Universidade do Minho
Escola de Engenharia

Filipe José Gomes da Silva

Aplicação de técnicas de *Data Mining* na avaliação da qualidade da carne de cordeiro

Tese de Mestrado
Mestrado em Engenharia e Gestão de Sistemas de Informação

Trabalho efectuado sob a orientação do
Professor Doutor Paulo Cortez

Outubro de 2010

Agradecimentos

Ao professor e orientador Doutor Paulo Cortez deixo um agradecimento muito especial pelas suas inestimáveis orientações e ensinamentos, amizade, confiança, incentivos e solidariedade demonstrados nos momentos mais difíceis desde trabalho.

A toda a equipa responsável pela recolha dos dados, em especial para o Doutor Vasco Cadavez, meu co-orientador, a quem agradeço todos os esclarecimentos, orientações e disponibilidade demonstrada.

A todos os meus colegas e amigos, em especial para o Miguel Martins, Ricardo Dinis, Ricardo Fernandes, José Mogollon, Afonso Tomás, Tânia Sousa e Paulo Nóvoa, pela infinita paciência e pelo inestimável apoio.

Por último, mas não menos importante, quero deixar um agradecimento a toda a minha família, em especial aos meus pais e irmãos, pelo apoio incondicional.

Aplicação de Técnicas de *Data Mining* na Avaliação da Qualidade da Carne de Cordeiro

Resumo

A composição corporal dos animais, de todas as espécies, varia consideravelmente consoante o estádio de crescimento, o plano nutricional e a base genética. Tendo isso em conta, e considerando que os vários tecidos que compõem a carcaça possuem valor económico diferente, é fácil admitir que o valor económico dos animais dependa da composição da carcaça. Assim, este trabalho visa utilizar uma abordagem de *Data Mining* para prever, utilizando como entradas as medições obtidas na linha de abate, a composição tecidual de carcaças de cordeiro. Cento e vinte e cinco cordeiros da raça Churra Galega Bragançana foram abatidos. Durante o quarteamento das carcaças, foi utilizado um paquímetro para realizar medições da gordura subcutânea entre a 12ª e 13ª costelas (C12), e entre a 1ª e 2ª vértebra lombar (C1). As proporções de Músculo (PM), Osso (PO), Gordura Subcutânea (PGS), Gordura Intermuscular (PGI), e Gordura Pélvica e Renal (PGPR) das carcaças de cordeiro foram colectadas numa base de dados. Utilizamos a biblioteca rminer da ferramenta R e comparamos três técnicas de regressão: Regressão Múltipla (RM), Redes Neuronais Artificiais (RNA) e Máquinas de Vectores de Suporte (MVS). O modelo de RM apresenta o Erro Relativo Absoluto (RAE) mais baixo para PM (RAE=59.4%, $P<0.05$) e para PGI (RAE=64.1%, $P<0.05$). O modelo MVS apresenta o RAE mais baixo para PO (RAE=46.1%, $P<0.05$), para PGPR (RAE=51.5%, $P<0.05$), e para PGS (RAE=42.2%, $P<0.05$). Para além disso, um procedimento de análise de sensibilidade revelou a medida C12 como a entrada mais importante para todos os cinco tecidos da carcaça.

Foi ainda apresentado um exemplo de um novo sistema de classificação de carcaças, desenvolvido através de um procedimento de *Clustering* e baseado em dados objectivos.

Palavras-Chave: Carcaça, Tecido, Regressão Múltipla, Redes Neuronais Artificiais, Máquinas de Vectores de Suporte, *Clustering*.

Applying Data Mining Techniques to Lamb Meat Quality Assessment

Abstract

The body composition of animals from all species varies considerably depending on the stage of growth, nutritional plan and genetic basis. Taking this into account, and considering that the various tissues that comprises the carcass have different economic value, it is easy to admit that the economic value of an animal depends of the carcass composition. Therefore, this study aims at applying a Data Mining approach to predict, using carcass measurements taken at slaughter line as predictors, the tissue composition of lamb carcasses. One hundred and twenty five lambs from the Churra Galega Bragançana breed were slaughtered. During quartering, a caliper was used to measure the subcutaneous fat depth between the 12th and 13th ribs (C12), and between the 1st and 2nd lumbar vertebrae (C1). The Muscle (PM), Bone (PO), Subcutaneous Fat (PGS), Inter-muscular Fat (PGI), and Kidney Knob and Channel Fat (PGPR) proportions were computed. We used the rminer R library and compared three regression techniques: Multiple Regression (RM), Artificial Neural Networks (RNA) and Support Vector Machines (MVS). The RM model presents the lower Relative Absolute Error (RAE) for PM (RAE=59.4%, $P<0.05$) and PGI (RAE=64.1%, $P<0.05$). The MVS model presents the lowest RAE for PO (RAE=46.1%, $P<0.05$), PGPR (RAE=51.5%, $P<0.05$), and PGS (RAE=42.2%, $P<0.05$). In addition, a sensitivity analysis procedure revealed the C12 measurement as the most important predictor for all five carcass tissues.

An example of a new carcass classification system was also presented, which was developed through a clustering procedure and based on objective data.

Keywords: Carcass, Tissue, Multiple Regression, Artificial Neural Networks, Support Vector Machines, Clustering.

Conteúdo

Notação

ACOB Associação Nacional de Criadores de Ovinos da Raça Churra Galega Bragançana

AD Árvores de Decisão

AP Agrupamento de Produtores

BI Business Intelligence

CE Comunidade Europeia

CRISP-DM CRoss-Industry Standard Process for Data Mining

DM Data Mining

DO Denominação de Origem

DOP Denominação de Origem Protegida

GPP Gabinete de Planeamento e Políticas

IG Indicação Geográfica

IGP Indicação Geográfica Protegida

MADRP Ministério da Agricultura, do Desenvolvimento Rural e das Pescas

MMG Mistura de Modelos Gaussianos

MVS Máquinas de Vectores de Suporte

OC Organismo Privado de Controlo e Certificação

OLAP On-Line Analytical Processing

PGI Proporção de Gordura Intermuscular

PGPR Proporção de Gordura Pélvica e Renal

PGS Proporção de Gordura Subcutânea

PM Proporção de Músculo

PO Proporção de Osso

PTG Proporção Total de Gordura

RAE Relative Absolute Error

REC Regression Error Characteristic

RM Regressão Linear Múltipla

RNA Redes Neuronais Artificiais

SAD Sistema de Apoio à Decisão

SEMMA Sample, Explore, Modify, Model, Assess

SI Sistemas de Informação

SQL Structured Query Language

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Motivação

O desenvolvimento de um método rápido e económico para prever a composição de carcaças terá aplicação para a classificação de carcaças na linha de abate (Cadavez et al., 1999), e para a definição de preços ao longo da cadeia comercial (Cadavez et al., 2002). Carcaças com uma composição óptima deverão ter um máximo de percentagem de carne magra, e óptimas propriedades organolépticas. Neste caso, a carcaça deverá ter um preço máximo e se a composição da carcaça se desviar deste óptimo o seu preço deverá sofrer penalizações.

Tradicionalmente, os produtores estimam a composição de carcaças de cordeiro através de métodos subjectivos, logo imprecisos, como a avaliação visual ou a palpação. No entanto, a metodologia para prever a composição de carcaças na linha de abate deve ser precisa, rápida e automatizada. As técnicas de *Data Mining* (DM) têm como objectivo extrair conhecimento de alto nível a partir de dados brutos (Witten and Frank, 2005) e podem representar uma alternativa interessante para prever a composição de carcaças, o que pode ser conseguido através da obtenção de parâmetros da carcaça na linha de abate.

Tipicamente, estes parâmetros são obtidos durante o processo de abate ou nas primeiras 24h após o abate. De facto, vários estudos adoptaram esta abordagem orientada aos dados baseados em modelos de Regressão Múltipla (RM) (Cadavez, 2009; Hopkins, 2008), usando como variável independente (ou de entrada) o peso da carcaça, em combinação com a profundidade da gordura subcutânea (Hopkins et al., 2008), profundidade do músculo *longissimus*, e espessura total de tecidos (Hopkins et al., 2008; Kirton et al., 1984). No entanto, estes modelos lineares podem falhar quando relações não-lineares estão presentes nos dados e quando a entrada sofre de múltipla colinearidade (Cadavez, 2009). Em tais cenários, existe a necessidade de técnicas de modelação alternativas, tais como as mais flexíveis Redes Neuronais Artificiais (RNA) ou Máquinas de Vectores de Suporte (MVS) (Hastie et al., 2008).

Nesta dissertação, seguimos uma abordagem DM para prever a composição tecidual de carcaças com base em medições não invasivas que podem ser facilmente obtidas após o abate. Em particular, comparamos três modelos de regressão (RM, RNA e MVS). Mais ainda, adoptamos um modelo de classificação via um *clustering* baseado numa mistura de modelos gaussianos, com o intuito de criar um sistema de classificação de carcaças de cordeiro que seja objectivo e adequado.

1.2 Objectivos

Este trabalho de investigação é relevante na medida em que nele se irão desenvolver novos métodos de aferição da composição tecidual de carcaças de cordeiro, com base em modelos gerados a partir de algoritmos de Regressão Múltipla, Redes Neuronais Artificiais e Máquinas de Vectores de Suporte. Assim, pretende-se ficar a conhecer quais as reais capacidades de cada técnica e qual a melhor forma de utilizá-las 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 (SAD) para operar em ambiente real, permitindo uniformizar critérios, dar resposta num curto espaço de tempo e reduzir custos. Assim, através da utilização de dados analíticos que poderão ser obtidos de forma fácil, rápida e económica no matadouro, pretende-se:

1. Obter modelos com uma boa capacidade de previsão da composição de carcaças de cordeiro;

2. Analisar os factores mais influentes nos modelos de previsão desenvolvidos;

3. Classificar a qualidade de carcaças de cordeiro com base na sua composição tecidual.

Os modelos de regressão a desenvolver terão como entradas valores físicos como o sexo do animal, peso da carcaça quente e várias medidas de profundidade de tecidos, referentes a um conjunto de dados disponibilizados no início

deste trabalho. O valor a prever será de acordo com as proporções de tecidos, obtidas por dissecação.

1.3 Organização

Será apresentado o contexto geral do trabalho, bem como as motivações que levaram ao desenvolvimento da investigação e os objectivos que se pretendem atingir. A dissertação é dividida em duas partes fundamentais. Na primeira (Capítulos 2 e 3) 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 realizado para aplicação dos objectivos inicialmente propostos, sendo também analisados e discutidos os resultados obtidos.

Em conformidade, esta dissertação possui cinco capítulos organizados da seguinte forma (excluindo o capítulo introdutório):

2 - *Business Intelligence* e *Data Mining* É apresentada uma revisão dos conceitos principais sobre *Business Intelligence* e *Data Mining*, explicando as várias etapas desse processo, bem como a metodologia utilizada e todos os procedimentos necessários à sua implementação.

3 - Avaliação da Qualidade da Carne de Cordeiro É apresentado o Cordeiro Bragançano e é analisada a influência da composição tecidual das carcaças sobre a qualidade da carne. São também apresentados alguns sistemas de classificação de carcaças mais relevantes e ainda alguns casos de estudo em que foram utilizados métodos de *Business Intelligence* para a previsão da composição tecidual de carcaças.

4 - Avaliação da Qualidade da Carne de Cordeiro via Técnicas de *Data Mining* Apresenta-se o trabalho prático realizado, incluindo uma descrição da ferramenta utilizada, uma descrição do trabalho realizado de acordo com a metodologia CRISP-DM, e por fim são apresentados e analisados os resultados obtidos.

5 - Conclusões Neste último capítulo, é feita uma síntese do trabalho realizado, são discutidas as conclusões mais importantes do trabalho desenvolvido, e são apresentadas as contribuições e recomendações para trabalhos futuros.

2 *Business Intelligence* e *Data Mining*

2.1 Introdução

As rápidas mudanças que se vivem no mercado actual fazem com que as empresas não possam adiar as decisões relacionadas com o negócio. É necessário contar com um sistema que represente o papel de suporte para a tomada de decisão, com resposta rápida e com a informação necessária para que a empresa usufrua das oportunidades que surjam: estar no lugar certo, no momento oportuno, com a informação correcta. Os sistemas orientados à tomada de decisão são definidos pelo termo *Business Intelligence* (BI). Estes permitem combinar a recolha de dados com ferramentas de análise, com vista a disponibilizar informação para a tomada de decisão (Santos and Ramos, 2006).

Estes sistemas incluem *Data Warehouses* (armazéns de dados), que são repositórios de informação organizacional. Incluem também os sistemas *On-Line Analytical Processing* (cubos OLAP), que providenciam uma análise multidimensional dos dados, e incluem ainda o uso de técnicas de *Data Mining* (DM) para extracção de conhecimento. DM é um termo utilizado para descrever a descoberta de conhecimento em bases de dados. É um processo que utiliza estatística, matemática, inteligência artificial e técnicas de aprendizagem para extrair e identificar informações úteis e consequente conhecimento de grandes bases de dados, envolvendo a utilização de algoritmos para a extracção e determinação de padrões observados nos dados. A escolha da melhor técnica está relacionada com o tipo de base de dados a estudar, com o conhecimento a extrair, com o problema a solucionar e principalmente com o objectivo do DM. Convém referir ainda que o DM passou a ser considerado pelas organizações como uma tecnologia crucial, a par das ferramentas OLAP, sendo mesmo considerada capaz de dotar as organizações de capacidade cognitiva (Cortes, 2005). A técnica de DM compreende diversas tarefas, algumas de previsão e outras de descrição. Nesta dissertação as tarefas de previsão utilizadas foram a Regressão e a Classificação. Como técnicas de descrição, foi utilizado um método de *clustering*, bem como árvores de de-

cisão. A metodologia CRISP-DM utilizada compreende as seguintes fases: **Compreensão do Negócio**, **Compreensão dos Dados**, **Preparação dos Dados**, **Modelação**, **Avaliação** e **Implementação**. Estas fases serão descritas pormenorizadamente ao longo deste capítulo.

2.2 *Business Intelligence*

O conceito de *Business Intelligence* (BI) vem ganhando cada vez mais força junto dos mais variados tipos de organização. O uso do conhecimento é um factor crítico de sucesso para qualquer organização, fazendo com que estas invistam cada vez mais em meios que as tornem mais eficientes nesse processo de produzir e disseminar conhecimento, transformando os seus **Sistemas de Informação (SI)** em peças chave para a definição das suas estratégias. O BI busca transformar a grande massa de dados da organização, produzindo conhecimento que possibilite auxiliar a tomada de decisão. O BI não é um sistema nem uma ferramenta mas sim um conceito que se aplica e que se vive no dia-a-dia de uma organização. Compreende qualquer ferramenta envolvida no ambiente organizacional que apresente dados que possam ser aproveitados pela organização das mais diversas formas, principalmente no que diz respeito à tomada de decisão. Os sistemas de BI têm aplicado a funcionalidade, escalabilidade e segurança dos actuais sistemas gestores de bases de dados para construir *Data Warehouses* que são analisados com técnicas de *On-Line Analytical Processing* e de *Data Mining*.

2.3 *Data Mining*

Existe actualmente uma quantidade enorme de dados armazenados, e essa quantidade cresce a cada segundo que passa. No entanto, à medida que o volume de dados aumenta, a proporção de dados que as pessoas percebem diminui, de forma alarmante. Escondida por entre todos estes dados existe informação, informação potencialmente útil, e que raramente é explicitada ou aproveitada (Witten and Frank, 2005).

Data mining é um processo que usa técnicas estatísticas, matemáticas, da inteligência artificial, e da aprendizagem de máquina para extrair e identificar

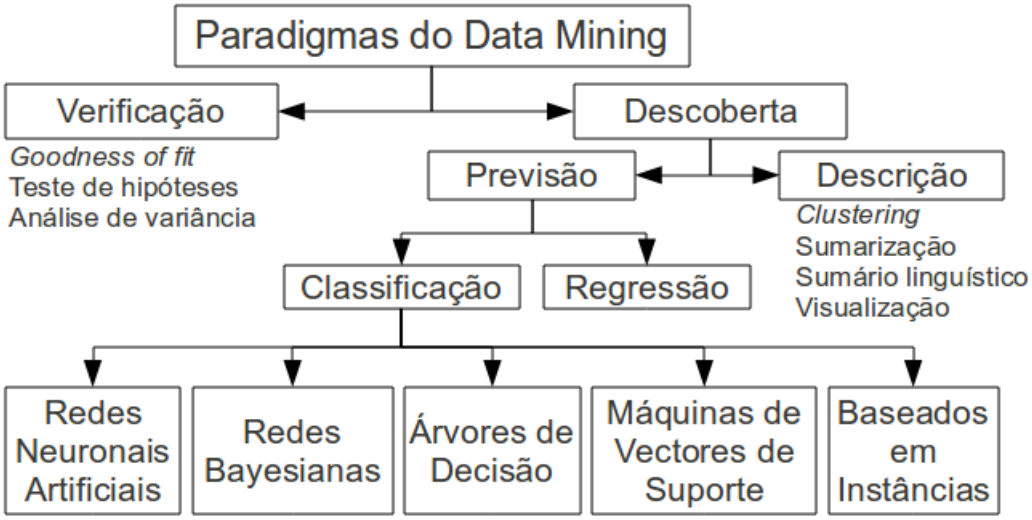

Figura 1: Taxonomia do *Data Mining* (Maimon and Rokach, 2005).

informação útil e subsequente conhecimento de grandes bases de dados. Isto é conseguido através da descoberta de padrões matemáticos, que podem ser regras, afinidades, correlações, tendências, ou modelos de previsão (Turban et al., 2007).

Três métodos são utilizados para identificar padrões nos dados (Witten and Frank, 2005):

- Modelos Simples (e.g. *Queries* baseadas em SQL, *On-Line Analytical Processing* [OLAP], julgamento humano)
- Modelos Intermédios (e.g. Regressões, Árvores de Decisão, *Clustering*)
- Modelos Complexos (e.g. Redes Neuronais Artificiais, outras induções por regras)

É útil distinguir dois tipos principais de DM: orientado à verificação (o sistema verifica as hipóteses do utilizador) e orientado à descoberta (o sistema descobre novas regras e padrões de forma autónoma) (Maimon and Rokach, 2005). A Figura 1 apresenta esta taxonomia.

Os métodos de descoberta são os que identificam padrões nos dados de forma automática. O ramo dos métodos de descoberta consiste em métodos

de previsão e métodos de descrição. Os métodos descritivos são orientados à interpretação dos dados, que se foca em perceber (e.g. por visualização) a forma como os dados subjacentes se relacionam com as suas partes. Métodos orientados à previsão têm como objectivo construir um modelo de comportamento, que obtém amostras novas e desconhecidas, e é capaz de prever o valor de uma ou mais variáveis relacionadas com a amostra. Também desenvolve padrões que formam a descoberta de conhecimento, de uma forma que é compreensível e fácil de utilizar como base de trabalho. Alguns métodos orientados à previsão podem também ajudar a perceber os dados (Maimon and Rokach, 2005).

A maioria das técnicas de DM orientadas à descoberta (as quantitativas em particular) são baseadas na aprendizagem indutiva, onde um modelo é construído, explicitamente ou implicitamente, por generalização de um número suficiente de exemplos de treino. A suposição por trás da abordagem indutiva é que o modelo de treino é aplicável a exemplos futuros, desconhecidos do modelo (Maimon and Rokach, 2005).

Por outro lado, os métodos de verificação lidam com a avaliação de hipóteses propostas por uma fonte externa (e.g. um especialista). Estes métodos são menos associados ao DM que os métodos orientados à descoberta, porque a maioria dos problemas de DM se preocupam em descobrir uma hipótese, em vez de testar uma hipótese conhecida (Maimon and Rokach, 2005).

Outra terminologia comum, utilizada pela comunidade da aprendizagem de máquina, refere-se aos métodos de previsão como métodos de aprendizagem supervisionada, contra a aprendizagem não-supervisionada (Maimon and Rokach, 2005).

A aprendizagem não-supervisionada refere-se principalmente a técnicas que agrupam instâncias sem um atributo dependente, pré-especificado. Assim, o termo "aprendizagem não-supervisionada" abrange apenas uma porção dos métodos descritivos da Figura 1. Por exemplo, abrange os métodos de *clustering* mas não os métodos de visualização (Maimon and Rokach, 2005).

Os métodos de aprendizagem supervisionada tentam descobrir a relação entre vários atributos de entrada (ou variáveis independentes) e o atributo de saída (ou variável dependente). A relação descoberta é representada numa

estrutura referida como um modelo. Normalmente, os modelos descrevem e explicam fenómenos que estão escondidos no conjunto de dados, e podem ser utilizados para prever o valor do atributo de saída conhecendo os valores dos atributos de entrada (Maimon and Rokach, 2005).

É útil distinguir entre dois grandes modelos supervisionados: modelos de **Classificação** e modelos de **Regressão**. Os modelos de Regressão mapeiam o espaço de entrada num domínio de valores reais. Por exemplo, um regressor pode prever a procura de um determinado produto dadas as suas características. Por outro lado, os classificadores mapeiam o espaço de entrada em classes predefinidas. Por exemplo, os classificadores podem ser utilizados para classificar os clientes com hipotecas como bons (pagam completamente a hipoteca a tempo) e maus (pagam atrasados), ou em quantas classes forem precisas. Existem muitas alternativas para representar os classificadores. Exemplos típicos incluem árvores de decisão ou máquinas de vectores de suporte (Maimon and Rokach, 2005).

2.4 Metodologia CRISP-DM

Se o processo de DM for enquadrado no contexto de uma metodologia, torna-se mais fácil de compreender, implementar e desenvolver (Santos and Azevedo, 2005). Existem duas metodologias principais de DM: *SEMMA - Sample, Explore, Modify, Model, Assess*, e o *CRISP-DM - CRoss-Industry Standard Process for Data Mining* (Santos and Azevedo, 2005). A metodologia CRISP-DM foi desenvolvida por um consórcio composto por *NCR Systems Engineering Copenhagen* (EUA e Dinamarca), *DaimlerChrysler AG* (Alemanha), *SPSS Inc* (EUA) e *OHRA Verzekeringen en Bank Groep B.V* (Chapman et al., 2000). Por sua vez, a SEMMA foi desenvolvida pela empresa *SAS*, cuja área de negócio é o BI e o Suporte à Decisão. O CRISP-DM é mais utilizado do que o SEMMA, sendo que neste trabalho também se optou por adoptar esta metodologia.

O desenvolvimento do CRISP-DM teve origem no interesse crescente e generalizado, por um lado do mercado de DM, e por outro, pelo consenso de que a indústria necessitava de um processo padronizado. A metodologia

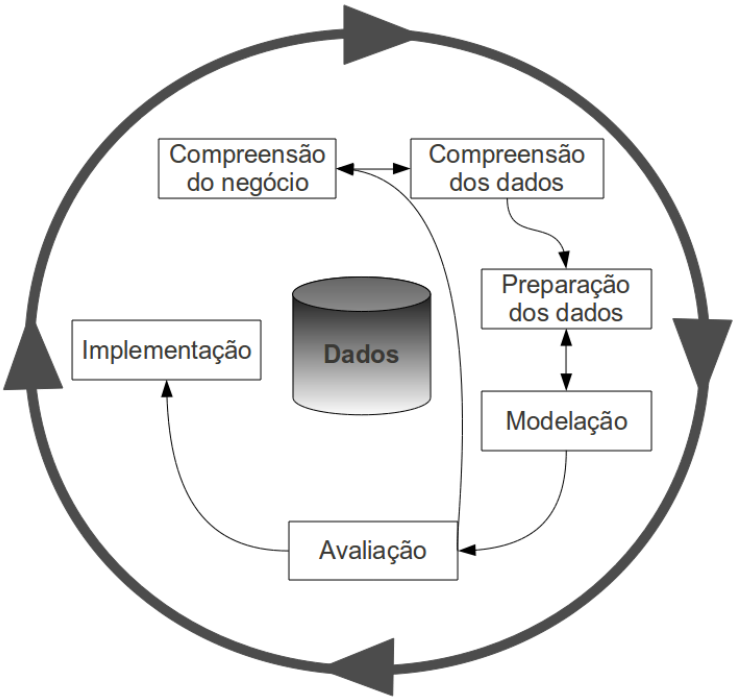

Figura 2: Ciclo de vida da metodologia CRISP-DM (Chapman et al., 2000).

CRISP-DM caracteriza-se como sendo de fácil dedução e tem como objectivo fazer com que grandes e até mesmo pequenos projectos de DM se tornem rápidos, baratos e simples de gerir. A metodologia CRISP-DM é neutra, não está associada à utilização de um *software,* sendo descrita em termos de um 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). O seu ciclo de vida desenvolve-se em seis Fases: **Compreensão do Negócio**, **Compreensão dos Dados**, **Preparação dos Dados**, **Modelação**, **Avaliação** e **Implementação** (Wirth and Hipp, 2000). As fases não têm sequência fixa, dependendo do resultado das outras fases ou das tarefas particulares de determinada fase. A Figura 2 mostra as seis fases da metodologia CRISP-DM.

De seguida, é apresentado uma breve descrição de cada uma das seis fases.

2.4.1 Compreensão do Negócio

A primeira fase da metodologia CRISP-DM é a compreensão do negócio, que se foca na percepção dos objectivos do projecto e dos requisitos do ponto de vista do negócio, convertendo depois este conhecimento na definição de um problema de DM e num plano preliminar desenhado para atingir os objectivos (Chapman et al., 2000).

2.4.2 Compreensão dos Dados

A fase de compreensão dos dados começa com uma recolha inicial de dados e continua com actividades que permitirão perceber os dados, identificar problemas de qualidade dos dados, ter uma percepção inicial das relações entre os dados ou detectar subconjuntos interessantes que permitam formar hipóteses para a informação escondida (Chapman et al., 2000).

2.4.3 Preparação dos Dados

A fase de preparação dos dados envolve todas as actividades necessárias para a construção do conjunto de dados final (dados que serão introduzidos na(s) ferramenta(s) de modelação) a partir dos dados brutos iniciais. É provável que as tarefas de preparação dos dados sejam efectuadas várias vezes e sem ser em nenhuma ordem prescrita. As tarefas incluem a selecção de tabelas, campos e registos, bem como a transformação e limpeza dos dados para as ferramentas de modelação (Chapman et al., 2000).

2.4.4 Modelação

Na fase de modelação, várias técnicas de modelação são seleccionadas e aplicadas, e os seus parâmetros são ajustados de forma a optimizar os resultados. Tipicamente, existem várias técnicas para o mesmo tipo de problema de DM. Algumas técnicas têm requisitos específicos na forma dos dados, por isso voltar à fase de preparação dos dados é muitas vezes necessário (Chapman et al., 2000). No âmbito deste trabalho, optou-se por adoptar as seguintes técnicas: **Regressão Linear/Múltipla (RM)**, **Redes Neuronais Arti-**

ficiais (RNA) e **Máquinas de Vectores de Suporte (MVS)**, para os modelos de regressão. Para a proposta de num novo sistema de classificação de carcaças, utilizou-se a técnica de *clustering* de **Mistura de Modelos Gaussianos (MMG)**, bem como **Árvores de Decisão (AD)** para explicar o conteúdo de cada grupo obtido pelo método MMG.

2.4.4.1 Regressão Múltipla A **Regressão Linear/Múltipla (RM)** é uma técnica estatística usada para analisar as relações entre uma única variável dependente e diversas variáveis independentes. A análise de RM pode ainda servir para verificar quais as variáveis independentes que mais influenciam a variável dependente. No entanto, as variáveis envolvidas devem ser numéricas (Hastie et al., 2008).

As relações entre duas variáveis: X considerada independente, e Y considerada dependente, podem ser representadas num diagrama de dispersão, com os valores de Y_i em ordenada e os de X_i em abcissa. Cada par de valores X_i e Y_i fornecerá um ponto e utilizando-se, por exemplo, o método dos desvios mínimos ao quadrado, pode-se calcular a equação de uma recta.

Para verificar se o peso específico pode ser previsto em função das outras variáveis e qual a sua ordem de importância nessa previsão, pode-se optar pela análise de regressão múltipla. Na regressão simples, utilizamos um modelo que relaciona a variável dependente (Y) com apenas um factor (X) através da equação: $Y = a + bx$. Já o modelo de Regressão Múltipla (RM) é definido pela equação (Hastie et al., 2008):

$$Y_i = \beta_0 + \sum_{i=1}^{n} \beta_i X_i, i = 1, 2, ..., n \quad (1)$$

onde: Y_i é o valor de saída no *iésimo* caso, X_i é o valor da variável independente no *iésimo* caso e β_i são os coeficientes de regressão.

Enquanto uma regressão simples de duas variáveis resulta na equação de uma recta, um problema de três variáveis implica um plano, e um problema de k variáveis implica um hiperplano.

2.4.4.2 Redes Neuronais Artificiais Segundo Dayhoff and DeLeo (2001), as **Redes Neuronais Artificiais (RNA)** são metodologias computacionais que realizam análises multi-factoriais. Inspirados por redes biológicas de neurónios, os modelos de redes neuronais artificiais contêm camadas de nodos computacionais simples que operam como dispositivos de soma não-lineares. Estes nodos estão ricamente interligados por linhas de conexão "pesadas", e os pesos são ajustados quando dados são apresentados à rede durante um processo de treino. De um treino bem sucedido podem resultar redes neuronais artificiais que realizam tarefas tais como prever um valor de saída, classificar um objecto, aproximar uma função, reconhecer um padrão em dados multi-factoriais, e completar um padrão conhecido. As redes neuronais artificiais são já utilizadas em muitos campos. São consideradas metodologias computacionais viáveis, multi-propósito e robustas, com suporte teórico sólido e com grande potencial para serem eficazes em qualquer disciplina.

Em particular, o *Multilayer Perceptron* é a arquitectura RNA mais popular, e pode ser definida como uma rede *feedforward* onde neurónios com poder de processamento são agrupados em camadas e interligados por conexões pesadas (Haykin, 1999).

2.4.4.3 Máquinas de Vectores de Suporte As **Máquinas de Vectores de Suporte (MVS)** são mais recentes do que as RNAs, tendo sido propostas por Cortes and Vapnik em 1995. De acordo com a teoria das MVS (Cortes and Vapnik, 1995), enquanto que as técnicas tradicionais para reconhecimento de padrões se baseiam na minimização do *risco empírico* - isto é, na tentativa de optimizar o desempenho do conjunto de treino -, as MVSs minimizam o *risco estrutural* - isto é, a probabilidade de classificar mal padrões ainda por descobrir segundo uma probabilidade de distribuição dos dados fixa mas desconhecida, o que confere às MVSs uma vantagem teórica em relação às RNAs (Pontil and Verri, 1998).

De forma semelhante às RNAs, o processo de treino das MVSs consiste na obtenção de valores para os pesos, de modo a minimizar uma função de custo. Quando aplicadas a um problema de Classificação com duas classes, as MVSs procuram uma superfície de decisão determinada por certos pontos

do conjunto de treino, denominados de vectores de suporte (Pontil and Verri, 1998).

Hearst et al. (1998) consideram que o método MVS se encontra entre a teoria e a prática da aprendizagem, na medida em que constrói modelos suficientemente complexos para responder aos desafios do mundo real, sendo, no entanto, simples o suficiente para serem analisados matematicamente.

Apesar de podermos pensar no método MVS como um algoritmo linear num espaço com n-dimensões, na prática, não envolve qualquer computação nesse espaço com n-dimensões. Ao recorrer a *kernels*, toda a computação necessária é realizada no espaço de entrada (Hearst et al., 1998). Neste trabalho, o *kernel* adoptado será o popular *kernel* gaussiano, que apresenta menos parâmetros do que outros *kernels* (e.g. polinominal).

2.4.4.4 Mistura de Modelos Gaussianos (MMG) A **Mistura de Modelos Gaussianos (MMG)** é um modelo probabilístico para estimação de densidade usando uma distribuição de *mistura*, e é considerado um tipo de *Clustering*. Uma *mistura* é um conjunto de k distribuições de probabilidade, representando k *clusters*, que governam os valores dos atributos para os membros daquele *cluster*. O desafio do *clustering* é pegar num conjunto de instâncias e num número pré-especificado de *clusters*, e encontrar a média e a variância de cada *cluster* e a distribuição da população entre eles (Witten and Frank, 2005).

No entanto, como no modelo MMG não se conhece a distribuição de onde cada instância de treino veio nem os parâmetros da mistura de modelos, é utilizado o algoritmo EM, de *expectativa-maximização*. O primeiro passo é a "expectativa", que consiste em calcular as probabilidades de cada *cluster* (quais são os valores "esperados" para cada classe); e o segundo passo, o cálculo dos parâmetros de distribuição, é a "maximização" da probabilidade das distribuições tendo em conta os dados (Witten and Frank, 2005).

2.4.4.5 Árvore de Decisão (AD) A **Árvore de Decisão (AD)** é uma ferramenta de apoio à decisão que utiliza um gráfico em forma de árvore para classificar dados num número finito de classes, com base no valor dos dados

de entrada. As AD são essencialmente compostas por uma hierarquia de proposições "se-então" pelo que são significativamente mais rápidas do que as RNAs. As AD são mais adequadas para dados categóricos ou divididos em intervalos porque incorporar dados contínuos numa AD pode ser difícil (Turban et al., 2007).

2.4.5 Avaliação

Nesta fase do projecto foi já construído um modelo (ou modelos) que aparenta ter uma qualidade elevada do ponto de vista da análise dos dados. Antes de proceder à implementação final do modelo, é importante avaliar cuidadosamente o modelo e rever os passos executados para a construção do mesmo, de forma a ter a certeza que se atingiu devidamente os objectivos do negócio. Um objectivo chave é determinar se não há alguma questão importante do negócio que não foi devidamente considerada. No final desta fase, deverá ser decidido o que fazer com os resultados do DM (Chapman et al., 2000). Independentemente do tipo de aprendizagem seleccionada para a determinação e avaliação dos resultados, utilizam-se técnicas de amostragem que nos permitem aferir a precisão dos modelos gerados. As técnicas de amostragem tipicamente separam os dados em dois conjuntos: conjunto de treino e conjunto de teste. O modelo de eleição deverá ser o que melhor generalize os dados treinados e o que melhor se identifique na aprendizagem de novos casos, os quais fazem parte do conjunto de teste.

2.4.5.1 Matriz de Confusão A matriz de confusão de um classificador indica o número de classificações correctas em comparação com o número de previsões efectuadas sobre um conjunto de exemplos (Witten and Frank, 2005). Esta matriz é uma das técnicas de avaliação mais utilizadas em problemas de classificação. No caso binário, cada exemplo é previsto como sendo Positivo ou Negativo (Santos and Azevedo, 2005). Daí que a tabela 2x2 possui quatro valores possíveis (Tabela 7).

Verdadeiros Positivos designados por TP, correspondem ao número de exemplos positivos correctamente classificados;

Tabela 1: Matriz de confusão de duas classes (Santos and Ramos, 2006).

Classe	Previsão C+	Previsão C-
Real C+	Verdadeiros Positivos (TP)	Falsos Negativos (FN)
Real C-	Falsos Positivos (FP)	Verdadeiros Negativos (TN)

Verdadeiros Negativos designados por TN, correspondem ao número de exemplos negativos efectivamente classificados como negativos;

Falsos Positivos corresponde ao número de exemplos positivos classificados como negativos (i.e. mal classificados), representados por FP;

Falsos Negativos número de exemplos negativos classificados como positivos (i.e. mal classificados), representados por FN.

Desta matriz podem ser derivadas muitas outras medidas, tais como (Santos and Azevedo, 2005): taxa de erro da classe C+, Taxa de erro da classe C-, Taxa de erro total, confiança positiva e negativa, nível de suporte, sensibilidade, especificidade e acuidade da previsão.

2.4.5.2 Métricas de Regressão Um conjunto de dados de regressão D é constituído por $k \in \{1, ..., N\}$ exemplos. Cada exemplo mapeia um vector de entrada $(x_1^k, ..., x_I^k)$ para uma dada saída y^k. O erro para um dado k é: $e_k = y_k - \hat{y}_k$, onde $\hat{y}_k$representa o valor previsto para o padrão de entrada k.

O desempenho dos modelos de regressão gerados no decorrer deste trabalho será avaliado recorrendo ao Erro Absoluto Relativo (RAE - *Relative Absolute Error*) (Witten and Frank, 2005):

$$RAE = 1/N \times \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{|y_i - \overline{y_i}|} \tag{2}$$

onde $\overline{y}$ é a média da variável de saída. O RAE é independente da escala dos valores da variável de saída, e valores próximos de 100% correspondem a um modelo que tem um desempenho similar ao do previsor médio *naïve* (i.e. $\hat{y}_i = \overline{y_i}$). Quanto menor o RAE, melhor é o modelo de regressão, pelo que o modelo de regressão ideal apresenta um valor próximo de 0%.

Para estimar a capacidade de generalização dos modelos de regressão, iremos utilizar um procedimento *10-fold cross-validation*. Com este procedimento é testado um subconjunto de cada vez e os restantes dados são usados para ajustar o modelo. O processo é repetido sequencialmente até todos os subconjuntos terem sido testados. Desta forma, todos os dados são utilizados para treino e teste.

Para avaliar e comparar os modelos de regressão será utilizada a técnica *Regression Error Characteristic* (REC) (Bi and Bennett, 2003). As curvas REC mostram a taxa de acerto global (eixo das ordenadas) para diversos valores de tolerância (T) de erro absoluto (eixo das abcissas). O uso das curvas REC facilita muito a avaliação da capacidade de previsão de um modelo, permitindo que até quem não tenha grandes conhecimentos de DM o faça. A precisão, ou taxa de acertos, é definida como a percentagem de pontos que se encaixam dentro da tolerância. Se a tolerância fosse zero, apenas os pontos de ajuste seriam considerados. Se escolhermos uma tolerância que exceda o erro máximo observado para o modelo, então todos os pontos considerados serão correctos. Assim, existe um *trade-off* claro entre a tolerância de erro e a precisão da função de regressão. O conceito de tolerância de erro é atraente porque muitas das vezes os dados de regressão são imprecisos devido por exemplo a erros de medição.

2.4.6 Implementação

A criação do modelo geralmente não é o final do projecto. Mesmo que o objectivo do modelo seja aumentar o conhecimento relativo aos dados, o conhecimento obtido tem que ser organizado e apresentado de forma a que o cliente o possa utilizar. Dependendo dos requisitos, a fase de implementação pode ser tão simples como gerar um relatório ou tão complexo como implementar um processo de DM repetitivo ao longo de toda a empresa (Chapman et al., 2000).

2.5 Sumário

Os sistemas de BI trazem vantagens para as organizações que operam num mercado competitivo, pois permitem combinar a recolha de dados com ferramentas de análise, com o objectivo de disponibilizar informação para a tomada de decisão. Compreende vários sistemas tais como: *Data Warehouse*, sistemas OLAP ou técnicas de DM para a extracção de conhecimento. Através das ferramentas de DM podem-se encontrar padrões nos dados, inferindo regras a partir destes. A descoberta de padrões divide-se em previsão e descrição. A previsão pode ser conseguida com métodos de regressão ou com métodos de classificação. Serão estes dois últimos que farão parte das experiências a realizar nesta dissertação.

A etapa de DM torna-se numa tarefa complexa, daí que tenham sido propostas diversas metodologias de DM, das quais se destaca o CRISP-DM. A metodologia CRISP-DM é extremamente completa e documentada uma vez que as suas fases estão devidamente organizadas, estruturadas e definidas, permitindo que o projecto possa ser facilmente compreendido ou revisto. É descrita em termos de um processo hierárquico, com um ciclo de vida que se desenvolve em seis fases: **Compreensão do Negócio**, **Compreensão dos Dados**, **Preparação dos Dados**, **Modelação**, **Avaliação** e **Implementação**. Na fase de Modelação foram descritas algumas técnicas a utilizar neste trabalho: RM, RNA, MVS, MMG e AD.

Existem diversas métricas para a avaliação dos modelos criados por técnicas de DM. Destacam-se as curvas REC e o erro RAE para avaliar a qualidade da regressão, pois serão utilizadas nas experiências a realizar nesta dissertação.

3 Avaliação da Qualidade da Carne de Cordeiro

3.1 Introdução

Em Portugal existem quatro produtos de origem ovina e caprina com Denominação de Origem Protegida e treze com Indicação Geográfica Protegida. A criação deste tipo de produtos pretendeu representar um incentivo aos produtos ditos de qualidade e, desta forma, contribuir para a preservação do mundo rural e da sua diversidade, através da valorização dos sistemas de produção e dos produtos típicos ou específicos das diversas regiões de Portugal e da União Europeia (Cadavez et al., 2004). No entanto, a qualidade de carcaças depende de muitos aspectos, entre os quais se encontra a sua composição tecidual. Tendo isso em consideração, vários países do mundo criaram sistemas de classificação que visam agrupar as carcaças de acordo com as suas características, de modo a formar lotes uniformes associados à procura do mercado e ao seu valor comercial (Price and Jones, 1995).

3.2 Cordeiro Bragançano, DOP

O Cordeiro Bragançano é um produto certificado com Denominação de Origem Protegida, referente a cordeiros de 3 a 4 meses de idade, filhos de animais inscritos no Livro Genealógico da Raça Churra Galega Bragançana, criados na região e alimentados exclusivamente com leite inteiro materno e pastagem natural.

O Cordeiro Bragançano obteve Designação de Origem Protegida no regulamento (CE) nº 1107/96, de 12 de Junho de 1996, sendo o *Agrupamento de Produtores de Cordeiros Bragançanos, Lda* a Entidade Gestora da DOP, e sendo a *Associação Nacional de Criadores de Ovinos da Raça Churra Galega Bragançana (ACOB)* o Organismo Privado de Controlo e Certificação (OC) responsável pelo controlo desta raça. No final de 2009, a ACOB contava com 9.700 fêmeas exploradas em linha pura, inscritas no livro de adultos (DGV, 2010).

3.2.1 Denominação de Origem e Indicação Geográfica

Consciente da importância dos produtos agro-alimentares típicos e tradicionais para o desenvolvimento das regiões europeias mais desfavorecidas, a União Europeia publicou em 1992, e novamente em 2006, legislação comunitária relativa à protecção do nome de produtos que, pela sua origem geográfica e/ou modos particulares de produção, possuem características particulares. De igual modo, também em Portugal foi publicada em 1997 legislação nacional a regulamentar estes produtos.

Como o resumo efectuado nesta subsubsecção se baseia nestes regulamentos, e de forma a facilitar uma consulta mais aprofundada, apresenta-se de seguida um enquadramento regulamentar comunitário e nacional:

Legislação comunitária:

- Reg. (CE) nº 510/2006 do Conselho - Relativo à protecção das indicações geográficas e denominações de origem dos produtos agrícolas e dos géneros alimentícios;
- Reg. (CE) nº 509/2006 do Conselho - Relativo às especialidades tradicionais garantidas dos produtos agrícolas e dos géneros alimentícios;
- Reg. (CE) nº 1898/2006 da Comissão - Estabelece as regras de execução do Reg. (CE) nº 510/2006;
- Reg. (CE) nº 1216/2007 da Comissão - Estabelece as regras de execução do Reg. (CE) nº 509/2006;
- Reg. (CE) nº 417/2008 da Comissão - Altera os anexos I e II do Reg. (CE) nº 510/2006;
- Reg. (CE) Nº 628/2008 da Comissão - Altera os pontos 1, 2 e 3 do Anexo V do Reg. (CE) nº 898/2006.

Legislação nacional:

- DN nº 47/97 de 11 de Agosto;
- DN nº 12/99 de 8 de Março;

- DN nº 32/2000 de 31 de Julho.

Denominação de Origem (DO), ou **Denominação de Origem Protegida (DOP)**, é o nome - reconhecido a nível nacional (DO) ou reconhecido a nível comunitário (DOP) - de uma região, de um local determinado ou, em casos excepcionais, de um país, que serve para designar um produto agrícola ou um género alimentício originário dessa região, desse local determinado ou desse país e cuja qualidade ou características se devem essencialmente ou exclusivamente ao meio geográfico, incluindo os factores naturais e humanos, e cuja produção, transformação e elaboração ocorrem na área geográfica delimitada. O nome pode, em casos excepcionais, não ser geográfico. No entanto, as DOP não abrangem produtos do sector vitivinícola, com excepção dos vinagres de vinho, nem as bebidas espirituosas.

Indicação Geográfica (IG), ou **Indicação Geográfica Protegida (IGP)**, é o nome - reconhecido a nível nacional (IG) ou reconhecido a nível comunitário (IGP) - de uma região, de um local determinado ou, em casos excepcionais, de um país, que serve para designar um produto agrícola ou um género alimentício originário dessa região, desse local determinado ou desse país e cuja reputação, determinada qualidade ou outra característica podem ser atribuídas a essa origem geográfica e cuja produção e/ou transformação e/ou elaboração ocorrem na área geográfica delimitada. No entanto, as IGP não abrangem produtos do sector vitivinícola, com excepção dos vinagres de vinho, nem as bebidas espirituosas.

Todos os produtos com DOP ou IGP devem ter associados um Agrupamento-Gestor da DOP/IGP e um Organismo Privado de Controlo e Certificação (OC).

O Agrupamento-Gestor é um Agrupamento de Produtores (AP) que tem como dever zelar pelo nome da DOP/IGP (cuja gestão lhe está legalmente confiada), indigitar o OC para realizar as acções sistemáticas de controlo e certificação, velar pelo cumprimento das normas constantes do caderno de especificações, autorizar o uso da DOP ou da IGP aos produtores e/ou transformadores que o solicitem, promover comercialmente o produto, e aplicar sanções aos produtores e/ou transformadores que cometam infracções.

O Organismo Privado de Controlo e Certificação (OC) deverá ser re-

conhecido pelo Ministério da Agricultura, do Desenvolvimento Rural e das Pescas (MADRP)[1] como cumprindo a Norma EN 45 011 (e a partir de 1/05/2010 formalmente acreditada) e deve dispor de meios técnicos e materiais, procedimentos escritos e planos de controlo aprovados, para proceder ao controlo das fileiras produtivas e à eventual certificação de produtos que podem usar as menções e símbolos relativos às DO e às IG.

Em termo de protecção jurídica, é conferido direito exclusivo de uso da DOP ou da IGP (em todo o território da União Europeia e em países terceiros[2]) para os produtores da área geográfica, que requeiram ao agrupamento de produtores gestor da DOP ou IGP, cumpram as regras constantes do caderno de especificações e se submetam ao controlo pelo OC reconhecido. São proibidas todas e quaisquer práticas que, sem direito, utilizem ou façam apelo à indicação registada, qualquer que seja o objectivo e, em particular, para poderem beneficiar do seu prestígio ou da sua reputação.

3.2.2 Churra Galega Bragançana

Em Portugal estão reconhecidas 15 raças autóctones de ovinos, classificadas consoante o seu tipo de lã em Merino (lã fina), Bordaleiro (lã média) ou Churro (lã grosseira). O grupo Churro inclui as raças Algarvia, Badana, Galega Bragançana, Galega Mirandesa, Mondegueira, Churra da Terra Quente e, de reconhecimento recente, a Churra do Minho e a Churra do Campo. A maioria é criada no Norte do País, uma no Centro (Churra do Campo) e uma no Sul (Algarvia). O número de fêmeas registadas varia entre 100 (Churra do Campo) e 27000 (Churra da Terra Quente) mas, à excepção da Churra da Terra Quente e da Galega Bragançana, todas têm efectivos em declínio e estão em risco de extinção (Santos-Silva et al., 2009).

O ovino Bragançano, como todos os outros englobados, tradicionalmente, nas raças churras autóctones, tem relações filogénicas com o *Ovis aries Studery* (Sobral, 1987), e está delimitado à área geográfica dos concelhos de Bra-

[1] Actualmente a entidade que reconhece os OC neste âmbito é o Gabinete de Planeamento e Políticas (GPP).

[2] Apenas naqueles países que solicitem protecção à CE para as suas indicações geográficas e denominações de origem.

gança e Vinhais, e parte dos concelhos de Macedo de Cavaleiros, Vimioso, Mirandela, Chaves e Valpaços. Segundo Sobral (1987), *"as particularidades da área que constitui o seu berço determinaram a formação dum tipo de animal bem diferenciado, com características genéticas que se transmite de geração para geração, bem ajustadas às condições ambientais dessa parcela da Terra Fria, influenciada pelas serras de Montezinho e Nogueira"*.

3.3 Qualidade da Carcaça

Kempster (1983) considerou a qualidade das carcaças produzidas como uma medida de produção primária e um critério chave no melhoramento genético das raças. O termo qualidade é, em si mesmo, subjectivo podendo ser interpretado de várias formas, mas a qualidade de uma carcaça deve, sem dúvida, ser associada à sua composição tecidual uma vez que esta é determinante na sua valorização comercial, por dois aspectos igualmente importantes: 1) rendimento em carne magra, e 2) características organolépticas[3] da carne a que dá origem. Assim, a qualidade das carcaças é condicionada por dois factores principais, que Teixeira et al. (1998) classificaram em: 1) factores intrínsecos ao animal: raça, idade e sexo; e 2) factores extrínsecos ao animal: sistema de produção, dieta e nível alimentar.

De uma forma geral, as carcaças devem possuir uma reduzida quantidade de gordura, mas esta deve ser suficiente para lhes garantir uma boa apresentação, conservação e protecção durante a refrigeração (Teixeira et al., 1992; Delfa and Teixeira, 1998). A gordura subcutânea e intermuscular desempenham um papel importante no isolamento das carcaças durante a refrigeração, protegendo-as do fenómeno vulgarmente conhecido por *cold-shortening* (encurtamento pelo frio). Por outro lado, algumas das características de qualidade da carne, como a ternura e a suculência, estão positivamente correlacionadas com o teor em gordura, pelo que desempenha um importante papel nas características organolépticas da carne (Wood, 1990, 1995). A presença a nível óptimo de gordura na carcaça é, pois, essencial para maximizar

[3]Chamam-se propriedades organolépticas às características dos objetos que podem ser percebidas pelos sentidos humanos, como a cor, o brilho, o sabor, o odor e a textura.

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 também acarreta custos (Cadavez et al., 2004).

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 e, sempre que isto acontece, a carcaça deve possuir um valor máximo. Sempre que a composição das carcaças se afaste do ideal, anteriormente estabelecido, o seu preço deve sofrer penalizações, pelo que os sistemas de classificação de carcaças desempenham um papel importante na definição de regras para as transacções comerciais (Cadavez et al., 2004).

3.4 Sistemas de Classificação de Carcaças de Ovino

A composição corporal dos animais, de todas as espécies, varia consideravelmente como resultado do estádio de crescimento, do plano nutricional e da base genética (Cadavez et al., 2004). A percentagem de músculo no corpo dos animais varia de 35% a próximo de 50% do peso corporal (Topel and Kauffman, 1998). Desta forma, é fácil admitir que o valor económico dos animais dependa da composição da sua carcaça (Topel and Kauffman, 1998; Forrest, 1995; Delfa and Teixeira, 1998), já que os diferentes tecidos que compõem a carcaça possuem valor económico diferente, dependendo da utilização que lhes é dada, bem como das exigências do consumidor (Forrest, 1995). Todavia, a composição das carcaças é condicionada pelo peso (Hall et al., 2001), pela raça (Kirton et al., 1995; Fogarty et al., 2000), pelo sexo (Hall et al., 2001), pelo ritmo de crescimento (Snowder et al., 1994) e pelas variações no ritmo de crescimento (Murphy et al., 1994; Hall et al., 2001).

Assim, desde longa data que são investigados métodos de estimativa da composição, com o objectivo de desenvolver sistemas de classificação de carcaças. A classificação visa agrupar as carcaças de acordo com as suas características, de modo a formar lotes uniformes associados à procura do mercado e ao seu valor comercial (Price and Jones, 1995). O estabelecimento destes lotes uniformes permitirá também direccionar os diferentes tipos de carcaças

para mercados com procura específica (Cadavez et al., 2004).

Os sistemas de classificação de carcaças desempenham essencialmente duas funções: 1) fornecer informações sobre as características relevantes para o mercado (Fisher, 1987; Price and Jones, 1995), regulando e facilitando a comercialização (Price and Jones, 1995; Kirton, 1998), através de uma linguagem comum entre produtores e comerciantes (Kirton, 1998) e 2) estabelecer uma base formal para o pagamento (Fisher, 1987; Kirton, 1998), que pode também funcionar como um incentivo à produção de carcaças com as características procuradas pelos consumidores (Fisher, 1987).

Na maioria dos países onde as carcaças de ovino são classificadas, o sistema de classificação é similar ao utilizado nos bovinos, representando apenas imitações fundamentadas que podem não ter aplicação nesta espécie pecuária (Cadavez et al., 2004). Por outro lado, as carcaças dos ovinos são, em geral, comercializadas inteiras, pelo que a utilização dos sistemas desenvolvidos para os bovinos tem pouca utilidade (Price and Jones, 1995).

São abordados de seguida alguns sistemas de classificação relevantes, que são actualmente utilizados em diversas regiões do globo, nomeadamente na União Europeia, na Austrália e Nova Zelândia, e nos Estados Unidos da América.

3.4.1 Sistema de Classificação Europeu

A legislação em vigor na União Europeia, referente à classificação de carcaças de ovinos, foi definida no regulamento (CE) nº 1234/2007. Este regulamento estabelece que a classificação deve basear-se no sistema SEUROP, que consiste na avaliação da conformação[4] (Tabela 2), tendo em consideração o desenvolvimento dos perfis da carcaça, nomeadamente das suas partes essenciais (coxa, dorso, pá), e na avaliação da camada de gordura (Tabela 3), tendo em consideração a quantidade de tecido adiposo no exterior da carcaça e na cavidade torácica. Ambos os critérios são avaliados por apreciação visual, pelo que se baseia em critérios de elevada subjectividade. A legislação autoriza, no entanto, que os estados membros definam outros sistemas de

[4]De acordo com vários autores, a conformação é um termo utilizado para a descrição visual da forma de uma carcaça, manifestando a espessura relativa de carne gorda e magra.

Tabela 2: Classificação Europeia de carcaças de ovino quanto à conformação.

Classe de conformação	Descrição
S (Superior)	Todos os perfis extremamente convexos; desenvolvimento muscular excepcional com duplos músculos
E (Excelente)	Todos os perfis convexos a superconvexos; desenvolvimento muscular excepcional
U (Muito boa)	Perfis em general convexos; forte desenvolvimento muscular
R (Boa)	Perfis em geral rectilíneos; bom desenvolvimento muscular
O (Média)	Perfis rectilíneos a côncavos; desenvolvimento muscular médio
P (Fraca)	Todos os perfis côncavos a muito côncavos; reduzido desenvolvimento muscular

classificação para as carcaças ligeiras, ou seja, para carcaças com menos de 13kg de peso. Com base neste regulamento, a Espanha desenvolveu um sistema de classificação para as carcaças ligeiras (as predominantes no mercado espanhol) com o principal objectivo de retirar da classificação SEUROP as carcaças ligeiras, sempre mal classificadas neste sistema, produzidas na zona Mediterrânea (Cadavez et al., 2004).

3.4.2 Sistema de Classificação Australiano e Neozelandês

A Austrália e Nova Zelândia foram pioneiros na implementação de um sistema de classificação objectivo de carcaças de ovino, descrito e estudado em vários trabalhos de investigação (Hopkins et al., 1993; Kirton et al., 1992, 1999; Safari et al., 2001). Este sistema de classificação é utilizado para formar lotes comerciais com base no peso da carcaça, distribuídos por cinco classes, e na espessura total dos tecidos avaliada pela medida GR[5] (Kirton

[5]A medida GR corresponde à medida da espessura total dos tecidos a 11cm da linha média dorsal ao nível da 12ª costela.

Tabela 3: Classificação Europeia de carcaças de ovino quanto à camada de gordura.

Classe de estado da gordura	**Descrição**
1 (Fraco)	Gordura de cobertura inexistente a muito fraca
2 (Leve)	Leve cobertura de gordura, com músculos quase sempre aparentes
3 (Médio)	Músculos quase sempre cobertos de gordura, com excepção dos das coxas e da pá; reduzidos depósitos de gordura na cavidade torácica
4 (Forte)	Músculos cobertos de gordura, mas ainda parcialmente visíveis ao nível da coxa e da espádua; alguns depósitos pronunciados de gordura no interior da cavidade torácica
5 (Muito forte)	Carcaça coberta por uma camada de gordura; depósitos substanciais de gordura na cavidade torácica

and Johnson, 1979), também com cinco classes (sendo 1 para a carcaça menos gorda e 5 para a mais gorda).

3.4.3 Sistema de Classificação Norte-americano

Nos Estados Unidos da América, o sistema de classificação de carcaças de ovino (*US yield grading standards*) é, também, efectuado com base em critérios objectivos, utilizando a medida da espessura da gordura subcutânea efectuada no centro do músculo *longissimus* ao nível da 12ª costela (USDA, 1992b,a; Snowder et al., 1994). As carcaças são classificadas com base em duas avaliações independentes das características de palatabilidade (*quality* - qualidade) e de rendimento em carne magra (*yield* - rendimento) (USDA, 1992a).

Este sistema utiliza quatro classes de qualidade: *prime*, *choice*, *good* e *utility*; e cinco classes de rendimento: *yield1* a *yield5*, com *yield1* representando as carcaças com maior rendimento em carne magra. Este sistema já inclui critérios de qualidade organoléptica da carne, apesar de avaliada de forma indirecta através do desenvolvimento da camada de gordura subcutânea. Para serem classificadas como *prime* ou *choice* as carcaças devem apresentar-se completamente revestidas por uma pequena camada de gordura subcutânea, o que corresponderá a uma medida de espessura da gordura subcutânea, ao nível da 12ª costela, de aproximadamente 2mm (Cadavez et al., 2004).

3.5 *Business Intelligence* para a Previsão da Composição Tecidual de Carcaças

Desde os estudos pioneiros de Palsson (1939) com medições de carcaças e tecidos, muitos outros têm sido desenvolvidos no sentido de avaliar o valor de diversas medidas de dimensão e de espessura dos tecidos da carcaça para estimar a sua composição. Vários modelos de estimativa têm sido desenvolvidos utilizando como variáveis independentes o peso da carcaça, conjuntamente com medidas de espessura da gordura subcutânea (Timon and Bichard, 1965; Cadavez et al., 2002; Delfa et al., 1996; Jones et al., 1992; Wood et al., 1980; Teixeira et al., 2006; Wolf et al., 2006; Cadavez, 2009), de profundidade do

músculo *longissimus* (Timon and Bichard, 1965; Jones et al., 1992; Delfa et al., 1996; Cadavez et al., 2002; Wolf et al., 2006; Teixeira et al., 2006; Cadavez, 2009), medições de espessura total de tecidos (Kirton et al., 1984; Delfa et al., 1996) e medidas de dimensão da carcaça (Stanford et al., 1997; Timon and Bichard, 1965; Wood et al., 1980; Wolf et al., 2006; Cadavez, 2009). A maioria dos modelos foram desenvolvidos recorrendo a regressões lineares múltiplas, onde não é avaliada a colinearidade entre variáveis independentes. No entanto, devem-se esperar problemas de colinearidade entre variáveis independentes, pois estão correlacionadas tanto geneticamente como fenotipicamente (Simm and Dingwall, 1989), e sabe-se que modelos baseados em variáveis multicolineares podem limitar as inferências e a acuidade de previsões (Chatterjee and Hadi, 2006).

No que diz respeito à utilização de técnicas de DM não lineares, foram utilizadas RNAs e MVS para prever a tenrura da carne de cordeiro em (Cortez et al., 2006), embora tal estudo não tenha abordado a composição dos tecidos da carne.

4 Avaliação da Qualidade da Carne de Cordeiro via Técnicas de *Data Mining*

4.1 Introdução

Devido à limitação dos especialistas humanos é necessário recorrer a ferramentas (semi-) automatizadas de DM/BI para analisar os dados em estado bruto e extrair informações de alto nível para os decisores (Turban et al., 2007). Muitos algoritmos de DM estão disponíveis para tarefas de supervisão. **Redes Neuronais Artificiais (RNA)** e **Máquinas de Vectores de Suporte (MVS)** são modelos flexíveis e não lineares que podem lidar com complexos mapeamentos, e que estão a ser cada vez mais utilizados no campo de DM (Turban et al., 2007; Hastie et al., 2008). As MVSs apresentam vantagens teóricas sobre as RNAs devido à ausência de mínimos locais na fase de aprendizagem e foram recentemente consideradas um dos mais influentes algoritmos de DM (Wu et al., 2008). Nesta dissertação, os algoritmos de RM, RNA e MVS serão utilizados para prever a composição tecidual de carcaças de cordeiro através do uso de medições que podem ser obtidas no matadouro até 24h após o abate dos animais. Será utilizada uma análise de sensibilidade, que permitirá identificar quais as medições mais relevantes para a previsão da percentagem de cada tecido em estudo. Por último, será proposto um novo sistema de classificação da qualidade de carcaças, com base num *clustering* via método de MMQ e descrição das características de cada grupo (*cluster*) via AD.

A ferramenta utilizada será o ambiente de programação R, que é um projecto aberto (*open source*), e que funciona em múltiplas plataformas (e.g. *Windows, Linux, Mac OS*). Trata-se de uma linguagem de programação poderosa com foco na estatística e análise de dados. Embora não seja especificamente orientada para o DM e BI, esta ferramenta inclui uma elevada variedade de algoritmos de DM (e.g. RM, RNA, MVS), sendo que na actualidade é adoptada por um elevado número de analistas. Por exemplo, um inquérito de DM realizado em 2008 relatou um aumento no uso da ferramenta R, com um total de 36% de respostas (Rexer, 2008). Em particular, o

código escrito utiliza a biblioteca *rminer*, que facilita o uso do R para tarefas de classificação e regressão, definindo um pequeno conjunto de funções coerentes. De referir que esta ferramenta tem sido utilizada com êxito em aplicações de domínios distintos, como por exemplo medicina (Silva et al., 2008), engenharia civil (Marques et al., 2009) e previsão da qualidade de vinho verde (Cortez et al., 2009a). Nesta secção descreve-se a ferramenta R, bem como todo o processo de DM conduzido nesta dissertação, com as diferentes fases da metodologia CRISP-DM.

4.2 Ferramentas Utilizadas

O ambiente de programação estatístico "R" é um ambiente *open source* e multi-plataforma (e.g. Windows, Linux, Mac OS), para estatística e análise de dados. Embora não seja orientada especificamente para o *data mining*, a ferramenta R inclui uma grande variedade de algoritmos de DM e é hoje utilizada por um grande número de analista de DM. Por exemplo, o estudo sobre DM de Rexer (2008) registou um aumento no uso do R, com 36% das respostas. Da mesma forma, a votação realizada pelo KDnuggets em 2009, relativa a ferramentas de DM utilizadas em projectos reais, elegeu o R como a segunda ferramenta *open source* mais utilizada e a sexta no total (Piatetsky-Shapiro, 2009). Quando comparado com ferramentas comerciais (e.g. Enterprise Miner da SAS) ou até com ambientes *open source* (e.g. WEKA), o R tem a vantagem de ser mais flexível e extensível por concepção, por isso a integração de estatística, programação e gráficos é mais natural (Miller, 2008). Para além disso, devido à sua disponibilidade *open source* e à actividade dos seus utilizadores, novos métodos de *data mining* são geralmente codificados mais depressa no R do que em ferramentas comerciais. A comunidade R é muito activa e novos pacotes estão continuamente a ser criados, com mais de 2544 pacotes disponíveis em http://www.r-project.org/.

Em especial, a biblioteca **rminer** é uma biblioteca que facilita o uso do R para resolver tarefas DM de regressão e classificação. Esta biblioteca é particularmente adaptada a RNAs e MVSs, técnicas de aprendizagem flexíveis e não-lineares que são promissoras devido ao seu desempenho em previsões. O

rminer é completamente escrito em R e apenas requer a instalação de alguns pacotes (e.g. kernlab) (Cortez, 2010).

4.3 Compreensão do Negócio

Neste trabalho realizou-se um estudo empírico, recorrendo a bases de dados e técnicas de DM, com o objectivo de construir modelos capazes de classificar carcaças de cordeiro de acordo com a sua composição tecidual. Os modelos desenvolvidos tiveram como entradas valores físicos como o sexo do animal, peso da carcaça quente e várias medidas de profundidade de tecidos, sendo que a saída desejada correspondia às percentagens de tecidos obtidas por dissecação. Para esse efeito, foram utilizados dados recolhidos pela Escola Superior Agrária do Instituto Politécnico de Bragança, referentes a 125 amostras de carcaças de cordeiro.

Pretendia-se que os modelos obtidos pudessem classificar de modo satisfatório as carcaças de cordeiro. Também interessava que os modelos de algum modo pudessem ajudar a identificar quais as medidas relevantes para a previsão da percentagem total de cada tecido. Como objectivo último, e caso os modelos obtidos fossem confiáveis, pretendia-se no futuro incorporá-los num sistema de apoio à decisão, de modo a auxiliar o processo de classificação de carcaças de cordeiro no matadouro.

Depois de elaborada a análise ao negócio e definição dos objectivos, torna-se indispensável converter esse conhecimento numa tarefa de DM capaz de traduzir os objectivos propostos. Uma vez que a determinação das proporções de tecidos corresponde a valores numéricos, neste trabalho optou-se por utilizar uma abordagem de **Regressão**, onde se tentará modelar uma função desconhecida (mapeamento) entre um conjunto de variáveis independentes (as medições à carcaça) e as variáveis dependentes (saídas, percentagem de cada tecido). No entanto, como o sistema de classificação de carcaças de ovinos em vigor na União Europeia classifica as carcaças com base em critérios subjectivos, também se irá utilizar uma abordagem de **Classificação**, onde se tentará propor um novo sistema de classificação de carcaça de cordeiro, a ser desenvolvido com base em técnicas de DM.

4.4 Compreensão e Preparação dos Dados

Este estudo, tal como já foi anteriormente mencionado, analisará carcaças de cordeiro da raça Churra Galega Bragançana, uma raça exclusiva da região Trás-os-montes de Portugal.

Foram utilizados cento e vinte e cinco cordeiros Bragançanos (42 fêmeas e 83 machos), escolhidos aleatoriamente do rebanho da Escola Superior Agrária de Bragança. Após 24h de jejum, os cordeiros foram abatidos no matadouro experimental da Escola Superior Agrária de Bragança, e as carcaças foram pesadas aproximadamente 30 minutos após o abate de forma a se obter o Peso da Carcaça Quente (PCQ). As carcaças foram então cortadas a meio pelo centro da coluna vertebral, e a Gordura Pélvica e Renal (GPR) foi removida e pesada. Durante o quarteamento, foram efectuadas medições de tecidos com um paquímetro, relativamente à profundidade do músculo *longissimus* (mm) e à espessura da gordura subcutânea (mm) entre a 12ª e 13ª costelas (B12 e C12, respectivamente), e a 1ª e 2ª vértebra lombar (B1 e C1, respectivamente). Os principais atributos do conjunto de dados são apresentados na Tabela 4.

Cada carcaça foi então dissecada em músculo, gordura subcutânea, gordura intermuscular, osso e restos (principais vasos sanguíneos, ligamentos, tendões, e tecidos conectores espessos associados aos músculos), e as proporções de Músculo (PM), Osso (PO), Gordura Subcutânea (PGS), Gordura Intermuscular (PGI) e Gordura Pélvica e Renal (PGPR) das carcaças de cordeiro foram computorizadas.

4.5 Modelação

Neste trabalho será utilizada uma abordagem de regressão e uma abordagem de classificação. No que diz respeito aos modelos de regressão, foram testados três algoritmos distintos: **Regressão Múltipla (RM)**, **Redes Neuronais Artificiais (RNA)** e **Máquinas de Vectores de Suporte (MVS)**. A configuração utilizada para estes algoritmos é descrita de seguida.

Como algoritmo de RNA, este estudo considerou uma arquitectura *Multilayer Perceptron* com uma camada escondida de H nodos escondidos, com

Tabela 4: Principais atributos do conjunto de dados.

Atributos	**Descrição**	**Domínio**
sexo	Sexo do cordeiro	$\{1, 2\}^a$
PCQ	Peso da Carcaça Quente (kg)	[5.3, 23.3]
C1	Gordura subcutânea na 1ª vértebra lombar (mm)	[0.4, 5.9]
C12	Gordura subcutânea na 12ª costela (mm)	[0.5, 7.1]
B1	Espessura do músculo *longissimus* na 1ª vértebra lombar (mm)	[14.9, 37.7]
B12	Espessura do músculo *longissimus* na 12ª costela (mm)	[13.6, 33.6]
PM	Proporção de músculo (fracção de massa)	[0.47, 0.68]
PO	Proporção de osso (fracção de massa)	[0.14, 0.26]
PGS	Proporção de gordura subcutânea (fracção de massa)	[0.02, 0.16]
PGI	Proporção de gordura intermuscular (fracção de massa)	[0.06, 0.16]
PGPR	Proporção de gordura pélvica e renal (fracção de massa)	[0.01, 0.11]

[a] 1- Macho, 2 - Fêmea

funções de activação logística e um nodo de saída com uma função linear (Hastie et al., 2008). Tendo em conta que a função de custo das RNAs é não-convexa (com múltiplos mínimos), $NR = 3$ treinos foi aplicado a cada configuração neural, e a RNA com menor erro de ajuste foi seleccionada. Sob esta configuração, o desempenho da RNA depende do valor de H. Se $H = 0$, então o modelo é equivalente à RM. Quando se aumenta H, um mapeamento mais complexo é efectuado, no entanto um valor excessivo de H irá sobreajustar os dados, levando à perda de generalização.

Na regressão por MVS, este estudo adopta o popular *kernel* gaussiano, que apresenta menos parâmetros do que outros *kernels* (e.g. polinominal): $K(x, x') = exp(-\gamma \left\| x - x' \right\|^2)$, $\gamma > 0$. Também será adoptado a frequentemente utilizada função de perda ϵ-insensível, que coloca um tubo insensível em redor dos resíduos e cujos pequenos erros dentro do tubo são descartados. Sob esta configuração, o desempenho da MVS é afectado por três parâmetros: γ, ϵ e C (um *trade-off* entre ajustar os erros e o nivelamento do mapeamento). Para reduzir o espaço de procura, os dois primeiros valores serão definidos usando a heurística (Cherkassky and Ma, 2004): $C = 3$ (para uma saída padronizada) e $\epsilon = \hat{\sigma}/\sqrt{N}$, onde $\hat{\sigma} = 1.5/N \times \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ e $\hat{y}$ é o valor previsto por um algoritmo *3-nearest neighbor*. O parâmetro do *kernel* (γ) produz o maior impacto no desempenho do MVS, com valores muito elevados ou muito baixos a levar a más previsões.

Para ajustar os hiperparâmetros dos RNA e MVS (e.g. H e γ) uma pesquisa em grelha (com $H \in 1, 2, ..., 8$ e $\gamma = 2^{-13}, 2^{-11}, ..., 2^{1}$, num total de 8 pesquisas por modelo) foi utilizada. Para seleccionar o melhor hiperparâmetro foi utilizado um *3-fold* interno (usando apenas dados de treino). Posteriormente, o melhor modelo foi de novo treinado com todos os dados de treino (tal como definido pelo esquema de validação *10-fold* externo).

A importância relativa dos previsores (ou entradas) para um dado modelo DM pode ser estimada utilizando procedimentos de análise de sensibilidade (Cortez et al., 2009b). Este procedimento mede como as respostas são afectadas quando todos os valores de entrada são mantidos nos seus valores médios excepto x_a, que varia por todo o seu domínio. O atributo x_a é considerado mais relevante se produzir uma variância mais elevada nas respostas.

Uma análise mais detalhada da influência dos valores de entrada será obtida com a curva *Variable Effect Characteristic* (VEC), que apresenta num gráfico os valores de x_a (eixo das abcissas) em comparação com as respostas (eixo das ordenadas).

No que diz respeito aos modelos de classificação, foi testado um algoritmo de *Clustering* de Mistura de Modelos Gaussianos (MMG), tal como definido na biblioteca **mclust** (Fraley and Raftery, 2006). Para a classificação, apenas foram utilizadas quatro variáveis de entrada, sendo elas a espessura da gordura subcutânea (mm) entre a 12ª e 13ª costelas (C12), a proporção de músculo (PM), a proporção de osso (PO), e a proporção total de gordura (PTG), consistindo esta última na soma das proporções de gordura subcutânea (PGS), gordura intermuscular (PGI) e gordura pélvica e renal (PGPR). Tendo em conta que a variável de entrada C12 se encontra em milímetros e as restantes são fracções de massa, antes de aplicar o algoritmo de *Clustering* todos os dados de entrada foram convertidos à mesma escala de valores através da função **scale** (Becker et al., 1988), de modo a evitar deturpações devido a diferenças de escala. Dessa forma, foi aplicado o algoritmo de *Clustering* ao conjunto de dados escalados, tendo o número de *clusters* sido definido como três. Após realizado o *Clustering*, os *clusters* obtidos foram modelados via AD, com vista à obtenção de regras para descrever os mesmos. Com esta AD, conseguiu-se obter uma matriz de confusão, de modo a classificar as carcaças com base nos valores previstos pelos modelos de regressão.

4.6 Avaliação

Para avaliar os resultados obtidos, foi utilizada a métrica RAE. Mais adiante, e de forma a comparar os modelos de regressão, será também utilizada uma análise via curvas REC (Bi and Bennett, 2003).

Para estimar a capacidade de generalização dos modelos de regressão, foi utilizado um procedimento *10-fold cross-validation*. Visto que os resultados podem depender da divisão aleatória utilizada para definir os 10 subconjuntos, também serão aplicadas 20 execuções a cada procedimento *10-fold*, num

Tabela 5: Valores de RAE (em %) para a previsão da composição de carcaças de cordeiro (resultados do conjunto de dados)

	RM	**RNA**	**MVS**
PM	**59.4**±0.3	63.0±1.6	60.1±0.7
PO	48.4±0.3	47.0±0.4	**46.1**±0.3
PGS	43.1±0.3	42.5±0.6	**42.2**±0.4
PGI	**64.1**±0.3	64.5±0.8	65.8±0.8
PGPR	53.1±0.5	57.7±2.0	**51.5**±0.5

total de $20 \times 10 = 200$ experiências para cada configuração de teste. A confiança estatística será obtida pelo teste *t-student* com o nível de confiança a 95%.

Na Tabela 5, os melhores valores estão a **negrito**, enquanto que os sublinhados denotam significância estatística ($P<0.05$) sob uma comparação por pares contra outros métodos.

O modelo de RM apresenta o RAE mais baixo para PM (RAE=59.4%, $P<0.05$) e para PGI (RAE=64.1%, $P<0.05$). O modelo MVS apresenta o RAE mais baixo para PO (RAE=46.1%, $P<0.05$), para PGPR (RAE=51.5%, $P<0.05$), e para PGS (RAE=42.2%, $P<0.05$). Através de uma análise por curvas REC, é fácil constatar o desempenho superior do modelo MVS para PGPR, tal como se pode constatar pela Figura 3.

No entanto, é importante notar que para as previsões PGS e PGI as diferenças entre modelos de regressão não são estatisticamente significativas. Os resultados de RAE mostram uma melhoria de aproximadamente 36% (PGI) a 58% (PGS) quando comparados com o previsor médio naïve. A modelação por MVS fornece as melhores previsões para PO, PGS e PGPR, enquanto a RM obtém o RAE mais baixo nas previsões de PM e PGI. O modelo RNA apenas tem melhor desempenho que o modelo RM na previsão de PO e PGS.

De forma a demonstrar a qualidade dos resultados atingidos, a Figura 4 apresenta os gráficos de dispersão para os melhores modelos de regressão, onde se pode comparar os valores observados com os valores previstos. Nos gráficos, a maioria dos pontos estão próximos da linha diagonal, que denota

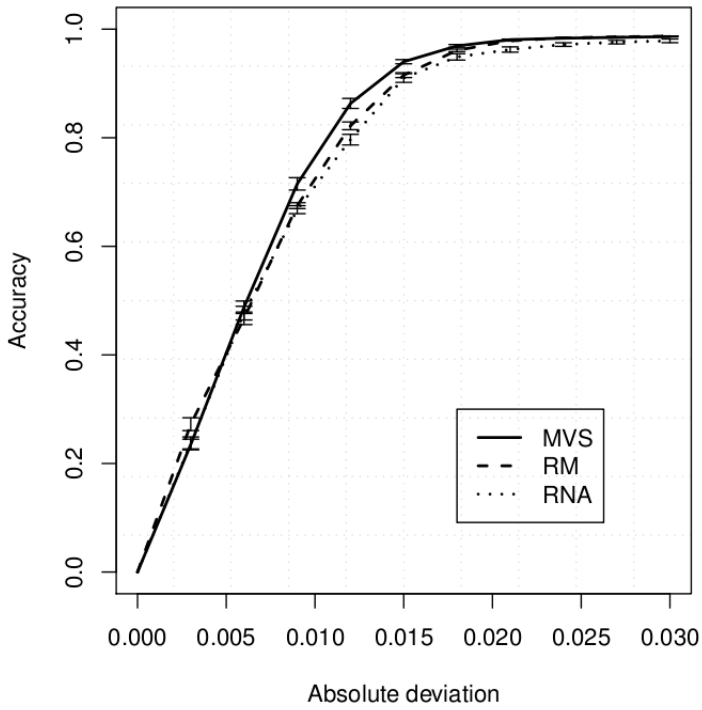

Figura 3: Curva REC para os modelos de previsão de PGPR.

a previsão perfeita. No entanto, em alguns casos existem erros elevados ao prever valores extremos. Por exemplo, as previsões de PGPR subestimam os valores esperados que estão próximos do valor máximo de PGPR. Também se pode observar uma tendência para sobrestimar os valores mais baixos de PGPR, PGS e PGI (Figura 4).

A importância relativa (como definida pela análise de sensibilidade) das variáveis de entrada na composição de carcaças, para os melhores modelos, é apresentada na Figura 5. A medida C12 é a variável de entrada mais importante para todas as tarefas de regressão, com uma influência que varia entre cerca de 25% (PO e modelo de RM) até 80% (PM e modelo RM). Estes resultados estão em conformidade com os obtidos por Cadavez (2009), onde as medidas de gordura dominaram os modelos para previsão da PM.

A importância relativa das restantes variáveis de entrada varia de tarefa para tarefa. Por exemplo, o sexo foi a segunda variável de entrada mais relevante para o modelo de previsão da PGPR (MVS), enquanto que foi a

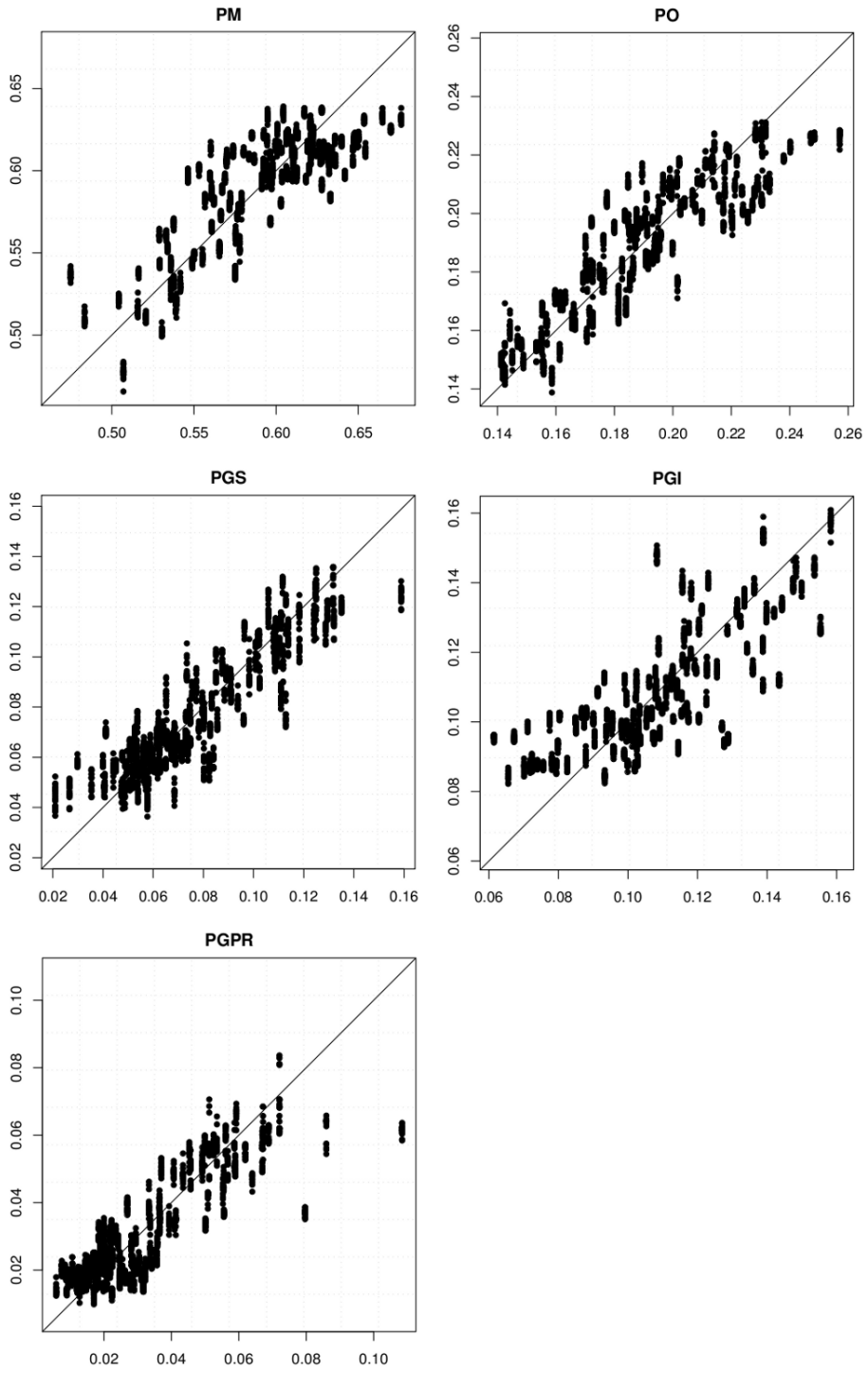

Figura 4: Gráficos de dispersão para os melhores modelos de regressão (eixo das abcissas - valores observados, eixo das ordenadas - previsões)

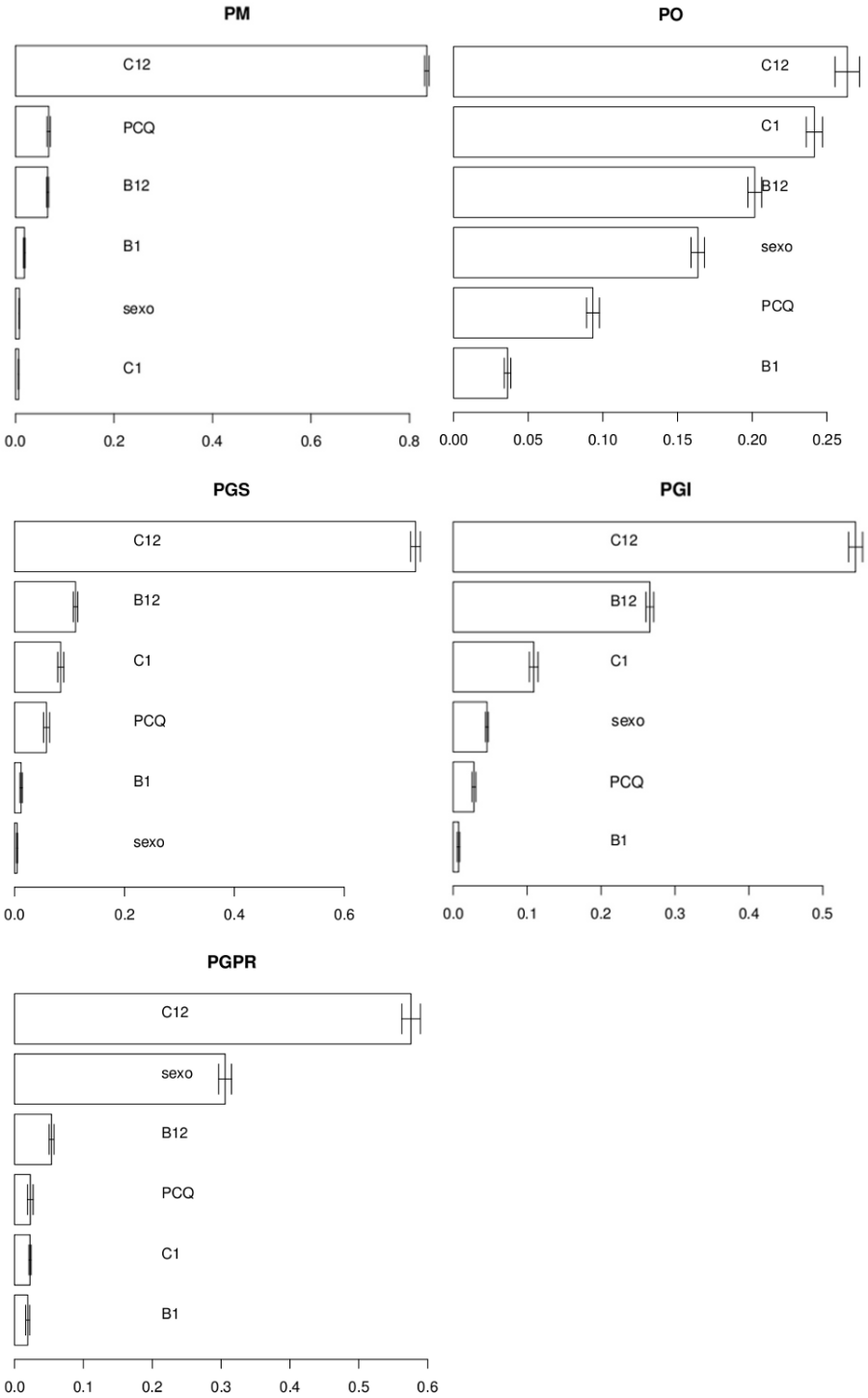

Figura 5: Importância relativa das variáveis de entrada

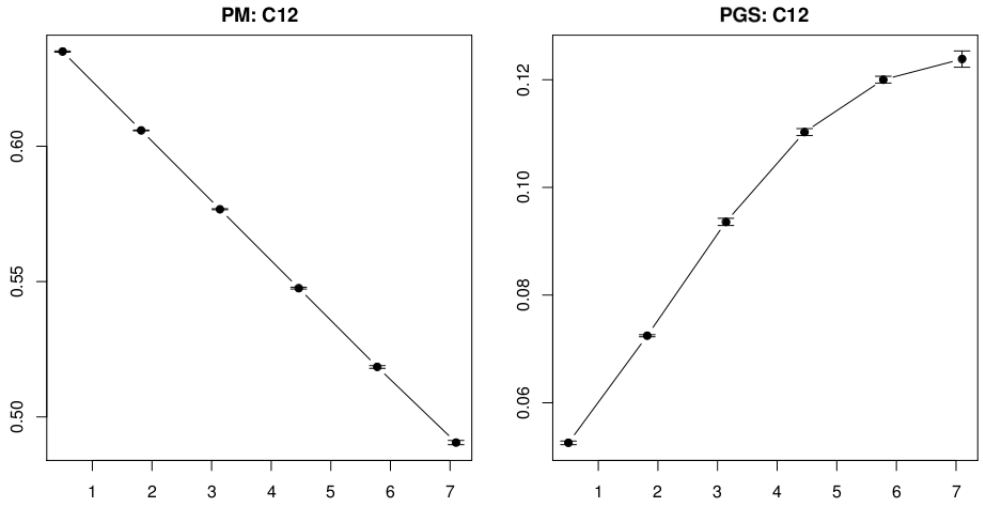

Figura 6: Curvas VEC apresentando a influência de C12 sobre os modelos de previsão da PM (esquerda) e da PGS (direita).

Tabela 6: Média de valores para cada classe, e respectivo desvio padrão

Classe	**C12**	**PM**	**PO**	**PTG**
1	4.79±0.97	0.54±0.02	0.16±0.02	0.30±0.02
2	2.40±0.53	0.59±0.01	0.19±0.01	0.21±0.02
3	1.29±0.44	0.62±0.01	0.21±0.01	0.17±0.02

variável de entrada menos importante para o modelo de previsão da PGS. A Figura 6 apresenta as curvas VEC para a variável de entrada C12 e os modelos PM e PGS. No primeiro gráfico, existe uma influência linear negativa de C12. Por outras palavras, o aumento da medida C12 leva a uma diminuição na PM das carcaças. Em relação ao segundo gráfico VEC, a influência de C12 na PGS é, em geral, positiva. Neste caso, o MVS mediu uma influência não linear (i.e. em forma de parábola).

Quanto ao modelo de classificação, este foi desenvolvido através do algoritmo MMG e definido para dividir os dados em três *clusters*. Na Figura 7 podemos observar a distribuição dos dados pelos três *clusters*, estando representado a preto a classe 1, a vermelho a classe 2 e a verde a classe 3. A Tabela 6 apresenta algumas estatísticas efectuadas a cada *cluster*, nomeadamente a média e o desvio padrão para cada atributo.

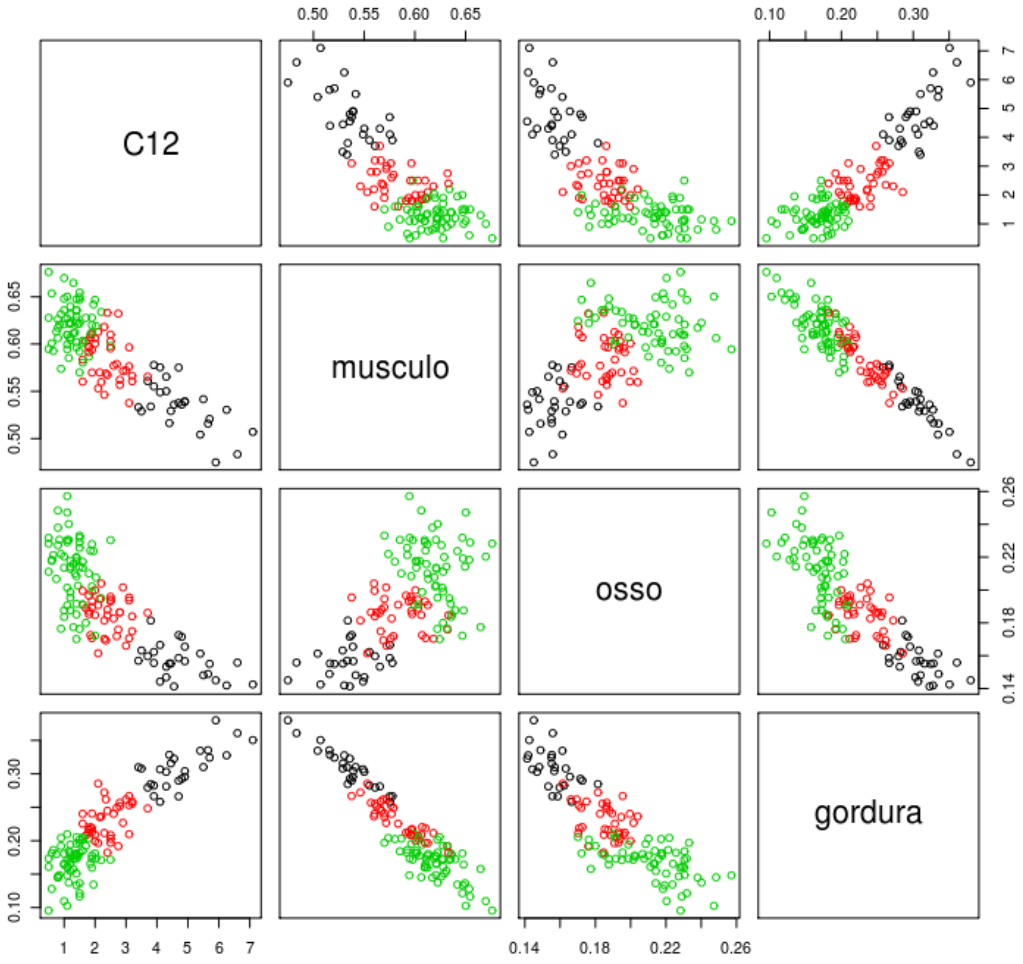

Figura 7: Representação gráfica dos *clusters*

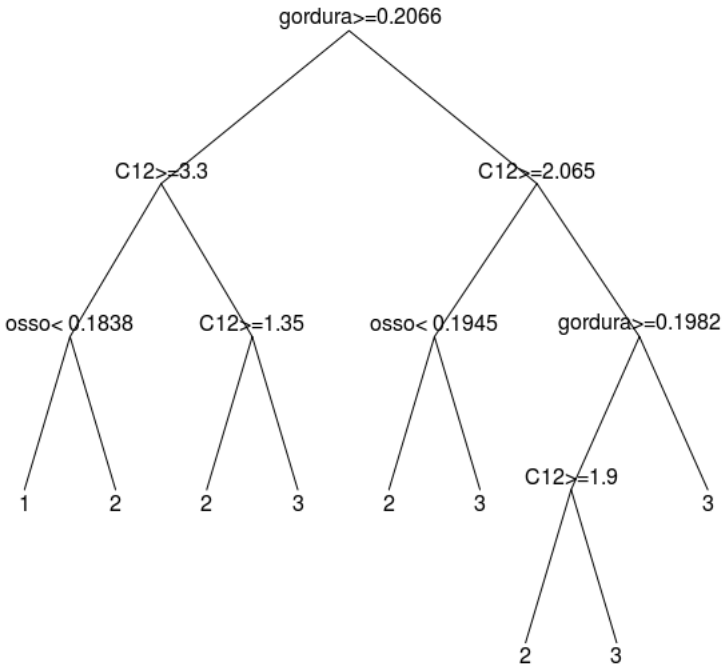

Figura 8: Árvore de Decisão para o modelo MMG.

Como se pode verificar na Figura 7 e na Tabela 6, a classe 1 é composta por carcaças de cordeiro com alto teor de gordura, com uma camada espessa de gordura subcutânea e com uma proporção de músculo relativamente baixa. Já a classe 2 é composta por carcaças de cordeiro com níveis intermédios de gordura subcutânea, músculo, osso e gordura. Por fim, a classe 3 define-se como uma classe com baixo teor de gordura e com proporções elevadas de músculo e osso, tendo, no entanto, níveis relativamente baixos de gordura subcutânea, o que poderá afectar a qualidade da carne. Tendo isso em conta, a classe mais desejável será a classe 2.

Na Figura 8 podemos observar a árvore de decisão obtida a partir do modelo MMG. Esta AD apresenta uma acuidade de 99.2% ao tentar mimetizar o modelo MMG.

As previsões dos melhores modelos de regressão foram de seguida intro-

Tabela 7: Matriz de confusão para a classificação das previsões dos melhores modelos de regressão através da AD.

	Classe Prevista		
Classe Real	**1**	**2**	**3**
1	26	0	0
2	0	28	9
3	0	1	61

duzidas na árvore de decisão gerada anteriormente, em que se observou que a AD classificou as previsões dos modelos de regressão com uma acuidade de 92%. A Tabela 7 apresenta a matriz de confusão obtida.

Este é um exemplo de como poderia ser proposto um novo sistema de classificação de carcaças, no entanto, por falta de tempo, não foi possível que estes resultados fossem analisados por um perito da área. Dessa forma, no futuro estes dados serão analisados cuidadosamente e, se necessário, serão realizadas novas experiências, na tentativa de se obter um sistema de classificação de carcaças eficaz. No Anexo A apresentam-se diversos exemplos do código R desenvolvido neste trabalho.

4.7 Implementação

Este trabalho de investigação é relevante na medida em que consiste numa oportunidade para se desenvolverem novos sistemas de classificação de carcaças de cordeiro com base em modelos gerados a partir de diversas técnicas de *Data Mining*. O objectivo final, caso se obtenham bons resultados, é que este estudo possa servir de base para o desenvolvimento de um Sistema de Apoio à Decisão para operar em ambiente real ao nível dos matadouros, permitindo uniformizar critérios, dar respostas num curto espaço de tempo e com custos reduzidos. No entanto, tal esforço encontra-se fora do âmbito desta dissertação.

4.8 Sumário

A primeira preocupação para a elaboração da parte prática deste trabalho foi adoptar uma metodologia que permitisse o correcto acompanhamento de um projecto de DM e que se ajustasse ao caso de estudo, pelo que se optou pela metodologia CRISP-DM. De seguida, surgiu a necessidade de escolher a ferramenta de DM a utilizar, pelo que, após analisar um conjunto de ferramentas disponíveis, se optou pelo ambiente de programação estatístico **R**, o qual satisfazia as necessidades pretendidas.

O objectivo proposto para este estudo passava pela modelação de um problema de regressão e classificação, onde se pretendia prever a composição de carcaças de cordeiro bragançanos a partir de um conjunto de medições passíveis de serem facilmente efectuadas ao nível dos matadouros. Durante o processo, houve necessidade de trocar impressões e ideias com um especialista da área e realizar experiências adicionais.

Nas primeiras experiências realizadas, os modelos RM e MVS obtiveram os melhores resultados, tendo sido possível identificar os atributos mais relevantes (e.g. profundidade da gordura subcutânea na 12ª costela). Por sua vez, quando se aplicou um modelo de classificação, este revelou possuir uma acuidade elevada, sendo bastante promissor. No entanto, por falta de tempo não foi possível aprofundar este método.

5 Conclusões

5.1 Síntese

A criação de um método de estimativa da composição tecidual de carcaças é um ponto fundamental para o desenvolvimento de sistemas objectivos de classificação. A classificação de carcaças visa agrupar carcaças de acordo com as suas características, de modo a formar lotes uniformes, regulando assim a sua comercialização e criando uma base formal para o pagamento. Este trabalho visa a previsão da composição tecidual de carcaças de cordeiro bragançano com base em medições que podem ser obtidas ao nível dos matadouros de forma rápida, económica e não destrutiva, e ainda propor um sistema de classificação de carcaças de cordeiro eficaz, baseado em técnicas de DM. Foi considerado um pequeno conjunto de dados (num total de 125 amostras) de cordeiros bragançanos da zona nordeste de Portugal.

Os sistemas de BI orientados à tomada de decisão permitem combinar a recolha de dados com ferramentas de análise, com o principal objectivo de disponibilizar informações para a tomada de decisão. De entre estes sistemas fazem parte as técnicas de DM para a extracção de conhecimento. O DM é um processo chave neste trabalho, tendo sido comprovada a sua importância pela aplicação de algoritmos de aprendizagem com vista à procura de padrões e subsequente descoberta de informações úteis. Para auxiliar na condução deste trabalho, recorreu-se à metodologia CRISP-DM e à ferramenta estatística **R**, em conjunto com a biblioteca *RMiner* (Cortez, 2010).

Dada a natureza das variáveis a modelar, optou-se por definir o objectivo de DM como sendo de **Regressão**, e, para facilitar a sua aplicação em ambiente real, optou-se por definir também uma tarefa de **Classificação**. Na fase de modelação foram adoptadas três técnicas de regressão: RM, RNAs e MVSs; e uma técnica de classificação: MMG. No decorrer de todo o processo, foram aprofundados os factores que influenciam a composição de carcaças de cordeiro bragançanos, através de uma análise de sensibilidade.

Após uma análise detalhada dos resultados de regressão obtidos, através de medições de erro, RAE e curvas REC, concluiu-se que se obtém resultados

com uma elevada qualidade de previsão ao utilizar todos os atributos e ao utilizar os modelos mais eficazes para cada tecido (RM para PM e PGI, e MVS para PO, PGS e PGPR). Por sua vez, após uma análise dos resultados de classificação obtidos, através de estatísticas e de uma matriz de confusão, concluiu-se que se obtém resultados com uma elevada acuidade, justificando um estudo mais detalhado no futuro.

5.2 Discussão

Ao terminar um projecto, é adequado confrontar os resultados obtidos com os objectivos inicialmente estabelecidos. É fundamental optar por uma postura crítica relativamente às limitações de que entretanto se foi tomando consciência. Em primeiro lugar, convém referir que o objectivo inicialmente proposto tem a sua complexidade, apesar de já terem sido desenvolvidos vários estudos similares sobre o tema da estimação da composição tecidual de carcaças. Importa também referir que foram efectuadas diversas experiências que não estão descritas neste trabalho (e.g. aumentar a acuidade dos modelos de regressão procurando modelos com acuidade elevada nas extremidades, onde se registou baixo desempenho nos modelos obtidos) e que não foram levadas até ao fim por não se terem atingido resultados satisfatórios e porque, em conversa com o orientador e co-orientador, se chegou à conclusão que seria melhor optar por outra via.

Os resultados obtidos têm utilidade. Nos modelos de regressão, os resultados de RAE para os melhores modelos mostram uma melhoria de aproximadamente 36% (PGI) a 58% (PGS) quando comparados com o previsor médio naïve. Por sua vez, o modelo de classificação apresenta uma acuidade de 92%, o que é muito bom.

Através de uma análise de sensibilidade, concluiu-se que o factor com maior influência na estimativa dos cinco tecidos em estudo é a espessura da gordura subcutânea entre a 12ª e 13ª costelas (C12). O resultado deste trabalho é assim relevante para o domínio da ciência animal pois ajuda na compreensão da relação entre algumas medições de carcaças e a proporção total de determinados tecidos.

Devido à legislação europeia, na fase de abate dos animais, estes devem ser classificados segundo o modelo europeu de classificação de carcaças. No entanto, este sistema de classificação de carcaças é baseado em factores subjectivos, como a avaliação visual da conformação. A abordagem proposta neste trabalho é baseada em medições objectivas, gerando modelos que podem ser integrados num sistema de apoio à decisão, o que permitirá auxiliar a velocidade e qualidade da classificação de carcaças ao nível do matadouro. Por outro lado, uma vez que algumas medições podem ser efectuadas durante a fase de crescimento dos ovinos, por exemplo por ultrassonografia, esta informação poderá ser utilizada para controlar o crescimento dos animais de forma a estes terem uma composição óptima quando forem abatidos. Por exemplo, a composição dos animais pode ser manipulada através da sua alimentação.

Por sua vez, a ferramenta R revelou-se adequada a este trabalho. Permite pré-processar os dados e aplicar diversas técnicas de DM com comandos simples. De destacar também que a liberdade de controlo oferecida pelo R permitiu que diversos critérios fossem utilizados na fase de avaliação. Ou seja, foi possível obter erros RAE e curvas REC (que não são muito comuns noutras aplicações de DM). Todavia, há que referir a curva de aprendizagem do R, que é maior do que outras aplicações gráficas, exigindo conhecimentos ao nível de programação por parte do utilizador.

Por último, importa realçar que este trabalho deu origem a uma publicação científica em conferência de âmbito internacional:

F. Silva, P. Cortez and V. Cadavez. Using Multiple Regression, Neural Networks and Support Vector Machines to Predict Lamb Carcasses Composition. **FOODSIM'2010 Conference, Bragança, Portugal, 2010 : proceedings**, pp. 41-45, Bragança, Portugal, June, 2010. EUROSIS. ISBN 978-90-77381-56-1.

5.3 Trabalho Futuro

Concluído este trabalho, importa indicar alguns caminhos de investigação na área da classificação de carcaças. Em particular indicam-se as seguintes vias:

1. Analisar cuidadosamente o exemplo de sistema de classificação obtido por *Clustering* e, se necessário, realizar novas experiências de forma a obter um sistema de classificação eficaz.

2. Implementar a solução proposta num ambiente real, através de um protótipo. A ideia é desenvolver um sistema de apoio à decisão amigável, que possa operar em tempo real, de forma a obter um *feedback* por parte dos matadouros;

3. Repetir as experiências com uma base de dados mais extensa;

4. Estender o estudo a outras raças de ovinos.

Bibliografia

Becker, R., Chambers, J., and Wilks, A. (1988). *The new S language: a programming environment for data analysis and graphics*. Wadsworth and Brooks/Cole Advanced Books & Software Monterey, CA, USA.

Bi, J. and Bennett, K. (2003). Regression error characteristic curves. In *Proceedings of the Twentieth International Conference on Machine Learning*, Washington, DC.

Cadavez, V. (2009). Prediction of lean meat proportion of lamb carcasses. *Archiva Zootechnica*, 12(4):46–58.

Cadavez, V., Rodrigues, S., Pereira, E., Delfa, R., and Teixeira, A. (2002). Predicción de la composición de la canal de cabritos por ultrasonografia in vivo. *ITEA*, 98A(1):39–50.

Cadavez, V., Teixeira, A., and de Trás-os-Montes e Alto Douro, U. (2004). *Ultra-sonografia para avaliar in vivo e ex vivo carcaças de ovinos. Estudos nas raças Churra Galega Bragançana e Suffolk*. PhD thesis, Tese de Doutoramento, Universidade de Trás-os-Montes e Alto Douro.

Cadavez, V., Teixeira, A., and Delfa, R. (1999). Utilización de ultrasonidos junto con el peso vivo y el peso de la canal caliente para la estimación del peso de las piezas de carnicería en corderos de raza churra Galega Bragançana: Comparación de sondas de 5 y 7, 5 MHz. *Producción Ovina y Caprina, SEOC, n^{o}*, 24:425–432.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc*, 78.

Chatterjee, S. and Hadi, A. (2006). *Regression analysis by example*. John Wiley and Sons.

Cherkassky, V. and Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks*, 17(1):113–126.

Cortes, B. (2005). Sistemas de suporte à decisão. *FCA-Editora Informática*.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cortez, P. (2010). Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. *Advances in Data Mining. Applications and Theoretical Aspects*, pages 572–583.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009a). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553.

Cortez, P., Portelinha, M., Rodrigues, S., Cadavez, V., and Teixeira, A. (2006). Lamb meat quality assessment by support vector machines. *Neural Processing Letters*, 24(1):41–51.

Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009b). Using data mining for wine quality assessment. In *Discovery Science*, pages 66–79. Springer.

Dayhoff, J. and DeLeo, J. (2001). Artificial neural networks. *CA A Cancer Journal for Clinicians*, 91(S8):1615–1635.

Delfa, R., González, C., and Teixeira, A. (1996). Use of cold carcass weight and fat depth measurements to predict carcass composition of Rasa Aragonesa lambs. *Small Ruminant Research*, 20(3):267–274.

Delfa, R. and Teixeira, A. (1998). Calidad de canal ovina. *Ovino de carne: Aspectos claves*, pages 373–400.

DGV (2010). Efectivos de Fêmeas Autóctones Exploradas em Linha Pura Inscritas no Livro de Adultos. Technical report, Ministério da Agricultura.

Fisher, A. V. (1987). Limitations of present classification schemes and new developments. *Winter Meeting of the British Society of Animal Science, Paper n.o 47*.

Fogarty, N., Hopkins, D., and van de, V. (2000). Lamb production from diverse genotypes. 2. Carcass characteristics. *Animal Science (Glasgow)*, 70(1):147–156.

Forrest, J. (1995). New techniques for estimation of carcass composition. *Quality and Grading of Carcasses of Meat Animals*, pages 157–172.

Fraley, C. and Raftery, A. (2006). MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical report, Citeseer.

Hall, D., Gilmour, A., Fogarty, N., Holst, P., and Hopkins, D. (2001). Growth and carcass composition of second-cross lambs. 1. Effect of sex and growth path on pre-and post-slaughter estimates of carcass composition. *Australian Journal of Agricultural Research*, 52(8):859–867.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, NY, USA, 2nd ed.

Haykin, S. (1999). *Neural networks: a comprehensive foundation*. Prentice Hall PTR Upper Saddle River, NJ, USA, 2 edition.

Hearst, M., Dumais, S., Osman, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent systems*, 13(4):18–28.

Hopkins, D. (2008). An industry applicable model for predicting lean meat yield in lamb carcasses. *Australian Journal of Experimental Agriculture*, 48(6-7):757–761.

Hopkins, D., Ponnampalam, E., and Warner, R. (2008). Predicting the composition of lamb carcasses using alternative fat and muscle depth measures. *Meat Science*, 78:400–405.

Hopkins, D. L., Roberts, A. H. K., and Pirlot, K. l. (1993). Estimation of mutton carcass components using two predictors. *Meat Science*, 33:293–299.

Jones, S., Jeremiah, L., Tong, A., Robertson, W., and Gibson, L. (1992). Estimation of lamb carcass composition using an electronic probe, a visual scoring system and carcass measurements. *Canadian Journal Animal Science*, 72:237–244.

Kempster, A. (1983). Carcass quality and its measurement in sheep. *Sheep production*, pages 59–74.

Kirton, A., Carter, A., Clarke, J., Sinclair, D., Mercer, G., and Duganzich, D. (1995). A comparison between 15 ram breeds for export lamb production. 1. Liveweights, body components, carcass measurements, and composition. *New Zealand journal of agricultural research*, 38:347–360.

Kirton, A. and Johnson, D. (1979). Interrelationships between GR and other lamb carcass fatness measurements. In *Proceedings of the New Zealand Society of Animal Production*, volume 39, pages 194–201.

Kirton, A., Mercer, G., and Duganzich, D. (1992). A comparison between subjective and objective (carcass weight plus GR or the Hennessy Grading Probe) methods for classifying lamb carcasses. In *Proceedings of the New Zealand Society of Animal Production*, volume 52, pages 41–41. New Zealand Society of Animal Prod Publ.

Kirton, A., Mercer, G., Duganzich, D., Clarke, J., and Woods, E. (1999). Composition of lamb carcasses and cuts based on the October 1983 to 1998 export lamb carcass classification standards in New Zealand. *New Zealand Journal of Agricultural Research*, 42:65–76.

Kirton, A., Woods, E., and Duganzich, D. (1984). Predicting the fatness of lamb carcasses from carcass wall thickness measured by ruler or by a total depth indicator (TDI) probe. *Livestock Production Science*, 11(2):185–194.

Kirton, A. H. (1998). Is carcass classification useful or necessary? In *Proceedings of the New Zealand Society of Animal Production*, volume 58, pages 211–213. New Zealand Society of Animal Prod Publ.

Maimon, O. and Rokach, L. (2005). *Data mining and knowledge discovery handbook*. Springer-Verlag New York Inc.

Marques, R., Correia, A., and Cortez, P. (2009). Data mining applied to compaction of geomaterials. In *Proc.of the 8th Int. Conf. on the Bearing Capacity of Roads, Railways and Airelds (BCR2A09)*.

Miller, S. (2008). R You Ready for Open Source Statistics? *Open BI Forum, Information Management*.

Murphy, T., Loerch, S., McClure, K., and Solomon, M. (1994). Effects of restricted feeding on growth performance and carcass composition of lambs. *Journal of Animal Science*, 72(12):3131–3137.

Palsson, H. (1939). Meat qualities in the sheep with special reference to Scottish breeds and crosses. *Journal of Agricultural Science*, 29:544–626.

Piatetsky-Shapiro, G. (2009). Data mining tools used poll. http://www.kdnuggets.com/polls/2009/data-mining-tools-used.htm.

Pontil, M. and Verri, A. (1998). Properties of support vector machines. *Neural Computation*, 10(4):955–974.

Price, M. A. and Jones, S. D. M. (1995). Development of carcass grading and classification systems. *Quality and Grading of Carcasses of Meat Animals*, pages 215–228.

Rexer, K. (2008). Second annual data miner survey. Technical report, Rexer Analytics.

Safari, E., Hopkins, D., and Fogarty, N. (2001). Diverse lamb genotypes 4. Predicting the yield of saleable meat and high value trimmed cuts from carcass measurements. *Meat Science*, 58(2):207–214.

Santos, M. and Azevedo, C. (2005). *Data Mining: Descoberta de conhecimento em bases de dados*. FCA-Editora de Informática.

Santos, M. and Ramos, I. (2006). Business Intelligence: tecnologias da informação na gestão de conhecimento.

Santos-Silva, F., Ivo, R., Sousa, M., Vicente, A., Carolino, I., Carolino, N., and Gama, L. (2009). Análise da Estrutura Genética de Populações Ovinas Churras Portuguesas. *Archivos de zootecnia*, 58(Supl 1):493–496.

Silva, Á., Cortez, P., Santos, M., Gomes, L., and Neves, J. (2008). Rating organ failure via adverse events using data mining in the intensive care unit. *Arificial Intelligence in Medicine*, 43(3):179–193.

Simm, G. and Dingwall, W. (1989). Selection indices for lean meat production in sheep. *Livestock Production Science*, 21(3):223–233.

Snowder, G. D., Field, R. A., and Busboom, J. R. (1994). The efficacy of the body wall thickness measure for estimating total commercially trimmed retail cuts of lamb. *Sheep research progress report, U.S. Sheep Experiment, Idaho*.

Sobral, M. (1987). *Recursos Genéticos, Raças autóctones, espécies ovina e caprina*. Direcção Geral da Pecuária, Portugal.

Stanford, K., Woloschuk, C., McClelland, L., Jones, S., and Price, M. (1997). Comparison of objective external carcass measurements and subjective conformation scores for prediction of lamb carcass quality. *Canadian Journal of Animal Science*, 72(2):217–223.

Teixeira, A., Delfa, R., and Alberti, P. (1998). Influence of production factors on the characteristics of meat from ruminants in Mediterranean area. *Publication - European Association for Animal Production*, 90:315–319.

Teixeira, A., Delfa, R., and González, C. (1992). El grado de engrasamiento. *Rev. Ovis*, 19:21–35.

Teixeira, A., Matos, S., Rodrigues, S., Delfa, R., and Cadavez, V. (2006). In vivo estimation of lamb carcass composition by real-time ultrasonography. *Meat Science*, 74(2):289–295.

Timon, V. and Bichard, M. (1965). Quantitative estimates of lamb carcass composition. 3. Carcass measurements and a comparison of predictive efficiency of sample joint composition, carcass specific gravity determinations and carcass measurements. *Animal Production*, 7:189–201.

Topel, D. and Kauffman, R. (1998). Live animal and carcass composition measurement. *Designing foods: Animal product options in the marketplace. Ed. National Academy Press. Washington, DC*, pages 258–272.

Turban, E., Sharda, R., Aronson, J. E., and King, D. (2007). *Business intelligence, a managerial approach.* Prentice Hall.

USDA (1992a). United states standards for grades of lamb, yearling mutton, and mutton carcasses. Washington, DC, USDA Agricultural Marketing Service Livestock and Seed Division.

USDA (1992b). United states standards for grades of slaughter lambs, yearlings, and sheep. Washington, DC, USDA Agricultural Marketing Service Livestock and Seed Division.

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pages 29–39.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, San Francisco, CA, USA.

Wolf, B., Jones, D., and Owen, M. (2006). In vivo prediction of carcass composition and muscularity in purebred Texel lambs. *Meat Science*, 74(2):416–423.

Wood, J. (1990). Consequences for meat quality of reducing carcass fatness. *Reducing fat in meat animals*, pages 344–397.

Wood, J., MacFie, H., Pomeroy, R., and Twinn, D. (1980). Carcass composition in four sheep breeds: the importance of type of breed and stage of maturity. *Animal production*, 30(1):135–152.

Wood, J. D. (1995). The influence of carcass composition on meat quality. *Quality and grading of carcasses of meat animals*, page 131Â151.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B., Yu, P., et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37.

Anexos A - Exemplos de código

Em seguida são apresentados os vários ficheiros de código que foram utilizados com vista à criação dos vários modelos de RM, RNA, MVS, MMG e AD, assim como o código de análise e avaliação dos respectivos modelos.

pre-processamento.R

```
# Carregar tabela para a variavel "d"
d=read.table("master.csv",sep=";",header=TRUE)
d1<-d[,c(2:8)] # BD 1 -> musculo
d2<-d[,c(2:7,9)] # BD 2 -> gsub
d3<-d[,c(2:7,10)] # BD 3 -> gint
d4<-d[,c(2:7,11)] # BD 4 -> osso
d5<-d[,c(2:7,12)] # BD 5 -> grenal
```

modelacao.R

```
source("pre-processamento.R") # Carregar BDs
library(rminer)
RUNS=20
### Data.frame1
# Regressao multipla (mr)
v=c("kfold",10)
MR1=mining(musculo~.,d1,model="mr",Runs=RUNS,method=v)
# Redes neuronais artificiais (mlp)
m=c(3,100,"kfold",3,"RAE"); s=seq(1,8,1)
NN1=mining(musculo~.,d1,model="mlpe",Runs=RUNS,method=v,mpar=m,
search=s,feat="s")
# Support Vector Machines (svm)
m=c(NA,NA,"kfold",3,"RAE"); s=2^seq(-13,1,2)
SV1=mining(musculo~.,d1,model="svm",Runs=RUNS,method=v,mpar=m,
search=s,feat="s")
# Mostrar os resultados
```

```
print(MR1); print(NN1); print(SV1)
# Guardar resultados
savemining(MR1,"imr1"); savemining(NN1,"inn1"); savemining(SV1,"isv1")
### Data.frame2
# Regressao multipla (mr)
v=c("kfold",10)
MR2=mining(gsub~.,d2,model="mr",Runs=RUNS,method=v)
# Redes neuronais artificiais (mlp)
m=c(3,100,"kfold",3,"RAE"); s=seq(1,8,1)
NN2=mining(gsub~.,d2,model="mlpe",Runs=RUNS,method=v,mpar=m,
search=s,feat="s")
# Support Vector Machines (svm)
m=c(NA,NA,"kfold",3,"RAE"); s=2^seq(-13,1,2)
SV2=mining(gsub~.,d2,model="svm",Runs=RUNS,method=v,mpar=m,
search=s,feat="s")
# Mostrar os resultados
print(MR2); print(NN2); print(SV2)
# Guardar resultados
savemining(MR2,"imr2"); savemining(NN2,"inn2"); savemining(SV2,"isv2")
### Data.frame3
# Regressao multipla (mr)
v=c("kfold",10)
MR3=mining(gint~.,d3,model="mr",Runs=RUNS,method=v)
# Redes neuronais artificiais (mlp)
m=c(3,100,"kfold",3,"RAE"); s=seq(1,8,1)
NN3=mining(gint~.,d3,model="mlpe",Runs=RUNS,method=v,mpar=m,
search=s,feat="s")
# Support Vector Machines (svm)
m=c(NA,NA,"kfold",3,"RAE"); s=2^seq(-13,1,2)
SV3=mining(gint~.,d3,model="svm",Runs=RUNS,method=v,mpar=m,
search=s,feat="s")
# Mostrar os resultados
print(MR3); print(NN3); print(SV3)
```

```
# Guardar resultados
savemining(MR3,"imr3"); savemining(NN3,"inn3"); savemining(SV3,"isv3")
### Data.frame4
# Regressao multipla (mr)
v=c("kfold",10)
MR4=mining(osso~.,d4,model="mr",Runs=RUNS,method=v)
# Redes neuronais artificiais (mlp)
m=c(3,100,"kfold",3,"RAE"); s=seq(1,8,1)
NN4=mining(osso~.,d4,model="mlpe",Runs=RUNS,method=v,mpar=m,
search=s,feat="s")
# Support Vector Machines (svm)
m=c(NA,NA,"kfold",3,"RAE"); s=2^seq(-13,1,2)
SV4=mining(osso~.,d4,model="svm",Runs=RUNS,method=v,mpar=m,
search=s,feat="s")
# Mostrar os resultados
print(MR4); print(NN4); print(SV4)
# Guardar resultados
savemining(MR4,"imr4"); savemining(NN4,"inn4"); savemining(SV4,"isv4")
### Data.frame5
# Regressao multipla (mr)
v=c("kfold",10)
MR5=mining(grenal~.,d5,model="mr",Runs=RUNS,method=v)
# Redes neuronais artificiais (mlp)
m=c(3,100,"kfold",3,"RAE"); s=seq(1,8,1)
NN5=mining(grenal~.,d5,model="mlpe",Runs=RUNS,method=v,mpar=m,
search=s,feat="s")
# Support Vector Machines (svm)
m=c(NA,NA,"kfold",3,"RAE"); s=2^seq(-13,1,2)
SV5=mining(grenal~.,d5,model="svm",Runs=RUNS,method=v,mpar=m,
search=s,feat="s")
# Mostrar os resultados
print(MR5); print(NN5); print(SV5)
# Guardar resultados
```

```
savemining(MR5,"imr5"); savemining(NN5,"inn5"); savemining(SV5,"isv5")
```

avaliacao.R

```
L=c("sex","HCW","c1","c12","b1","b12")
## Modelo 1
# Carregar os modelos
MR1=loadmining("imr1");NN1=loadmining("inn1");SV1=loadmining("isv1")
print(t.test(mmetric(NN1,"RAE"),mmetric(SV1,"RAE")))
print(meanint(mmetric(SV1,"RAE")))
M=vector("list",3);M[[1]]=SV1;M[[2]]=NN1;M[[3]]=MR1
# Gerar gráfico REC
mgraph(M,graph="REC",leg=c("SVM","NN","MR"),PDF="rec1")
# Gerar gráfico de Importancia
mgraph(SV1,graph="IMP",leg=L,xval=0.3,PDF="imp1")
# Gerar gráfico VEC
mgraph(SV1,graph="VEC",leg=L,PDF="vec1")
## Modelo 2
# Carregar os modelos
MR2=loadmining("imr2");NN2=loadmining("inn2");SV2=loadmining("isv2")
print(t.test(mmetric(NN2,"RAE"),mmetric(SV2,"RAE")))
print(meanint(mmetric(SV2,"RAE")))
M=vector("list",3);M[[1]]=SV2;M[[2]]=NN2;M[[3]]=MR2
# Gerar gráfico REC
mgraph(M,graph="REC",leg=c("SVM","NN","MR"),PDF="rec2")
# Gerar gráfico de Importancia
mgraph(SV2,graph="IMP",leg=L,xval=0.3,PDF="imp2")
# Gerar gráfico VEC
mgraph(SV2,graph="VEC",leg=L,PDF="vec2")
## Modelo 3
# Carregar os modelos
MR3=loadmining("imr3");NN3=loadmining("inn3");SV3=loadmining("isv3")
print(t.test(mmetric(NN3,"RAE"),mmetric(SV3,"RAE")))
```

```
print(meanint(mmetric(SV3,"RAE")))
M=vector("list",3);M[[1]]=SV3;M[[2]]=NN3;M[[3]]=MR3
# Gerar gráfico REC
mgraph(M,graph="REC",leg=c("SVM","NN","MR"),PDF="rec3")
# Gerar gráfico de Importancia
mgraph(SV3,graph="IMP",leg=L,xval=0.3,PDF="imp3")
# Gerar gráfico VEC
mgraph(SV3,graph="VEC",leg=L,PDF="vec3")
## Modelo 4
# Carregar os modelos
MR4=loadmining("imr4");NN4=loadmining("inn4");SV4=loadmining("isv4")
print(t.test(mmetric(NN4,"RAE"),mmetric(SV4,"RAE")))
print(meanint(mmetric(SV4,"RAE")))
M=vector("list",3);M[[1]]=SV4;M[[2]]=NN4;M[[3]]=MR4
# Gerar gráfico REC
mgraph(M,graph="REC",leg=c("SVM","NN","MR"),PDF="rec4")
# Gerar gráfico de Importancia
mgraph(SV4,graph="IMP",leg=L,xval=0,PDF="imp4")
# Gerar gráfico VEC
mgraph(SV4,graph="VEC",leg=L,PDF="vec4")
## Modelo 5
# Carregar os modelos
MR5=loadmining("imr5");NN5=loadmining("inn5");SV5=loadmining("isv5")
print(t.test(mmetric(NN5,"RAE"),mmetric(SV5,"RAE")))
print(meanint(mmetric(SV5,"RAE")))
M=vector("list",3);M[[1]]=SV5;M[[2]]=NN5;M[[3]]=MR5
# Gerar gráfico REC
mgraph(M,graph="REC",leg=c("SVM","NN","MR"),PDF="rec5")
# Gerar gráfico de Importancia
mgraph(SV5,graph="IMP",leg=L,xval=0.3,PDF="imp5")
# Gerar gráfico VEC
mgraph(SV5,graph="VEC",leg=L,PDF="vec5")
```

mclust.R

```
## Preparacao dos dados
d=read.table("master.csv",sep=";",header=TRUE)
dA<-d[,c(5,8,11)]
# Somar gorduras:
for(i in 1:nrow(d)){ dA[i,4]=d[i,9]+d[i,10]+d[i,12]}
colnames(dA)<-c("C12","musculo","osso","gordura")
sdA=scale(dA)
#####################################
## Preparacao dos dados de previsao
library(rminer)
M1=loadmining("Modelos/Modelos/imr1") # musc
M2=loadmining("Modelos/Modelos/isv2") # gsub
M3=loadmining("Modelos/Modelos/isv3") # gint
M4=loadmining("Modelos/Modelos/imr4") # osso
M5=loadmining("Modelos/Modelos/isv5") # grenal
dT<-matrix(nrow=125,ncol=5)
a<-matrix(nrow=125,ncol=20)
for(j in 1:20){a[,j]<-unlist(M1$pred[j])} # valores previstos numa matriz
for(j in 1:nrow(d)){dT[j,1]=mean(a[j,])} # Media dos valores previstos
for(j in 1:20){a[,j]<-unlist(M2$pred[j])} # valores previstos numa matriz
for(j in 1:nrow(d)){dT[j,3]=mean(a[j,])} # Media dos valores previstos
for(j in 1:20){a[,j]<-unlist(M3$pred[j])} # valores previstos numa matriz
for(j in 1:nrow(d)){dT[j,4]=mean(a[j,])} # Media dos valores previstos
for(j in 1:20){a[,j]<-unlist(M4$pred[j])} # valores previstos numa matriz
for(j in 1:nrow(d)){dT[j,2]=mean(a[j,])} # Media dos valores previstos
for(j in 1:20){a[,j]<-unlist(M5$pred[j])} # valores previstos numa matriz
for(j in 1:nrow(d)){dT[j,5]=mean(a[j,])} # Media dos valores previstos
dB<-matrix(nrow=125,ncol=4)
dB[,1]<-d[,5]
dB[,c(2:3)]<-dT[,c(1:2)]
# Somar gorduras:
```

```
for(i in 1:nrow(d)){ dB[i,4]=dT[i,3]+dT[i,4]+dT[i,5]}
colnames(dB)<-c("C12","musculo","osso","gordura")
#####################################
## Modelacao
# Model Based Clustering
library(mclust)
fit1<-Mclust(sdA,G=3)
y=factor(fit1$classification);d1<-data.frame(dA,y)
library(rminer)
# melhorar o detalhe da arvore, para ter melhor informacao e precisao:
RP=rpart.control(minsplit=4, cp=0.0005)
MM1=fit(y~.,d1,model="dt",control=RP)
# para 3 clusters:
print(MM1@object)
PMM1=majorClass(predict(MM1,d1),L=levels(d1$y))
print(mmetric(d1$y,PMM1,"ACC"))
# verificar as previsoes:
dB=data.frame(dB)
P3=majorClass(predict(MM1,dB),L=levels(d1$y))
# accuracy:
print(mmetric(PMM1,P3,"ACC"))
# matriz de confusao:
print(mmetric(PMM1,P3,"CONF"))
```