras:

- **Regra Gradiente Descendente** – O erro de saída de um neurónio é obtido comparando o resultado por ele calculado com o resultado desejado. O erro actua assim como mecanismo de controlo, onde esses ajustamentos vão melhorando a resposta;
- **Regra Baseada na Memória** – Procura um padrão aproximado de entrada, sendo que a selecção do padrão aproximado da região vizinha tem por base as experiências passadas;
- **Regra de *Hebb*** – Esta regra estabelece que, quando um neurónio da célula *A* está suficientemente perto para excitar uma célula *B* e, repetidamente ou persistentemente, ocorre um disparo, algum processo de crescimento ou mudança metabólica acontece em uma ou ambas as células, tal que a eficiência de *A*, como uma das células disparadores de *B*, é aumentada. Esta regra foi proposta por Hebb em 1949;
- **Regra Competitiva** – Na aprendizagem competitiva, usada nas populares redes de Kohoen, neurónios são inibidos por outros neurónios de modo a que a competição entre eles, leve a apenas um a acabar excitado. Neste tipo de aprendizagem, somente um neurónio de saída fica activado;
- **Regra de Boltzmann** – Este algoritmo de aprendizagem realiza o ajuste dos pesos, baseando-se na probabilidade e na mecânica estatística. A *RNA* que utiliza esta regra é denominada máquina de Boltzmann.

3.10 – As Redes *Perceptrão*

O *perceptrão*, introduzido por Rosenblatt, em 1958, demonstrou algumas aplicações práticas. O *perceptrão* é uma forma simples de *RNAs*, em que a sua principal aplicação é a classificação de padrões, ou seja, "*o processo pelo qual um padrão/sinal recebido é atribuído a uma classe de entre um número predeterminado de classes (categorias)*" [Haykin, 1999]. De seguida serão abordados os dois tipos de camadas existente, a saber, *perceptrão* de uma única camada e *perceptrão multicamada*.

3.10.1 – *Perceptrão* de Camada Única

O modelo mais simples de uma *RNA*, no qual várias unidades de processamento estão conectadas unicamente a uma única camada de saída através de pesos sinápticos (Figura 3.6). O neurónio de saída é do tipo binário, em que pode ser utilizado como classificador ou como representação de funções booleanas (verdadeiro/falso) [Brío e Molina, 2001].

O *perceptrão* é uma rede de uma camada, na qual se pode treinar pesos vinculados a padrões de entrada que são fornecidos à rede, além do uso de um conceito de controlo de neurónio (*bias*), para se obter uma saída correcta. O conceito *bias* é uma forma de controlar o *limite*[19] do neurónio. O *limite* é o valor a partir do qual o somatório dos pesos convencionais determinará se o neurónio estará activo (1) ou inactivo (0), ou seja, se pertence ou não à classe que representa [Russel e Norvig, 1995].

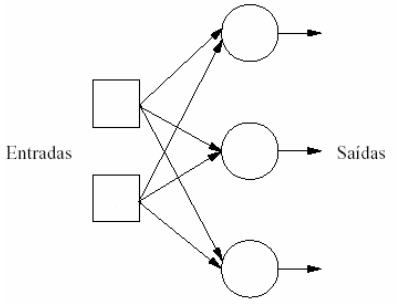


Figura 3.6 – *RNA* de uma só camada [Cortez, 2002].

[19] Tradução do inglês *Threshold*.

Apesar de na época o *perceptrão* ter causado grande euforia, este não teve a vida muito longa, já que foram muitas as criticas às suas limitações, levando mesmo ao desinteresse pela área. Esta visão pessimista da capacidade do *perceptrão* e das *RNAs* mudou com as descrições da rede de Hopfield (*multicamadas*) em 1982 e o algoritmo de *retropropagação* em 1986 [Costa, 2003].

3.10.2 – *Perceptrão Multicamada*

Como o *perceptrão* de uma única camada só é capaz de classificar padrões linearmente separáveis, o que acontece somente num número limitado de aplicações, torna-se necessário recorrer ao *perceptrão* de *multicamada*. Esta arquitectura constitui o modelo neuronal artificial mais utilizado e conhecido actualmente. Tipicamente, consiste num conjunto de dados de entrada, uma ou mais camadas de intermédias (ou escondidas) e uma camada de saída. Esta arquitectura será estudada mais pormenorizadamente na secção seguinte.

3.11 – *Perceptrão Multicamada* e o Algoritmo de *Retropropagação*

Arquitecturas neuronais são tipicamente organizadas em camadas, como ilustrado na Figura 3.7, com unidades que podem estar conectadas às unidades da camada superior. Existe um conjunto de dados que são apresentados à rede, que serão posteriormente tratados pela camada intermédia (oculta) e por último, depois de processados enviados para a camada de saída. Cada camada pode ter de 1 a n neurónios artificiais [Dhar e Stein, 1997].

A camada de saída recebe os estímulos da camada intermédia e constrói o padrão que será a resposta. As camada(s) intermédia(s) funciona(m) como extractora(s) de características, os seus pesos são uma codificação de características apresentadas nos padrões de entrada e permitem que a rede crie a sua própria representação, mais rica e complexa do problema. Entre uma camada e outra existe uma matriz de pesos, sendo a regra de propagação dada pela combinação entre as saídas de cada unidade e a matriz. Esta é construída através da soma ponderada de cada sinal que chega, via conexões, pelo respectivo peso. O estado de activação assume valores contínuos e devido a isto a

regra de activação das unidades utiliza em geral a função a logística [Yamamoto e Nikiforuk, 2000].

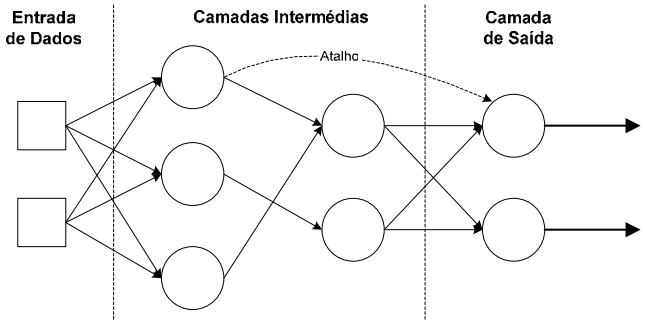


Figura 3.7 – Rede *Perceptrão Multicamada*.

Apenas com o desenvolvimento do algoritmo de treino de *retropropagação*, por Rumelhart e seus colaboradores em 1986, se mostrou que é possível treinar de modo eficiente redes com camadas intermédias, resultando no modelo de *RNAs* mais utilizado actualmente.

3.11.1 – Algoritmo de *Retropropagação*

Trata-se do mais popular método para a aprendizagem de *RNAs* com *multicamadas*. Consiste em dois passos. No primeiro, um padrão é apresentado às unidades da entrada e, a partir daí as unidades intermédias calculam a sua resposta que é produzida na camada de saída, sendo então o erro calculado. No segundo passo, este é propagado[20] a partir da camada de saída até à camada de entrada, e os pesos das conexões das camadas internas vão sendo modificados utilizando a regra do gradiente descendente. Com este processo o erro vai sendo progressivamente diminuído.

Estas regras de treino, especificam como os pesos podem ser adaptados durante a aprendizagem, com o objectivo de melhorar a performance da rede. A alteração dos parâmetros de uma rede só é permitida durante a fase de treino, permanecendo estáticos durante as fases de testes e execução.

O algoritmo de *retropropagação* baseia-se em duas fases bem distintas [Riedmiller, 1994; Bose e Liang, 1996]:

[20] Daí o termo de *retropropagação*.

1. Propagação Frontal – Sendo o vector de entrada (x^k), para o exemplo k, fornecido aos neurónios de entrada, propagando-se em frente, camada por camada, sendo s_0=1 (valor de entrada da conexão de *bias*), $s_1=x^k_1$, . . ., $s_E=x^k_E$ (para E neurónios de entrada). De seguida, aplica-se a função de erro:

$$\varepsilon = \frac{1}{2}\sum_{i \in s}(t_i^k - s_i^k)^2 \qquad \textbf{(2)}$$

2. *Retropropagação* – O erro é propagado para trás, desde a saída até aos neurónios de entrada. Os pesos vão sendo ajustados segundo a regra de Widrow-Hoff . Para um único peso tem-se que:

$$\Delta w_{ij}(t) = -\eta \frac{\delta \varepsilon}{\delta w_{ij}}(t) \qquad \textbf{(3)}$$

em que t representa a ordem da iteração e η a taxa de aprendizagem. As derivadas parciais são calculadas aplicando as seguintes fórmulas:

$$\frac{\delta E}{\delta w_{ij}} = \frac{\delta \varepsilon}{\delta s_i}\frac{\delta s_i}{w_{ij}} \qquad \textbf{(4)}$$

onde

$$\frac{\delta s_i}{w_{ij}} = f'(u_i)s_j \qquad \textbf{(5)}$$

Para se obter $\frac{\delta \varepsilon}{\delta s_i}$, ou seja, a influência da saída s_i do neurónio i no erro global δ, é necessário atender ao tipo do neurónio:

$$\frac{\delta \varepsilon}{\delta s_i} = \left\{ \begin{array}{c} -(t_i - s_i), i \in s \\ \sum_{j \in succ(i)} \frac{\delta \varepsilon}{\delta s_i} f'(u_j) w_{ji}, i \notin s \end{array} \right\} \qquad \textbf{(6)}$$

Antes de iniciar o treino de uma *RNA*, devem ser definidos os valores iniciais dos pesos. Em geral, estes devem ser gerados de uma forma aleatória, dentro do intervalo [-0.7,0.7] [Hastie *et al.*, 2001]. Também é necessário definir uma taxa de aprendizagem, entre [0,0 e 1,0], embora, habitualmente se utilize valores próximos de zero (*e.g.* 0,1). Quanto mais baixos forem os valores da taxa de aprendizagem, mais lento se torna a

convergência do treino, mas menores valores de erro serão obtidos. Uma outra variável que pode ser incluída na actualização de pesos é o *momentum*. Este novo termo permite aumentar a velocidade de aprendizagem, ou seja, ajusta os pesos, após verificar todos os casos de treino [Haykin, 1999]

3.11.2 – Regra Delta Generalizada

A regra delta generalizada, é a regra de aprendizagem mais utilizada para o treino de *redes multicamada*. Esta regra é baseada no método do gradiente descendente (Secção 3.9.2), que significa encontrar nos pesos uma direcção para a mudança do peso, com o intuito de reduzir o valor do erro.

O algoritmo de *retropropagação* é responsável pelo cálculo das funções de erro (Equação 2). O objectivo da fase de treino é diminuir constantemente o valor desse erro e, para tal, o valor dos pesos devem ser ajustados a cada nova iteração. A regra de *retropropagação* faz com que os pesos da camada de saída sejam os primeiros a ser ajustados e, posteriormente, os pesos das restantes camadas, da frente para trás.

Note-se que todos os cálculos envolvidos nas correcções dos pesos das conexões neuronais são baseadas no sinal de erro, e esse sinal é obtido através da informação externa de uma saída desejada. Quando, porém, o neurónio está localizado numa camada oculta, não se tem previamente determinado uma resposta desejável para esse neurónio. Neste caso, o erro deve então ser calculado recursivamente, baseado nos sinais de erro de todos os neurónios da camada imediatamente posterior, aos quais o neurónio oculto está directamente ligado.

3.11.3 – Problemas e Limitações do Algoritmo de *Retropropagação*

As *RNAs* que utilizam algoritmos de *retropropagação*, assim como muitos outros tipos de *RNAs*, podem ser vistas como "caixas pretas", das quais pouco se sabe, a razão pelo qual a rede chega a um determinado resultado, uma vez que os modelos não apresentam justificativas para suas respostas. Neste sentido, muitas pesquisas vêm sendo realizadas visando a extracção de conhecimento de *RNAs*, e na criação de procedimentos explicativos, onde se tenta compreender o comportamento da rede em determinadas situações [Beal e Smith, 2000; Setiono, 2003].

Uma outra limitação refere-se ao tempo de aprendizagem das *RNAs* que pode ser demorado. Contudo, existe uma activa investigação para minorar este efeito, através do uso de computação paralela e do desenvolvimento de algoritmos de treino mais eficientes. De facto, convém referir que o treino de *RNAs* pode ser modelado como um problema de optimização numérica. Por conseguinte, existe toda uma panóplia algoritmos que podem ser aplicados, tais como métodos de *Newton*, *Quasi-Newton* e *Gradientes Conjugados* [Sarle, 2005].

É também difícil definir qual a arquitectura ideal da rede. Não existem regras claras para definir quantas unidades devem existir nas camadas intermédias, ou como devem ser as conexões entre essas unidades. Para resolver este tipo de problema, soluções sofisticadas têm vindo a ser apresentadas, tais como os algoritmos de corte, construtivos e genéticos [Rocha *et al.*, 2005].

3.12 – Mínimos Locais e *Conjuntos de Modelos*

Um *mínimo local* é o valor mais reduzido numa dada vizinhança. Quando uma função é convexa, apresenta-se como um grande "vale", pelo que o mínimo global é mais facilmente encontrado por métodos de minimização. Quando tal não acontece, ou seja, se existir uma série de "montes" e "vales", surgirão diversos mínimos locais. Ora, os métodos baseados no gradiente tendem a cair em mínimos locais, pois a função de erro é não convexa [Hastie *et al.*, 2001]. Assim que um mínimo local é encontrado, o algoritmo de treino consegue melhorias muito diminutas, e em geral, deve-se parar a aprendizagem por falta de progresso. Este fenómeno pode ocorrer independentemente de uma alteração da taxa de aprendizagem e/ou do termo *momentum* [Freeman, 1992]. Como resultado, a solução final obtida é dependente da escolha dos pesos iniciais.

Para solucionar este inconveniente, uma das soluções passa por usar vários treinos para o mesmo modelo de rede, sendo os pesos gerados de modo aleatório. No final, escolhe-se a solução que obteve um erro mais baixo. Todavia, uma melhor alternativa passa pelo uso de um *Conjunto de Modelos*, onde a previsão final é dada por uma combinação de saídas de cada modelo, neste caso, das diversas *RNAs* [Dietterich, 2000]. De seguida será feita uma breve introdução aos *Conjuntos de Modelos (CM)*.

Ao invés de se utilizar somente um modelo individual, é possível treinar diversos modelos e combinar as suas saídas numa única saída global. Primeiro, há que definir

como se chega a cada um dos modelos individuais (construção do *CM*) e depois como se combinam os vários modelos numa única saída (função de combinação do *CM*). Nos últimos anos, tem sido dedicada uma especial atenção ao uso de *CM* pela comunidades do *Data Mining* e *Aprendizagem Automática*. Isto porque em geral, um *CM* apresenta melhores resultados do que um modelo individual [Dietterich, 2000].

Diferentes formas de combinar as saídas de um *CM* têm sido propostas, distinguindo-se essencialmente, pelo tipo de problema. Nos problemas de classificação consideram-se os seguintes métodos: *i) Votação*, a saída é dada por aquela que constitui a solução oferecida pela maioria dos modelos individuais e; *ii) Confiança*, a saída é decidida pelo modelo que apresentar uma maior confiança associada à classe que propõe [Liu *et al.*, 2000]. Para a regressão destacam-se os seguintes métodos: *i) Médias Simples*, calculando-se a média das saídas propostas para cada modelo e; *ii) Médias Pesadas*, semelhante ao anterior, mas sendo atribuído um peso a cada modelo.

3.13 – *Sobre-Ajustamento* e Generalização

As *RNAs* utilizam um conjunto de dados de treino para ajustar os pesos da rede. Uma vez treinada, a rede deve ser usada para prever dados desconhecidos, isto é, dados que nunca foram usados como casos de treino. Uma *RNA* pode apresentar uma fraca capacidade de generalização se for demasiado complexa, ou seja, quando há muitas unidades de processamento, num fenómeno designado por *sobre-ajustamento*[21]. O que acontece é que ao ter uma elevada capacidade, a rede fixa-se em demasia aos casos de treino e perde capacidade para generalizar. Por outro lado, com um número demasiado reduzido de unidades de processamento pode ocorrer *sub-ajustamento*. Assim, a determinação do número ideal de unidades de processamento que pertencem às camadas internas deve ser definida de forma empírica, e normalmente depende da distribuição dos padrões de treino e teste da rede [Sarle, 1995].

Acontece *sobre-ajustamento* quando o modelo da rede obtém um desempenho quase perfeito nos dados de treino, mas um desempenho pobre nos novos dados. Neste caso, a rede cria um modelo que descreve os dados de treino, ao invés de criar um modelo de generalização. Isto pode dever-se à possibilidade de existir ruído nos dados ou à

[21] Tradução adoptada para o termo *Overfitting*.

existência de um conjunto de dados inadequado ou insuficiente. A solução para este tipo de problema, que é considerado crítico com as *RNAs*, poderá passar por [Sarle, 1995]: *i)* usar um conjunto de dados de treino com uma elevada cardinalidade; *ii)* usar uma *paragem antecipada*, terminando-se o treino quando o erro obtido num conjunto de *validação* (com exemplos não utilizados durante o treino) sobe; *iii)* uso de métodos de *regularização*, que penalizam modelos complexos [Hastie et al., 2001].

Neste trabalho, será usada uma *constante de decaimento*[22], uma forma de regularização que evita que certas unidades de processamento fiquem com pesos elevados [Bartlett, 1998]. Assim, ocorrerá uma pressão para a redução do valor dos pesos, fazendo com que alguns se mantenham (os importantes), enquanto outros se aproximarão do valor nulo. O valor desta constante (λ) varia entre o intervalo [0,1], embora em geral tenha valores mais próximos de zero. Em [Hastie et al., 2001], afirma-se que é mais importante a procura do valor ideal de decaimento do que o número ideal de neurónios intermédios.

3.14 – Avaliação de Modelos

Para comparar modelos, é necessário medir qual a sua capacidade de generalização, uma vez, que em geral, o erro de treino não é suficiente. Existem várias técnicas de estimação da capacidade de generalização de modelos. Neste trabalho serão apresentadas as mais comuns: a *Divisão da Amostra* e *Validação Cruzada*. Cada uma destas técnicas tem vantagens e desvantagens do ponto de vista estatístico. Do ponto de vista computacional a Validação Cruzada é mais exigente, pelo que na prática é comum usar-se quando o conjunto de dados tem uma dimensão reduzido.

3.14.1 – Divisão da Amostra

Um método popular para estimação do erro é a Divisão da Amostra[23]. Consiste na divisão dos dados em dois blocos, como demonstrado na Figura 3.8, o primeiro chamado de exemplos de treino, sendo usado para produzir o modelo; e o segundo de exemplos de teste, usado para medir a capacidade de generalização do modelo.

[22] Na terminologia anglo-saxónica *Decay*.

[23] Conhecido em inglês por *hold-out*.

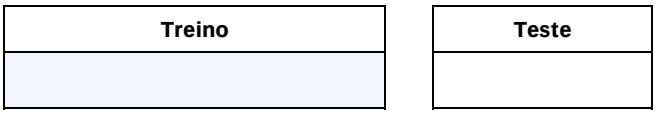


Figura 3.8 – Exemplo de divisão da amostra.

Este processo de validação tem como pontos fortes a sua simplicidade e rapidez. Em contrapartida tem as desvantagens de reservar boa parte dos registos (sendo também necessário decidir qual a percentagem) para teste, levando a que estes registos sejam desperdiçados do ponto de vista da modelação, uma vez que, não são usados no procedimento de aprendizagem.

3.14.2 – Validação Cruzada

Trata-se de um melhoramento do método anterior, que permite utilizar todos os exemplos disponíveis. Na *validação cruzada k-desdobrável*[24], os exemplos dados (E) são divididos aleatoriamente em K subconjuntos mutuamente exclusivos ($E_1, E_2, \ldots, E_K$) de cardinalidades idênticas (Figura 3.9). Sequencialmente, será testado um bloco diferente, sendo os restantes dados usados para ajustar os pesos da rede. No final dos K treinos, o modelo neuronal foi testado em todos os dados de treino, sendo a estimativa dada pelo erro médio calculado ao longo dos K conjuntos de teste. Em geral, K toma o valor de 10, embora possa variar entre 2 e o N, o número total de exemplos. Convém referir que um valor pequeno de K origina uma má estimativa, enquanto um valor elevado aumenta consideravelmente o esforço computacional.

Treino	Treino	Treino	Teste
Treino	Treino	Teste	Treino
Treino	Teste	Treino	Treino
Teste	Treino	Treino	Treino

Figura 3.9 – Exemplo de uma validação cruzada para $K = 4$.

A comparação mais comum de desempenho de cada tipo de algoritmo é feita por meio da medida de erro durante as N iterações (diferença entre o resultado conseguido pelo algoritmo de aprendizagem e o resultado esperado). Um valor pequeno para N origina

[24] Na terminologia anglo-saxónica *k-fold crossvalidation*.

uma má estimativa, enquanto um valor elevado aumenta consideravelmente o esforço computacional.

A validação cruzada é considerada mais eficiente do que a simples divisão da amostra, uma vez que todos os registos disponíveis são usados no processo de avaliação do modelo [Efron, 1983]. Todavia, tal acontece à custa de um esforço computacional considerável, visto que em vez de um são necessários *K* treinos [Prechelt, 1998].

3.15 – Discussão

A utilização de *RNAs* tem sido intensificada em diversos domínios nos últimos anos. No entanto, todo o poder oferecido pelas *RNAs* esbarra num problema: a dificuldade para explicar de forma compreensível as suas respostas ou decisões. Este problema é um factor de motivação para as várias pesquisas relacionadas com o desenvolvimento de técnicas de extracção de conhecimento de *RNAs*. Essas técnicas têm a finalidade de fornecer uma certa capacidade de explicação [Setiono, 2003].

Definitivamente, as *RNAs* não são mais do que um modelo artificial e simplificado do cérebro humano, que é o exemplo mais perfeito de que dispomos para um sistema que é capaz de adquirir conhecimento através da experiência. Uma *RNA* é assim um novo sistema para o tratamento da informação, cuja a unidade básica de processamento está inspirada numa célula fundamental do sistema nervoso humano: o neurónio.

Grande parte do tempo necessário para a construção de *RNAs* é gasto em duas tarefas, a saber, processo de identificação da melhor topologia e treino. Por isso, o binómio topologia *vs* treino, é o factor que hoje se apresenta como a barreira a ser transposta para o sucesso de uma solução baseada em *RNAs*. Assim, é necessário que se tenha uma ideia, ainda que genérica, sobre os algoritmos de aprendizagem disponíveis para treino e das diversas topologias de redes existentes.

Um modelo demasiadamente bem ajustado perde capacidade de generalização (*sobre-ajustamento*). De forma, a evitar este tipo de problema, deverão ser colocadas restrições ao processo de aprendizagem (regularização). Para tal, recorre-se à utilização de técnicas tais como a *constante de decaimento*.

Poder-se-á assim concluir que as *RNAs* consistem em unidades de processamento não linear que trocam dados e informação. Utilizam-se para reconhecimento de padrões,

incluindo imagens, manuscritos e sequências de tempo, tendências financeiras, etc., com a virtuosidade de aprender e melhorar o seu funcionamento de uma forma automática.

Capítulo 4

Qualidade da Carne de Cordeiro

Apresentação do problema objecto de estudo. São analisadas e comentadas as características essenciais para a qualidade da carcaça e da carne dos animais, assim como, as principais características da análise instrumental e sensorial da qualidade da carne. Finalmente, são apresentadas algumas abordagens clássicas e casos de estudo em que foram utilizadas Redes Neuronais Artificiais.

4.1 – Introdução

Desde longos anos que são investigados métodos de estimativa da composição das carcaças de animais. O produtor tenta estimar a composição da carcaça, na tentativa de programar o maneio alimentar e a venda dos animais, através de métodos subjectivos e relativamente pouco precisos, tais como: a apreciação visual e a palpação [Cadavez, 2004]. Na investigação científica, a estimativa da qualidade da carcaça tem sido realizada por métodos destrutivos, como são exemplos a dissecação e as análises químicas de amostras homogéneas do corpo dos animais. No entanto estas técnicas, embora precisas, são laboriosas, complicadas, muito dispendiosas em termos financeiros [Cadavez, 2004]. Por outro lado, foram já efectuados trabalhos nos quais se demonstram técnicas capazes de prever, *in vivo*, qual será a composição da carcaça com alguma precisão, como é o exemplo dos ultrasons [Stanford *et al.*, 1995; Delfa *et al.*, 1998; Delfa *et al.*, 1999; Cadavez *et al.*, 2000] e das medidas de dimensão da carcaça [González *et al.*, 1996; Stanford *et al.*, 1997].

Kempster [1983] considerou que a qualidade das carcaças produzidas como uma medida primária de produção é um critério chave no melhoramento genético das raças. Uma carcaça ideal deve apresentar uma composição que maximize o rendimento em carne magra e as características organolépticas da mesma. Sempre que isso acontece, a carcaça possui um valor máximo. Quando não se verifica, o seu preço sofre

penalizações, pelo que, os sistemas de classificação de carcaças desempenham um papel fundamental na definição de regras para as transacções comerciais [Cadavez, 2004].

O sistema de produção ovina europeu é caracterizado pelo elevado número de raças, de vários tamanhos, explorados em diversos sistemas de produção, pelo que no mercado é disponibilizado uma grande variedade de produtos [Colomer-Rocher, 1993], que, de uma forma geral, vão de encontro aos gostos e costumes de cada região. Assim, é necessária a existência de sistemas de classificação que tenham em conta essas diferenças de modo a não penalizar algumas das carcaças, nomeadamente as ligeiras produzidas nos países da região mediterrânea.

Devido às exigências de qualidade por parte dos consumidores e à ausência de legislação, começaram a surgir aproveitamentos oportunistas e fraudulentos que, tirando partido da desorientação dos consumidores, levam a provocar uma concorrência desleal. Só quando surge legislação que passa a definir regras que orientam os produtos animais e os produtos destinados à alimentação humana é que se verifica algum rigor, controlo e respeito pelas normas. Daqui surge a necessidade de certificar e qualificar os produtos que apresentem qualidades e que cumpram os parâmetros exigidos por essa mesma qualificação. A Denominação de Origem Protegida e a Indicação Geográfica Protegida são duas dessas formas de certificação de produtos.

Para que um produto possa beneficiar de uma ***Denominação de Origem Protegida*** *(DOP)* tem que demonstrar ter origem no local que lhe dá o nome e ter uma forte ligação com essa mesma região. De facto, a qualidade do produto é influenciada pelos solos, pelo clima, pelas raças animais ou variedades vegetais e pelo saber fazer das pessoas dessa área [Agricultura.pt, 2004].

A ***Indicação Geográfica Protegida*** *(IGP)*, tem que demonstrar que pelo menos uma parte do ciclo produtivo tem origem no local que lhe dá o nome e que tem uma "reputação" associada a essa mesma região, de tal forma que é possível ligar algumas das características do produto aos solos, ou ao clima, ou às raças animais, ou às variedades vegetais, ou ao saber fazer das pessoas dessa área [Agricultura.pt, 2004].

De um modo geral, uma *DOP* ou uma *IGP* é o nome de uma região ou de um local determinado que serve para designar um produto: *i)* originário dessa região ou desse local e; *ii)* cuja qualidade ou característica se deve ao meio geográfico (*DOP*) ou cuja reputação ou determinada qualidade podem ser atribuídas a essa origem geográfica

(*IGP*). Em Portugal, existem quatro produtos de origem ovina e caprina com *DOP* e treze com *IGP* [Cadavez, 2004].

A classificação e denominação dos produtos de qualidade pretendem representar um incentivo à produção deste tipo de produtos, através da valorização dos sistemas de produção e dos produtos típicos ou específicos das diversas regiões de Portugal e da União Europeia. Evita-se a vulgarização da denominação, isto é, que não se torne um nome comum de uma série de produtos semelhantes, impedindo que apareçam marcas registadas com reputação e notoriedade, tais que, o registo de denominação possa induzir o consumidor em erro quanto à origem do produto.

Sempre que uma *DOP* ou uma *IGP* sejam reconhecidas como tal, estas denominações ficam a nível nacional e comunitário, protegidas contra qualquer utilização comercial directa ou indirecta, de uma denominação registada para produtos não abrangidos pelo registo. Apesar da comercialização de produtos de qualidade, o mercado tradicional de animais/carcaças, utiliza o peso vivo da carcaça como variável principal para o estabelecimento do preço de comercialização, bem como para estabelecer os lotes e categorias comerciais [Cadavez *et al.*, 2002].

4.2 – Abordagem Histórica

Como consequência da 2ª guerra mundial, em que estiveram envolvidos grande parte dos países europeus, os governadores de diversos países intensificaram a produção agrícola e animal, como forma de solucionar os problemas relacionados com a necessidade de alimentar as suas populações famintas e com o território destruído. Nos anos oitenta a intensificação foi de tal ordem, que os países europeus começaram a deparar-se com uma situação insustentável. Havia demasiados alimentos, o que causava dificuldades de armazenamento dos excedentes, provocando graves problemas ambientais [Rodrigues, 2003].

Inicialmente, o nível de exigência por parte dos consumidores era apenas em quantidade, descurando a qualidade. Com o decurso do tempo, a situação inverteu-se, e o grau de exigência de produtos com qualidade começou a impor-se, ficando a dever-se ao excesso de oferta e à má qualidade de alguns produtos encontrados no mercado. Devido a esta nova necessidade, os agricultores que inicialmente haviam sido

incentivados para a produção em quantidade e subestimando a qualidade, começaram a ter dificuldades no escoamento dos seus produtos estandardizados e indistintos.

Nessa mesma altura, a Europa começou a assistir ao aparecimento no mercado de produtos com referência, a modos de produção não intensiva, mais preocupada com a qualidade do produto e satisfação dos consumidores. Nestes sistemas dava-se ênfase à preservação do ambiente, à preservação de raças animais e variedades vegetais, ao bem-estar do animal e à saúde do consumidor [Rodrigues, 2003].

4.3 – Qualidade da Carcaça

O conceito de carcaça pode sofrer variações de região para região, tratando-se de um conceito importante na definição dos preços. O Regulamento (CEE) nº 2137/92 define, para efeitos de classificação, carcaça padrão como: "*o corpo inteiro de um animal abatido, tal como se apresentam após sangria, evisceração e esfola, e sem cabeça, pés, cauda, úbere, órgãos genitais, fígado e fessura. Os rins e respectiva gordura são incluídos na carcaça"*. No entanto, os estados membros podem ser autorizados a aceitar outras apresentações que não a de referência.

A qualidade de uma carcaça deve ser associada à sua composição tecidual, uma vez que, esta é determinante na sua valorização comercial, por dois aspectos igualmente importantes: *i)* rendimento em carne magra, e *ii)* características organolépticas da carne a que dá origem. Assim, a qualidade da carne é condicionada por dois grupos principais de factores: *i)* factores intrínsecos ao animal: raça, idade e sexo; e *ii)* factores extrínsecos ao animal: sistema de produção, dieta e nível alimentar [Teixeira *et al.*, 1998].

A recente criação e desenvolvimento de produtos cárnicos ovinos com *DOP* ou *IGP*, são um incentivo à produção de produtos de qualidade, cujas características devem corresponder às características dos consumidores. Assim, torna-se necessário estudar os factores que condicionam a qualidade das carcaças e da carne, para a sua caracterização e padronização [Teixeira *et al.*, 2003].

Basicamente, a qualidade da carcaça está relacionada com a eventual preferência do consumidor por determinadas características. Todavia, inclui também, aspectos de adaptabilidade a todas as etapas de processamento, distribuição e comercialização desde o produtor até ao consumidor [Rodrigues, 2002]. Segundo os autores Boccard e Dumont

[1976], as definições de qualidade podem variar de acordo com o interveniente: **produtor**, **talhante** ou **consumidor**.

O **produtor**, deve interessar-se pela qualidade das carcaças, para orientar a produção, de modo a tirar o maior lucro da sua exploração. Deverá também ter em atenção às modificações que faz no seu sistema de produção, de forma a melhorar a qualidade dos seus produtos, tendo sempre a preocupação de analisar que alterações financeiras essas modificações técnicas provocam, quando se propôs fazer esse melhoramento.

O **talhante**, ainda segundo os mesmos autores, deve interessar-se pela qualidade, pois, normalmente esta deve traduzir as suas necessidades, ou seja, as suas e a dos seus consumidores. Através da carcaça, o talhante obterá as peças e tecidos vendíveis aos consumidores.

Finalmente, quanto ao **consumidor**, não está interessado na carcaça, mas sim na sua carne. A maioria exige a carne não seja dura, não tenha excesso de gordura ou seja difícil de cozinhar. Em particular, a resistência dos consumidores à carne com excesso de gordura, deve-se essencialmente a dois factores: *i) desperdício*, pois esta não é consumida e; *ii) saúde*, no que diz respeito a doenças cardiovasculares.

O produtor necessita então de produzir um produto que seja tenro, suculento, possua baixo teor em gordura e osso e seja agradável ao paladar. De modo a produzir tal animal deve ter em atenção os factores que condicionam a qualidade da carcaça, tal como referiu Johnston [1983]. São vários os autores que referem factores dos quais depende a qualidade da carcaça e, são também muitos, os factores referidos. Para Kempster [1983], o valor das carcaças depende: do peso, da conformação, da proporção dos principais tecidos (músculo, gordura e osso), da distribuição desses tecidos na carcaça, da espessura do músculo e da qualidade da carne. O autor refere ainda que, o peso e o tamanho da carcaça têm a maior influência.

Por outro lado, Fraser e Stamp [1989] consideram que a conformação da carcaça é a principal característica na avaliação da qualidade. De modo semelhante, Colomer-Rocher [1993] indica o peso da carcaça, o grau de engorda, a morfologia ou conformação e a composição regional, tecidual e química da carcaça, como o conjunto de factores específicos ou fundamentais.

De seguida, são apresentados os vários factores que poderão condicionar a qualidade da carcaça, começando pelo peso da mesma, seguido do seu rendimento, o terceiro aspecto

é a conformação da carcaça, associada à sua forma, o quarto aspecto relaciona-se com a composição da carcaça e, finalmente, a distribuição dos seus tecidos [Rodrigues, 2002].

1. **Peso da Carcaça** – Trata-se de um factor quantitativo, facilmente mensurável com erros mínimos, e cuja variação origina os diferentes tipos de carcaças produzidas, condicionando de modo implícito o seu valor económico, de acordo com os gostos dos diferentes mercados de carne. Para Sañudo e Sierra [1986], a importância do peso da carcaça pode valorizar-se sob diversos pontos de vista (experimental, técnico, comercial, prático, biológico e político). Cada um destes conceitos conduz a diferentes pesos de abate óptimos e, por conseguinte, distintos pesos de carcaça, consoante os objectivos;

2. **Rendimento da Carcaça** – Diz respeito à percentagem de carcaça obtida relativamente ao peso vivo do animal [Sañudo e Sierra, 1986; Fraser e Stamp, 1989]. De um modo geral, os factores de variação do rendimento da carcaça são bastantes complexos e variados, destacam-se entre eles, os factores intrínsecos e extrínsecos aos animais, como anteriormente referido. O rendimento da carcaça tem um grande interesse para o produtor e comprador de animais vivos. Nem sempre as carcaças de maior rendimento são as melhores, pois, este facto, pode estar associado a um excessivo estado de engorda, à idade e ao peso elevados, sendo estas situações negativas na avaliação da qualidade e no preço;

3. **Conformação da Carcaça** – Pode considerar-se a espessura do músculo, gordura subcutânea e gordura intermuscular, relativamente às dimensões do esqueleto [Rodrigues, 2002]. As carcaças bem conformadas são aquelas que apresentam um aspecto compacto, ou seja, são curtas e largas, com pernas globosas, planos musculares desenvolvidos e um predomínio geral dos perfis convexos [Sañudo e Sierra, 1986]. A maioria dos esquemas de classificação inclui a conformação como factor, sendo as carcaças com boa conformação, aquelas que normalmente comandam os preços mais elevados [Kempster, 1983];

4. **Composição da Carcaça** – A proporção de peças da carcaça, assim como a quantidade de músculo, gordura e osso que cada uma das peças proporciona, são os aspectos de qualidade mais importantes a avaliar nas carcaças. Ainda, segundo a autora, uma carcaça com uma composição ideal, seria aquela que tivesse a maior percentagem de peças de primeira categoria e com maior quantidade de músculo, a mínima de osso e a necessária de gordura para conferir à carne um sabor idóneo;

5. **Distribuição de Tecidos na Carcaça** – É a distribuição do músculo, do osso e da gordura na carcaça, ou seja, é a composição tecidular das peças que se obtêm a partir do corte da carcaça, sendo o seu conhecimento bastante útil para a avaliação do valor comercial individual de cada peça.

Assim, uma carcaça com composição de referência, ou ideal, deve apresentar uma composição que maximize o rendimento em carne magra e as características organolépticas da mesma. Sempre que isso acontece, a carcaça deverá possuir um valor máximo. Na impossibilidade de se verificar esta situação, o preço para as transacções comerciais deve sofrer uma penalização.

4.4 – Classificação de Carcaças

Desde sempre, o homem teve que ordenar, comparar e classificar tudo que o rodeia. Esta qualidade habitual do ser humano abarcou também os animais de talho. Com o passar do tempo, os caracteres quantitativos e qualitativos das carcaças, têm sido estudados utilizando diferentes métodos ou sistemas operacionais, com o objectivo de realizar uma avaliação, o mais completa e exacta possível, do seu valor comercial. A principal razão para o desenvolvimento de esquemas de classificação das carcaças, é fornecer uma melhor comunicação dos requisitos do consumidor ao produtor [Kempster, 1983].

Assim, o objectivo principal de um sistema de classificação de carcaças é facilitar a comercialização [Jeremiah, 1998], através da caracterização das carcaças de uma forma compreensível para todos os intervenientes no mercado. Os descritores utilizados, como o peso, estado de engorda, desenvolvimento muscular e, eventuais lesões, devem corresponder às expectativas de todos os intervenientes: produtores, talhantes e consumidores. Kempster [1983] considerou que a essência da classificação é permitir que indivíduos diferentes escolham carcaças diferentes, de acordo com as suas preferências. Desta forma, a classificação é vista como uma forma de agrupar as carcaças com características semelhantes, tornando possível direccioná-las para mercados específicos, mas também para atribuição do seu valor comercial.

A organização comum do mercado do sector das carnes de ovino e caprino, obriga à criação de normas de classificação das carcaças de ovino e caprino em todos os países da união europeia. Esta classificação deverá ser realizada por pessoal suficientemente

qualificado, cuja fiabilidade deve ser verificada por uma inspecção efectiva, de forma a garantir uma aplicação uniforme das normas de classificação pelos estados membros [Cadavez, 2004].

A escala de classificação descrita nos Regulamentos (CEE) nº 2137/92 e 461/93 utiliza seis classes de conformação, sistema vulgarmente designado por SEUROP, no qual as letras correspondem a: S – Superior, E – Excelente, U – Muito boa, R – Boa, O – Relativamente boa, e P – Medíocre. Este sistema é considerado como padrão, no entanto, a elevada variedade observada no sistema de produção europeu, levou à necessidade de autorizar a aplicação de diferentes critérios de classificação, desde que os estados membros o considerem necessário. Desta forma, o Regulamento (CEE) nº 2137/92 autoriza a utilização de outros critérios para a classificação de carcaças com menos de 13 kg de peso. Como se observa pela sua descrição, os dois sistemas de classificação de carcaças de ovino, autorizados pela união europeia, baseiam-se em critérios subjectivos, como são a avaliação do estado de engorda e da conformação por apreciação visual.

As vantagens de um sistema de classificação de carcaças são claras e foram enumeradas por Kirton e seus colaboradores [1992]: *i)* consistência na classificação das carcaças entre diferentes classificadores dentro de um matadouro, bem como entre matadouros; *ii)* implementação de um sistema de mensuração contínuo que permita às empresas seleccionar carcaças para satisfazer pedidos bem determinados, e *iii)* com um sistema automático de ordenação, as carcaças podem ser agrupadas em lotes com características bem determinadas.

4.5 – Qualidade da Carne

Considera-se um produto de qualidade, aquele que serve perfeitamente, de forma aceitável, acessível, segura e no momento em que for solicitado, às necessidades e aos anseios do consumidor. A qualidade da carne está estreitamente ligada à qualidade da carcaça. Apesar da complexidade de tecidos que compõem uma carcaça, a composição tecidual vê-se reduzida, na prática, à quantidade de gordura, músculo e osso [Rodrigues, 2002]. A presença, a nível óptimo, de gordura na carcaça, é um dos aspectos fulcrais para optimizar as características organolépticas da carne que dela se obtém. O excesso de gordura é indesejável pois tem custos de produção elevados e obriga, também, o

talhante a proceder à sua remoção aquando da venda da carne, o que acarreta ainda mais custos [Cadavez, 2004].

Nos últimos anos a qualidade da carne tem vindo a ganhar muita importância, quer pelos aspectos sanitários que abalaram o sector, quer pela necessidade de promover produtos, que de alguma forma, estão associados a marcas de qualidade. Este conhecimento pode contribuir para a satisfação do consumidor e, desta forma, revelar-se importante para a conservação das raças autóctones através da sua valorização ao nível do consumidor [Rodrigues, 2002].

Hofmann [1990] atribui à expressão qualidade da carne duas interpretações, uma objectiva e outra subjectiva. Assim, quando interpretada de uma forma objectiva, a qualidade da carne é afectada por múltiplos factores passíveis de serem descritos ou medidos objectivamente: factores nutricionais (composição química, digestibilidade, valor biológico), sensoriais (aparência, sabor, aroma, *flavour*[25], suculência e *tenrura*), higiénicos e toxicológicos (estrutura, actividade da água, pH, cor), muitos deles interligados e pertencentes a mais de uma das categorias referidas.

Quando interpretada de uma forma subjectiva, tem a ver com a apreciação do produto por todos os membros da cadeia[26] de comercialização da carne, ou seja, a aceitabilidade, preferência, utilidade, o lucro, o valor social, entre outros critérios. Assim, de uma forma objectiva, o termo 'Qualidade da Carne', é relativo às propriedades mensuráveis do produto, e 'Carne de Qualidade' é a forma subjectiva, que depende da apreciação do produto [Silva, 1996].

Vários autores [Berian, 1998; Sañudo *et al.*, 1998], referem que as condições durante a fase de abate (transporte, condições de recepção e repouso, sangria, condições higiénicas, tipo de insensibilização), bem como a fase posterior ao abate (refrigeração, acondicionamento, embalagem), e finalmente o tratamento culinário a que é sujeita a carne podem condicionar significativamente a qualidade da mesma.

Existem vários métodos para a análise das propriedades da carne, podendo ser agrupadas da seguinte forma: *i) métodos de avaliação sensorial*, como são os casos de

[25] Combinação da percepção das sensações olfactivas e gustativas em conjunto com outras mais complexas, antes, durante e após a mastigação e deglutição de um alimento.

[26] Os membros da cadeia são todos os intervenientes que participam no processo, que vão desde o produtor até ao consumidor.

provas sensoriais com um painel de provadores (interpretação subjectiva); *ii) métodos instrumentais*, como por exemplo os testes de avaliação de força, como o *Warner-Bratzler*[27] (interpretação objectiva) e; *iii) métodos indirectos* (ex: determinação de teor em colagénio) [Kamdem e Hardy, 1995]. Dada a sua importância, serão abordadas a análise sensorial, que inclui os factores de variação da aparência, do odor e sabor, da suculência e da *tenrura*, e a análise instrumental, isto é, a avaliação da cor, do pH e da textura (dureza), da qualidade da carne, ambas intimamente relacionadas e englobadas no que se pode denominar qualidade organoléptica da carne.

4.5.1 – Qualidade Organoléptica da Carne

Hoje em dia, os parâmetros que avaliam a qualidade da carne ovina e, logo, o seu valor comercial, estão relacionados com a carcaça. Estes parâmetros são facilmente medidos no circuito de abate, mas possuem grande subjectividade. Assim, esses parâmetros deveriam ser complementados com outros mais relacionados com as características sensoriais da carne que, apesar da sua subjectividade, são de grande importância, uma vez que, permitem a consumidores, produtores, retalhistas e investigadores definirem as suas preferências. No controlo da qualidade sensorial, os humanos são utilizados como instrumentos de medida [Baptista, 2004].

Segundo Stone [1999], a análise sensorial da carne, é a ciência que mede, analisa e interpreta as reacções dos sentidos (visão, olfacto, audição, gosto e textura) aos alimentos. De acordo com Angulo [2001], é o ramo da ciência utilizado para obter, medir, analisar e interpretar as reacções a determinadas características dos alimentos e materiais, tal como são percebidos pelos sentidos da visão, olfacto, gosto, tacto e audição. Desde sempre, e de uma forma natural, o homem faz uma elaboração sensorial dos produtos que ingere, aceitando ou recusando os alimentos de acordo com as sensações que experimenta ao consumi-los.

No que diz respeito às carcaças de ovinos, as preferências dos consumidores não se encontram totalmente definidas, isto motivado por hábitos sensoriais e culturais que criam entrave à aceitação do produto, este tipo de hábito do consumidor só será alterado

[27] Máquina ou célula de corte com características especificas, para avaliar a tenrura da carne. Mede a força necessária para cortar e esmagar uma amostra de carne de 1 cm de espessura. A *Warner-Bratzler Shear Force* é a medida instrumental (da tenrura da carne) mais popular e precisa.

com a apresentação de um produto com elevada qualidade e confiança. Daí a indústria alimentar recorrer à avaliação sensorial, de forma a usar este tipo de análise para a selecção de matérias-primas e, consequentemente, avaliação do produto final, bem como, no desenvolvimento de novos produtos [Baptista, 2004].

Isto leva à necessidade de treinar e testar grupos de provadores humanos, que farão a análise sensorial. Esses grupos são utilizados como instrumentos de medida para quantificar as sensações percebidas pelos órgãos dos sentidos aquando das provas. As alternativas à técnica de medição como é a avaliação sensorial, são os métodos químicos, físicos ou microbiológicos de avaliação da qualidade.

A análise sensorial pode ser aplicada em empresas alimentares, podendo ser utilizada com resultados satisfatórios na produção, vendas, controlo de qualidade e desenvolvimento de novos produtos, embora os principais estudos se debruçam sobre a avaliação, análise e controlo [Costell e Durán, 1981]. Outra utilidade da análise sensorial são as possíveis modificações que provocam no produto, quer por eliminação, substituição ou adição de um novo ingrediente. Pode ainda originar a modificação de um processo de elaboração [Angulo, 2001], assim como análise sensorial de produtos concorrentes do mercado.

Em jeito de conclusão, é de senso comum a importância da análise sensorial na investigação da indústria alimentar, permitindo a comparação de resultados sensoriais com resultados instrumentais analíticos. Mas para que os resultados da análise sensorial sejam correctos, é necessário uma formação adequada ao painel de participantes, assim como o desenvolvimento de uma terminologia descritiva, técnicas de avaliação sensorial e ensaios físico-químicos que ajudem a caracterizar sensorialmente o alimento [Angulo, 2001].

4.5.2 – Factores Determinantes da Qualidade da Carne

Na Europa a inclusão de características sensoriais da carne na valorização comercial das carcaças foi sugerida por Berian [1998]. No entanto, a utilização deste critério necessita de investigação para identificar os factores que determinam a composição tecidual da carcaça, bem como associá-la com as preferências do consumidor, aspectos até ao momento muito pouco estudados e, portanto, pouco claros [Cadavez, 2004]. Existem duas principais características na análise sensorial da carne, são elas a qualidade visual,

onde se inclui obviamente a aparência, que atrai ou afasta o consumidor na hora da compra, e a qualidade gustativa, que é o caso da *tenrura*, sabor, aroma, *flavour* e suculência, só percebida após a sua prova.

Num estudo efectuado por Huffman [1997], foi analisado o nível de preferência dos consumidores relativamente a atributos sensoriais, como a *tenrura*, sabor e suculência. Os consumidores foram convidados a indicar qual deles tinha maior importância na determinação da sua preferência, aquando da análise da qualidade da carne. Os resultados mostraram que 51% dos consumidores consideraram a *tenrura* o atributo que mais procuram na carne. O sabor, foi o mais importante para 39% dos consumidores e os restantes 10% deram preferência para a suculência. Existem vários atributos de avaliação sensorial da carne, de seguida apenas são abordados os mais referenciados pelos diversos estudiosos da área.

4.5.2.1 – pH

O pH está relacionado com os processos bioquímicos de transformação do músculo em carne o que o torna o principal factor objectivo da qualidade da carne [Berian, 1998]. A sua evolução durante o período *postmortem* e o valor final do mesmo influenciam as características organolépticas da carne, podendo modificar quer a *tenrura* quer a cor da carne. De acordo com o mesmo autor, o pH da espécie ovina depende de factores intrínsecos (raça, sexo e peso vivo) e de factores extrínsecos (alimentação e stress) ao animal. Destaca-se o stress antes do abate que pode produzir, como exemplo, uma diminuição anormalmente rápida do pH e conduzir à obtenção de carnes com grande exsudação de água, coloração muito clara e uma textura muito branda.

4.5.2.2 – Aparência

Este tipo de análise pode ser o primeiro contacto que o consumidor tem com o produto, logo, será a primeira característica sensorial a ser analisada, tratando-se de uma percepção visual, deverá causar a impressão de que o produto é fresco e saudável. A aparência avalia-se principalmente em relação à cor e ao marmoreado da carne.

4.5.2.2.1 – Cor

No momento da compra o atributo que mais impressiona (caso o produto lhe seja desconhecido) o consumidor, é sem dúvida, a cor da carne. A cor da carne aceitável é um vermelho vivo e brilhante e sensação de frescura.

Verificam-se diferenças de cor entre espécies diferentes [Sañudo *et al.*, 1998]. Existem diferenças na cor muscular entre machos e fêmeas com peso de carcaça semelhante e idades diferentes. Quando comparadas à mesma idade, as fêmeas têm uma carne mais escura. O nível alimentar pode afectar a cor, através de uma maior deposição de gordura intramuscular, conduzindo à percepção de uma carne mais clara [Smulders *et al.*, 1991] e citado por Silva [1996].

4.5.2.2.2 – Marmoreado

O marmoreado, nome corrente para a gordura intramuscular, refere-se normalmente à gordura visível nas superfícies de corte das carnes. Os autores Gracey e Collins [1992], referem que para uma carne ter uma boa qualidade, deve conter aproximadamente 3,5% de gordura intramuscular, relativamente à carne fresca. Estudos têm demonstrado a preferência por carnes mais macias, embora, os aspectos de qualidade visual da carne crua como cor do músculo e da gordura, marmoreado, firmeza do tecido muscular e textura visual, sejam determinantes na hora da compra [Warkup, 1997].

4.5.2.3 – Sabor, Aroma e *Flavour*

Depois da *tenrura*, estas são as características mais importantes da carne. O sabor ou o gosto, o aroma ou o odor e o *flavour* que é uma combinação das características do sabor e aroma. O sabor detecta as quatro sensações gustativas básicas (doce, salgado, ácido e amargo), enquanto que o aroma detecta bastantes mais, tanto directamente, como é o caso do odor, como por via retronasal, quando o produto se encontra na boca [Wong, 1995]. O *flavour* é a facilidade do composto em se evaporar no ar, e percepcionado directamente, logo após a amostra se encontrar na boca, ou após a sua mastigação, dependendo da respectiva volatilidade dos componentes odoríferos [Baptista, 2004].

Para diversos autores, a raça, o sexo e a idade não são muito importantes para o sabor, aroma e *flavour*. Normalmente, para a análise destas características é frequente a associação com a idade do animal, no entanto não é referida qual a idade mais aceitável [Carmack *et al.*, 1995]. O mesmo autor, apenas refere que o aumento da intensidade do

sabor, aroma e *flavour*, com a idade parece dever-se a alterações da composição muscular.

Das três características, os consumidores consideram o *flavour* como a principal propriedade sensorial, que o ajuda na sua tomada de decisão, seleccionando, aceitando e ingerindo um alimento [Vergara e Gallego, 1999]. O verdadeiro sabor, aroma e *flavour* da carne desenvolve-se quando a carne é cozinhada, o facto de estas características serem mais intensas na carne cozinhada, do que na carne crua, é consequência do método culinário, na sua duração e temperatura, bem como do tipo de carne e do seu tratamento antes de ser cozinhada [Cross, 1994].

4.5.2.4 – Suculência

A percepção da suculência da carne cozinhada pode dividir-se em dois componentes. O primeiro refere-se à impressão de humidade durante as primeiras mastigações, produzida pela libertação rápida de fluidos. O segundo componente persiste durante mais tempo e deve-se ao potencial efeito estimulador da gordura na produção da saliva [Cross, 1994]. Tendo em conta estas características e ao facto do segundo componente durar mais tempo, a maioria dos estudos mostram a existência de uma estreita correlação entre a suculência e a quantidade de gordura, e não com a quantidade de fluidos libertados pela mastigação [Baptista, 2004]. Existe uma maior relação de suculência com animais mais novos, com uma primeira impressão de suculência e posteriormente uma impressão de secura, devido à pouca gordura de animais mais jovens [Silva, 1996].

A exemplo do atributo anterior, também na confecção da carne, existem processos que permitem reter uma maior quantidade de água e menos perda de gordura, tal como, grelhar e carne mal passada, permitindo que se obtenha uma carne mais suculenta, já que o aumento do tempo de tratamento está inversamente relacionado com a suculência [Smulders *et al.*, 1991]. A suculência e a *tenrura* da carne estão fortemente interligadas pois uma menor textura origina uma libertação mais rápida dos sucos ao mastigar. Para carnes com maior grau de dureza, a suculência é maior e mais uniforme, dado que a libertação de suco e gordura é lenta [Cross, 1994; Cañeque e Sañudo, 2000].

4.5.2.5 – *Tenrura*

A *tenrura* ou textura da carne é determinante da sua qualidade. É a mais importante característica sensorial, daí ser o atributo que se pretende obter como resultado final neste trabalho. Poder-se-á afirmar, que a *tenrura* de um alimento é uma manifestação devida à mastigação e resistência à aplicação de uma força. Esta propriedade sensorial é a primeira qualidade avaliada entre todas as outras, quando se mencionam os aspectos qualitativos procurados na carne.

A *tenrura* da carne poderá ser obtida de duas formas: *i)* quando se utilizam medidas físicas de resistência ao corte (*Warner-Bratzler*); *ii)* quando se utiliza a resistência encontrada na mastigação da carne por análise sensorial (painel de provadores). Devido ao elevado custo em avaliar a *tenrura* através do uso de um painel sensorial, a resistência à força de corte da *Warner-Bratzler*, é usada frequentemente, como uma medida para a *tenrura* da carne [Baptista, 2004]. No entanto, é importante alertar para o facto que a medida fornecida pelo *instron* indica a força máxima necessária para cortar um núcleo cilíndrico de carne aquecido em água, enquanto que a percepção da *tenrura* pelo consumidor é um resultado de morder e mastigar da carne cozinhada [Oeckel *et al.*, 1999].

Num trabalho realizado por Safari e seus colaboradores [2001], que analisa resultados de medidas instrumentais de qualidade e os resultados dos painéis sensoriais, conclui que a medida instrumental *Warner-Bratzler* pode ser utilizada com sucesso como critério de determinação da *tenrura* da carne e aceitabilidade. É importante salientar, que os dados para análise da *tenrura* da carne neste trabalho, foram obtidas por ambas as abordagens.

Segundo Cross [1994], a impressão geral da dureza da carne resulta de três fontes: *i)* facilidade com que os dentes cortam a carne no processo de mastigação; *ii)* facilidade com que a carne se rompe em fragmentos e; *iii)* quantidade de resíduo que fica após a mastigação. Uma revisão bibliográfica sobre a influência do pH na qualidade da carne [Silva, 1996] indica que relativamente ao pH final, a maioria dos autores está de acordo que um valor final elevado está associado com uma carne mais tenra, o que resulta em parte da relação estreita entre a *tenrura* e a água retida no músculo.

Deve também ser analisada a velocidade de crescimento no período anterior ao abate, uma vez que tem maior importância na determinação da *tenrura* do que o período total de tempo em que os animais são mantidos num regime alimentar elevado.

4.6 – Abordagens Clássicas do Problema

Pretende-se, nesta secção, apresentar uma revisão bibliográfica de abordagens clássicas que foram feitas para a análise da qualidade da carcaça e da carne. Na literatura, aparecem alguns dados sobre a precisão relativa das técnicas de estimativa, normalmente avaliada pela correlação entre as variáveis estimadoras e as variáveis da previsão da carcaça [Cadavez, 2004]. A informação existente sobre a aplicação de técnicas de estimação em diferentes raças é escassa, ou seja, a robustez dos modelos de estimativa não é avaliada [Kempster *et al.*, 1986]. Contudo, existem vários trabalhos que comparam a precisão de vários estimadores da composição da carcaça [Cadavez *et al.*, 2000; Santos *et al.*, 2000; Cadavez *et al.*, 2002]. No entanto, Kempster e seus colaboradores [1986] discordam desta abordagem, uma vez que, as várias medidas utilizadas como estimadores têm previsões muito parecidas.

Os modelos utilizados para a previsão da qualidade da carcaça e da carne, baseiam-se em procedimentos de **regressão linear simples** e **múltipla**, avaliando a qualidade do ajuste dos modelos pelo coeficiente de determinação e pelo desvio padrão residual. Estes modelos não analisam as inter-relações entre as variáveis independentes de forma a conseguirem explorar toda a informação. O desenvolvimento de modelos de regressão linear múltipla, utilizando variáveis correlacionadas, pode apresentar limitações na inferência e na precisão dos mesmos. Nesta situação, a estimativa dos coeficientes de regressão é efectuada de forma instável e dependente das variáveis independentes já presentes no modelo [Cadavez, 2004].

Vários autores, têm procurado alternativas aos modelos de regressão linear múltipla para dados enviesados, tendo a análise de **factores comuns** sido apontada como uma forma alternativa. As novas variáveis, produzem uma aproximação unificada, permitindo obter informação sobre a multicolinearidade entre as variáveis. Os factores comuns, têm ainda a vantagem de poderem ser utilizados como variáveis independentes, no desenvolvimento de modelos de regressão linear múltipla, pelo que, servem como base a uma técnica alternativa de estimativa com dados enviesados, produzindo um

ajuste adequado aos dados e, consequente explicação das associações entre variáveis [Chatterjee *et al.*, 2000].

Assim, a análise de factores comuns poderá ser usada para: *i)* compreender a dimensão dos dados, ou seja, saber o que cada uma das variáveis mede; *ii)* proceder à substituição das variáveis originais, por um grupo reduzido de novas variáveis (valores), que descrevam os dados originais, e/ou; *iii)* reduzir o número de variáveis originais objecto de estudo, para tal, eliminam-se as variáveis supérfluas e/ou redundantes [Cadavez, 2004]. No entanto, a análise de factores comuns apresentou uma melhor alternativa, do que a tradicional regressão linear, já que permite que modelos com boa qualidade de ajuste aos dados sejam obtidos, e estima os coeficientes de regressão com uma maior estabilidade. No entanto, a análise de factores comuns tem na sua interpretação um dos seus problemas, já que cada um deles é a mistura de vários originais.

4.7 – O Uso de *RNAs* na Análise do Problema

Como foi devidamente referenciado no início deste capítulo, as técnicas para análise da qualidade da carne, apesar de precisas, são caras e muito trabalhosas, surgindo a necessidade de reduzir esse trabalho e consequentemente os custos. Na opinião de Kempster [1989], a principal barreira à implementação de sistemas de classificação objectivas tem sido a inexistência de técnicas precisas e baratas.

É aqui que surge este novo paradigma, o uso de *RNAs*, de forma a executar o mesmo papel que os métodos de regressão. A partir de uma amostra (nas técnicas de *aprendizagem automática* é denominado de conjuntos de treino) sintetizam uma função (normalmente não linear), que estimará o valor da variável que queremos prever, em função dos restantes atributos.

Alguns estudos efectuados por diversos autores relatam o bom desempenho das técnicas tradicionais, confrontando mesmo os resultados obtidos com o uso de *RNAs* [Li *et al.*, 2000]. No entanto, apenas se verificam esses resultados quando as variáveis apresentam um comportamento linear. Quando o conjunto de variáveis tem um comportamento não linear, as *RNAs* apresentam resultados bastantes superiores [Li *et al.*, 1999]. Na bibliografia pesquisada sobre a aplicação de *RNAs* ao do problema em questão, foram encontrados diversos artigos que relatam o seu uso em dados de origem animal [Berg *et*

al., 2001; Lu e Tan, 2003; Tan, 2004]. No entanto, são escassos os trabalhos dedicados à previsão da *tenrura* da carne.

Foi encontrado um artigo que descreve o uso de *RNAs*, usando medidas de carcaças para a previsão da *tenrura* da carne [Hill *et al.*, 2000]. O objectivo era a classificação das carcaças em quatro categorias de *tenrura* da carne do Canadá: tenra, provavelmente tenra, provavelmente dura e dura.

A rede construída continha um total de onze entradas, dez das quais sobre atributos relacionados com o abate dos animais após 24h., sendo a outra entrada relativa à forma de ser cozinhada. Foram utilizados dados recolhidos entre os anos de 1985 e 1995 de 1452 animais de várias raças canadianas. Os dados relativos à *tenrura* da carne foram obtidos utilizando um texturómetro equipado com uma célula de corte *Warner-Bratzler* medindo a força de corte.

A taxa de insatisfação dos consumidores da carne canadiana era de aproximadamente 23% e pretendia-se uma redução para valores próximos dos 10-12%. Sendo a *tenrura*, o factor que mais influência tem na aceitabilidade da carne [Huffman *et al.*, 1996], o objectivo desse estudo passa pelo desenvolvimento de um modelo de *RNAs*, para: *i)* Prever a *tenrura* da carne e *ii)* Classificar as carcaças em categorias de *tenrura*. Os limites estabelecidos para cada uma das categorias foram: tenra (<5.6 Kg); provavelmente tenra (5.6-7.84 Kg); provavelmente dura: (7.85-9.6 Kg) e dura: (>9.6 Kg)

Depois de analisados os dados pela *RNA*, as taxas de precisão encontradas para a classificação das carcaças nas quatro categorias, foram respectivamente de: tenra – 64%, provavelmente tenra – 40%, provavelmente dura – 29% e dura – 79%. As *RNAs*, apresentaram assim excelente precisão na classificação da carne tenra e dura, mesmo quando comparadas com os valores de precisão apresentados por Park e seus colaboradores [1994], quando utilizaram a ultra-sonografia para classificar a *tenrura* das carcaças.

Um outro estudo mais recente, embora não tão elaborado nem detalhado, avalia as características da *tenrura* da carne de animais de pastoreio a partir de carne crua [Tian *et al.*, 2002]. O objectivo desta pesquisa, era estudar as amostras de carne crua de animais de pastoreio da Nova Zelândia, de forma a analisar a possibilidade de usar as características da carne crua para estimar a *tenrura* da carne cozinhada. Para tal, foram

usadas uma análise estatística linear e *RNAs*. As características dos dados acerca da *tenrura* da carne foram obtidas por um painel sensorial devidamente treinado.

A *RNA* construída, uma rede *perceptrão multicamada*, foi treinada com o algoritmo de *retropropagação*. No total, a rede tinha vinte variáveis de entrada devidamente normalizadas, seleccionadas a partir de sessenta e seis variáveis disponíveis. Como resultado final, e devido à natureza da não linearidade do problema, a *RNA* apresentou uma melhor previsão (62% contra os 58% apresentado pela regressão linear múltipla). É ainda importante referir que a literatura que reporta o uso de *RNAs* para a previsão da *tenrura* da carne, apesar do seu bom desempenho, é muito escassa, o que poderá trazer uma mais valia a este trabalho.

4.8 – Discussão

No contexto actual, é importante que a carne disponibilizada para o mercado corresponda às qualidades desejadas. Verifica-se uma tendência do consumidor em dar prioridade à qualidade e segurança dos produtos que consome, em detrimento da quantidade e do preço, focando a sua procura em produtos certificados e reclamando por informações e esclarecimentos sobre a origem e produção do mesmo, permitindo uma melhor escolha, com a garantia de confiança e consequente satisfação. No contexto actual da indústria, é importante que a carne disponibilizada para o mercado corresponda às qualidades desejadas pelos consumidores. Sendo a *tenrura* da carne, o atributo mais importante e consequentemente o factor determinante na sua aquisição, torna-se necessário criar mecanismos que consigam determinar esse factor.

Convém referir que a qualidade da carne e a satisfação do cliente depende, também, de outros factores que podem influenciar a qualidade da carne, tais como, abate, transporte e tratamento culinário. Outros factores que não estão directamente ligados com o produto, podem condicionar a satisfação do consumidor [Huffman *et al.*, 1997], de entre eles destacam-se, o rendimento, a idade e sexo de cada um. O rendimento económico também poderá afectar a análise do produto, uma vez que, quanto maior for o rendimento da pessoa maior é o nível de exigência e maior é a critica em relação a produtos de baixa qualidade. Apesar da idade não ser um factor importante na análise da qualidade da carne, existem alguns grupos etários em que o seu nível de exigência é mais evidente. Finalmente, se o nível de *tenrura* for relativamente baixo, o nível de

exigência do sexo feminino é muito menor quando comparado com o sexo masculino. Estes são algumas conclusões dos vários especialistas na área [Tian *et al.*, 2002].

Além da indispensável precisão, é também importante que a implementação de um sistema de classificação não origine trabalho extra no matadouro, não interfira com a sequência de tarefas de abate e não altere as condições físicas e sanitárias da carcaça. Para Hopkins e seus colaboradores [1995], a rapidez de classificação é de primordial importância para aplicação em linhas de abate, pelo que métodos rápidos, capazes de classificar uma carcaça em segundos, apresentam um especial interesse.

Por outro lado, o estabelecimento de níveis de aceitabilidade para a *tenrura* da carne, para cada um dos mercados, pode levar à construção de novos planos de marketing, de forma a direccionar esses níveis de *tenrura* às necessidades exigidas por cada um desses mercados. Estimar a qualidade da carne é um processo difícil de efectuar por meios estatísticos, existindo a necessidade de recorrer a painéis de provadores ou a texturómetros, o que implica o abate de animais e consequentes análises do produto, originando custos avultados.

O uso de *RNAs* para a previsão da *tenrura* da carne e posterior selecção de mercados alvo, é comummente referido por toda a comunidade. No entanto, a utilização deste paradigma, necessita de investigação para identificar quais os atributos mais importantes a analisar, bem como, associa-los às preferências dos consumidores, aspectos até ao momento muito pouco estudados e, portanto, pouco claros.

Capítulo 5

Previsão da *Tenrura* da Carne de Cordeiro

Fornece-se uma descrição do trabalho prático realizado. É apresentada a ferramenta utilizada, sendo também descrito todo o processo de Descoberta de Conhecimento em Bases de Dados, que foi conduzido com base na metodologia CRISP-DM, compreendendo as fases: estudo do negócio, estudos dos dados, preparação dos dados, modelação e avaliação. Por último, são apresentados e analisados os resultados obtidos.

5.1 – Introdução

Este capítulo apresenta os resultados de simulações, onde o objectivo primordial é tentar prever a *tenrura* da carne de cordeiro recorrendo a *RNAs*. O processo de *DCBD/DM* de suporte a este trabalho prático foi desenvolvido segundo a metodologia *CRISP-DM* (ver Secção 2.5), compreendo as fases: **Estudo do Negócio**, **Estudo dos Dados**, **Preparação dos Dados**, **Modelação** e **Avaliação**, que serão devidamente explanadas nas próximas secções. A ferramenta escolhida para o suporte computacional foi o ambiente de programação estatístico **R**, o qual será descrito na secção seguinte. Uma vez que o código escrito para as simulações é deveras extenso, foi decidido disponibilizá-lo no Anexo A.

5.2 – Ferramenta Utilizada (R)

A ferramenta **R** é uma plataforma integrada de *software* para manipulação de dados, cálculo e visualização gráfica, que se caracteriza pela sua versatilidade e pela constante participação de uma comunidade de investigadores no seu desenvolvimento [Ihaka e Gentleman, 1996]. Trata-se de uma aplicação de distribuição gratuita e de código aberto

(http://cran.r-project.org), sendo que nos últimos anos se converteu numa ferramenta amplamente utilizada em várias áreas de conhecimento, entre as quais se destaca a *estatística*, o *DM* e as *RNAs* (objecto de estudo neste trabalho). O **R** foi desenvolvido a partir da linguagem S[28], um projecto GNU[29] criado nos laboratórios AT&T Bell no final dos anos 80 por Rick Becker, John Cambers e Allan Walkins. Em 1995, Ross Ihaka e Robert Gentleman, dois professores de estatística da Universidade de Auckland, na Nova Zelândia, iniciaram o "projecto R", com o intuito de desenvolver um programa estatístico poderoso e de domínio público.

O código fonte foi escrito essencialmente em linguagem *C* e algumas rotinas em *Fortran*. Está disponível para a família *Unix*, *MAC OS* e ainda para as diversas versões da família *Windows*. Os ficheiros necessários para a sua instalação, são distribuídos no *site* da *CRAN*[30], em ficheiros binários pré-compilados, juntamente com as instruções de instalação. As actualizações à ferramenta são feitas com bastante regularidade, como se comprova pelas várias versões em que este trabalho foi desenvolvido: 1081, 2001, 2010 e actualmente 2011. A ferramenta também traz alguns pacotes de base que fornecem as principais funcionalidades para a análise de dados. No entanto, dado que o **R** é uma aplicação de código aberto, outras funcionalidades vão sendo desenvolvidas e disponibilizadas para todos os utilizadores (no que é designado por um *package*), sendo possível retirá-las do *CRAN*, através de ficheiros *zip*, ou mesmo de actualizações efectuadas directamente através do **R**.

O **R** é uma linguagem que manipula determinadas estruturas de dados, designados por objectos, daí se tratar de uma linguagem orientada a objectos. Por outras palavras, as variáveis, dados, funções e resultados ficam guardados na memória activa do computador na forma de objectos, com um determinado nome [Grunsky, 2002]. O utilizador pode modificar ou manipular estes objectos com operadores (aritméticos, lógicos e de comparação) ou funções (que por sua vez também são objectos). Assim, o **R** contém uma linguagem simples e flexível, sendo interpretada tal como a linguagem *Java* (e não compilada como a linguagem *C*). Tal significa que os comandos escritos no

[28] A linguagem S-Plus é usada numa versão comercial (para mais informação consultar: www.insightful.com/products/splus/default.html).

[29] Acrónimo recursivo do inglês *Gnu is Not Unix*. Designa o primeiro projecto de partilha gratuita de *software* (www.gnu.org).

[30] Comprehensive R Archive Network (cran.r-project.org). Colecção de *sites* contendo o mesmo material: ficheiros de distribuição, extensões à aplicação, documentação e ficheiros binários, localizados em vários países.

teclado são executados directamente sem necessidade de construir executáveis. Além das vantagens anteriormente referidas, acrescenta-se o facto da sintaxe do **R** ser intuitiva, sendo semelhante à linguagem matemática [Paradis, 2003].

Actualmente, contribuem para o projecto **R** vários investigadores, que se comunicam pela *Internet* através de uma lista de discussão de acesso livre. Deste modo, não só erros de programação são detectados e corrigidos, como também novos módulos, contendo métodos estatísticos recentemente implementados, são regularmente disponibilizados e actualizados na rede.

O **R** possui diversas formas de obter ajuda sobre o programa em geral, ou ainda sobre funções e conjuntos de dados disponíveis. Uma característica muito útil do **R** é a sua ajuda baseada em *HTML*, ou na introdução da função precedida do sinal de interrogação. Dada a crescente comunidade que utiliza a ferramenta, a documentação de suporte é muito vasta, podendo grande parte dela ser obtida no *CRAN*.

Em geral, utiliza-se a ferramenta **R** num ambiente de consola, via comandos, seguindo o modelo de "pergunta-resposta". Também existe uma biblioteca (*Rcmdr*) que dá acesso a uma janela com menus, permitindo o acesso a algumas funções essencialmente gráficas, evitando a digitação de código. Todavia, o ambiente de consola permite um maior controlo por parte do utilizador, embora o seu processo de aprendizagem se torne mais custoso, quando comparado com outros pacotes de *software* comerciais. Uma vez superada esta etapa, o seu uso torna-se numa grande vantagem, pois o **R** permite uma grande flexibilidade em relação às funções pré-existentes, isto é, as funções são editáveis e adaptáveis ao nível das necessidades de cada utilizador/programador [RDCT, 2004].

Assim, são inúmeras as suas vantagens, embora estas nem sempre sejam óbvias na fase inicial de aprendizagem. De modo resumido, destacam-se:

- é um *software* livre e gratuito, podendo ser livremente copiado, instalado e distribuído entre os utilizadores;
- tem um código aberto, permitindo uma rápida correcção de erros e respectiva actualização[31];

[31] Obviamente, esta correcção está dependente da participação em massa dos seus utilizadores e mediante gestão de um grupo denominado "*Grupo Nuclear de Desenvolvimento do R*" (http://cran.r-project.org/contributors).

- código interpretado, permitindo grande flexibilidade na sua programação, transparência de funções e a entrada de dados é altamente didáctica; e
- multi-plataforma, disponível para diversas plataformas, tais como, *Unix*, *Linux*, *Macintosh* e *Windows*.

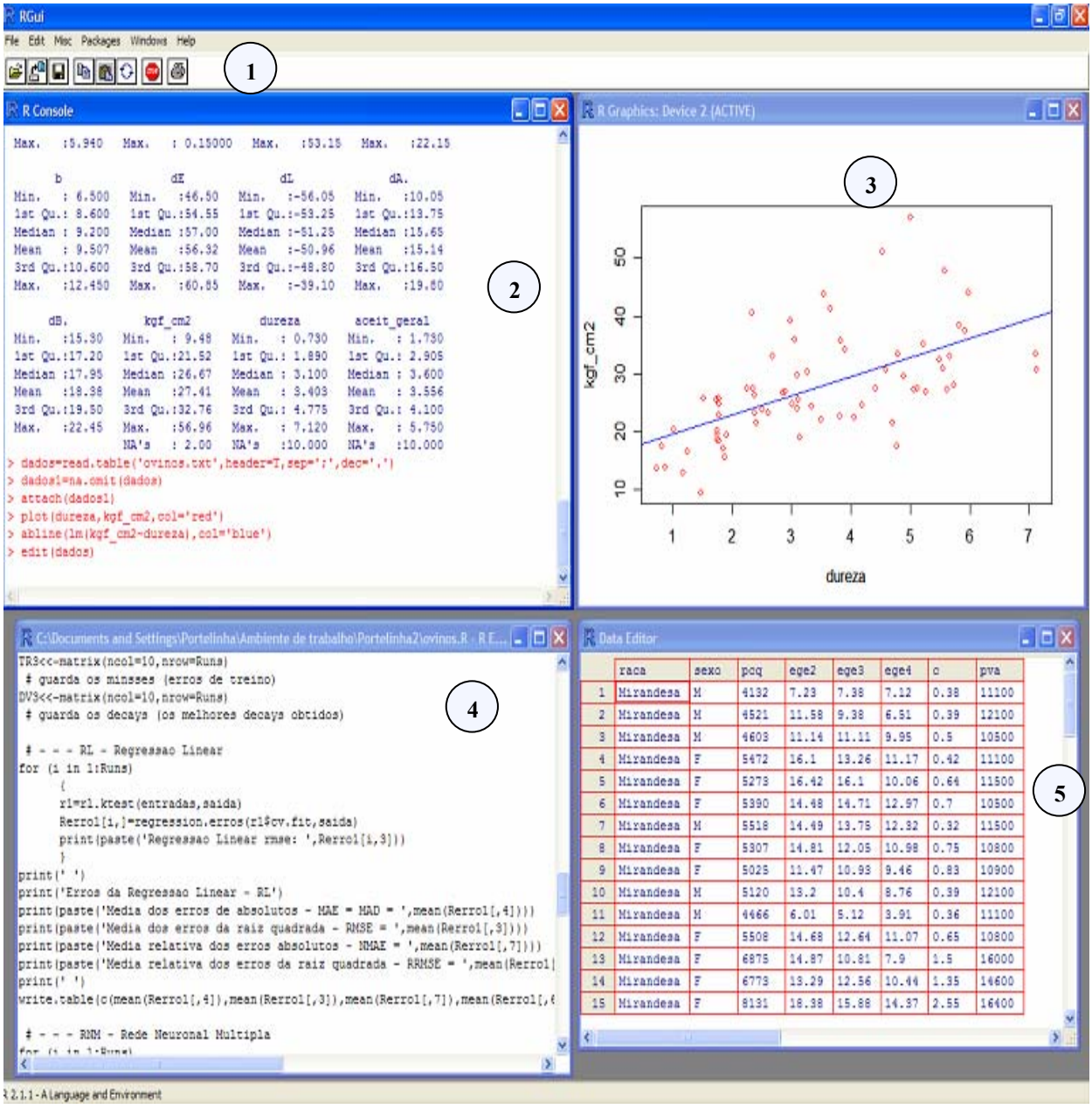


Figura 5.1 – Exemlo do ambiente de programação R.

Como se pode verificar na Figura 5.1, podem existir várias janelas no ambiente de trabalho do **R**, consoante a necessidade da análise dos dados. De seguida, será apresentada uma breve descrição de algumas janelas do **R**:

1. Janela principal do **R**, contém alguns menus que permitem o acesso a operações básicas, tais como, ajuda, actualizações, opções de procura, etc;

2. Consola do **R**, sendo aqui onde são introduzidos todos os comandos que permitem a manipulação de dados e acesso a visualizações gráficas;

3. Representação gráfica da regressão linear;

4. Ficheiro em **R** com código;

5. Tabela com dados.

5.3 – Estudo do Negócio

Para definir o objectivo do processo de *DCBD* é necessário um estudo do domínio da aplicação, de forma a conseguir identificar as informações relevantes disponíveis, bem como as necessidades do utilizador final. Por sua vez, para a execução do processo, torna-se necessário traduzir as necessidades em várias etapas/fases.

Na parte prática deste trabalho foi realizado um estudo, recorrendo ao uso de *RNAs*, onde se pretendeu construir modelos que permitam caracterizar a *tenrura* de carne de cordeiro. Utilizaram-se dados de animais de duas raças certificadas do Nordeste Transmontano Português, de ambos os sexos, com diferentes pesos, cujas amostras, retiradas do músculo *longissimus thoracis*, foram avaliadas quanto à *tenrura*, através de duas abordagens distintas [Arvanitoyannis e Houwelingen-Koukaliaroglou, 2003]:

- um **método instrumental** – Avaliação de força de corte, com a célula *Warner-Bratzler*;
- um **método sensorial** – Como são os casos de provas sensoriais com um painel de provadores.

É importante realçar que ambas as estratégias medem aspectos diferentes relacionados com a *tenrura*: o *instron* mede a força máxima necessária para cortar um núcleo cilíndrico de carne aquecido em água, enquanto que a percepção da *tenrura* pelo consumidor é o resultado de morder e mastigar a carne cozinhada [Oeckel *et al.*, 1999].

Depois de analisado o negócio, definidos os objectivos e as suas necessidades, é necessário converter esse conhecimento numa tarefa de *DM* e num plano inicial para consecução dos objectivos. Assim, torna-se necessário um conhecimento aprofundado

do domínio e dos dados em questão e, então, seleccionar os dados relevantes para posterior construção dos modelos. Esta sequência de passos será analisada de seguida.

5.4 – Estudo dos Dados

Utilizaram-se dados recolhidos pelos investigadores da Escola Superior Agrária do Instituto Politécnico de Bragança relativos à *tenrura* da carne de cordeiro. A base de dados inclui diversos atributos relevantes à estimação da *tenrura*, relativos à carcaça (*e.g.* sexo, peso ou estado de engorda) e à carne (*e.g.* pH ou cor).

Este estudo considerou somente cordeiros com o certificado de *Denominação de Origem Protegida (DOP)*, da região do Nordeste de Portugal, conhecida como Nordeste Transmontano. Os dados foram recolhidos durante o período de um ano, compreendido entre Novembro/2002 e Novembro/2003. Cada registo ou instância da base de dados contém as leituras obtidas para cada animal abatido. Convém referir que a base de dados é de dimensão reduzida, contendo um total de 81 exemplos. Isto deve-se ao facto de o abate de cada animal representa custos elevados (cerca de 6€/kg por carcaça no produtor) mais custos laboratoriais e de transporte.

Dos 81 cordeiros, 45 (24 machos e 21 fêmeas) são da raça autóctone *Churra Galega Bragançana* e 36 (18 machos e 18 fêmeas) da raça autóctone *Churra Galega Mirandesa* (Figura 5.2). Os animais foram criados num sistema tradicional de produção da raça e logo que atingido o peso pré-determinado, eram transportados para o matadouro experimental da Escola Superior Agrária de Bragança, onde eram pesados (**pva**), sendo verificada a raça (**raca**) e o sexo (**sexo**) e permanecendo 24 horas em repouso e jejum, antes do abate. Após o abate, exfoliação e evisceração, as carcaças foram pesadas de forma a obter o peso da carcaça quente (**pcq**), e 1h após o abate foi medido o pH (**pH1**). Seguidamente, as carcaças foram refrigeradas durante 24h após as quais foi medido de novo o pH (**pH24**). À excepção do pH24, que foi medido no laboratório, as restantes medidas foram recolhidas no matadouro.

As carcaças são depois seccionadas sagitalmente, ao longo da coluna vertebral, obtendo-se duas metades. Com a metade direita foi realizada a análise sensorial, enquanto que a metade esquerda foi utilizada para recolher todas as medidas instrumentais. Esta parte da carcaça foi então submetida a 72h de maturação e dividida em 8 peças comerciais. Durante este procedimento foram retiradas, entre outras, as

seguintes medidas: espessura da gordura esternal (**ege2, ege3, ege4**), a medida c (**c**), e as medidas respeitantes à cor (**L**, **a**, **b**, **dE**, , **dA**, **dB**, **dL**). Estes atributos foram todos recolhidos em laboratório.

5.4.1 – Análise Intrumental (*Warner-Bratzler*)

A *tenrura* (**kgf_cm2**/**WBS**) foi determinada utilizando o *instron* equipado com a célula *Warner-Bratzler*, tendo sido realizadas várias repetições para cada amostra. Esta medida apenas pode ser obtida em laboratório, usualmente em carne cozinhada, não antes de 72 horas após o abate. Assim, da metade esquerda foi removido o músculo *longissimus thoracis* e colocado no interior de um saco, pré-aquecido em banho-maria a 70º C. Após o músculo atingir a temperatura interna de 70ºC, foi arrefecido e procedeu-se ao corte do músculo em cubos com 2x1x1cm de aresta, que posteriormente se colocavam na célula *Warner-Bratzler* para avaliação. Nesta máquina efectuaram-se cortes perpendiculares ao sentido das fibras.

5.4.2 – Análise Sensorial

Para a obtenção dos valores sensoriais foi idealizado um esquema mais elaborado, sendo necessárias as seguintes etapas: *i)* **formação do painel** de provadores e *ii)* **preparação** e **apresentação das amostras**. A formação do painel de provadores, que neste estudo ficou constituído por 12 elementos efectivos e 3 suplentes, consistiu na realização das seguintes fases: **recrutamento**, (efectuado por contacto directo e aleatório de pessoas pertencentes ao Instituto Politécnico de Bragança, aos quais foi apresentado o objecto de estudo, bem como a necessidade de disponibilidade e interesse), **selecção** (foram efectuados testes descritivos e determinantes de modo a seleccionar as pessoas mais aptas para este tipo de análises) e **treino** (após a selecção das pessoas, foram efectuados testes gerais e específicos, de modo a adaptar e aperfeiçoar as pessoas relativamente aos parâmetros de estudo).

Relativamente à **preparação** e **apresentação das amostras**, foi utilizado o músculo *longissimus thoracis*. Um dia antes da realização da prova, colocaram-se as amostras a descongelar a 4º C. Seguidamente os músculos foram envolvidos por folha de alumínio, evitando assim a secagem da superfície e preparados num forno até atingir uma temperatura interna de 70 – 80 ºC. Imediatamente após se atingir a temperatura

desejada, o músculo foi cortado em cubos de 2x2x0,5cm de aresta, perpendicularmente ao sentido das fibras musculares, envolvidos em papel de alumínio e colocados em estufas, reguladas para 100ºC, para a manutenção da temperatura das amostras.

A codificação das amostras foi aleatória, com números de três dígitos, variando de avaliador para avaliador e de sessão para sessão de modo a prevenir influências (acidentais ou deliberadas). Os provadores avaliaram as amostras de acordo com a ordem estabelecida pelo coordenador das provas. Foram informados da necessidade de limpar a boca no início e entre as várias amostras da sessão com água e pedaços de maçã do tipo *golden*. As condições ambientais envolventes das provas eram idênticas de provador para provador e de sessão para sessão.

5.4.3 – Descrição dos Dados

A base de dados obtida após a preparação dos dados, contém um total de 20 atributos e 81 registos. O conjunto total dos dados está disponível no Anexo B. A Tabela 5.1, apresenta a totalidade dos atributos da base de dados de cordeiro e respectivo grupo de análise.

Atributos Gerais	**pH**	**Cor**	**Análise Instrumental** *(Warner-Bratzler)*	**Análise Sensorial**
raca	pH1	L	kgf_cm2	dureza
sexo	pH24	a		
pcq	dpH	b		
ege2		dE		
ege3		dL		
ege4		dA		
c		dB		
pva				

Tabela 5.1 – Atributos da base de dados dos cordeiros.

De seguida, para cada um desses atributos e grupos de análise, será feita uma breve descrição:

- **raca** – Foram analisadas 2 tipos de raças autóctones "Churra Galega Bragançana" e "Churra Galega Mirandesa". São raças da região do Nordeste Transmontano;
- **sexo** – Macho ou Fêmea;
- **pcq** – Peso da Carcaça Quente, sendo o peso da carcaça obtido logo após o abate. O abate foi efectuado por degola, após insensibilização do animal com choupa. De seguida procedeu-se à esfola e evisceração. Finalmente foi determinado o peso da carcaça quente. Valor em gramas;
- **ege2, ege3** e **ege4** – Espessura da Gordura Esternal ao nível da segunda, terceira e quarta esternebra, medida obtida no laboratório com um paquímetro. Valor obtido em mm;
- **c** – Medida C, relativa à profundidade da gordura subcutânea logo acima do músculo *longissimus thoracis*, sensivelmente ao meio deste e no seu desenvolvimento. É realizada na superfície da secção da meia carcaça, efectuada entre a primeira e a segunda vértebra lombar. Medida obtida no laboratório por um paquimetro. Valor obtido em mm;
- **pva** – Peso Vivo do Animal, sendo este pesado numa balança antes do abate. Valor em Kg.

O **pH** é o logaritmo negativo da concentração de protões de uma solução. O seu valor expressa-se numa escala de 0 (ácido) a 14 (básico). A sua evolução durante o período *postmortem* e o valor final do mesmo influenciam as características organolépticas da carne, podendo modificar quer a *tenrura* quer a cor da carne. O valor é obtido por um medidor de pH, ao nível da 12ª costela, Assim tem-se:

- **pH1** – Nível de Acidez, medido 1 hora após o abate, no matadouro;
- **pH24** – Nível de Acidez, medido 24 horas após o abate, no laboratório;
- **dpH** – Diferença do Nível de Acidez, entre o pH1 e o pH24.

A **Cor** da carne é um dos factores mais relevantes nos quais se fixa o consumidor, já que é o primeiro que pode perceber e, às vezes, a única informação de que dispõe para emitir o seu juízo com respeito a ela, apesar da correlação com a qualidade geral ser limitada. A *CIE* (*Comission International de L' Eclairage*) define a cor percebida como

o atributo visual que se compõe de uma combinação qualquer de conteúdos cromáticos e acromáticos. A cor de um produto resulta da capacidade de reflexão pela matéria das diferentes radiações do espectro visível. Os seus atributos são L, a, b, dE, dL, dA, dB, sendo seguidamente descritos:

- **L** – Luminosidade julgada em relação com a luminosidade de outro estímulo que aparece como branco ou transparente. As variações de L vão do branco (100) ao negro (0);
- **a** – Índice de Vermelho, correspondendo ao atributo da sensação visual segundo o qual o estímulo aparece similar a uma das cores percebidas, vermelho (para valores positivos) e verde (para valores negativos), sem limite numérico específico;
- **b** – Índice de Amarelo, atributo da sensação visual segundo o qual o estímulo aparece similar a uma das cores percebidas, Amarelo (para valores positivos) e Azul (para valores negativos), sem limite numérico específico;
- **dE** – Diferença de Cor Total, sendo um valor único que tem em conta as diferenças entre o "L" o "a" e "b" da amostra e o padrão (branco);
- **dL** – Diferencial de Luminosidade, indica a diferença em termos de luminosidade entre a amostra e o branco. Caso o dL seja positivo significa que a amostra é mais luminosa do que o branco;
- **dA** – Diferencial de Índice de Vermelho, indica a diferença em termos de índice de vermelho entre a amostra e o branco. Caso o dA seja positivo significa que a amostra é mais vermelha do que o branco;
- **dB** – Diferencial de Índice de Amarelo, indica a diferença em termos de índice de amarelo entre a amostra e o branco. Caso o dB seja positivo significa que a amostra é mais amarela do que o branco.

No que diz respeito às variáveis de saída:

- **WBS** – força de corte *Warner-Bratzler*. Trata-se do índice de maior objectividade para medir a *tenrura* da carne. A força de corte WBS regista a força (em kg) requerida para cortar uma amostra de carne com 1cm de espessura. Os valores mais baixos sugerem carne tenra, enquanto leituras elevadas sugerem dureza;

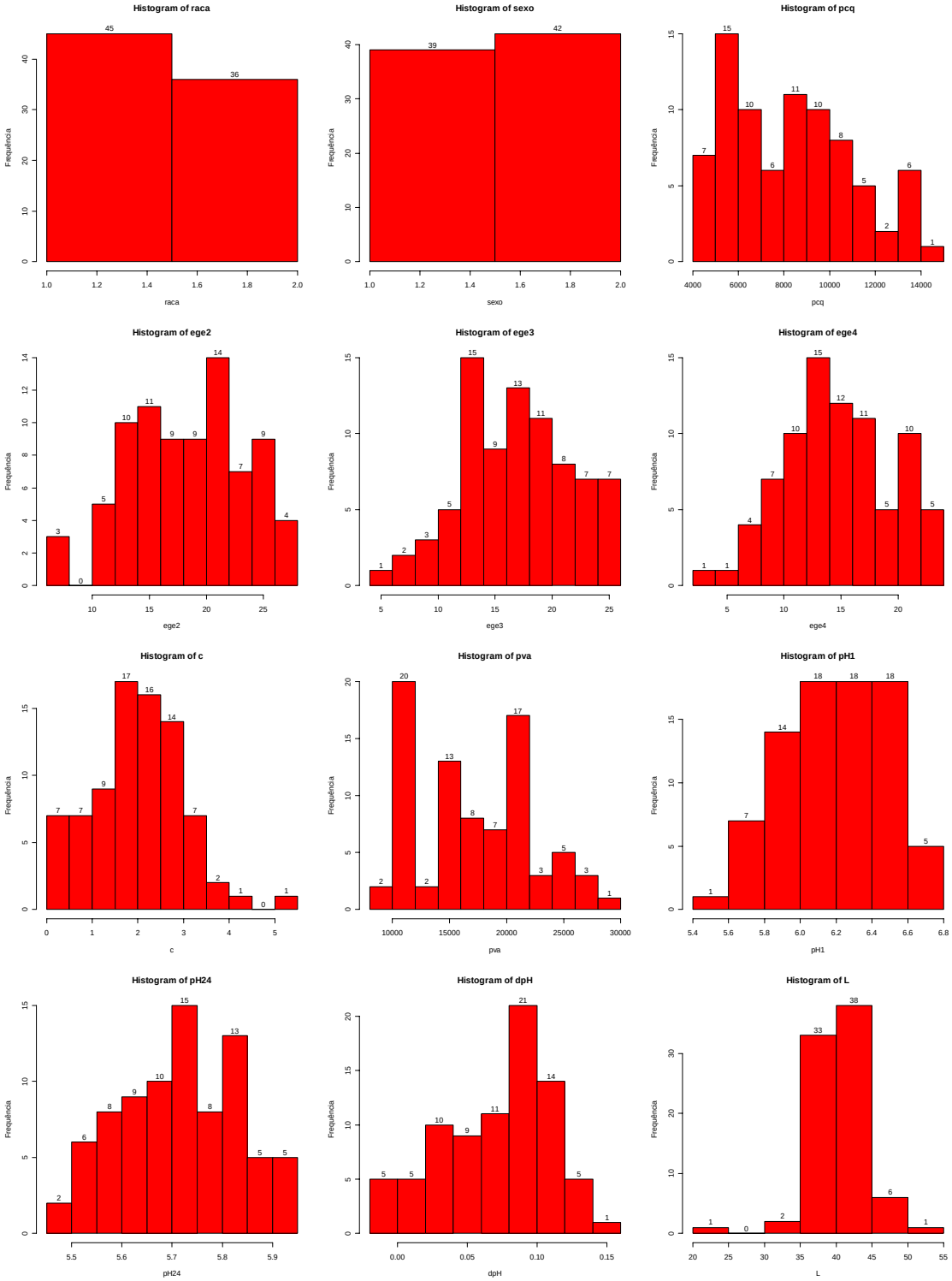
- **PAS** – Painel de Análise Sensorial. A *dureza* é um dos primeiros critérios determinantes na qualidade para o consumidor e pode definir-se como a capacidade da carne para deixar-se cortar e mastigar. Cada amostra foi avaliada de 0 (a mais tenra) até 10 (a mais dura), conforme o processo descrito anteriormente. O **PAS** foi medido como a média dos valores dados pelo painel.

Na Figura 5.2 é apresentado um breve sumário dos dados. Entre outras descrições são apresentados os valores mínimos, máximos, médios e os valores em falta de todos os atributos dos dados da carne de cordeiro.

```
> summary(dados[1:20])
         raca     sexo         pcq              ege2             ege3             ege4
 Bragançano:45   F:39   Min.   : 4132   Min.   : 6.01   Min.   : 5.12   Min.   : 3.91
 Mirandesa :36   M:42   1st Qu.: 5921   1st Qu.:14.48   1st Qu.:12.89   1st Qu.:11.07
                        Median : 8303   Median :18.71   Median :17.31   Median :14.27
                        Mean   : 8333   Mean   :18.27   Mean   :16.83   Mean   :14.78
                        3rd Qu.:10125   3rd Qu.:21.82   3rd Qu.:20.93   3rd Qu.:17.98
                        Max.   :14845   Max.   :27.81   Max.   :25.56   Max.   :23.87

       c               pva             pH1             pH24             dpH
 Min.   :0.320   Min.   : 9800   Min.   :5.540   Min.   :5.490   Min.   :-0.02000
 1st Qu.:1.350   1st Qu.:11600   1st Qu.:6.000   1st Qu.:5.640   1st Qu.: 0.05000
 Median :2.030   Median :16600   Median :6.210   Median :5.720   Median : 0.09000
 Mean   :1.995   Mean   :17151   Mean   :6.202   Mean   :5.719   Mean   : 0.07605
 3rd Qu.:2.620   3rd Qu.:21100   3rd Qu.:6.450   3rd Qu.:5.810   3rd Qu.: 0.10000
 Max.   :5.080   Max.   :28400   Max.   :6.790   Max.   :5.940   Max.   : 0.15000

       L               a               b               dE              dL
 Min.   :23.20   Min.   :11.50   Min.   : 6.500   Min.   :46.50   Min.   :-56.05
 1st Qu.:38.55   1st Qu.:15.10   1st Qu.: 8.600   1st Qu.:54.55   1st Qu.:-53.25
 Median :40.86   Median :17.25   Median : 9.200   Median :57.00   Median :-51.25
 Mean   :40.97   Mean   :16.71   Mean   : 9.507   Mean   :56.32   Mean   :-50.96
 3rd Qu.:43.65   3rd Qu.:18.15   3rd Qu.:10.600   3rd Qu.:58.70   3rd Qu.:-48.80
 Max.   :53.15   Max.   :22.15   Max.   :12.450   Max.   :60.85   Max.   :-39.10

      dA.             dB.           kgf_cm2          dureza
 Min.   :10.05   Min.   :15.30   Min.   : 9.48   Min.   : 0.730
 1st Qu.:13.75   1st Qu.:17.20   1st Qu.:21.52   1st Qu.: 1.890
 Median :15.65   Median :17.95   Median :26.67   Median : 3.100
 Mean   :15.14   Mean   :18.38   Mean   :27.41   Mean   : 3.403
 3rd Qu.:16.50   3rd Qu.:19.50   3rd Qu.:32.76   3rd Qu.: 4.775
 Max.   :19.80   Max.   :22.45   Max.   :56.96   Max.   : 7.120
                                 NA's   : 2.00   NA's   :10.000
```

Figura 5.2 – Sumário dos dados da carne de cordeiro.

Na Figura 5.3 podem ver-se os histogramas dos dados respeitantes à carne de cordeiro, dos atributos previamente explanados.


Histogram of raca
raca
Frequência
Histogram of sexo
sexo
Frequência
Histogram of pcq
pcq
Frequência
Histogram of ege2
ege2
Frequência
Histogram of ege3
ege3
Frequência
Histogram of ege4
ege4
Frequência
Histogram of c
c
Frequência
Histogram of pva
pva
Frequência
Histogram of pH1
pH1
Frequência
Histogram of pH24
pH24
Frequência
Histogram of dpH
dpH
Frequência
Histogram of L
L
Frequência

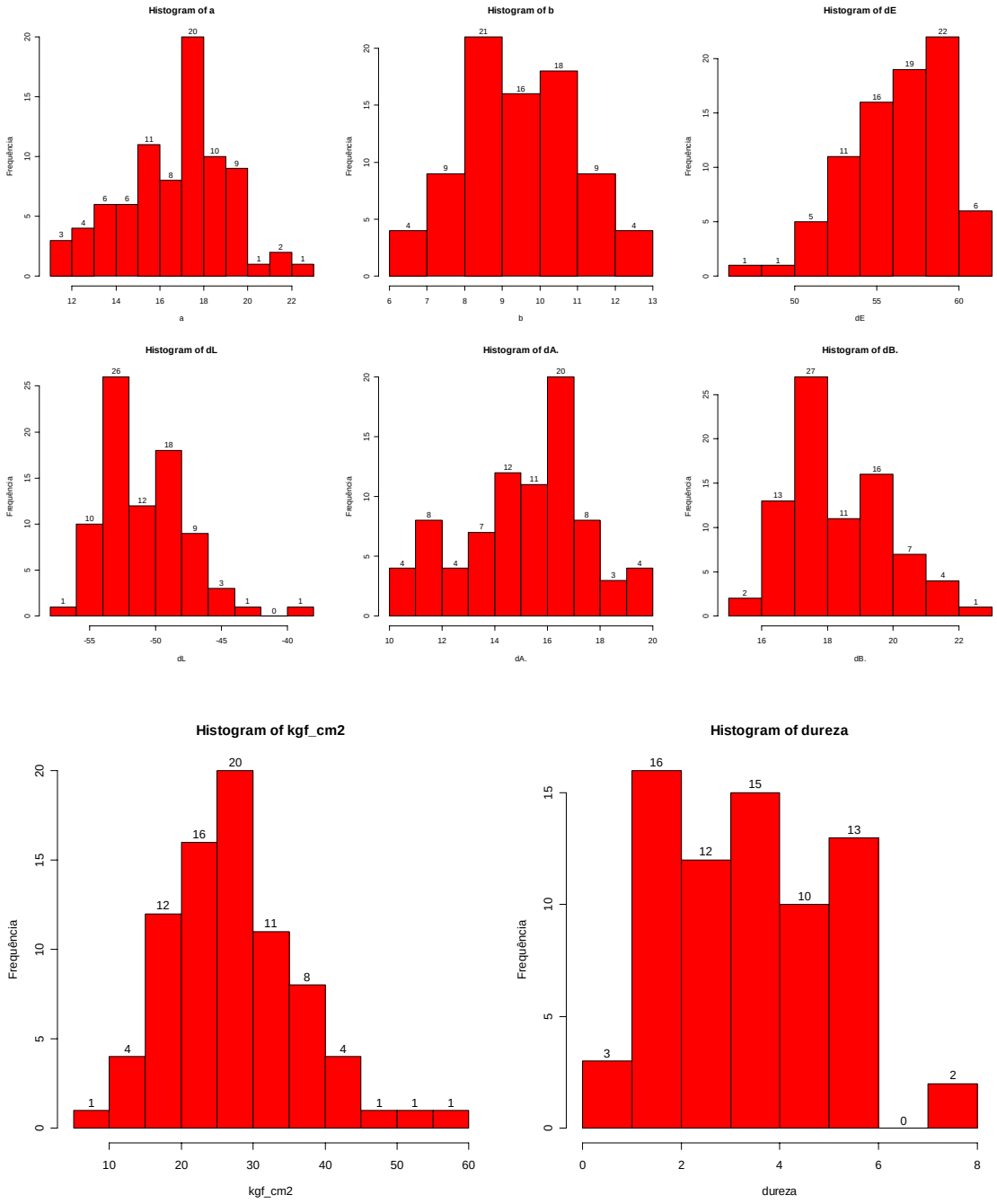


Figura 5.3 – Histogramas dos atributos da carne de cordeiro.

5.5 – Preparação dos Dados

A preparação dos dados é uma etapa de grande importância para todo o processo de *DCBD*, consistindo num conjunto de actividades destinadas a obter a base de dados final, englobando essencialmente a **Selecção dos Dados**, **Limpeza** e **Transformação**.

Para o sucesso deste processo é necessário que os dados tenham sido correctamente seleccionados, corrigidos e transformados.

5.5.1 – Selecção dos Dados

Nesta fase, existiu o cuidado de discutir com os especialistas da área sobre a quantidade/qualidade dos dados, uma vez que se trata de um factor de extrema importância. Para uma solução satisfatória da *DCBD* é necessário uma amostragem de casos do problema, ou seja, é necessário garantir a existência de exemplos que validem o modelo para o efectivo funcionamento do processo. Além do volume de dados, o seu conjunto deve conter apenas os dados relevantes, que permitam determinar o padrão de comportamento. Estes requisitos foram garantidos por especialistas da Escola Superior Agrária do Instituto Politécnico de Bragança.

A base de dados inicial foi disponibilizada no formato "*.xls*" do *Microsoft Excel* e continha uma quantidade deveras elevada de atributos, encontrando-se separados por várias folhas, cada uma referente a um grupo de informação da análise instrumental da carne: medidas, pH, cor e medidas do *instron*. Posteriormente foram disponibilizados os dados da análise sensorial, os quais continham uma estrutura idêntica aos dados instrumentais.

5.5.2 – Limpeza dos Dados

Qualquer base de dados pode conter vários problemas na qualidade dos seus dados. Por isso, para o perfeito funcionamento do processo de *DCBD*, é necessário assegurar que os dados utilizados no processo estejam correctos.

Neste caso de estudo apenas foram removidas algumas inconsistências, como era o caso de atributos com valores totalmente distintos, que foram prontamente explicados pelos especialistas da área e transformados para valores compreensíveis. Procedeu-se ainda à eliminação de registos duplicados e uniformização, resultando num conjunto de dados relevantes para aplicação dos modelos.

5.5.3 – Transformação dos Dados

Nesta secção irá proceder-se à apresentação dos resultados obtidos após a aplicação de alguns tratamentos estatísticos sobre os mesmos. Pretende-se obter resultados que permitam tirar conclusões acerca dos dados, bem como, acerca da forma de como estes se relacionam entre si. As duas tabelas de dados foram integradas, numa única tabela. Ainda nesta fase, e dado que a quantidade de atributos era considerável, foram calculadas as correlações lineares entre os vários atributos, sendo que aqueles que continham elevadas correlações foram descartados.

Na interpretação de correlações há dois aspectos que podem complicar a explicação de resultados: *i)* a existência de valores extremos e; *ii)* a ocorrência de duas ou mais sub-populações [Ferreira, 2000]. Nos dados deste trabalho, não se verificou nenhuma destas situações, daí que o passo seguinte foi o cálculo das correlações entre todos os atributos da base de dados de cordeiros. Para não comprometer os resultados finais, foram apenas descartados os atributos com uma correlação superior a 90%. O código criado encontra-se no Anexo C. Este permite obter dois tipos de análises: percentagem de correlação entre todos os atributos apresentado (Tabela 5.2); e apresentação dos atributos que têm um correlação acima dos 90% (Tabela 5.3). Um exemplo do comando utilizado pelo R para o cálculo das correlações é: **`cor(pcq,pva)`**.

	raca	sexo	pcq	ege2	ege3	ege4	c	pva	pH1	pH24	dpH	L	a	b	dE	dL	dA.	dB.	kgf_cm2	dureza
raca	***100%***	3%	-22%	-14%	-17%	-18%	-35%	-2%	49%	14%	50%	7%	-10%	-25%	-11%	6%	-4%	-19%	-14%	-40%
sexo	3%	***100%***	-15%	-14%	-10%	-11%	-16%	2%	13%	17%	8%	16%	-22%	13%	-32%	27%	-14%	6%	1%	-3%
pcq	-22%	-15%	***100%***	85%	86%	87%	81%	**94%**	-21%	16%	-32%	-36%	33%	-29%	54%	-53%	29%	-23%	44%	22%
ege2	-14%	-14%	85%	***100%***	**97%**	**93%**	72%	83%	-18%	15%	-29%	-43%	36%	-29%	56%	-56%	34%	-20%	40%	20%
ege3	-17%	-10%	86%	**97%**	***100%***	**97%**	75%	83%	-19%	16%	-30%	-41%	38%	-28%	56%	-55%	35%	-18%	41%	23%
ege4	-18%	-11%	87%	**93%**	**97%**	***100%***	75%	82%	-18%	16%	-29%	-35%	41%	-25%	56%	-56%	37%	-16%	37%	19%
c	-35%	-16%	81%	72%	75%	75%	***100%***	73%	-16%	23%	-29%	-37%	25%	-26%	48%	-50%	21%	-29%	40%	30%
pva	-2%	2%	**94%**	83%	83%	82%	73%	***100%***	-4%	31%	-20%	-38%	23%	-40%	48%	-52%	24%	-33%	49%	17%
pH1	49%	13%	-21%	-18%	-19%	-18%	-16%	-4%	***100%***	50%	**90%**	4%	-23%	-13%	-24%	15%	-18%	-25%	9%	4%
pH24	14%	17%	16%	15%	16%	16%	23%	31%	50%	***100%***	8%	-27%	-3%	-28%	13%	-24%	3%	-43%	28%	19%
dpH	50%	8%	-32%	-29%	-30%	-29%	-29%	-20%	**90%**	8%	***100%***	18%	-27%	0%	-36%	31%	-24%	-7%	-2%	-2%
L	7%	16%	-36%	-43%	-41%	-35%	-37%	-38%	4%	-27%	18%	***100%***	-36%	41%	**-93%**	73%	-41%	39%	-32%	-18%
a	-10%	-22%	33%	36%	38%	41%	25%	23%	-23%	-3%	-27%	-36%	***100%***	24%	68%	-49%	**91%**	28%	-23%	-20%
b	-25%	13%	-29%	-29%	-28%	-25%	-26%	-40%	-13%	-28%	0%	41%	24%	***100%***	-27%	48%	13%	77%	-40%	-15%
dE	-11%	-32%	54%	56%	56%	56%	48%	48%	-24%	13%	-36%	**-93%**	68%	-27%	***100%***	68%	66%	-26%	22%	5%
dL	6%	27%	-53%	-56%	-55%	-56%	-50%	-52%	15%	-24%	31%	73%	-49%	48%	68%	***100%***	-48%	51%	-33%	-12%
dA.	-4%	-14%	29%	34%	35%	37%	21%	24%	-18%	3%	-24%	-41%	**91%**	13%	66%	-48%	***100%***	21%	-19%	-16%
dB.	-19%	6%	-23%	-20%	-18%	-16%	-29%	-33%	-25%	-43%	-7%	39%	28%	77%	-26%	51%	21%	***100%***	-48%	-24%
kgf_cm2	-14%	1%	44%	40%	41%	37%	40%	49%	9%	28%	-2%	-32%	-23%	-40%	22%	-33%	-19%	-48%	***100%***	59%
dureza	-40%	-3%	22%	20%	23%	19%	30%	17%	4%	19%	-2%	-18%	-20%	-15%	5%	-12%	-16%	-24%	59%	***100%***

Tabela 5.2 – Percentagem de correlação entre todos os atributos da base de dados.

Como se pode verificar na Tabela 5.2 existem algumas correlações fortes entre os dados. Esses atributos foram então retirados da base de dados, uma vez que foi considerado que não iriam acrescentar uma mais valia ao trabalho. Na Tabela 5.3 são apresentados os atributos que foram eliminados e respectivas correlações.

	ege3	ege4	pva	dpH	L	dA.
pcq			**94%**			
ege2	**97%**	**93%**				
ege3		**97%**				
ege4	**97%**					
pH1				**90%**		
a						**91%**
dE					**-93%**	

Tabela 5.3 – Atributos com correlações acima dos 90%.

A Tabela 5.4 apresenta os atributos considerados no estudo de *DM*, mostrando ainda a descrição dos atributos considerados neste trabalho, assim como o respectivo domínio. Daqui emerge a primeira vantagem desta análise estatística: em vez dos 20 atributos iniciais, a base de dados ficou reduzida a apenas 14, o que permitiu a eliminação de 6 atributos.

Atributo	***Descrição***	***Domínio***
raca	Raça do animal	{1, 2}[a]
sexo	Sexo do animal	{1, 2} [b]
pcq	Peso da carcaça Quente (kg)	[4.1, 14.8]
ege2	Espessura da Gordura Esternal ao nível da 2ª esternebra (mm)	[6.0, 27.8]
c	Espessura da gordura subcutânea acima do músculo longissimus (mm)	[0.3, 5.1]
pH1	pH medido 1 hora após abate	[5.5, 6.8]
pH24	pH medido 24 horas após abate	[5.5, 5.9]
a	Índice de vermelho	[11.5, 22.2]
b	Índice de amarelo	[6.5, 12.5]
dE	Diferença de cor total (engloba L, a e b)	[46.5, 60.9]
dL	Diferencial de luminosidade quando comparado com um padrão	[-56.1, -39.1]
db	Diferencial do índice de amarelo	[15.3, 22.5]
WBS	*Warner-Bratzler Shear force* (kg/cm2) – Força de corte	[9.5, 57.0]
PAS	Painel de Análise Sensorial	[0.7, . . . , 7.1]

[a] 1 - Churra Galega Bragançana, 2 - Churra Galega Mirandesa

[b] 1 - Macho, 2 - Fêmea

Tabela 5.4 – Atributos dos dados objecto de estudo.

Convém referir que a base de dados original continha 2 valores em falta na variável **kgf_cm2/WBS** e 10 valores na variável **dureza/PAS**. Assim, e para finalizar a fase de Preparação dos Dados foram criadas duas novas bases de dados, retirando os registos com valores desconhecidos (**`dados1=na.omit(dados)`**). A primeira contém 79 linhas e será usada para prever os valores do *instron* (**WBS**), enquanto a segunda tem 71 exemplos, sendo usada para a previsão sensorial (**PAS**).

Outras transformações foram efectuadas nos dados, mas essas serão descritas em pormenor na secção dedicada à explanação dos modelos de *RNAs*. Finalizada a etapa de preparação dos dados, obteve-se um conjunto de dados prontos para serem analisados pelos algoritmos de *DM*.

5.6 – Modelos de Aprendizagem

Nesta fase, são escolhidos os modelos mais apropriados aos objectivos pretendidos para a descoberta de conhecimento. A escolha do modelo representa a fase central da metodologia e consiste em aplicar uma técnica de modelação sobre o conjunto de dados. Como alguns modelos têm requisitos específicos na formatação dos dados, houve a necessidade de regressar à fase anterior para efectuar a normalização e a conversão dos atributos não numéricos.

Seja D uma base de dados de regressão com $k \in \{1,...,N\}$ exemplos, cada um correspondendo um vector de entrada $\left(x_1^k,....,x_I^k\right)$ com um dado objectivo y_k. O erro, para um determinado k, é então dado por: $e_k = y_k - \hat{y}_k$, onde $\hat{y}_k$, representa o valor previsto para o padrão de entrada k. O desempenho de uma regressão é calculado por métrica global, nomeadamente: *Erro Absoluto Médio (EAM)*, *Erro Absoluto Médio Relativo (EAMR)*, *Raiz da Média do Quadrado dos Erros (RMQE)* ou *Raiz da Média do Quadrado dos Erros Relativo (RMQER)*. Estes podem ser calculados de acordo com as equações:

$$\begin{aligned} EAM &= 1/N \times \sum_{i=1}^{N} |y_i - \hat{y}_i| \\ EAMR &= 1/N \times EAM / \sum_{i=1}^{N} |y_i - \bar{y}_i| \times 100(\%) \\ RMQE &= \sqrt{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 / N} \\ RMQER &= RMQE / \sqrt{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2 / N} \times 100(\%) \end{aligned} \qquad \textbf{(7)}$$

- *Erro Absoluto Médio (EAM)* – É a diferença entre os valores da saída real (observado) e estimado (previsto pelo modelo);
- *Erro Absoluto Médio Relativo (EAMR)* – Rácio entre o *EAM* obtido pelo modelo de regressão e o *EAM* relativo à média dos valores da saída;
- *Raiz da Média do Quadrado dos Erros (RMQE)* – Raiz quadrada da soma de quadrados dos resíduos a dividir pelo total de resíduos;
- *Raiz da Média do Quadrado dos Erros Relativo (RMQER)* – Representa a proporção entre o *RMQE* do modelo com o *RMQE* obtido via a média dos valores da saída.

Em todos estes erros, valores mais baixos correspondem a melhores modelos preditivos, já que, quanto menor for a proporção de erro ou resíduo, melhor é o ajuste dos dados em relação ao modelo. As medidas *EAMR* e *RMQER*, têm a vantagem de ser independentes da escala, onde 100% significa que o método de regressão tem desempenho similar a uma previsão simples via a média dos valores.

Um modelo de *Regressão Múltipla* é definido pela equação [Hastie *et al.*, 2001]:

$$\hat{y} = \beta_0 + \sum_{i=1}^{I} \beta_i x_i \qquad \textbf{(8)}$$

onde $\{x_1, \ldots, x_I\}$ denota o conjunto das variáveis de entrada (quando $I = 1$ é conhecida como *Regressão Linear*) e $\{\beta_0,...,\beta_1\}$ o conjunto de parâmetros a ser ajustado, usualmente aplicando o algoritmo dos mínimos quadrados. Devido à sua natureza aditiva, este modelo é fácil de interpretar e tem sido largamente utilizado em aplicações de regressão.

Por outro lado, as *Redes Neuronais* denotam um conjunto de modelos conexionistas inspirados no comportamento do sistema nervoso central dos seres vivos. Em particular, o *perceptão multicamada* é a arquitectura mais popular, onde os neurónios são agrupados em camadas e apenas existem conexões unidireccionais [Haykin, 1999]. Neste estudo, serão utilizadas redes deste tipo com conexões de *bias*, uma camada intermédia com H unidades de processamento (ou neurónios) e funções de activação logísticas. Para o único neurónio de saída foi adoptada a função de activação linear, uma solução que é comum em tarefas de regressão, uma vez que a variável de saída pode ter valores fora do contradomínio da função logística ([0, 1]) [Hastie *et al.*, 2001]. Assim,

cada tarefa de regressão (**WBS** e **PAS**) será modelada por uma rede diferente, sendo o modelo geral dado pela equação:

$$\hat{y} = w_{o,0} + \sum_{j=I+1}^{o-1} f\left(\sum_{i=1}^{I} x_i w_{j,i} + w_{j,0}\right) w_{o,i} \qquad \textbf{(9)}$$

onde $w_{j,i}$ denota o peso da conexão desde o neurónio j até à unidade i (se j=0 então é uma conexão *bias*), o corresponde à unidade de saída, f à função logística $\left(\frac{1}{1+e^{-x}}\right)$, e I ao número de neurónios de entrada. A Figura 5.4 descreve a arquitectura neuronal escolhida.

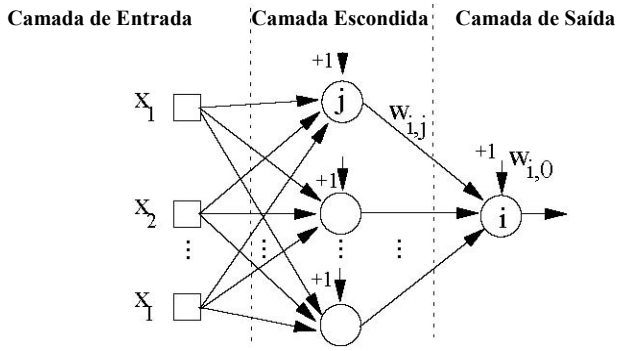


Figura 5.4 – Um *Perceptrão Multicamada* genérico com uma camada intermédia.

A aprendizagem supervisionada é alcançada por um ajustamento iterativo dos pesos das conexões da rede (o algoritmo de treino), de modo a minimizar uma função de erro (tipicamente a soma dos erros quadrados), calculada sobre os *exemplos de treino* (ou *casos*). Antes de alimentar as *Redes Neuronais*, os dados necessitam de ser pré-processados. Neste estudo, todos os atributos foram padronizados para uma média "zero" e desvio padrão "um", aplicando a normalização:

$$z = \frac{x - \bar{x}}{s_x} \qquad \textbf{(10)}$$

onde $\bar{x}$ denota a média da variável no conjunto de treino e S_x o desvio padrão correspondente.

A arquitectura neuronal precisa de ser estabelecida *a priori*, antes do treino. Por conseguinte, o desempenho será sensível a esta escolha: uma pequena rede fornecerá capacidades limitadas de aprendizagem, enquanto que uma rede de elevada dimensão tenderá a sobre-ajustar os dados, perdendo capacidade de generalização. Para resolver

este obstáculo, é comum usar um elevado número de neurónios intermédios (*H*), sendo a *Rede Neuronal* treinada com uma *constante de decaimento* [Hastie *et al.*, 2001]. Trata-se de um procedimento de regularização, onde um valor constante (λ) funciona como uma penalização que reduz os valores dos pesos da rede. Em particular, os pesos mais baixos tenderão a aproximar-se do valor nulo. Sob este esquema, o parâmetro crucial é a escolha do valor de λ, que permite controlar a complexidade da rede. Neste trabalho, este hiper-parâmetro será ajustado de modo automático por um procedimento de *procura em grelha*.

Para uma determinada rede, os pesos iniciais serão estabelecidos aleatoriamente dentro do intervalo [-0.7, +0.7] [Hastie *et al.*, 2001]. A seguir, procede-se ao treino da rede, sendo este parado quando o declive do erro de treino estiver perto de zero ou após um máximo de Max_{it} iterações [Prechelt, 1998]. Após o treino, a saída é desnormalizada, aplicando o inverso da Equação 10. Depois, é realizada a *Análise de Sensibilidade*, sendo medida como a variância (V_a) produzida na saída ($\hat{y}$) quando o atributo de entrada (a) é movido através da sua amplitude total [Kewley *et al.*, 2000]:

$$\begin{aligned} V_a &= \sum_{i=1}^{L} \left(\hat{y}_i - \bar{\hat{y}}\right) \Big/ (L-1) \\ R_a &= V_a \Big/ \sum_{j=1}^{I} V_j \times 100(\%) \end{aligned} \qquad \textbf{(11)}$$

onde I denota o número de atributos de entrada e R_a a importância relativa do atributo. A saída $\hat{y}_i$, é obtida mantendo todas as variáveis de entrada nos seus valores médios. A excepção é x_a, que varia ao longo da sua escala de valores possíveis, com L níveis. Neste trabalho, L foi estabelecido em 2 para os atributos binários e 7 para as entradas contínuas (**`Levels<-c(1,1,6,6,6,6,6,6,6,6,6,6)`**).

Uma vez que a função de erro da *Rede Neuronal* é não convexa, podem existir diversos mínimos locais, pelo que a qualidade da rede treinada dependerá da escolha dos pesos iniciais. Uma forma de contornar este problema, consiste em aplicar R execuções do algoritmo de treino para cada configuração neuronal, sendo esta inicializada de modo aleatório. No final deste processo, é escolhida a rede com o menor erro penalizado. Esta configuração será designada de *Rede Neuronal Múltipla (RNM)*. Outra opção consiste em usar um *Conjunto de Redes Neuronais (CRN)*, onde R redes também são treinadas a partir de pesos aleatórios iniciais. No entanto, a previsão final passa a ser dada pela média das previsões individuais de cada rede. De realçar que este método de construção

de *Conjuntos de Modelos* é conhecido pelo termo *Injecting Randomness* [Dietterich, 2000].

5.7 – Avaliação dos Modelos

Os modelos escolhidos na fase anterior são agora aplicados, de forma a garantir que vão ao encontro dos objectivos definidos inicialmente. Vários modelos foram avaliados e revistos, com a finalidade de encontrar factores importantes que tenham sido omitidos nas fases anteriores. Para aferição dos resultados alcançados foram utilizadas as métricas *EAM*, *EAMR*, *RMQE* e *RMQER* (Equação 7), de uso comum em problemas de previsão.

Todas as experiências deste trabalho foram conduzidas usando o ambiente estatístico **R** (Figura 5.1) [RDCT, 2004]. Relativamente à configuração da *Rede Neuronal*, o ambiente R usa o algoritmo de treino *BFGS*, da família do método *quasi-Newton*. Após algumas experiências preliminares, o número máximo de *iterações* de treino foi estabelecido em $\boldsymbol{Max_{It}}$ **= 10** (**`maxit=10`**). Valores mais elevados aumentam o esforço computacional sem aumento no desempenho. Uma vez que o número de neurónios intermédios não é um factor crítico, este foi fixado no valor de $\boldsymbol{H}$ **= 24** (duas vezes o número de atributos de entrada) (**`Nh<<-24`**). Quanto ao número de *Execuções/Conjunto de Modelos* (3.12), foi estabelecido o valor de $\boldsymbol{R}$ **= 5** (**`RUNS=5`**).

O parâmetro mais importante (λ) foi ajustado através de uma *procura em grelha*[32] em dois níveis. Embora o λ possa variar teoricamente dentro do intervalo [0.0, 1.0], na prática é usual apresentar valores próximos de zero. Neste caso, testar todas as combinações é impossível, pois teoricamente são infinitos os valores existentes entre 0 e 1. Uma forma de lidar com este problema reside na utilização de uma *procura em grelha*, onde em vez de se testarem todos os valores, usa-se uma procura discreta. Assim, o primeiro nível da rede procura todos os valores discretos dentro da amplitude {0.00, 0.01, …., 0.20}, sendo seleccionada a configuração com o erro de previsão mais baixo (λ_1). Depois, prossegue o segundo nível com uma afinação mais fina, dentro da

[32] Tradução adoptada para o termo *Grid-Search*.

gama $\lambda_2 \in \{\lambda_1\text{-}0.005, \ldots, \lambda_1\text{-}0.001, \lambda_1\text{+}0.001, \ldots, \lambda_1\text{+}0.004\} \wedge \lambda_2 \geq 0$. Assim, o número de procuras é igual a 21 + 9 = 30 (ou 21 + 5 = 26 se $\lambda_1 = 0$).

Para estimar a qualidade de previsão da *Rede Neuronal* para a *procura em grelha*, foi adoptada uma *validação cruzada com 10-desdobramentos* [Kohavi, 1995] (**`res=crossval(x,y,alfa.fit,alfa.predict,ngroup=10)`**) (Secção 3.14.2), onde o conjunto de treinos é dividido em 10 subconjuntos de igual tamanho. Sequencialmente, foi testado um subconjunto diferente (com 10% dos dados), sendo os restantes dados usados para ajustar os pesos da rede. No final dos 10 treinos, o modelo foi testado em todos os dados de treino, sendo a estimativa dada pelo *RMQE* (Equação 7) médio calculado ao longo dos 10 conjuntos de treino. A Figura 5.5 esquematiza um exemplo da evolução do erro na procura em grelha de dois níveis, para a tarefa **WBS**. Neste caso, a *constante de decaimento* com melhor qualidade (*RMQE* = 6.75) foi encontrada para $\lambda = 0.097$. Convém referir que após a obtenção da melhor *constante de decaimento*, as *Redes Neuronais* são novamente treinadas com todos os dados do conjunto de treino.

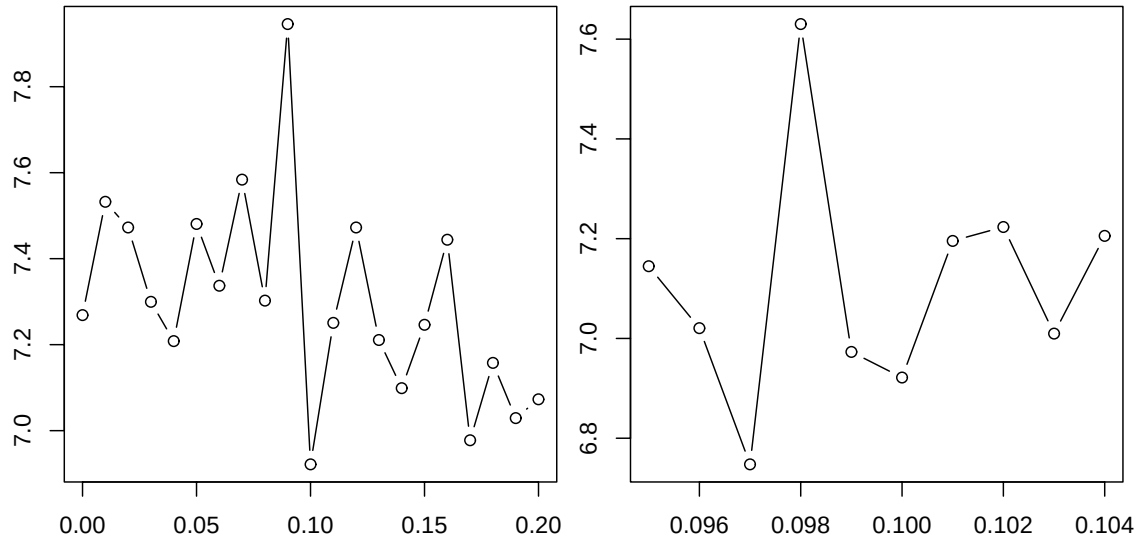


Figura 5.5 – Exemplo de uma *procura em grelha* para a *constante de decaimento* (eixo do x) e do *RMQE* (eixo y), valores para o primeiro nível (esquerda) e segundo nível (direita).

A um nível superior, e para comparar os diferentes modelos de aprendizagem, foram realizadas 5 *execuções* de uma *validação cruzada 10-desdobravél* (calculada sobre todos os dados disponíveis). Isto significa que em cada uma destas 50 experiências, 90% dos dados são usados para aprendizagem e 10% para teste. Os resultados são

mostrados na Tabela 5.5, em termos da média dos erros de teste obtidos sobre as 50 experiências.

Tarefa	*Modelo*	*EAM*	*EAMR*	*RMQE*	*RMQER*
WBS	RM	9.17	134.7%	11.64	130.4%
	RNM	6.14	90.1%	8.14	91.2%
	CRN	5.89	86.6%	7.79	87.3%
	CRNAS	**5.54**	**81.4%**	**7.46**	**83.7%**
PAS	RM	1.64	119.3%	2.13	131.7%
	RNM	1.37	99.9%	1.68	104.1%
	CRN	1.27	92.4%	1.56	96.7%
	CRNAS	**1.17**	**84.9%**	**1.45**	**89.5%**

Tabela 5.5 – Resultados da regressão para a carne de cordeiro.

Os resultados da *Regressão Múltipla (RM)* são pobres, apresentando valores de *EAMR* e *RMQER* maiores do que 100%, para ambas as tarefas. Assim, a *RM* é pior do que a previsão trivial via a média dos valores. As diferenças entre os métodos da *RM* e das *Redes Neuronais* sugerem que ambas as tarefas apresentam uma não linearidade. A abordagem *RNM* funciona melhor do que a *RM*, embora seja pior do que a *CRN*. De facto, existe uma vantagem no uso de *Conjuntos de Modelos*, uma vez que a melhoria apresentada pelo *CRN* é de 3.5% (3.9%) em termos dos valores do *EAMR* (*RMQER*) para a tarefa **WBS**, e é ainda maior para a tarefa **PAS** (7.5% e 7.4%). Para este último conjunto de dados, o desempenho do método *RNM* é pior, sem melhoria sobre a previsão simples da média de valores. Dado que ambas as estratégias (*RNM* e *CRN*) requerem o mesmo esforço computacional, a última será privilegiada neste estudo.

Para a análise das características dos atributos, foi decidido seleccionar os atributos de entrada com uma importância relativa maior ou igual a 3%. A Tabela 5.6 mostra a importância relativa média (Equação 11) das características das entradas para a abordagem *CRN*.

Tarefa	*Atributos*											
	raca	***sexo***	***pcq***	***ege2***	c	***pH1***	***pH24***	***a***	***b***	***dE***	***dL***	***db***
WBS	4.30	0.08	5.80	7.62	1.51	1.20	1.74	50.29	1.70	11.07	5.50	8.48
PAS	41.01	1.48	2.88	5.13	2.61	6.63	1.05	22.58	2.46	7.46	2.91	3.79

Tabela 5.6 – Importância relativa de todas as variáveis de entrada da carne de cordeiro (valores percentuais).

A regra ≥ 3% permite uma redução das entradas para cerca de metade. De facto, a Tabela 5.7 mostra que 7 atributos contribuem para 93.1% da influência da saída **WBS** e 6 entradas afectam a saída **PAS** com uma importância de 86.6%. Apesar da diferença dos valores percentuais, as características seleccionadas são bastante similares para ambos os problemas. As excepções são a não inclusão do **pH1** no caso da **WBS** e do **pcq** e **dL** na tarefa **PAS**. É também interessante notar que o atributo **Sexo** é o factor menos relevante, com influências de 0.08% (**WBS**) e 1.48% (**PAS**). Aparentemente, isto contrasta com o conhecimento de que o sexo afecta a *tenrura*. Contudo, as fêmeas geralmente apresentam um maior peso e estado de engorda, assim, a informação sobre o sexo pode ser indirectamente representada nas variáveis **pcq** e **ege2**.

Tarefa	*Modelo*	*Atributos*								*Influência*
		raca	***pcq***	***ege2***	***pH1***	***a***	***dE***	***dL***	***db***	
WBS	CRN	4.3	5.8	7.6		50.3	11.1	5.5	8.5	93.1
	CRNAS	2.8	21.4	7.7		41.7	11.7	6.2	8.5	100
PAS	CRN	41.0		5.1	6.6	22.6	7.5		3.8	86.6
	CRNAS	36.8		20.1	9.3	22.4	9.7		1.7	100

Tabela 5.7 – Importância relativa dos atributos de entrada seleccionados via a regra ≥ 3% (valores percentuais).

Uma vez que entradas não relevantes podem afectar o desempenho da *RN*, foi criado outro conjunto de dados, considerando apenas as entradas mais importantes da Tabela 5.7. Esta configuração será chamada *Conjunto de Redes Neuronais baseado na Análise de Sensibilidade* (*CRNAS*). Com efeito, este novo modelo conseguiu obter melhores

resultados (Tabela 5.5), quando comparado com o modelo *CRN*, especialmente para a tarefa **PAS**, com aumento de 7.5% e 7.2% em termos dos valores de *EAMR* e *RMQER*.

Na Tabela 5.7, os valores de sensibilidade são também apresentados para este último modelo. Para a previsão do **WBS**, o índice de vermelho (**a**) parece ser o atributo mais importante, seguido pelo peso (**pcq**) e o diferencial da cor total (**dE**). Relativamente ao problema **PAS**, as características mais relevantes são a raça (**raca**), o índice de vermelho (**a**) e a espessura da gordura esternal (**ege2**). As diferenças obtidas entre as duas tarefas podem ser explicadas por factores psicológicos. Por exemplo, a importância da **Raça** aumentou de 2.8% (**WBS**) para 36.8% (**PAS**). Após este estudo, foram contactados alguns peritos do painel, tendo-se concluído que os cordeiros *Mirandeses* eram menos fibrosos e tinham odor mais intenso do que os cordeiros *Bragançanos*. Isto pode ser justificado pelo stress dos animais durante o abate, embora seja necessária mais investigação sobre este assunto.

Como exemplo da importância e da qualidade da solução obtida, a Figura 5.6 compara os valores previstos com os valores observados para a tarefa **WBS**. Nos gráficos apresentados, a linha diagonal denota a previsão perfeita. A aproximação *CRNAS* (direita) apresenta claramente mais previsões ao longo da linha do que o método da *Regressão Múltipla (RM)* (esquerda).

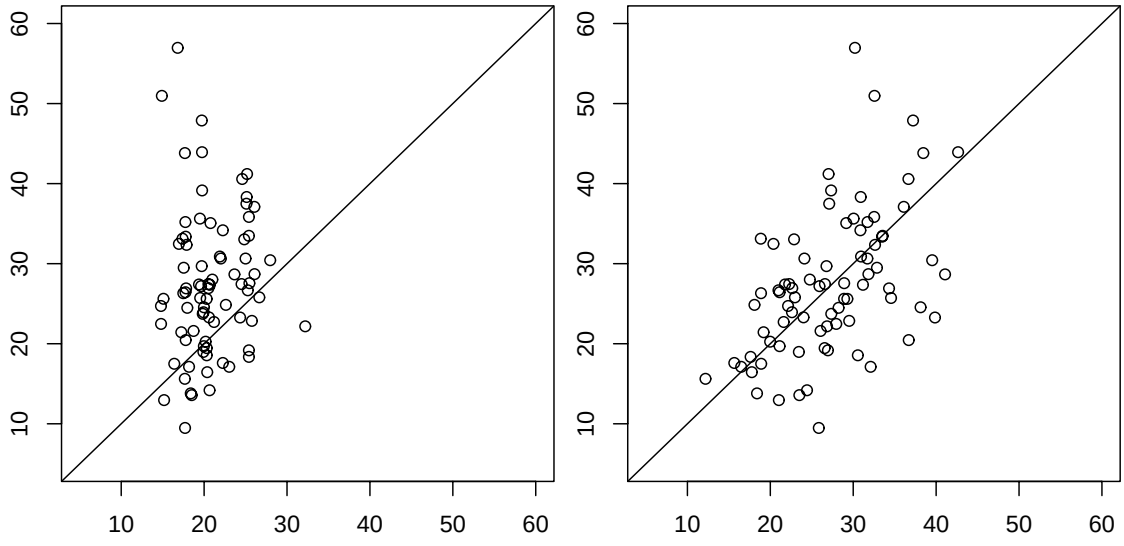


Figura 5.6 – Exemplo de gráficos de regressão dos valores previstos (eixo x) e observados (eixo y) para a *RM* (esquerda) e a abordagem proposta (direita).

5.8 – Discussão

A primeira preocupação para a elaboração da parte prática deste trabalho, foi de seguir uma metodologia que permitisse um correcto acompanhamento de projectos de *DCBD/DM* e que se ajustasse ao caso objecto de estudo. Assim, optou-se por escolher a metodologia *CRISP-DM*, compreendendo as fases: **Estudo do Negócio**, **Estudo dos Dados**, **Preparação dos Dados**, **Modelação** e **Avaliação**.

Depois de analisadas algumas ferramentas de construção de *Redes Neuronais*, optou-se por escolher o ambiente de programação estatístico **R**, tendo-se verificado que esta ferramenta satisfazia as necessidades pretendidas. Quanto ao processo de *DCBD/DM*, este foi iterativo e interactivo. À medida que se iam fazendo diversos testes e experiências com a ferramenta, foram trocadas opiniões com os especialistas da área da *Produção Animal*, a fim de se conseguir fazer um correcto estudo do negócio e dos objectivos pretendidos. Seguidamente foram recebidos os dados, e realizados estudos e análises nos mesmos. Após esta análise foram preparados os dados, tendo como principal objectivo a selecção, limpeza e transformação dos dados. Por sua vez, estes foram modelados via diversos algoritmos de *Redes Neuronais* e de *Regressão Múltipla*. Finalmente, os modelos obtidos foram avalidados, para se poder verificar a viabilidade do projecto.

Em resumo, neste trabalho é proposto um *Conjunto de Redes Neuronais baseado numa Análise de Sensibilidade* (*CRNAS*), com o objectivo de prever uma tarefa do mundo real. Foram consideradas duas abordagens para medir a *tenrura* da carne de cordeiro: uma objectiva, baseada numa máquina de corte *Warner-Bratzler (WBS)*, e outra subjectiva, baseada num *Painel de Análise Sensorial (PAS)*. Para ambas as tarefas de regressão, a estratégia *CRNAS* obteve melhores resultados do que outras *Redes Neuronais* e a *Regressão Múltipla*.

Capítulo 6

Conclusões

É feita uma síntese do trabalho realizado, são discutidas as conclusões mais importantes relativas ao trabalho desenvolvido, assim como são apresentadas as contribuições e recomendações para trabalhos futuros.

6.1 – Síntese

Um factor prioritário no sucesso da indústria da carne assenta na capacidade de fornecer especialidades que satisfaçam os gostos dos consumidores. Em particular, avaliar a qualidade é importante para as explorações de carne de cordeiro, especialmente se o objectivo passar por centrar-se em determinados nichos de mercado para diferenciação dos seus produtos. Assim, tem sido dedicado um elevado esforço na procura de estimadores fiáveis da qualidade.

Segundo Angulo [2001], a qualidade de um alimento é um termo complexo, difícil de definir e de carácter multi-dimensional. Berian [1998] afirma que o conceito de qualidade varia em função da cadeia de produção e de comercialização, mas também da necessidade de satisfazer as exigências do mercado específico a que se destina. Dos vários factores que influenciam a qualidade da carne, a *tenrura* é comummente aceite como o atributo mais importante na influência da percepção alimentar [Huffman *et al.*, 1997]. Para a avaliação da *tenrura*, duas grandes abordagens têm vindo a ser propostas [Arvanitoyannis e Houwelingen-Koukaliaroglou, 2003]: análises instrumentais e sensoriais. A primeira medida é obtida através da medição da célula de corte *Warner-Bratzler*, enquanto a segunda é realizada por um painel sensorial.

Ora, uma alternativa passa pelo uso de medidas de análise da carcaça, que são baratas e não invasivas, podendo ser obtidas nas primeiras 24 horas após o abate, tal como leituras do pH e da cor. A ideia é obter um modelo que com base nestas entradas seja capaz de prever, com elevada precisão, qual a *tenrura* de uma dada amostra. Na

aproximação clássica da ciência animal a este problema, têm-se utilizado sobretudo modelos de *Regressão Múltipla* [Arvanitoyannis e Houwelingen-Koukaliaroglou, 2003], usando as características da carne como variáveis independentes (entradas) e a medida instrumental *Warner-Bratzler* ou a medida *sensorial* como variáveis dependentes (saída). Contudo, estes modelos lineares falham claramente quando estão presentes nos dados fortes relações não lineares. Nestes casos, uma melhor opção poderá passar pelo uso de *Redes Neuronais Artificiais*, que apresentam vantagens tais como uma aprendizagem não linear e tolerância ao ruído.

No passado, vários estudos abordaram *Redes Neuronais Artificiais* para avaliar a qualidade da carne (*e.g.* bovinos, suínos, aves ou salsichas) [Arvanitoyannis e Houwelingen-Koukaliaroglou, 2003]. Contudo, no que respeita à previsão da *tenrura*, a literatura parece escassa, sendo principalmente orientada para a carne de bovino. Num estudo efectuado por Li e seus colaboradores [1999], foram utilizadas redes do tipo *perceptrão multicamada* para mapear imagens da textura com valores de *tenrura* sensorial. Em outro estudo, Hill e seus colaboradores [2000] utilizaram o mesmo tipo de *Redes Neuronais* para a predição da força de corte *Warner-Bratzler* em seis bases de dados diferentes, obtendo melhores resultados do que com a *Regressão Múltipla*.

No processo de *Descoberta de Conhecimento em Bases de Dados*, estão envolvidas várias etapas que vão desde a selecção das bases de dados sobre as quais será realizado o processamento, até à disponibilização do conhecimento descoberto [Michalski e Kaufman, 1998]. Dentro deste enquadramento, as *Redes Neuronais* estão a ganhar atenção crescente, em especial nas áreas do *Data Mining* e da *Aprendizagem Automática*, devido ao seu desempenho potencial em termos de conhecimento preditivo [Mitra *et al.*, 2002]. Outro tópico de investigação, que se tem revelado promissor nestas áreas, é baseado no uso de *Conjunto de Modelos*, onde vários modelos são combinados de alguma maneira de modo a produzir uma única resposta [Dietterich, 2000]. Este interesse surgiu devido à descoberta de que os *Conjuntos de Modelos* são em geral mais precisos do que modelos simples.

Por outro lado, em aplicações de *Data Mining*, para além de se obter uma elevada performance predictiva, é usualmente útil fornecer um conhecimento explicativo acerca do que o modelo aprendeu. Em particular, a importância de uma dada entrada é relevante no domínio da *tenrura* da carne. Uma vez que as medidas da carcaça são geralmente correlacionadas, no passado foi proposto o uso da *Análise de Componentes*

Principais para reduzir a dimensão da entrada [Park *et al.*, 2002; Arvanitoyannis e Houwelingen-Koukaliaroglou, 2003]. Contudo, os componentes principais são variáveis comprimidas e não representam um significado directo para o utilizador da carne. Por conseguinte, uma alternativa reside na adopção da *Análise de Sensibilidade* [Kewley *et al.*, 2000], que obteve melhores resultados do que outros métodos, tais como o *forward selection* ou *algoritmos genéticos*. De facto, a abordagem que obteve melhores resultados nesta dissertação baseia-se no uso de *Conjuntos de Redes Neuronais*, sendo o número de entradas reduzidas a metade via um procedimento de *Análise de Sensibilidade*.

6.2 – Discussão

Ao concluir-se um projecto, é tido por adequado salientar-se as conclusões que foram alcançadas, confrontando os resultados obtidos com os objectivos estabelecidos, à data de início. Considera-se também fundamental uma postura crítica, relativamente às limitações de que entretanto se foi tomando consciência. No sentido de ultrapassar essas limitações, bem como de tentar promover alguns dos pressupostos fixados no trabalho em causa, devem-se apontar perspectivas de trabalho futuro. As soluções propostas nesta tese, esperam que sejam um passo em frente, no sentido de apresentar modelos de previsão e, contribuir com os conceitos que possibilitem o aprofundar e evoluir do conhecimento científico na área das *Redes Neuronais Artificiais*.

Uma das primeiras etapas do processo conducente à elaboração deste trabalho consistiu essencialmente na prospecção e familiarização com as abordagens do processo de *DCBD* existentes na literatura sobre esta temática. Em especial, no decurso deste trabalho foi adoptada a metodologia *CRoss-Industry Standard Process for Data Mining* (*CRISP-DM*). De todas as fases que a metodologia encerra, foi atribuído especial relevo à preparação de dados. Esta postura assentou no facto de a passagem de dados aos algoritmos de *Data Mining* dever ser precedida de uma preparação adequada e criteriosa, assegurando-se assim a capacidade destes gerarem conhecimento com mais elevado nível de fiabilidade.

Para o caso de estudo em causa, as técnicas e modelos seleccionados no processo de *Descoberta de Conhecimento em Bases de Dados*, revelaram-se eficientes em função do volume de dados disponíveis. Basicamente, este trabalho propõe um *Conjunto de Redes*

Neuronais coadjuvadas com um procedimento de selecção de atributos, recorrendo a uma *Análise de Sensibilidade (CRNAS)*, aplicados no processo de previsão da *tenrura* da carne de cordeiro, baseada em medidas instrumentais e sensoriais. Esta estratégia foi testada em dados animais, colectados no Nordeste Transmontano, sendo comparada não só com a *Regressão Múltipla*, mas também com outras abordagens de *Redes Neuronais*.

Os resultados obtidos revelaram-se bastante satisfatórios, no sentido em que o sistema de *CRNAS* superou outros modelos neuronais, assim como a *Regressão Múltipla* (Tabela 5.5). Por outro lado, a solução proposta é consideravelmente mais simples, pois necessita somente de 6 entradas para a tarefa de regressão da *Análise Sensorial* e 7 para a tarefa de *Análise Instrumental* (em contraponto com as 12 utilizadas pelas restantes abordagens). Mais ainda, esta abordagem permite que se obtenham previsões 24 horas após o abate, contrariamente às 72 h apresentadas pelo método instrumental e aos vários dias exigidos pelas análises sensoriais. Convém referir que a necessidade das 24 horas é consequência do uso do atributo **pH24**, pois este só pode ser recolhido 24 horas após o abate. Acresce a tudo isto o facto de as *Redes Neuronais* serem mais baratas[33] e não invasivas. Todo este conjunto de vantagens abre caminho a que se possa implementar um sistema de apoio à decisão.

Outro aspecto interessante tem a ver com a importância das variáveis de entrada. De facto, foram obtidos alguns resultados que contradizem a teoria da ciência animal (ver capítulo anterior), como a relevância do atributo raça. Embora isto possa ser justificado por factores psicológicos, também poder-se-á dever ao stress durante o abate. Assim, este trabalho levantou algumas questões de investigação em ciência animal e que deverão ser investigadas no futuro.

Uma desvantagem pode ser a precisão obtida, que actualmente é somente 16.3% (WBS) e 10.5% (PAS) melhor quando comparada com a previsão simples da média de valores. Contudo, como referido por Díez e seus colaboradores [2004], a modelação de preferências de aprendizagem, e em particular a análise de dados sensoriais, é considerada uma tarefa de regressão muito difícil. Pelo que se conhece, esta é a primeira vez que a *tenrura* da carne de cordeiro é aproximada por modelos de regressão neuronais, sendo ainda necessário uma investigação exploratória mais aprofundada.

[33] Pode mesmo afirmar-se gratuitas, no caso do uso de *software* livre como o **R**.

6.3 – Contribuições

Ao nível das contribuições, o presente trabalho proporcionou descobertas relevantes sobre os dados da carne de cordeiros, apresentando um estudo das técnicas comummente utilizadas para *Descoberta de Conhecimento em Bases de Dados*. Pelo que se conhece, modelos de regressão neuronais nunca tinham sido utilizados para previsão da *tenrura* da carne de cordeiro. Consequentemente, é razoável considerar-se a aplicação e validação destas técnicas como um contributo original.

Por outro lado, efectuou-se uma comparação entre diversos métodos de *Redes Neuronais* e métodos tradicionais, como a *Regressão Múltipla*. Convém realçar que estes modelos foram testados num problema real, sendo que este levantou um conjunto de obstáculos. Em primeiro lugar, porque ao contrário de outras aplicações de *DM*, aqui os dados são escassos. Por ouro lado, como referido anteriormente, as duas tarefas de regressão são de difícil resolução. Por conseguinte, o sucesso de algumas soluções aqui apresentadas (*e.g.* uso da *Análise de Sensibilidade* e *Conjunto de Modelos*) poderá ser visto como uma contribuição para a área dos *Sistemas de Informação*. Isto porque estas estratégias poderão ser úteis para a resolução de outros problemas de *DM* com características semelhantes.

Outro aspecto deste trabalho foi o estudo das relações entre as variáveis envolvidas, permitindo concluir quais as mais importantes no processo de predição da *tenrura* da carne de cordeiro. Esta actividade e seus resultados poderão também ser apontados como uma contribuição, uma vez que os resultados obtidos se revelaram surpreendentes, inclusivé para os investigadores da área.

No presente trabalho, a necessidade de se lidar com duas áreas científicas distintas (área dos sistemas de informação, particularizada na vertente da *Descoberta de Conhecimento em Bases de Dados* e a área da *Produção Animal*), influiu decisivamente nas dificuldades encontradas ao longo de toda a dissertação. Pese embora o facto de se entenderem os resultados obtidos com este trabalho como um pequeno contributo nesta temática, a simbiose efectuada entre as duas áreas científicas referidas, apresenta-se como uma postura de índole inovadora, nomeadamente ao nível da produção animal, potenciando o aparecimento de possibilidades de trabalho futuro.

6.4 – Trabalho Futuro

Neste momento, importa indicar pistas para novos caminhos de investigação na área da qualidade da carne via processos de *Descoberta de Conhecimento em Bases de Dados*. Em particular, apontam-se três vias:

1. Aumentar o número de registos da base de dados, de forma a consolidar os resultados obtidos;

2. Aplicar outros tipos de algoritmos de *Data Mining*, dando continuidade a esta pesquisa. Como possibilidades interessantes, podem-se testar outras técnicas não lineares, tais como as *Radial Basis Functions* [Broomhead e Lowe, 1988] e/ou as *Support Vector Machines* [Cortes e Vapnik, 1995];

3. Implementar a solução proposta num ambiente real, através de uma rápida prototipagem. Neste estudo foi efectuada uma aprendizagem *off-line*, ou seja, o processo de *Descoberta de Conhecimento em Bases de Dados* ocorreu somente depois dos dados serem colectados. A ideia é desenvolver um sistema amigável de apoio à decisão, que possa operar em tempo real num laboratório e/ou matadouro [Turban *et al.*, 2004]. Assim, ao fim de algum tempo, poder-se-á obter um *feedback* valioso sobre a validade da proposta apresentada. Por último, convém referir que este protótipo deveria ser desenvolvido num ambiente *Web*, de modo a permitir um acesso remoto ao sistema.

Anexo A – Código Escrito em R

O ambiente de programação estatístico em **R** possui um conjunto de funcionalidades para lidar com *Redes Neuronais Artificiais*. Em seguida são apresentados os vários ficheiros que foram executados no Capítulo 5 com vista à criação dos vários modelos de *Redes Neuronais*, do modelo da Regressão Múltipla, assim como todo o código de análise e avaliação dos respectivos modelos.

A.1 – Ficheiro Principal (Ovinos.txt)

Trata-se do ficheiro principal da aplicação criada. Aqui são invocados os programas para a criação dos vários modelos: *Regressão Múltipla (RM)*, *Rede Neuronal Múltipla (RNM)*, *Conjunto de Redes Neuronais (CRN)* e *Conjunto de Redes Neuronais com Análise de Sensibilidade (CRNAS)*. Também são avaliados os modelos testados e efectuada a *Análise de Sensibilidade*.

```
library(bootstrap)
 # libraria para proceder a um k-fold crossvalidation
library(nnet)
 # libraria de redes neuronais

 # - ler ficheiros
source('erros.R')
 # funcoes de erros
source('rl.R')
 # regressao linear
source('pmc.R')
 # ter acesso ao RNM (Multi RN) sem ensemble
source('pmcens.R')
 # ter acesso a RNE (Ensembles)

dados=read.table('ovinos.txt',header=T,sep=';',dec='.')
 # ler os dados do ficheiro original
summary(dados)
dados$raca=as.numeric(dados$raca)
dados$sexo=as.numeric(dados$sexo)
 # converte atributos não numericos em numericos (por causa de
histogramas, tratamentos, etc.)

 ### ---
print('Da informacoes estatisticas de um atributo')
print('introduza => est(dados$atrib) <= em que atrib, e o atributo a
testar')
est<-function(y)
      {
      testar=list()
      testar$maximo=max(y,na.rm=T)
      testar$minimo=min(y,na.rm=T)
```

```
      testar$desvio_padrao=sd(y,na.rm=T)
      testar$media=mean(y,na.rm=T)
      testar$mediana=median(y,na.rm=T)
      hist(y,col='red',main='Grafico de ocorrencias')
      return(testar)
      }
 ### --- funcao para dar algumas informacoes estatisticas e o
histograma de um atributo

dados1=cbind(dados$raca,dados$sexo,dados$pcq,dados$ege2,dados$c,dados$
pH1,dados$dpH,dados$a,dados$b,dados$dE,dados$dL,dados$dB.,dados$kgf_cm
2)
dados2=cbind(dados$raca,dados$sexo,dados$pcq,dados$ege2,dados$c,dados$
pH1,dados$dpH,dados$a,dados$b,dados$dE,dados$dL,dados$dB.,dados$dureza
,dados$aceit_geral)
 # os atributos que tinham uma correlaccao acima dos 80% foram
eliminados
 # sao necessarios 2 datasets, ja que os NAs estao em sitios
diferentes
dados1=na.omit(dados1)
dados2=na.omit(dados2)
 # elminar os dados omissos, que estao nas saidas
nrow(dados1) # 79/81, 2 eliminados
nrow(dados2) # 71/81, 10 eliminados
entradas1=NULL
for (i in 1:12) entradas1=cbind(entradas1,dados1[,i])
entradas2=NULL
for (i in 1:12) entradas2=cbind(entradas2,dados2[,i])
 # selecciona as entradas
kgf=dados1[,13]
dureza=dados2[,13]
 # selecciona as saidas

 # - - - parametros que tem de ser definidos com antecedencia
Nh<<-24
 # ou ncol(ent1)*2 - o numero de nodos intermedios e 2 vezes o numero
de entradas
Sen<<-matrix(nrow=10,ncol=NCOL(entradas1))
 # matriz para guardar as sensibilidades dos testes
Si<<-1
 # indice de sensibilidade, comecar sempre em 1
Levels<<-c(1,1,6,6,6,6,6,6,6,6,6,6)
 # (binario -> 1 = 2 niveis) (real -> 6 = 7 niveis)

 # ---------- EXPERIENCIAS COM O kgf  E dureza ------------
saida=kgf
entradas=entradas1
 # para o "kgf"

#saida=dureza
#entradas=entradas2
 # para a "dureza", para testar a dureza retirar estes comentários e
comentar as linhas para o kgf

#entradasRNEAS=cbind(entradas[,1],entradas[,3],entradas[,4],entradas[,
8],entradas[,10],entradas[,11],entradas[,12])
#entradas=entradasRNEAS
 # serve para testar a RN Ensemble com Analise de Sensibilidade do
KGF, ou seja, com os melhores atributos
```

```
#entradasRNEAS=cbind(entradas[,1],entradas[,4],entradas[,6],entradas[,
8],entradas[,10],entradas[,12])
#entradas=entradasRNEAS
 # serve para testar a RN Ensemble com Analise de Sensibilidade do
DUREZA, ou seja, com os melhores atributos

 # ---- selecciona entradas, mediante o modelo pretendido

Runs=1
 # numero de Runs excutado pela rede, valor testado para 5 x 10fold
superior para obter uma estimativa da performance de cada metodo
(NNE,MNN,MR,NNESA)

 # - todos os erros de regressao
RE=9
 # reajustar conforme a funcao definida em "errors.R"
Rerro1=matrix(ncol=RE,nrow=Runs)
 # MR - regressao linear
Rerro2=matrix(ncol=RE,nrow=Runs)
 # MNN - RN multipla
Rerro3=matrix(ncol=RE,nrow=Runs)
 # NNE - RN ensemble

 # - todas as sensibilidades da rede neuronal
Sen2=array(dim=c(10,ncol=NCOL(entradas),Runs))
Sen3=array(dim=c(10,ncol=NCOL(entradas),Runs))

TR3<<-matrix(ncol=10,nrow=Runs)
 # guarda os minsses (erros de treino)
DV3<<-matrix(ncol=10,nrow=Runs)
 # guarda os decays (os melhores decays obtidos)

 # - - - RL - Regressao Linear
for (i in 1:Runs)
      {
      r1=rl.ktest(entradas,saida)
      Rerro1[i,]=regression.erros(r1$cv.fit,saida)
      print(paste('Regressao Linear rmse: ',Rerro1[i,3]))
      }
print(' ')
print('Erros da Regressao Linear - RL')
print(paste('Erro Absoluto Medio - EAM = MAD = ',mean(Rerro1[,4])))
print(paste('Raiz da Media Quadrada dos Erros - RMQE =
',mean(Rerro1[,3])))
print(paste('Erro Absoluto Medio Relativo - EAMR =
',mean(Rerro1[,7])))
print(paste('Raiz da Media Quadrada dos Erros Relativo - RMQER =
',mean(Rerro1[,6])))
print(' ')
write.table(c(mean(Rerro1[,4]),mean(Rerro1[,3]),mean(Rerro1[,7]),mean(
Rerro1[,6])),'erro1.xls',quote=F,sep=';',eol =
"\n",col.names=T,row.names=T,dec='.')

 # - - - RNM - Rede Neuronal Multipla
for (i in 1:Runs)
      {
       # parametros que tem de ser definidos com antecedencia
      DV<<-vector(length=0)   # guardar os decays
      TR<<-vector(length=0)   # guardar os minsses
      Sen<<-matrix(nrow=10,ncol=NCOL(entradas))   # guardar as
sensibilidades
```

```
      Si<<-1    # indice de sensibilidade, comecar sempre em 1
       # Erros de previsao da rede neuronal
      print('RNM kfold para obter o melhor DECAY (espere) . . .')
      r2=pmc.ktest(entradas,saida)
      Rerro2[i,]=regression.erros(r2$cv.fit,saida)
      print(paste('RNM treinos: ',mean(TR),' DECAY: ',mean(DV),' RMSE:
',Rerro2[i,3]))
      Sen3[,,i]=Sen   # guardar todas as sensibilidades
      TR3[i,]=TR    # guardar todos os treinos
      DV3[i,]=DV    # guardar todos os decays
      write.table(TR3,'tr.xls',quote=F,sep=';',eol =
"\n",col.names=T,row.names=T,dec='.')
      write.table(DV3,'dv.xls',quote=F,sep=';',eol =
"\n",col.names=T,row.names=T,dec='.')
      write.table(Sen,'sen.xls',quote=F,sep=';',eol =
"\n",col.names=T,row.names=T,dec='.')
       # gerar um grafico
      #postscript('RN.eps',paper='special',horizontal=FALSE,width=4,he
ight=4)
      #par(mar=c(2.2,2.2,1.0,0.4))
      #plot(r2$cv.fit,saida,xlim=c(5,60),ylim=c(5,60),ann=FALSE)
      #abline(0,1)
      #dev.off()
      }
print(' ')
print('Erros da Rede Neuronal Multipla - RNM')
print(paste('Erro Absoluto Medio - EAM = MAD = ',mean(Rerro2[,4])))
print(paste('Raiz da Media Quadrada dos Erros - RMQE =
',mean(Rerro2[,3])))
print(paste('Erro Absoluto Medio Relativo - EAMR =
',mean(Rerro2[,7])))
print(paste('Raiz da Media Quadrada dos Erros Relativo - RMQER
',mean(Rerro2[,6])))
write.table(c(mean(Rerro2[,4]),mean(Rerro2[,3]),mean(Rerro2[,7]),mean(
Rerro2[,6])),'erro2.xls',quote=F,sep=';',eol =
"\n",col.names=T,row.names=T,dec='.')
print(' ')

 # - - - RNE - Rede Neuronal Ensemble
for (i in 1:Runs)
      {
      # parametros a ser definidos com antecedencia
      DV<<-vector(length=0)    #guardar os decays
      TR<<-vector(length=0)    #guardar os minsses
      Sen<<-matrix(nrow=10,ncol=NCOL(entradas))    # guardar as
sensibilidades
      Si<<-1    # indice de sensibilidade, comecar sempre em 1
       # Erros de previsao da rede neuronal
      print('RNE kfold para obter o melhor DECAY (espere) . . .')
      r3=pmcens.ktest(entradas,saida)
      Rerro3[i,]=regression.erros(r3$cv.fit,saida)
      print(paste('RNE treinos: ',mean(TR),' DECAY: ',mean(DV),' RMSE:
',Rerro3[i,3]))
      Sen3[,,i]=Sen   # guardar todas as sensibilidades
      TR3[i,]=TR    # guardar todos os treinos
      DV3[i,]=DV    # guardar todos os decays
      write.table(TR3,'tr2.xls',quote=F,sep=';',eol =
"\n",col.names=T,row.names=T,dec='.')
      write.table(DV3,'dv2.xls',quote=F,sep=';',eol =
"\n",col.names=T,row.names=T,dec='.')
```

```
      write.table(Sen,'sen2.xls',quote=F,sep=';',eol =
"\n",col.names=T,row.names=T,dec='.')
      }
print(' ')
print('Erros da Rede Neuronal Ensemble - RNE')
print(paste('Erro Absoluto Medio - EAM = MAD = ',mean(Rerro3[,4])))
print(paste('Raiz da Media Quadrada dos Erros - RMQE =
',mean(Rerro3[,3])))
print(paste('Erro Absoluto Medio Relativo - EAMR =
',mean(Rerro3[,7])))
print(paste('Raiz da Media Quadrada dos Erros Relativo - RMQER =
',mean(Rerro3[,6])))
write.table(c(mean(Rerro3[,4]),mean(Rerro3[,3]),mean(Rerro3[,7]),mean(
Rerro3[,6])),'erro3.xls',quote=F,sep=';',eol =
"\n",col.names=T,row.names=T,dec='.')
print(' ')
```

A.2 – Ficheiro de Análise dos Erros de Treino (Erros.txt)

Este bloco de código permite obter os erros de treino. São calculadas e apresentadas as funções de erros: *Erro Absoluto Médio (EAM)*, *Erro Absoluto Médio Relativo (EAMR)*, *Raiz da Média do Quadrado dos Erros (RMQE)* ou *Raiz da Média do Quadrado dos Erros Relativo (RMQER)*.

```
# Funcoes de erros
# x - predicao, y - valores

regression.erros=function(x,y)
      {
      xsize=NROW(x)
      e=vector(length=xsize)
       # cria um vector com a dimensao dos dados
      e=y-x
       # calcula a diferenca entre as entradas e a saida [yi-^yi]
      MAD=sum(abs(e))/xsize
       # aplica a formula do MAD/MAE [(Syi-^yi)/N]
      SSE=sum(e*e)
       # quadrado das diferencas [(yi-^yi)2]
      MSE=SSE/xsize
       # quadrado das diferencas [((yi-^yi)2)/N]
      RMSE=sqrt(MSE)
       # raiz do quadrado das diferencas [((yi-^yi)2)/N]
      e=y-mean(y)
       # calcula a diferenca entre a saida e a media da saida [yi-^yi]
      MSEMEAN=sum(e*e)/xsize
      MADMEAN=sum(abs(e))/xsize
      RootRelativeSquaredError=100*RMSE/sqrt(MSEMEAN)
       # RRMSE , definido no software de data mining WEKA
      RelativeAbsoluteError=100*MAD/MADMEAN
       # RMAE , definido no software de data mining WEKA
      NMSE=100*MSE/MSEMEAN
      suppressWarnings(r2<-cor(x,y)^2)
      return(c(SSE,MSE,RMSE,MAD,NMSE,RootRelativeSquaredError,Relative
AbsoluteError,r2))
```

```
      }

MAE=function(x,y)
      {
       # Mean Absolute Deviation
       # Mean Absolute Error MAE
      xsize=NROW(x)
      e=vector(length=xsize)
      e=y-x
      MAD=sum(abs(e))/xsize
      return(MAD)
      }

NME=function(x,y)
      {
       # Normalized Mean Error 1996 based stat results
       # x - predictions, y - values
      xsize=NROW(x)
      e=vector(length=xsize)
      e=y-x
      return(100*sum(abs(e))/sum(y))
      }

SSE=function(x,y)
      {
      xsize=NROW(x)
      e=vector(length=xsize)
      e=y-x
      SSE=sum(e*e)
      return(SSE)
      }

RMSE=function(x,y)
      {
      xsize=NROW(x)
      e=vector(length=xsize)
      e=y-x
      SSE=sum(e*e)
      RMSE=sqrt(SSE/xsize)
      return(RMSE)
      }

NMSE=function(x,y)
      {
       # mean square error normalized by variance
      xsize=NROW(x)
      e=vector(length=xsize)
      e=y-x
      SSE=sum(e*e)
      return(100*SSE/(var(y)*xsize))
      }

RRMSE=function(x,y)
      {
       # mean square error normalized by variance
      xsize=NROW(x)
      e=vector(length=xsize)
      e=y-x
      RRMSE=sqrt((sum(e*e)/xsize))
      e=y-mean(y)
      RMSEMEAN=sqrt((sum(e*e)/xsize))
```

```
      return(100*RMSE/RMSEMEAN)
      }

sse=function(x,y)
      {
      xsize=NROW(x)
      e=vector(length=xsize)
      e=y-x
      SSE=sum(e*e)
      return(SSE)
      }

rmse=function(x,y)
      {
      xsize=NROW(x)
      e=vector(length=xsize)
      e=y-x
      SSE=sum(e*e)
      RMSE=sqrt(SSE/xsize)
      return(RMSE)
      }

meanint=function(x)
      {
      #ttt=t.test(x)
      mmm=mean(x)
      iii=mmm
      #iii=mean(x)-ttt$conf.int[1]
      DIGITS=2
      return(round(mmm,DIGITS))
      #return(paste(round(mmm,DIGITS),"$pm$",round(iii,DIGITS)))
      }

meanint2=function(x)
      {
      #ttt=t.test(x)
      mmm=mean(x)
      iii=mmm
      #iii=mean(x)-ttt$conf.int[1]
      DIGITS=2
      return(round(mmm,DIGITS))
      #return(paste(round(mmm,DIGITS),"$pm$",round(iii,DIGITS)))
      }
```

A.3 – Regressão Múltipla (RM.txt)

Código que permite calcular a *Regressão Múltipla*, o método usado na área da Produção Animal para previsão da *tenrura* da carne de cordeiro.

```
rl.adjust=function(x,y)
      {
      lm(y~x)
      }

 # K-fold validation auxiliar functions
gamma.fit=function(x,y)
      {
      rl.adjust(x,y)
```

```
      }

gamma.predict=function(fit,x)
      {
      x1=data.frame(x)
      predict(fit,x1)
      }

rl.ktest=function(x,y)
      {
      suppressWarnings(res<-
crossval(x,y,gamma.fit,gamma.predict,ngroup=10))
       # gera uma mensagem de aviso
       # x-entrada
       # y-saida
       # theta.fit-faz a validacao cruzada
       # theta.predict-faz a previsao para theta.fit
       # ngroup-n. de grupos formados
      return(res)
      }
```

A.4 – *Rede Neuronal* Múltipla (rnm.txt)

Código que permite criar e treinar uma rede *Perceptrão Multicamada*. Treina e devolve a Rede com menor erro de treino para cada uma das saídas pretendidas.

```
# criar e treinar uma rede Perceptrao Multicamada (MLP)
# versao corrente: multiple starts
# treino NetRuns MLP's e devolve a MLP com menor erro de treino

pmc.rede=function(x,y)
      {
      sx=scale(x)
       # normalizar entradas
      Cx<<-attr(sx,"scaled:center")   # get or set specific attributes
of an object
       # coloca as colunas de entrada escalonadas ao centro (media)
      Sx<<-attr(sx,"scaled:scale")
       # coloca as colunas de entradas escalonadas
      sy=scale(y) # normalizar a saida
       # normaliza saida
      Cy<<-attr(sy,"scaled:center")   # get or set specific attributes
of an object.
      # coloca a coluna de saida escalonada ao centro (media)
      Sy<<-attr(sy,"scaled:scale")
       # coloca a coluna de saida escalonada
      NetRuns=5
       # numero de vezes que a rede vai ser executada
      MinSSE=1e999
       # numero elevado, mas tem de ser maior do que o minimo do sse
(somatorio do quadrado dos erros)
      for (i in 1:NetRuns)
            {

      Net=nnet(sx,sy,size=Nh,linout=T,trace=F,maxit=10,decay=Decay,ski
p=F)
            # lineout-saida linear; trace-nao colocar mensagem de
erro; skip-switch to add skip-layer connections from input to output
```

```
            if (MinSSE>Net$value)
                  {
                  MinSSE=Net$value
                  MinNet=Net
                  }
            }
      return (MinNet)
      }

 # --- validacao k-fold, funcoes auxiliares
 # Parametros: Nh-camadas intermedias; decay- peso constante de
decaimento
alfa.fit=function(x,y)
      {
      pmc.rede(x,y)
      }

alfa.predict=function(fit,x)
      {
      predict(fit,scale(x,Cx,Sx))*Sy+Cy
       # obter previsao desnormalizada
      }

 # d: termo do peso do decay, especifica o quanto o valor do peso
antigo é reduzido, testar valores entre 0.005 e 0.3
 # obtem o melhor valor do decay usando um k-fold de procura em grelha
pmc.best.decay=function(x,y)
      {
       # - - - procura do melhor decay - 1. nivel
      minsse=1e999
      Tune=vector(length=21)
      for (i in 1:21) Tune[i]=0.01*(i-1)
      rd=vector(length=length(Tune))
      for (j in 1:NROW(Tune))
            {
            Decay<<-Tune[j]
            res=crossval(x,y,alfa.fit,alfa.predict,ngroup=10)
            rd[j]=rmse(res$cv.fit,y)
            if (minsse>rd[j])
                  {
                  minsse=rd[j]
                   mindecay=Decay
                  }
            }
       # - - - ajustamento mais fino do melhor decay - 2. nivel
      if(mindecay==0) Tune<-c(0.001,0.002,0.003,0.004,0.005) else {
            Tune<-c(mindecay-0.001,mindecay-0.002,mindecay-
0.003,mindecay-0.004,mindecay-
0.005,mindecay+0.001,mindecay+0.002,mindecay+0.003,mindecay+0.004)
      }
      print(paste("PMC - Nivel 1. Decay:", Decay, "RMSE=",rd[j], "
Minsse:", minsse, " Decay Minimo:",mindecay))
      rd=vector(length=length(Tune))
      for (j in 1:NROW(Tune))
            {
            Decay<<-Tune[j]
            res=crossval(x,y,alfa.fit,alfa.predict,ngroup=10)
            rd[j]=rmse(res$cv.fit,y)
            if (minsse>rd[j])
                  {
                  minsse=rd[j]
```

```
                  mindecay=Decay
                  }
            print(paste("PMC - Nivel 2. Decay:", Decay, "RMSE=",rd[j],
" Minsse:", minsse, " Decay Minimo:",mindecay))
            }
      TR<<-c(TR,minsse)
       # guarda os minsses
      return (mindecay)
      }

theta.fit=function(x,y)
      {
      print(paste("Fold: ", Si, "  Obter Melhor Decay (RNM) . .
.",format(Sys.time(),"%X")))
      Decay<<-pmc.best.decay(x,y)
      DV<<-c(DV,Decay)
       # guarda os Decays
      print(paste("Decayyyyyyy=",Decay))
      MLP=pmc.rede(x,y)
      Sen[Si,]<<-pmc.sensitivity(MLP,x,Levels,Cx,Sx,Cy,Sy)
      print(Sen[Si,])
      Si<<-Si+1
      return (MLP)
      }

theta.predict=function(fit,x)
      {
      predict(fit,scale(x,Cx,Sx))*Sy+Cy
       # obter previsao desnormalizada
      }

pmc.ktest=function(x,y)
      {
      res=crossval(x,y,theta.fit,theta.predict,ngroup=10)
      return (res)
      }

 # model - modelo para testar a sensibilidade
 # x - entrada
 # center - vector com os dados centrados
 # scale - vector com os dados escalonados
 # sy - saida escalonada
 # sx - saida centrada
pmc.sensitivity=function(model,x,levels,center,scale,cy=0,sy=1)
      {
      Xsize=NCOL(x)
      v=vector(length=Xsize)
      min=vector(length=Xsize)
      max=vector(length=Xsize)
      for (i in 1:Xsize)
            {
            if(is.vector(x)) v[i]=(mean(x)-center[i])/scale[i] else
v[i]=(mean(x[,i])-center[i])/scale[i]
            if(is.vector(x)) min[i]=(min(x)-center[i])/scale[i] else
min[i]=(min(x[,i])-center[i])/scale[i]
            if(is.vector(x)) max[i]=(max(x)-center[i])/scale[i] else
max[i]=(max(x[,i])-center[i])/scale[i]
            }
      z=vector(length=Xsize)
      Sv=vector(length=Xsize)
      Sum=0
```

```
      for (i in 1:Xsize)
            {
            z=v
            y=vector(length=levels+1)
            for (j in 0:levels[i])
                  {
                  z[i]=(max[i]-min[i])*(j)/levels[i]+min[i]
                  y[j+1]=predict(model,z)*sy+cy
                  }
            Sv[i]=sd(y)**2
             # variancia
            Sum=Sum+Sv[i]
            }
      Sv=Sv/Sum
      return (Sv)
      }
```

A.5 – Conjunto de Redes Neuronais com Análise de Sensibilidade (crnas.txt)

Este código permite criar e treinar um *Conjunto de Redes Neuronais*. De realçar que a previsão final passa a ser dada pela média das previsões individuais de cada rede. Ainda neste código é calculada a *Análise de Sensibilidade*.

```
# cria e treina uma rede Perceptrao Multicamada (MLP) - com acesso aos
ensembles
# versao corrente: multiple starts

pmcens.rede=function(x,y)
      {
      sx=scale(x)
       # normalizar entradas
      Cx<<-attr(sx,"scaled:center")   # get or set specific attributes
of an object
       # coloca as colunas de entrada escalonadas ao centro (media)
      Sx<<-attr(sx,"scaled:scale")
       # coloca as colunas de entradas escalonadas
      sy=scale(y)
       # normalizar a saida
      Cy<<-attr(sy,"scaled:center")   # get or set specific attributes
of an object
       # coloca a coluna de saida escalonada ao centro (media)
      Sy<<-attr(sy,"scaled:scale")
       # coloca a colunas de saida escalonada

      Rede1=nnet(sx,sy,size=Nh,linout=T,trace=F,maxit=10,decay=Decay)
       # lineout-saida linear; trace-nao colocar mensagem de erro
      Rede2=nnet(sx,sy,size=Nh,linout=T,trace=F,maxit=10,decay=Decay)
      Rede3=nnet(sx,sy,size=Nh,linout=T,trace=F,maxit=10,decay=Decay)
      Rede4=nnet(sx,sy,size=Nh,linout=T,trace=F,maxit=10,decay=Decay)
      Rede5=nnet(sx,sy,size=Nh,linout=T,trace=F,maxit=10,decay=Decay)
      Ens=list(Rede1,Rede2,Rede3,Rede4,Rede5)
       # cria uma lista com os dados das redes
      return (Ens)
      }
```

```
 # --- validacao k-fold, funcoes auxiliares
 # Parametros: Nh-camadas intermedias; decay- peso constante de
decaimento
alfaens.fit=function(x,y)
      {
      pmcens.rede(x,y)
      }

alfaens.predict=function(fit,x)
      {
      pred=vector(length=length(NROW(x)))
      for (i in 1:length(pred)) pred[i]=0
      for (i in 1:length(fit))
            {
            pred=pred+predict(fit[[i]],scale(x,Cx,Sx))*Sy+Cy
            }
      pred=pred/length(fit)
      return (pred)
      }

 # d: termo do peso do decay, especifica o quanto o valor do peso
antigo é reduzido, testar valores entre 0.005 e 0.3
 # obtem o melhor valor do decay usando um k-fold de procura em grelha
best.decay=function(x,y)
      {
       # - - - procura o melhor decay - 1. nivel
      minsse=1e999
      Tune=vector(length=21)
      for (i in 1:21) Tune[i]=0.01*(i-1)
      rd=vector(length=length(Tune))
      for (j in 1:NROW(Tune))
            {
            Decay<<-Tune[j]
            res=crossval(x,y,alfaens.fit,alfaens.predict,ngroup=10)
            rd[j]=rmse(res$cv.fit,y)
            if (minsse>rd[j])
                  {
                  minsse=rd[j]
                   mindecay=Decay
                  }
            print(paste("PMCens - Nivel 1. Decay:", Decay,
"RMSE=",rd[j], " Minsse:", minsse, " Decay Minimo:",mindecay))
            }

      # - - - ajustamento mais fino do melhor decay - 2. nivel
      if (mindecay==0) Tune=c(0.001,0.002,0.003,0.004,0.005)
      else
            {
            Tune=c(mindecay-0.001,mindecay-0.002,mindecay-
0.003,mindecay-0.004,mindecay-
0.005,mindecay+0.001,mindecay+0.002,mindecay+0.003,mindecay+0.004)
            }
      rd=vector(length=length(Tune))
      for (j in 1:NROW(Tune))
            {
            Decay<<-Tune[j]
            res=crossval(x,y,alfaens.fit,alfaens.predict,ngroup=10)
            rd[j]=rmse(res$cv.fit,y)
            if (minsse>rd[j])
                  {
                  minsse=rd[j]
```

```
                mindecay=Decay
                }
            print(paste("PMCens - Nivel 2. Decay:", Decay,
"RMSE=",rd[j], " Minsse:", minsse, " Decay Minimo:",mindecay))
            }
      TR<<-c(TR,minsse)
      return (mindecay)
      }

thetaens.fit=function(x,y)
      {
      print(paste("Fold: ", Si, " Obter Melhor Decay (RNE) . .
.",format(Sys.time(),"%X")))
      Decay<<-best.decay(x,y)
      DV<<-c(DV,Decay)
       # guarda os Decays
      print(paste("Decay=",Decay))
      MLPENS=pmcens.rede(x,y)
      Sen[Si,]<<-pmcens.sensitivity(MLPENS,x,Levels,Cx,Sx,Cy,Sy)
      print(Sen[Si,])
      Si<<-Si+1
      return (MLPENS)
      }

thetaens.predict=function(fit,x)
      {
      pred=vector(length=length(NROW(x)))
      for (i in 1:length(pred)) pred[i]=0
      for (i in 1:length(fit))
            {
            pred=pred+predict(fit[[i]],scale(x,Cx,Sx))*Sy+Cy
            }
      pred<-pred/length(fit)
      return (pred)
      }

pmcens.ktest=function(x,y)
      {
      res=crossval(x,y,thetaens.fit,thetaens.predict,ngroup=10)
      return (res)
      }

 # model - modelo para testar a sensibilidade
 # x - entrada
 # center - vector com os dados centrados
 # scale - vector com os dados escalonados
 # sy - saida escalonada
 # sx - saida centrada
pmcens.sensitivity=function(model,x,levels,center,scale,cy=0,sy=1)
      {
      Xsize=NCOL(x)
      v=vector(length=Xsize)
      min=vector(length=Xsize)
      max=vector(length=Xsize)
      for (i in 1:Xsize)
            {
            if(is.vector(x)) v[i]=(mean(x)-center[i])/scale[i] else
v[i]=(mean(x[,i])-center[i])/scale[i]
            if(is.vector(x)) min[i]=(min(x)-center[i])/scale[i] else
min[i]=(min(x[,i])-center[i])/scale[i]
```

```
            if(is.vector(x)) max[i]=(max(x)-center[i])/scale[i] else
max[i]=(max(x[,i])-center[i])/scale[i]
            }
      z=vector(length=Xsize)
      Sv=vector(length=Xsize)
      SSv=vector(length=Xsize)
      for (i in 1:Xsize) SSv[i]=0
      for (k in 1:length(model))
            {
            Sum=0
            for (i in 1:Xsize)
                  {
                  z=v
                  y=vector(length=levels+1)
                  for (j in 0:levels[i])
                        {
                        z[i]=(max[i]-min[i])*(j)/levels[i]+min[i]
                        y[j+1]<- predict(model[[k]],z)*sy+cy
                        }
                  Sv[i]=sd(y)**2
                   # variancia
                  Sum=Sum+Sv[i]
                  }
            Sv=Sv/Sum
            SSv=SSv+Sv
            }
      SSv=SSv/length(model)
      return (SSv)
      }
```

Anexo B – Base de dados da Carne de Cordeiro

Neste anexo é apresentada a Base de Dados inicial, a qual foi sujeita a um pré-processamento de forma a ser utilizada para a geração dos modelos para a previsão da *tenrura* da carne de Cordeiro.

				Medidas						pH				Cor							Dureza	
abate	animal	raca	sexo	engorda	pcq	ege2	ege3	ege4	C	pva	pH1	pH24	dpH	L	a	b	dE	dL	dA	dB	Análise Sensorial	Kgf_cm2
25-11-2002	80	Mirandesa	M		4132	7,23	7,38	7,12	0,38	11100	6,60	5,85	0,11	43,80	16,60	9,65	54,95	-49,10	15,00	19,35	0,81	17,49
25-11-2002	89	Mirandesa	M		4521	11,58	9,38	6,51	0,39	12100	6,40	5,70	0,11	42,50	16,50	10,20	55,80	-49,90	15,20	19,95	2,37	26,32
25-11-2002	489	Mirandesa	M		4603	11,14	11,11	9,95	0,50	10500	6,34	5,82	0,08	37,75	17,35	7,50	59,40	-54,50	16,30	17,15	0,73	13,58
25-11-2002	490	Mirandesa	F		5472	16,10	13,26	11,17	0,42	11100	6,07	5,55	0,09	46,60	14,60	10,90	51,75	-45,05	13,70	21,40	0,87	13,80
25-11-2002	491	Mirandesa	F		5273	16,42	16,10	10,06	0,64	11500	6,19	5,50	0,11	40,60	15,80	9,80	56,35	-50,85	14,00	18,80	2,37	23,29
25-11-2002	492	Mirandesa	F		5390	14,48	14,71	12,97	0,70	10500	6,20	5,56	0,10	45,00	16,25	10,60	53,55	-47,80	12,15	19,85	1,17	12,96
25-11-2002	493	Mirandesa	M		5518	14,49	13,75	12,32	0,32	11500	6,20	5,57	0,10	45,80	17,00	11,70	53,40	-46,55	14,45	20,45	1,25	16,45
25-11-2002	494	Mirandesa	F		5307	14,81	12,05	10,98	0,75	10800	5,92	5,57	0,06	43,50	15,80	9,90	54,00	-48,40	14,10	19,50	1,75	20,26
25-11-2002	495	Mirandesa	F		5025	11,47	10,93	9,46	0,83	10900	6,01	5,64	0,06	43,85	18,25	10,05	54,35	-47,50	17,25	20,00	4,77	17,59
25-11-2002	496	Mirandesa	M		5120	13,20	10,40	8,76	0,39	12100	6,47	5,55	0,14	47,30	17,25	11,30	52,20	-45,10	15,50	21,15	1,87	15,62
25-11-2002	497	Mirandesa	M		4466	6,01	5,12	3,91	0,36	11100	6,54	5,57	0,15	53,15	11,50	11,80	46,50	-39,10	11,25	22,45	1,85	17,12
25-11-2002	498	Mirandesa	F		5508	14,68	12,64	11,07	0,65	10800	6,29	5,53	0,12	49,50	11,55	10,40	49,05	-43,70	10,15	19,85	3,10	
17-12-2002	75	Mirandesa	F	2-	6875	14,87	10,81	7,90	1,50	16000	6,77	5,93	0,12	38,55	17,30	8,55	59,60	-55,45	15,00	16,05	2,50	23,74
17-12-2002	81	Mirandesa	F	1	6773	13,29	12,56	10,44	1,35	14600	6,21	5,75	0,07	41,60	17,65	8,90	57,10	-52,00	16,35	17,35	2,85	26,67
17-12-2002	461	Mirandesa	F	3	8131	18,38	15,88	14,37	2,55	16400	5,89	5,66	0,04	40,75	19,45	9,10	58,80	-53,20	17,75	17,25	1,47	9,48
17-12-2002	462	Mirandesa	M	3	7706	17,18	18,90	15,83	2,24	15000	6,45	5,89	0,09	44,75	17,80	10,50	54,85	-48,80	16,25	19,00	4,70	21,44
17-12-2002	463	Mirandesa	M	2+	7216	15,73	14,41	13,68	2,05	14200	6,21	5,49	0,12	43,80	12,55	7,90	53,60	-49,65	11,35	16,60	1,01	20,45
17-12-2002	464	Mirandesa	F	2+	7526	14,06	12,89	10,10	1,20	15000	6,30	5,66	0,10	44,25	15,00	9,15	54,35	-49,90	12,90	17,25	4,59	30,64
17-12-2002	467	Mirandesa	F	2	6882	12,31	10,80	9,73	1,75	15000	6,45	5,73	0,11	40,60	15,90	8,75	57,55	-53,20	13,95	17,10	3,49	22,17
17-12-2002	468	Mirandesa	M	2	7592	20,33	16,59	14,88	1,69	17500	6,58	5,89	0,10	41,60	16,60	9,20	56,30	-51,25	15,65	17,25	1,75	18,55
17-12-2002	469	Mirandesa	M	1+	6529	13,47	12,29	13,00	1,71	15500	6,54	5,91	0,10	42,25	15,60	10,40	56,05	-51,10	14,45	17,95	2,25	27,45

17-12-2002	475	Mirandesa	F	1+	6640	13,60	12,73	11,44	1,12	14500	6,61	5,85	0,11	41,65	17,65	8,90	57,15	-51,95	16,35	17,35	1,75	19,70
17-12-2002	476	Mirandesa	M	2	6958	12,27	13,45	12,50	1,71	16300	6,32	5,75	0,09	40,86	17,25	8,85	58,15	-53,10	16,50	16,95	1,51	25,79
17-12-2002	479	Mirandesa	M	2	6456	14,49	12,95	12,48	1,71	15500	6,59	5,77	0,12	45,40	13,70	8,85	51,85	-47,75	11,50	16,45	1,77	22,86
14-01-2003	59	Bragançano	M	1+	8416	16,82	16,00	13,92	1,72	18200	6,14	5,64	0,08	42,45	14,45	8,65	54,00	-49,75	13,05	16,45	7,12	30,64
14-01-2003	76	Bragançano	F	1	4607	10,40	9,05	8,44	0,97	10500	6,72	5,88	0,13	44,25	13,00	8,00	51,10	-47,05	11,35	16,25	5,61	27,36
14-01-2003	80	Bragançano	M	1+	8303	20,62	19,02	15,78	1,81	16600	6,00	5,69	0,05	42,60	16,00	9,55	56,30	-50,80	16,05	18,00	5,26	26,91
14-01-2003	83	Bragançano	M	2	10664	19,67	17,81	17,16	1,17	21200	5,79	5,80	0,00	42,55	19,05	10,85	57,60	-52,50	15,70	17,85	3,34	24,48
14-01-2003	84	Bragançano	F	2	10711	21,82	18,78	17,25	2,11	20100	5,72	5,72	0,00	37,95	18,05	9,10	59,95	-55,05	16,50	17,20	2,33	40,58
14-01-2003	86	Bragançano	F	2	9132	19,47	17,51	16,35	3,27	17000	6,00	5,81	0,03	39,85	19,35	11,00	60,50	-55,25	17,15	17,70	4,41	27,56
14-01-2003	89	Bragançano	F	1+	5448	14,51	14,02	12,70	1,70	11200	6,57	5,79	0,12	44,95	11,55	9,20	51,00	-46,60	10,45	17,85	5,54	30,89
14-01-2003	90	Bragançano	M	2-	9588	20,98	19,49	17,09	2,78	19200	6,55	5,79	0,12	44,55	17,80	11,50	55,70	-49,50	16,45	19,45	3,09	29,69
14-01-2003	97	Bragançano	F	1+	5921	16,47	16,08	13,14	1,66	11800	6,19	5,81	0,06	43,25	17,30	10,80	54,95	-49,35	15,05	18,90	2,89	26,94
14-01-2003	98	Bragançano	F	1+	8714	18,71	15,88	13,99	1,80	17000	6,12	5,69	0,07	37,50	17,45	7,75	59,15	-53,65	18,15	17,10	5,82	38,33
14-01-2003	207	Bragançano	M	2	10097	20,01	17,76	15,74	3,48	21100	6,02	5,84	0,03	45,85	13,05	9,45	50,55	-45,70	11,90	17,95	4,89	29,49
14-01-2003	216	Bragançano	M	2	12936	26,34	25,38	22,79	3,53	24800	5,88	5,72	0,03	38,25	15,55	7,60	57,45	-53,15	14,30	16,45		28,66
14-01-2003	226	Bragançano	M	2-	9202	18,73	18,43	16,00	2,35	18600	6,61	5,73	0,13	44,50	16,30	9,65	53,25	-47,55	15,20	18,50	1,73	25,62
21-01-2003	75	Bragançano	F	2+	11455	23,58	22,73	20,41	2,79	20700	6,01	5,71	0,05	39,90	19,90	10,05	58,80	-52,85	16,15	20,15	5,06	27,19
21-01-2003	79	Bragançano	F	2+	11852	23,35	22,93	20,71	4,04	21400	6,52	5,86	0,10	39,75	19,60	11,05	59,30	-53,25	16,05	20,50	5,73	28,02
21-01-2003	102	Bragançano	F	2	10323	25,29	22,50	20,23	2,52	19300	5,81	5,68	0,02	37,40	15,45	8,05	58,30	-53,55	13,30	18,65	4,78	33,48
21-01-2003	104	Bragançano	M	2-	9160	20,88	18,16	17,81	3,43	18000	5,88	5,72	0,03	38,35	16,75	8,25	57,55	-52,40	14,65	18,90		
21-01-2003	221	Bragançano	M	3	14845	26,16	22,26	20,02	2,95	28400	6,18	5,92	0,04	40,10	14,70	8,05	56,75	-52,50	11,90	17,95		37,10
28-01-2003	1415	Bragançano	M	1	6641	13,45	12,16	9,60	1,68	14600	6,09	5,65	0,07	41,50	15,10	9,65	55,05	-49,60	13,60	19,45	5,92	37,49
28-01-2003	1436	Bragançano	F	2	8306	21,46	18,55	14,31	3,14	15600	5,77	5,63	0,02	38,45	15,80	7,85	58,20	-53,70	15,00	16,75	2,61	23,28
28-01-2003	1441	Bragançano	F	2-	7934	16,85	16,19	14,16	1,99	14900	6,42	5,71	0,11	39,50	17,70	10,45	57,90	-52,05	16,40	19,35	4,05	22,47
28-01-2003	1461	Bragançano	F	1	5508	12,44	12,76	10,87	2,08	10300	5,67	5,60	0,01	39,45	18,55	8,70	58,50	-53,25	16,50	17,75	1,77	24,85
28-01-2003	1462	Bragançano	M	1	5178	10,89	9,61	5,54	1,27	10800	6,33	5,78	0,09	23,20	17,25	10,95	54,55	-48,50	16,15	19,00	3,81	22,71
28-01-2003	1463	Bragançano	M	1	4527	7,71	7,87	6,22	1,70	9800	6,36	5,75	0,10	44,85	14,10	10,10	52,35	-47,55	12,65	17,95	5,49	32,47
28-01-2003	1465	Bragançano	F	2	8153	21,03	17,31	16,55	2,39	15500	6,15	5,51	0,10	39,35	19,80	11,70	59,15	-52,95	17,85	19,35	3,13	18,97
28-01-2003	1474	Bragançano	M	1+	5929	12,01	12,31	9,43	1,21	11200	5,54	5,63	-0,02	42,30	18,70	12,30	57,00	-50,25	16,40	21,30	1,77	18,36
28-01-2003	1481	Bragançano	F	1+	5379	14,93	14,84	12,77	2,17	10100	6,21	5,65	0,09	38,80	16,85	10,20	57,60	-52,75	14,70	17,80	5,12	27,46
28-01-2003	1485	Bragançano	M	1+	5644	17,75	15,69	13,99	1,46	10900	6,10	5,63	0,08	42,45	18,15	12,45	56,15	-49,80	16,30	20,15	5,66	33,12
28-01-2003	1492	Bragançano	M	1+	6095	13,46	13,04	11,04	1,10	11100	5,88	5,65	0,04	42,35	13,65	8,45	55,60	-50,65	14,30	17,60	7,11	33,37

28-01-2003	1494	Bragançano	M	1+	6128	18,08	15,44	13,56	1,57	11600	6,22	5,69	0,09	43,40	18,30	11,80	55,45	-49,40	15,00	19,55	3,09	23,94
02-04-2003	57	Bragançano	F	3	13642	24,10	22,76	21,77	5,08	25200	5,80	5,54	0,04	37,15	19,55	9,10	60,45	-54,05	19,35	19,05		28,69
02-04-2003	65	Bragançano	M	2-	10663	21,79	21,71	19,55	3,19	22200	6,03	5,71	0,05	37,75	17,75	9,65	58,90	-53,40	15,90	19,00	3,83	35,63
02-04-2003	67	Bragançano	F	3-	13407	22,33	21,87	20,83	2,67	25000	6,54	5,63	0,14	39,35	22,15	12,10	60,85	-54,45	18,85	19,65		19,18
02-04-2003	88	Bragançano	F	2	10870	20,24	18,05	16,92	2,84	21600	5,92	5,51	0,07	39,15	17,60	9,00	58,85	-53,95	15,25	17,90	1,78	25,71
02-04-2003	96	Bragançano	M	2	13756	26,66	24,86	22,27	1,90	26400	5,97	5,67	0,05	38,85	18,40	10,70	58,95	-53,00	16,90	19,50		35,07
02-04-2003	99	Bragançano	F	2	11440	21,34	21,21	20,16	3,16	22000	6,04	5,59	0,07	39,15	18,00	9,35	60,10	-55,10	15,85	17,95	4,52	50,97
02-04-2003	201	Bragançano	M	2	9890	24,48	22,54	19,15	2,03	22000	5,63	5,71	-0,01	41,40	19,05	12,30	58,75	-51,55	18,25	21,45	2,40	21,59
02-04-2003	208	Bragançano	M	1	7782	16,41	16,37	13,18	2,93	16200	5,97	5,76	0,04	44,30	13,35	8,75	53,55	-49,85	10,05	16,80	3,89	34,17
02-04-2003	212	Bragançano	M	1	9392	17,72	17,37	13,91	2,28	20500	5,86	5,85	0,00	39,85	12,10	7,60	55,15	-51,20	11,05	17,25	3,27	30,43
02-04-2003	214	Bragançano	F	2+	13455	21,53	18,87	17,98	3,55	25800	5,88	5,56	0,06	41,40	21,55	11,65	58,15	-50,55	19,80	20,80		14,19
02-04-2003	220	Bragançano	M	2	12762	27,81	25,56	23,87	2,51	26300	6,33	5,77	0,09	39,15	18,65	10,25	59,30	-53,25	17,30	19,60		17,12
02-04-2003	222	Bragançano	M	2-	13197	24,98	24,75	22,29	2,11	26000	6,22	5,67	0,09	41,70	20,45	11,15	55,85	-48,10	19,65	20,35		26,43
02-04-2003	223	Bragançano	F	3-	11210	25,78	24,14	21,79	2,83	21200	5,72	5,59	0,02	38,85	21,50	6,50	60,45	-53,60	19,40	20,15	1,91	19,46
02-04-2003	225	Bragançano	M	2+	13638	25,62	24,43	23,08	2,40	27200	5,83	5,73	0,02	40,00	17,20	8,95	54,70	-48,80	16,10	18,80		32,35
02-04-2003	230	Bragançano	F	1	4522	14,35	12,10	11,06	0,89	9900	6,04	5,81	0,04	44,10	14,55	10,90	53,10	-47,25	13,75	19,95	3,01	24,72
02-11-2003	88	Mirandesa	F	2+	9379	21,24	18,98	18,97	2,76	20100	6,18	5,81	0,06	38,45	19,15	8,80	57,20	-51,50	17,05	18,20	2,35	27,40
02-11-2003	441	Mirandesa	F	2	10125	21,35	18,70	15,67	2,62	21100	6,36	5,73	0,10	37,95	18,00	9,15	57,10	-51,15	17,40	18,35	3,11	25,60
02-11-2003	442	Mirandesa	F	2	9969	19,07	17,71	14,27	2,03	19700	6,37	5,70	0,11	43,65	12,80	8,85	55,00	-48,50	10,60	16,55	5,00	56,96
02-11-2003	443	Mirandesa	F	3	10131	23,95	21,29	17,47	2,82	19900	6,51	5,91	0,09	37,70	17,80	8,50	57,65	-52,50	16,55	17,30	2,98	39,15
02-11-2003	444	Mirandesa	M	2-	9410	19,03	17,91	14,27	2,07	21300	6,24	5,71	0,08	43,10	15,90	9,00	53,20	-48,25	14,55	17,05	2,68	33,04
02-11-2003	466	Mirandesa	M	2-	8568	23,60	21,74	19,45	1,88	21000	6,47	5,83	0,10	34,95	17,10	7,00	55,50	-55,90	15,85	16,55	3,05	35,84
02-11-2003	470	Mirandesa	M	2-	8752	23,52	20,93	17,75	2,56	21550	6,43	5,94	0,08	36,20	15,05	7,15	58,10	-53,55	15,70	16,15	3,54	43,82
02-11-2003	472	Mirandesa	F	2	8597	19,99	17,92	15,62	1,80	19700	6,55	5,83	0,11	34,05	14,00	6,55	59,50	-56,05	12,90	15,30	5,57	47,88
02-11-2003	473	Mirandesa	M	2	8491	25,27	20,92	16,47	2,19	20800	6,39	5,82	0,09	36,35	13,70	6,75	57,70	-54,30	11,75	15,50	5,98	43,93
02-11-2003	474	Mirandesa	M	2-	9421	24,55	24,29	20,99	2,28	22300	6,79	5,89	0,13	36,85	18,10	7,70	58,60	-53,10	17,70	17,30	5,22	35,20
02-11-2003	477	Mirandesa	M	2	8680	22,01	21,48	19,23	2,28	21000	6,46	5,83	0,10	37,05	18,45	8,60	58,70	-53,60	16,70	17,20	3,65	41,20
02-11-2003	500	Mirandesa	F	3	11413	25,90	23,96	20,62	3,16	22300	6,26	5,80	0,07	35,20	17,85	8,10	60,05	-55,20	16,80	16,70	4,18	24,56

Anexo C – Cálculo das Correlações Lineares

Neste anexo é apresentado o código que permitiu o cálculo das correlações lineares entre os vários atributos, sendo que aqueles que continham correlações acima dos 90% foram descartados, uma vez que foi considerado que não iriam acrescentar uma mais valia ao trabalho.

C.1 – Cálculo e Apresentação de Todas as Correlações

Neste bloco de código são calculados os relacionamentos de todos os atributos da base de dados e são apresentadas as percentagens de correlação.

```
dados=read.table('ovinos.txt',header=T,sep=';',dec='.')
dados$raca=as.numeric(dados$raca)
dados$sexo=as.numeric(dados$sexo)
dados=na.omit(dados)
tabela=data.frame()
tabela=data.frame(raca=0,sexo=0,pcq=0,ege2=0,ege3=0,ege4=0,c=0,pva=0,p
H1=0,pH24=0,dpH=0,L=0,a=0,b=0,dE=0,dL=0,dA.=0,dB.=0,kgf_cm2=0,dureza=0
,aceit_geral=0)

#todos os correlacionamentos
for (i in 1:ncol(dados))
      {
      for (j in 1:ncol(dados))
            {
            aaa=cor(dados[,i],dados[,j])
            tabela[i,j]=aaa
            cat(i,',',j,'->',aaa,'\n')
            if((aaa>0.9) & (aaa<-0.9)) print('Correlacionado')
            }
      }
write.table(tabela,"per_correlacoes.txt",quote=F,sep="
",row.names=F,dec=",")
```

C.2 – Cálculo e Apresentação dos Atributos com mais de 90% de Correlação

De forma a simplificar o processo de eliminação dos atributos com grande percentagem de correlação, foi criado um pequeno código que apenas apresenta os atributos com uma correlação acima dos 90%.

```
#Apenas os relacionamentos acima de 90%
tabela2=data.frame(raca=0,sexo=0,pcq=0,ege2=0,ege3=0,ege4=0,c=0,pva=0,
pH1=0,pH24=0,dpH=0,L=0,a=0,b=0,dE=0,dL=0,dA.=0,dB.=0,kgf_cm2=0,dureza=
0,aceit_geral=0)
for (i in 1:ncol(dados))
      {
      for (j in 1:ncol(dados))
            {
            aaa=cor(dados[,i],dados[,j])
            if ((aaa>=0.9 | aaa<=-0.9) & (aaa<=0.9999))
                  {
                  cat(i,',',j,'->',aaa,'\n')
                  tabela2[i,j]=aaa
                  }
            }
      }
write.table(tabela2,"per_correlacoes2.txt",quote=F,sep="
",row.names=F,dec=",")
```

Glossário de Termos

Termo	Descrição
Análise Sensorial	É a ciência que mede, analisa e interpreta as reacções dos sentidos (visão, olfacto, audição, gosto e textura) aos alimentos, recorrendo ao uso de um painel sensorial.
Aparência	Percepção visual do produto.
Aprendizagem Automática	Subárea da inteligência artificial que pesquisa métodos computacionais relacionados com a aquisição de novos conhecimentos, novas habilidades e novas formas de organizar o conhecimento já existente.
Aprendizagem Supervisionada	São apresentados à rede conjuntos de padrões de entrada e os correspondentes padrões de saída.
Aroma	É a análise do sabor e do odor da amostra.
Associação	Responsável pela descoberta de relações ou correlações entre os atributos de um conjunto de dados.
Avaliação Instrumental	Medidas físicas de resistência ao corte de carne.
Base de Dados	Sistemas onde estão armazenados os dados que podem ser objecto de análise.
Carcaça	É o corpo inteiro de um animal abatido, tal como se apresentam após sangria, evisceração e esfola, sem cabeça, pés, cauda, úbere, órgãos genitais, fígado e fessura.
Carcaça Ideal	Quando é maximizado o rendimento em carne magra e as características organolépticas.
Carne de Qualidade	É uma análise subjectiva, depende da apreciação do produto.
Célula	É o processador do neurónio, recebe sinais de entrada através das dendrites e soma esses sinais.

Classificação de Carcaças	Forma de agrupar as carcaças com características semelhantes, tornando possível direccioná-las para mercados específicos, mas também para atribuição do seu valor comercial.
Composição da Carcaça	Proporção de peças da carcaça, assim como, a quantidade de músculo, gordura e osso que cada uma das peças proporciona.
Conformação da Carcaça	Espessura do músculo, gordura subcutânea e gordura intermuscular, relativamente às dimensões do esqueleto.
Conjunto de Modelos	A previsão final é dada por uma combinação das saídas de diversos modelos.
Cor	É a cor do produto. A cor da carne aceitável é um vermelho vivo e brilhante e sensação de frescura.
CRISP-DM	Metodologia que consiste num conjunto de fases e processos padrões para desenvolver projectos de *DCBD/DM*.
Data Mining (DM)	Etapa do processo de *DCBD*, em que se aplicam algoritmos de aprendizagem com vista à descoberta de padrões úteis nos dados.
Denominação de Origem Protegida	Um produto tem que demonstrar ter origem no local que lhe dá o nome e ter uma forte ligação com essa mesma região.
Descoberta de Conhecimento em Bases de Dados (DCBD)	Extrair de grandes bases de dados, sem nenhuma formulação prévia de hipóteses, informações genéricas, relevantes e previamente desconhecidas, que podem ser utilizadas para a tomada de decisões.
Divisão da Amostra	Consiste na divisão dos dados em dois blocos (treino e teste).
Factores Extrínsecos	Sistema de produção, dieta e nível alimentar.
Factores Intrínsecos	Raça, idade e sexo.

Ferramentas de Visualização	Permitem melhorar a compreensão dos resultados obtidos num processo de *DCBD* e a comunicação entre os utilizadores envolvidos no processo.
Flavour	Combinação das características do sabor com o aroma. É a facilidade do composto se evaporar no ar, e percepcionado directamente, logo após a amostra se encontrar na boca.
Generalização	Apresentam-se exemplos sobre um determinado problema para que a rede seja capaz de generalizar quando situações similares se apresentarem.
Impressão de Dureza	Resulta de três fontes: *i)* facilidade com que os dentes cortam a carne *ii)* Facilidade com que a carne se rompe em fragmentos e; *iii)* quantidade de resíduo que fica após a mastigação.
in vivo	Análises em animais vivos.
Indicação Geográfica Protegida	Um produto tem que demonstrar que pelo menos uma parte do ciclo produtivo tem origem no local que lhe dá o nome e que tem uma "reputação" associada a essa mesma região.
Interpretação de Resultados	Fase da *DCBD* onde se interpretam os resultados e se avalia a veracidade do conhecimento descoberto.
Marmoreado	Gordura visível nas superfícies de corte das carnes.
Métodos Destrutivos	Métodos usados para a análise da qualidade da carcaça, são exemplos a dissecação e as análises químicas.
Mínimo Local	É o valor mais reduzido numa dada vizinhança.
Não Linearidade	Capacidade de modelar funções não lineares.
Neurónio ou Unidade de Processamento	Permite três funções básicas: entrada, processamento e saída de sinais.
Painel Sensorial	Grupo de pessoas que são utilizadas como instrumentos de medida para quantificar as sensações percebidas pelos

	órgãos dos sentidos aquando das provas.
Perceptrão de Camada Única	*RNA* unidireccional com uma camada.
Perceptrão Multicamada	Trata-se do tipo mais popular de rede neuronal onde só existem conexões unidireccionais. Os nodos estão organizados por camadas existindo uma ou mais camadas intermédias e uma camada de saída.
pH	Nível de acidez;
Pré-Processamento dos Dados	São removidas as inconsistências nos dados e integrados, visando adequá-los aos algoritmos do processo de *DCBD*.
Produto de Qualidade	Aquele que serve perfeitamente, de forma aceitável, acessível, segura e no momento em que for solicitado, às necessidades e anseios do consumidor.
Qualidade da Carcaça	Medida primária de produção e um critério chave no melhoramento genético das raças.
Qualidade da Carne	São as propriedades mensuráveis do produto, ou seja, aquela que é tenra, suculenta, possui baixo teor em gordura e osso e seja agradável ao paladar.
Redes Neuronais Artificiais (RNA)	Sistema conexionista que tem a capacidade de aprender a partir de dados, de forma a atingir um objectivo específico.
Regra de Activação	Calcula o valor de activação de um neurónio num determinado instante.
Regressão	Tarefa de previsão que pretende modelar uma função de mapeamento entre um conjunto de entradas, designadas por variáveis independentes, e uma saída numérica, chamada variável dependente.
Rendimento da Carcaça	Percentagem de carcaça obtida, relativamente ao peso vivo do animal.

Sabor	Detecta as quatro sensações gustativas básicas (doce, salgado, ácido e amargo).
Selecção dos Dados	Identificação das bases de dados relevantes para a análise do processo.
Sistemas de Classificação de Carcaças	Definem as regras para as transacções comerciais.
Sobre-Ajustamento	A rede fixa-se em demasia nos casos de treino.
Suculência	Trata-se de uma impressão de humidade durante as primeiras mastigações e o efeito estimador da gordura na produção de saliva.
Sumariação	Tarefa que visa obter uma descrição completa de um conjunto de dados, que os distinga de outros.
Tecidos da Carcaça	É a composição tecidular das peças que se obtêm a partir do corte da carcaça.
Técnicas Estatísticas	É a área da matemática que estuda a colecção, organização e interpretação de dados.
Tenrura	É a manifestação à mastigação e resistência à aplicação de uma força.
Transformação dos Dados	Conjunto de operações diversas que modificam o modo de representação dos atributos.
Validação Cruzada k-desdobrável	Conjunto de dados é divido em k sub-conjuntos de igual tamanho. Sequencialmente, será testado um subconjunto diferente, sendo os restantes dados utilizados para ajustar o modelo de aprendizagem. No final dos k treinos, o modelo foi testado em todos os dados de treino, sendo a estimativa de generalização dada pela média do erro ao longo dos k conjuntos de teste.
Warner-Bratzler	Máquina ou célula de corte para analise da *tenrura* da carne.

Bibliografia

Agricultura.pt, M. d. [2004]. *AgroPortal - Qualidade Alimentar*, http://www.agroportal.pt/.

Ambrósio, P. E. [2002] *Redes neurais artificiais no apoio ao diagnóstico diferencial de lesões intersticiais pulmonares*. Tese de Mestrado, Universidade de São Paulo.

Angulo, E. B. [2001], Introducción al análises sensorial. *Análisis sensorial de alimentos. Métodos e aplicaciones*, Barcelona, **pp:** 1-11, Zootec 2001.

Arvanitoyannis, I. e M. Houwelingen-Koukaliaroglou [2003], Implementation of chemometrics for quality control and authentication of meat and meat products. *Critical Reviews in Food Science and Nutrition,* **43**(2): 173-218.

Baptista, S. M. G. [2004] *Caracterização sensorial da carne das raças autóctones de ovinos da região da terra fria de trás-os-montes*. Tese de Mestrado, Universidade Técnica de Lisboa.

Barreto, J. M. [2002]. Introdução às redes neuronais artificiais. Florianópolis, Universidade Federal de Santa Catarina**:** 57.

Bartlett, W. [1998], The sample complexity of pattern classification with neural networks: the size of the weights is more Important than the size of the network. *IEEE transactions on information theory,* **44**(2): 525-536.

Beal, F. D. e M. C. Smith [2000], Temporal difference learning for heuristic search and game. *Information Sciences,* **122**(1): 3-21.

Berg, E. P., B. A. Engel e J. C. Forrest [2001], Pork carcass composition derived from a neural network model of electromagnetic scans. *American society of animal science,* **76**(1): 18-22.

Berian, M. [1998], Calidad de la carne ovina. *Ovino de carne: aspectos claves*, España, **pp:** 404-418, C. Mundi-prensa.

Berry, M. J. A. e G. S. Linoff [2000]. *Mastering data mining*. Wiley computer publishing, Canada.

Bigus, J. P. [1991]. *Data mining with neural networks*. Mcgraw-Hill, USA.

Boccard, R. e B. L. Dumont [1976], La qualite des carcasses ovines. *Croissance, engraissement et qualité des carcasses d'agneauxe et de chevreaux*, Paris, **pp:** 44-78, Deuximiémes journées de la recherve ovine et caprine INTRA-ITO-VIC.

Bose, N. e P. Liang [1996]. *Neural networks fundamentals with graphs, algorithms and applications*. McGraw-Hill, USA.

Brachman, R. e T. Arnand [1996]. *The process of knowledge discovery in databases*. AAAI/MIT Press, USA.

Bradley, P., U. Fayyad e O. Mangasarian [1998]. Data mining: overview and optimazation opportunities. *Techinical report MSR-TR-98-04*. Redmond, WA, Microsoft research report.

Braga, A. P. e A. P. L. F. Carvalho [2000]. *Redes neuronais artificiais: teoria e aplicações*. LTC, Belo Horizonte.

Brío, B. M. e A. S. Molina [2001]. *Redes neuronales y sistemas barrosos*. Rama Editorial, Madrid.

Broomhead, D. e D. Lowe [1988], Multivariable functional interpolation and adaptative networks. *Complex systems,* **2**(1): 321-355.

Cabena, P., P. Hadjnian, R. Stadler, J. Verhees e A. Zanasi [1997]. *Discovering data mining from concept to implementation*. Prentice Hall PTR, USA.

Cadavez, V., S. Rodrigues, E. Pereira, R. Delfa e A. Teixeira [2000], Predicción de la composición de la canal de cabritos por ultrasonografia in vivo. *ITEA,* **1**(1): 39-50.

Cadavez, V. A. P. [2004] *Ultra-sonografia para avaliar in vivo e ex vivo carcaças de ovinos. Estudos nas raças Churra Galega Bragançana e Suffolk*. Tese de Doutoramento, Universidade de Trás-os-Montes e Alto Douro.

Cadavez, V. A. P., S. Rodrigues, E. Pereira, R. Delfa e A. Teixeira [2002], Predicción de la composición de la canal de cabritos por ultrasonografia *in vivo*. *ITEA. Producción animal*, Espanha, **pp:** 39-50, Consorci de biblioteques universitàries de catalunya.

Cadavez, V. A. P., A. Teixeira, R. Delfa e S. Rodrigues [2000], Utilización de ultrasonidos y el peso de la canal caliente para la prediccion de la composición de la canal en corderos. *XXV jornadas científicas y IV internacionales dela sociedad Española de ovinotecnia y caprinotecnia*, Teruel, **pp:** 169-172, Producción ovina y caprina, SEOC.

Cañeque, V. e C. Sañudo [2000], Condiciones y técnicas para controlar la calidad del producto. *Metodologia parael estudio de la calidad de la canal y de la carne en rumiantes*, Madrid, **pp:** 17-41, Instituto nacional de investigacion y tecnología agraria y alimentaria.

Carmack, C. F., C. L. Kastner, M. E. Dikeman, J. R. Schwenlke e L. M. G. Zepeda [1995], Sensory evaluation of beef-flavour-intensity, tenderness and juiciness among major muscles. *Meat Science,* **39**(1): 143-147.

Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer e R. Wirth [2000]. *Crisp-DM 1.0: Step-by-step data mining methods*, http://www.crisp-dm.org.

Chatterjee, S., A. S. Hadi e B. Price [2000]. *Regression analysis by example*. John Willey & Sons, Inc., New York.

Chen, M. S., J. Han e P. S. Yu [1996], Data Mining: an overview from a database perspective. *IEEE Transactions on knowledge and data engineering*: 866-883.

Colomer-Rocher, F. [1993]. Producción de canales ovinas frente al mercado común europeo. Interés de la denominación de origen del ternasco aragonés. Zaragoza, Institución Fernando el Católico.

Cortes, C. e V. Vapnik [1995], Support vector networks. *Machine learning,* **20**(1): 273-297.

Cortez, P. A. R. [2002] *Modelos inspirados na natureza para a previsão de séries temporais*. Tese de Doutoramento, Universidade do Minho.

Costa, P. T. d. [2003] *Uma análise do consumo de energia em transportes nas cidades portuguesas utilizando redes neuronais artificiais*. Tese de Mestrado, Universidade do Minho.

Costell, E. e L. Durán [1981], El análisis sensorial en el control de calidad de los alimentos. *Rev. Agrquímica y Tecnología de Alimentos,* **21**(2): 149-166.

Cottrell, G. W. [1985]. A connectionist approach to word sense disambiguation. USA, University of Rochester.

CRISP-DM [1999]. *CROSS industry standard process for data mining*, http://www.crisp-dm.org/.

Cross, H. R. [1994], Caracteristicas organolepticas de la carne. *Ciencia de la carne e de los produtos carnicos*, Zaragoza, **pp:** 279-298, PRICE, J.F.

Delfa, R., C. González, L. Torrano e J. Vaderrabano [1999], Utilización de ultrasonidos en cabritos vivos de raza blanca eltibérica, como predictores de la composición tisular de sus canales. *Archivos de Zootecnia,* **48**(182): 123-134.

Delfa, R., A. Teixeira e C. González [1998], El peso vivo matadero y ultrasonidos como predictores de la calidad de la canal y del reparto de la grasa en cabras adultas. *Revista Portuguesa de Zootecnia,* **2**(1): 1-16.

Dhar, V. e R. Stein [1997]. *Seven methods for transforming corporate data into business intelligence*. Prentice-Hall, New Jersey.

Dietterich, T. [2000], Ensemble methods in machine learning. *Springer,* **1857**(2000): 1-15.

Diéz, J., G. Bayón, J. Quevedo, J. Coz, O. Luaces, J. Alonso e A. Bahamonde [2004], Discovering relevancies in very difficult regression problems: applications to sensory data analysis. *Proceedings of the european conference on artifficial intelligence (ECAI 04)*, **pp:** 993-994,

Efron, B. [1983], Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the american statistical association,* **78**(1): 316-331.

Falas, T. [1995], Neural networks in empirical accounting research: an alternative to statistical models. *Neural network world,* **5**(4): 419-432.

Fayyad, U. M. e G. Piattetsky-Shapiro [1996], Knowledge discovery and data mining: Towards a unyfying framework. *Second international conference on knowledge discovery and data mining*, Portland, Oregon, **pp:** 37-54, American association for artificial intelligence.

Fayyad, U. M., G. Piattetsky-Shapiro, P. Smyth e R. Uthurusamy [1996]. *Advances in knowledge discovery and data mining*. AAAI Press,

Ferreira, A. M. P. J. [2000] *Dados geoquímicos de base de sedimentos fluviais de amostragem de baixa densidade de portugal continental: estudo de factores de variação regional*. Tese de Doutoramento, Universidade de Aveiro.

Fraser, A. e J. T. Stamp [1989]. *Ganado ovino - producción y enfermedades*. Ediciones Mundi-Prensa, Madrid.

Frawley, W., G. Piatetsky-Shapiro e C. Matheus [1991]. *Knowledge discovery in databases: An overview*. AAAI Press, USA.

Freeman, W. T. [1992] *Steerable filters and local analysis of image structure*. Tese de Doutoramento, Massachusetts Institute of Technology.

Gardner, S. R. [1998], Building the Data Warehouse. *Communications of the ACM,* **41**(9): 52-60.

Glymour, C., D. Mandigan, D. Pregibon e P. Smyth [1997], Data Mining and knowledge discovery. *Statistical themes and lessons for data mining,* **1**(1): 11-28.

González, C., R. Delfa e E. Vijil [1996]. The conformation as predictor of carcass quality of adult Blanca Celtibérica goats, In: 47th Annual Meeting of the EAAP**:** 273-328.

Gracey, J. F. e D. S. Collins [1992], Meat hygiene. *Bailliére tindall*, London, **pp:** 416-419, Bailliére Tindall.

Grunsky, E. C. [2002], R: a data analysis and statistical programming environment - an emerging tool for geosciences. *Computers & Geosciences*, Canada, **pp:** 1219:1222, Elsevier Science Ltd.

Han, J. e M. Kamber [2001]. *Data mining, concepts and techniques*. Morgan Kaufmann publishers, USA.

Hand, D., H. Mannila e P. Smyth [2001]. *Principles of data mining*. The MIT Press, USA.

Hastie, T., R. Tibshirani e J. Friedman [2001]. *The elements of statistical learning: data Mining, inference, and prediction*. Springer-Verlag, USA.

Haykin, S. [1999]. *Neural networks: A comprehensive foundation*. Prentice-Hall. Second edition, New Jersey.

Haykin, S. [2001]. *Kalman filtering and neural networks*. A Wiley-Interscience Publications - John Wiley & Sons, Inc., New York.

Hill, B. D., S. D. M. Jones, W. M. Robertson e I. T. Major [2000], Neural network modeling of carcass measurements to predict beef tenderness. *Can. J. Anim. Sci.,* **80**(2): 311-318.

Hinton, G. E. e T. J. Sejnowski [1986]. *Learning and relearning in Boltzmann machines*. MIT Press, Cambridge.

Hofmann, K. [1990], Definition of meat quality. *Proceedings of the 37th international congress of meat science and technology*, Cambridge, **pp:** 941-954, MIT.

Hopfield, J. J. [1982], Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, USA, **pp:** 2554-2558, Proc. Nat. Academic Science.

Hopfield, J. J. [1984], Neurons with graded response have collective computational properties like those of two-state neurons. *The national academy of sciences*, EUA, **pp:** 3088-3092, Proc. Nat. Academic Science.

Hopkins, D. L., M. A. Anderson, J. E. Morgan e D. G. Hall [1995]. A probe to measure GR in lamb carcasses at chain speed. Canada, Meat science**:** 159-165.

Huffman, K. L., M. F. Miller, L. C. Hoover, C. K. Wu, H. C. Brittin e C. B. Ramsey [1996], Effect of beef tenderness on consumer satisfaction with steaks consumed in the home and restaurant. *J. Animal Science*: 91-97.

Huffman, K. L., M. F. Miller, L. C. Hoover, C. K. Wu, H. C. Brittin e C. B. Ramsey [1997], Effect of beef tenderness on consumer satisfaction with steaks consumed in the home and restaurant. *Journal Animal Science,* **74**(1): 91-97.

Ihaka, R. e R. Gentleman [1996], R: a language for data analysis and graphics. *Journal of computational and graphical statistics,* **5**(1): 299-314.

Inmon, W. H. [1996], The data warehouse and data mining. *Communications of ACM,* **39**(11): 49-50.

Jeremiah, L. E. [1998], Development of a quality classification system for lamb carcasses. *Agriculture and agri-food*, Canada, **pp:** 211-223, Meat Science.

Johnston, R. G. [1983]. *Introduction to sheep farming*. Granada publishing limited - Technical books division, Great Britain.

Kamdem, A. T. K. e J. Hardy [1995], Grinding as a method of meat texture evaluation. *Meat Science,* **4**(2): 225-236.

Kempster, A. J. [1983], Carcass quality and its measurement in sheep. *Sheep Production*, London, **pp:** 59-74, Science Direct.

Kempster, A. J. [1989], Carcass and meat quality research to meet market needs. *Animal Production,* **43**(1): 483-496.

Kempster, A. J., D. W. Jones e B. T. Wolf [1986], A comparison of alternative methods for predicting the carcass composition of crossbred lambs of different breeds and crosses. *Meat Science,* **18**(1): 89-110.

Kewley, R., M. Embrechts e C. Breneman [2000], Data strip mining for the virtual design of pharmaceuticals with neural networks. *IEEE Transactions on Neural Networks,* **11**(3): 668-679.

Kirton, A. H., G. J. K. Mercer e D. M. Duzanzich [1992], A comparison between subjective and objective (carcass weight plus GR or the hennessy probe) methods for classifying lamb carcass. *MAF technology, ruakura agricultural centre*, New Zealand, **pp:** 41-44, Proceedings of the New Zealand society os animal production.

Kohavi, R. [1995], A study of cross-validation and bootstrap for accuracy estimation and model selection. *Conference on artificial intelligence*, Canada, **pp:** 33-42, In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI).

Kwon, O., B. Golden e E. Wasil [1995], Estimating teh lenght of the optimal TSP tour: an empirical study using regression and neural networks. *Computers & operations research,* **22**(10): 1047-1056.

Li, J., J. Tan, F. A. Martz e H. Heymann [1999], Image texture features as indicators of beef tenderness. *Meat Science,* **53**(1): 17-22.

Li, J., J. Tan e P. Shatadal [2000], Classification of tough and tender beef by image texture analysis. *Meat Science,* **80**(12): 341-346.

Liu, Y., X. Yao e T. Higuchi [2000], Evolutionary ensembles with negative correlation learning. *IEEE transactions on evolutionary computation,* **4**(4): 380-387.

Lu, W. e J. Tan [2003], Analysis of image-based measurements and USDA characteristics as predictors of beef lean yield. *Meat Science,* **65**(2): 483-491.

Manilla, H. [1994], Finding interesting rules from large sets of discovered association rules. *3rd international conference on information and knowledge management*, **pp**,

Mannila, H. [1996], Data mining: Machine learning, statistics, and databases. *Eight international conference on scientific and statical database management*, Helsinki, **pp:** 28-36, Department of Computer Science.

Michalski, R. e K. Kaufman [1998]. *Data mining and knowledge discovery: a review of issues and multistrategy approach*. Eds. New York: Wiley, USA.

Minsky, M. L. e S. A. Papert [1969]. *Perceptrons: an introduction to computational geometry*, Massachussets.

Mitchell, T. [1997]. *Machine Learning*. Mc Graw Hill, USA.

Mitra, S., S. Pal e P. Mitra [2002], Data mining in soft computing framework: a survey. *IEEE Trans. on Neural Networks,* **13**(1): 3-14.

Oeckel, M. J. V., N. Warnants e C. V. Boucqué [1999], Pork tenderness estimation by taste panel, Warner-Bratzler shear force and on-line methods. *Meat Science,* **53**(1): 259-267.

Paradis, E. [2003]. R para principiantes. Montepellier, Institut des Sciences de l'Évolution - Universit Montpellier II**:** 60.

Park, B., Y. Chen, W. Hruschka, S. Shackelford e M. Koohmaraie. [2002], Principal component regression of near-infrared Reflectance spectra for beef tenderness prediction. *Transactions of the ASAE,* **44**(3): 609-615.

Park, B., Y. R. Chen, A. D. Whittaker, R. K. Miller e D. S. Hale [1994], Neural network modeling for beef sensory evaluation. *ASAE,* **4**(2): 1547-1553.

Prechelt, L. [1998]. *Early Stopping - but when?* Springer Verlag, Heidelberg.

Pyle, D. [1999]. *Data preparation for data mining*. Morgan Kaufmann, S. Francisco CA, USA.

RDCT, R. D. C. T. [2004]. *R: A language and environment for statistical computing*. R foundation for statistical computing, Vienna, Austria.

Rich, E. e K. Knight [1993]. *Inteligência Artificial*. FCA - Editora de Informática, Portugal.

Riedmiller, M. [1994], Supervised learning in multilayer perceptrons - from backpropagation to adaptive learning techniques. *Computer standards and interfaces*: 1-16.

Rocha, C. A. J. [1999] *Redes bayesianas para extracção de conhecimento em base de dados, considerando a incorporação de conhecimento de fundo e o tratamento de dados incompletos*. Tese de Doutoramento, Universidade de São Paulo.

Rocha, M., P. Cortez e J. Neves [2005], Simultaneous evolution of neural network topologies and weights for classification and regression. *Lecture Notes in Computer Science,* **3512**(1): 59-66.

Rodrigues, A. M. [2003], O selo ecológico na união europeia. *Anais V1-palestras ZOOTEC 2003*, Brasil, **pp:** 162-179, ZOOTEC 2003.

Rodrigues, S. S. Q. [2002] *Estudo da qualidade da carcaça de cordeiros das raças Churra Galega Bragançana e Suffolk*. Tese de Mestrado, Universidade de Trás-os-Montes e Alto Douro.

Rumelhart, D. E. e J. M. Clelland [1986]. *Parallel distributed processing: exploration the microstructure of cognition 1,2,3*. Bradford Book - MIT Press, Cambridge.

Rumelhart, D. E., G. E. Hinton e R. J. Williams [1986], Learning representations by backpropagation errors. *Nature,* **323**(1): 533-536.

Russel, S. e P. Norvig [1995]. *Artificial intelligence - A modern approach*. Prentice-Hall, New Jersey.

Safari, E., N. M. Fogarty, G. R. Ferrier, L. D. Hopkins e A. Gilmour [2001], Diverse lamb genotypes. Eating quality and the relationship between its objective measurement and sensory assessment. *Meat Science,* **83**(2): 65-69.

Santos, M. Y. C. A. [2001] *Um sistema de descoberta de conhecimento em bases de dados geo-referenciadas*. Tese de Doutoramento, Universidade do Minho.

Santos, V., S. Silva, J. Azevedo e E. Gomes [2000], Estimativa da composição da carcaça de ovinos da raça Ile-de-France a partir de medidas obtidas por ultrasons ao nível da 13ª vértebra dorsal e entre as 3ª e 4ª vértebras lombares. *Revista Portuguesa de Zootecnia,* **1**(1): 91-104.

Sañudo, C., G. R. Nute, M. M. Campo, G. María, A. Baker, I. Sierra, M. E. Enser e J. D. Wood [1998], Influence of weaning on carcass quality, fatty acid composition and meat quality in intensive lamb production systems. *Journal of Animal Science,* **66**(1): 175-187.

Sañudo, C. e I. Sierra [1986], Calidad de la canal en la especie ovina. *In Ovino*, Barcelona, **pp:** 127-153, E.S.A One, ed.

Sarle, W. [1995], Stopped training and other remedies for overfitting. *27 th. symposium interface of computer science and statistics*, **pp:** 352-360,

Sarle, W. [2005]. *Neural network frequently asked questions*, ftp://ftp.sas.com/pub/neural/FAQ.html.

Setiono, R. [2003], Techniques for extracting classification and regressiona rules from artificial neural networks. *Computational intelligence: the experts speak*, IEEE, Piscataway NJ, USA, **pp:** 99-114, D. Fogel and C. Robinson, eds.

Shannon, C. E. e E. McCarty [1956]. *Automata Studies*. Princeton University Press, New Jersey.

Silva, M. M. M. F. d. [1996] *Crecimento, características da carcaça e qualidade da carne de raças bovinas nacionais*. Tese de Doutoramento, Universidade Técnica de Lisboa.

Smulders, F. J. M., R. L. J. M. V. Laack e G. Eikelenbbom [1991], Muscle and meat quality: Biological basis, processing, preparation. *European meat industry in the 1990's*, Utrecht, **pp**,

Stanford, K., A. T. McAllister, M. MacDougall e D. R. C. Bailey [1995], Use of ultrasound for the prediction of carcass characteristics in alpine goats. *Small Ruminant Research,* **75**(1): 195-201.

Stanford, K., C. M. Woloschuk, L. A. McClelland e S. D. M. Jones [1997], Comparison of objective external carcass measurements and subjective conformation scores for prediction of lamb carcass quality. *Canadian Journal Animal Science,* **77**(1): 217-223.

StatSoft, E. T. [2005]. *STATISTICA is a trademark of StatSoft, Inc.*, http://www.statsoft.com/textbook/stathome.html.

Stone, H. [1999], Sensory evaluation: Science and mythology. *Food technology,* **53**(10): 124-133.

Subramanian, V., M. S. Hung e M. Y. Hu [1993], An experimental evaluation of neural networks for classification. *Computers & operations research,* **20**(7): 769-782.

Tan, J. [2004], Meat quality evaluation by computer vision. *Journal of Food Engineering,* **61**(1): 27-35.

Teixeira, A., V. Cadavez, M. S. Bueno, S. Baptista, S. Rodrigues e R. Delfa [2003], Características da carcaça e da carne de cordeiros das raças churra galega bragançana e churra galega mirandesa. *Produzir qualidade em segurança*, Évora, **pp:** 7, XIII Congresso de Zootecnia.

Teixeira, A., R. Delfa e P. Alberti [1998], Influence of production factors on the characteristics of meat from ruminants in mediterranean area. *Proceedings of the international symposium on basis of the quality of typical mediterranean animal products*, Badajoz, **pp:** 315-319, CIRVAL.

Tian, Y. Q., j. Tan, D. G. McCall e P. Gong [2002], Evaluating beef tenderness of grazing animals from raw meat surface features. *Meat Science,* **42**(3): 125-130.

Turban, E., J. Aronson e T. Liang [2004]. *Decision support systems and intelligent systems*. Prentice Hall, UK.

Vergara, H. e L. Gallego [1999], Effect of type of suckling and length of lactation period on carcass and meat quality in intensive lamb production systems. *Meat Science,* **53**(3): 211-215.

Warkup, C. [1997], The development of "blueprint" specifications for improved meat quality. *Internation congress of meat science and technology*, Auckland, **pp:** 26, Meat science and technology.

Wasserman, P. D. [1989]. *Advanced methods in neural computing*. Van Nostrand Reinhold, New York.

Wong, B. K., V. S. Lai e J. Lam [2000], A bibliography of neural network business applications research. *Computers & Operations Research,* **27**(11-12): 1045-1076.

Wong, D. W. S. [1995]. *Quimica de los alimentos. Mecanismos e teoria*. Editorial Acribia, S. A, Zaragoza.

Yamamoto, Y. e P. Nikiforuk [2000], A new supervised learning algorithm for multilayered and interconnected neural networks. *IEEE transactions on neural networks,* **11**(1): 1313-1335.