Instituto Universitário de Lisboa

Departamento de Ciências e Tecnologias da Informação

Técnicas de *Data Mining* aplicadas à melhoria de gestão de medicamentos: estudo de uma farmácia comunitária

Hugo Alexandre Velho Vilalva Sena

Trabalho de projecto submetido como requisito parcial para obtenção do grau de

Mestre em Gestão de Sistemas de Informação

Orientador:

Doutor Paulo Alexandre Ribeiro Cortez, Professor Associado
Universidade do Minho

Co-orientador:

Doutor Pedro Nogueira Ramos, Professor Associado
ISCTE-Instituto Universitário de Lisboa

Outubro, 2012

Agradecimentos

Ao longo da elaboração desta dissertação, contei com o apoio e colaboração de diversas pessoas. Deste modo, gostaria de apresentar os meus mais sinceros agradecimentos a todos os que, directa e indirectamente, colaboraram para a realização da mesma:

Ao meu orientador Prof. Doutor Paulo Cortez que esteve sempre disponível em ajudar, orientar, e mesmo apesar da distância física conseguiu ter sempre uma presença activa e uma palavra de motivação, pelo que sem o próprio esta dissertação não seria possível;

Ao Prof. Doutor Pedro Ramos, co-orientador, que sempre esteve disponível e pelo que gostaria de deixar uma palavra de apreço;

À Dra. Matilde e à Dra. Patrícia pela disponibilização de todos os dados necessários e criação das condições para que este trabalho pudesse seguir em frente;

Aos meus amigos que estiveram sempre presentes e ajudaram com a sua disponibilidade e amizade sempre incondicional;

Aos meus Pais e irmão, que estiveram sempre presentes durante o meu percurso de vida;

Especialmente à Rita, a mulher da minha vida, que muito sofreu ao meu lado, sempre acreditou em mim e nunca me abandonou nas horas mais difíceis e foi sempre solidária até mesmo nas longas horas sem dormir. A ti espero agradecer todos dias da minha vida.

O meu *muito obrigado*!

Resumo

Com a crescente regulamentação, nacional e comunitária, no sector farmacêutico e a generalização de medicamentos genéricos, bem como as pressões externas e internas que o país atravessa, é cada vez mais reduzida a margem de lucro neste sector, pondo em causa a sustentabilidade de algumas farmácias, pelo existe, actualmente, a necessidade de repensar e dinamizar esta área de negócio. É desta forma que a área de *Business Intelligence* pode apresentar novos mecanismos de forma a trazer mais-valias para as empresas na generalidade, assim como para este tipo de negócios, que apesar de apresentarem características de natureza muito específica, visam igualmente o lucro e a satisfação dos seus públicos-alvo, nomeadamente os utentes. Assim, este trabalho visa utilizar uma abordagem de *Data Mining,* pretendendo-se a partir de um conjunto de dados extrair conhecimento válido e útil no sentido de apoiar a gestão e controlo de *stocks* de medicamentos, produtos de saúde e outros produtos e serviços comercializados numa farmácia comunitária.

Esta dissertação versa sobre um estudo de caso de uma farmácia comunitária, com instalações em território nacional, tendo como base a análise de séries temporais, com periodicidade mensal e semanal, de uma selecção de três produtos, nomeadamente através das vendas realizadas entre o período de 2003 e 2012. Do trabalho realizado resulta a comparação entre dois métodos populares de previsão: o alisamento exponencial (incluindo múltiplas das suas variantes) e a metodologia ARIMA, também conhecida por método de *Box-Jenkins*. Pretende-se, assim, determinar qual o método que melhor se adapta à previsão deste tipo de produtos farmacêuticos, com o intuito de melhor apoiar na tomada de decisões redireccionando assim esforços para uma melhor gestão e controlo de *stocks*. Embora existisse a intenção original de exploração de um número mais alargado de métodos de previsão (e.g., Redes Neuronais e Máquinas de Vectores de Suporte), tal não foi possível devido a limitações temporais, deixando-se essa análise para trabalho futuro.

Palavras-chave: Gestão e controlo de *stocks*, Farmácia Comunitária*, Business Intelligence, Knowledge Discovery in Data, Decision Support Systems*, CRISP-DM.

Abstract

With increasing regulation in the pharmaceutical area, at national and community level, and due to the generalization of generic drugs and also increase of economic pressures that the Portuguese country is facing, the profit is shrinking in this area, jeopardizing the sustainability of some pharmacies. Thus, there is a current need to rethink and optimize the pharmaceutical business. One way is to use Business Intelligence technologies, which can provide new mechanisms in order to bring added value to businesses in general, and also for this type of business, which despite showing some characteristics of a very specific nature, also aimed profit and the satisfaction of final consumers. In particular, this study aims to use a Data Mining approach, by extracting useful and valid knowledge from raw pharmaceutial data, in order to support stock management and control of drugs, health products and other products and services sold in a pharmacy.

This dissertation focus on a case-study about a Portuguese pharmacy, by analyzing time series of a selection of three products, through sales between the period 2003 and 2012, with monthly and weekly frequency. From this research, results the comparison between two popular forecasting methods: exponential smoothing (including several of its variants) and the ARIMA methodology, also known as *Box-Jenkins* method. It is also intended to determine which method best fits the prediction of this kind of products, in order to better support decision making and to redirect efforts for better management and stock control. While we intended to also explore more forecasting methods (e.g. Neural Networks and Support Vector Machines), it was not possible to test such methods due to time restrictions. Nevertheless, we intend to experiment such methods in future work.

Keywords: Management and stock control, Pharmacy*, Business Intelligence, Knowledge Discovery in Data, Decision Support Systems*, CRISP-DM.

Índice

Agradecimentos ..... **i**

Resumo ..... **ii**

Abstract ..... **iii**

Índice de Figuras ..... **vi**

Índice de Tabelas ..... **vii**

Lista de Abreviaturas ..... **viii**

1. Introdução **1**

1.1 Enquadramento e Motivação 1

1.2 Objectivos 2

1.3 Organização 3

2. Quadro teórico de referência **4**

2.1 Dados, Informação, Conhecimento, Inteligência e Sabedoria 4

2.1.1 Conceitos e Definições 4

2.1.2 Valor e Custo da Informação 7

2.1.3 Importância da Informação na Organização 8

2.2 *Business Intelligence* 8

2.3 KDD 9

2.3.1 Conceitos e Definições 9

2.3.2 Processo de KDD 10

2.4 Data Mining 14

2.4.1 Conceitos e definições 14

2.4.2 Evolução do *Data Mining* 16

2.4.3 Métodos de *Data Mining* 18

2.4.4 Metodologias de *Data Mining* 20

3. Aplicação de Técnicas de *Data Mining* em Farmácias Comunitárias **24**

4. Metodologia **28**

4.1 Contextualização 28

4.2 Planeamento 30

4.3 Ferramentas Utilizadas 31

4.4 Técnicas de *Data Mining* 33

4.4.1 Método de Alisamento Exponencial 33
4.4.2 Método de *Box-Jenkins* - ARIMA 35
4.5 Medidas de Desempenho da Previsão 37
5. Gestão de Produtos Farmacêuticos via Técnicas de *Data Mining* 39
5.1 Compreensão do Negócio 39
5.2 Compreensão dos Dados 40
5.3 Preparação dos Dados 42
5.4 Modelação 44
5.4.1 Aplicação do Método de Alisamento Exponencial 44
5.4.2 Aplicação do Método Box-Jenkins 51
5.5 Avaliação 57
5.6 Implementação 59
6. Conclusões 60
6.1 Síntese 60
6.2 Discussão 62
6.3 Limitações e Trabalho Futuro 63
Bibliografia 64

Índice de Figuras

Figura 1 - Transformação de dados em informação, adaptado de (Davis, 1974) 5

Figura 2 – Relacionamento entre dados, informação e conhecimento (Boisot & Canals, 2004) . 6

Figura 3 - A visão convencional da hierarquia do conhecimento, adaptado de (Tuomi, 1999).... 7

Figura 4 – Fases do processo de KDD, adaptado de (Fayyad, et al., 1996)..... 11

Figura 5 – O processo de KDD (Adriaans & Zantinge, 1996)..... 12

Figura 6 – Data Mining como um passo na descoberta de conhecimento. 14

Figura 7 – Taxonomia do *Data Mining*, adaptado de (Maimon & Rokach, 2010)..... 19

Figura 8 – Fases da metodologia CRISP-DM (Chapman, et al., 2000)..... 21

Figura 9 - Fases da metodologia DMAIC..... 22

Figura 10 - Fases da metodologia SEMMA 22

Figura 11 – Utilização da ferramenta R, adaptado de (Rexer, et al., 2011) 31

Figura 12 – Metodologia de *Box-Jenkins* (Box, et al., 1994) 37

Figura 13 – Correlograma dos resíduos para o Produto A..... 45

Figura 14 – Histograma dos erros de previsão para o produto A..... 46

Figura 15 – Correlograma dos resíduos, Produto B 48

Figura 16 – Histograma dos erros de previsão para o produto B..... 48

Figura 17 – Correlograma dos resíduos para o produto C 50

Figura 18 – Histograma dos erros de previsão para o produto C..... 50

Figura 19 – Teste aumentado de *Dickey-Fuller* para o produto A..... 51

Figura 20 – Avaliação do modelo ARIMA, Produto A..... 53

Figura 21 - Teste aumentado de *Dickey-Fuller* para o produto B..... 54

Figura 22 - Avaliação do modelo ARIMA para o produto B..... 55

Figura 23 - Teste aumentado de *Dickey-Fuller* para o produto C..... 55

Figura 24 - Avaliação do modelo ARIMA para o produto C..... 56

Figura 25 - Comparação e avaliação através de métricas de desempenho 58

Figura 26 – Previsão das séries temporais..... 61

Índice de Tabelas

Tabela 1 - Tipos de dados analisados entre Junho de 2010 e Junho de 2011 18

Tabela 2 – Comparação entre as metodologias SEMMA e CRISP-DM 23

Tabela 3 - Relacionamento de regras e factores de suporte e confiança 25

Tabela 4 - Classificador induzido pelo algoritmo PRISM com selecção “Local de Compra” 27

Tabela 5 – Descrição das séries temporais 43

Tabela 6 – Resultados do método de Alisamento Exponencial para o produto A 45

Tabela 7 – Resultados do método de Alisamento Exponencial para o produto B 47

Tabela 8 – Resultados do método de Alisamento Exponencial para o produto C 49

Tabela 9 – Modelos ARIMA identificados para o produto A 52

Tabela 10 - Modelos ARIMA identificados para o produto B 54

Tabela 11 - Modelos ARIMA identificados para o produto C 56

Tabela 12 – Comparação dos métodos de previsão através de métricas de previsão 57

Lista de Abreviaturas

ADF - *Augmented Dickey-Fuller*

AIC - *Akaike Information Criterion*

ARIMA - *AutoRegressive Integrated Moving Average*

BI - *Business Intelligence*

BIC - *Bayesian Information Criterion*

CRISP-DM - *CRoss Industry Standard Process for Data Mining*

DM - *Data Mining*

DMAIC - *Define, Measure, Analyze, Improve, Control*

DW - *Data Warehouse*

FAC - Função de Autocorrelação.

FACP - Função de Autocorrelação Parcial

KDD - *Knowledge Discovery in Databases*

MAE - *Mean Absolute Error*

MAPE - *Mean Absolute Percentage Error*

MSE - *Mean Squared Error*

OLAP - *Online Analytical Processing*

PDF - *Portable Document Format*

SEMMA - *Sample, Explore, Modify, Model, Assessment*

SQL - *Structured Query Language*

1. Introdução

1.1 Enquadramento e Motivação

Cada vez mais o mercado empresarial é mais agressivo e vinga quem utiliza estratégias que permitem reunir não apenas o máximo de informação possível, mas também informação com maior qualidade, ou seja a informação certa no momento e no local certo.

Dados os avanços nas tecnologias de informação, existe a necessidade de transformar os dados, previamente armazenados, em informação e essa informação em conhecimento útil, que possa suportar a tomada de decisões. É nessa óptica que surge o *Business Intelligence*, que agrega um conjunto vasto de tecnologias. Em particular, as técnicas de *Data Mining* que pretendem extrair, a partir de dados em bruto, padrões e tendências que possam ser utilizadas para melhorar a tomada de decisões numa organização.

Pretende-se com esta dissertação efectuar um estudo de numa farmácia comunitária, na tentativa de identificar padrões relevantes para apoio à gestão, considerando que actualmente existem muitas debilidades ao nível da gestão e na aposta de determinadas marcas/produtos que poderão fomentar vendas e tornar mais eficiente a gestão de *stocks*. Para esse efeito, pretende-se extrair e analisar um conjunto de informação, desde a criação da farmácia, relativamente às compras, vendas, sazonalidade, tendências de mercado e outros factores de interesse, de forma a melhor orientar as compras e apostar em determinados produtos em detrimento de outros.

Pretende-se, igualmente, mostrar a aplicação de técnicas de *Data Mining* numa empresa desta natureza, assim como avaliar a viabilidade e dificuldades sentidas no âmbito deste estudo de caso. A presente dissertação irá ser elaborada mediante a aplicação das técnicas supra mencionadas e seguindo a metodologia CRISP-DM.

Toda esta problemática motivou o autor em diversos momentos, na medida em que o tema, na área de *Business Intelligence* e aplicação de técnicas de *Data Mining,* associadas à previsão de vendas, suscitaram vontade de aprender mais e de estudar problemas relacionados com o mesmo. O interesse não reside somente no estudo directamente relacionado com a área da saúde, bastante presente na vida do autor, mas que seja um meio para aplicar e ampliar este conhecimento a outras áreas.

1.2 Objectivos

Desde a criação de qualquer empresa que esta necessita obrigatoriamente de armazenar e gerir um conjunto de dados de elevada dimensão, por exemplo relativos a clientes, encomendas ou vendas. Esses dados muitas das vezes limitam-se a figurar no arquivo não sendo aproveitados nem filtrados para acrescentar valor, apoiar na tomada de decisões e consequentemente permitir uma gestão mais eficiente e com lucros mais elevados.

Assim, existe a necessidade de transformar esses dados em informação e essa informação em conhecimento útil, que possa apoiar na tomada de decisões. É nessa óptica que surge o *Business Intelligence*, que agrega um conjunto vasto de tecnologias. Em particular, as técnicas de *Data Mining* pretendem extrair, a partir de dados em bruto, padrões e tendências que possam ser utilizadas para melhorar a tomada de decisões numa organização.

Pretende-se com esta dissertação efectuar um estudo de caso numa farmácia comunitária, na tentativa de identificar padrões relevantes para apoio à gestão, considerando que actualmente existem muitas debilidades ao nível da gestão e na aposta de determinadas marcas/produtos que poderão fomentar vendas e tornar mais eficiente a gestão de *stocks*. Para esse efeito, pretende-se extrair e analisar um conjunto de informação desde a criação da farmácia relativamente às compras, vendas, sazonalidade, tendências de mercado e outros factores de interesse, de forma a melhor orientar as compras e apostar em determinados produtos em detrimento de outros.

Esta proposta assenta na necessidade, neste sector, de se explorar novos mercados e apostar em novas áreas, tais como a puericultura, cosmética, produtos ortopédicos, perfumaria/aromaterapia, veterinária, dietética, fitoterapia, dispositivos médicos e homeopatia. Aliada a toda esta panóplia de serviços, existe ainda a disponibilização de um conjunto de serviços, onde se destacam a vacinação e determinação de parâmetros bioquímicos. Importa, ainda, ressalvar, que actualmente uma farmácia pode disponibilizar um conjunto restrito de serviços e venda de medicamentos *online*.

Pretende-se, igualmente, mostrar a aplicação de técnicas de *Data Mining* numa empresa desta natureza, com fins tão específicos, assim como efectuar um levantamento dos pontos críticos para aplicação destas técnicas no âmbito deste estudo de caso. Para a elaboração da presente dissertação irá ser utilizada a metodologia CRISP-DM, dado ser mais completa e neutra em relação à ferramenta de DM a adoptar.

1.3 Organização

Esta dissertação encontra-se organizada tendo por início a presente introdução e desenrola-se por mais cinco capítulos.

No capítulo 2 será efectuado o levantamento do quadro teórico de referência na área de *Business Intelligence,* desde as suas origens até aos dias de hoje, bem como a aplicação de técnicas de *Data Mining* e as suas diversas vertentes e aplicações. Segue-se, no capítulo 3, o levantamento e a descrição da aplicação de *Data Mining* associado ao sector farmacêutico. Para o efeito, tentou-se identificar casos de estudos levados a cabo noutros países, tendo-se verificado ser uma área bastante diversificada e com aplicações bastante diferentes tendo em consideração o modo de actuação nesses países.

Após esta explanação, sentiu-se necessidade de criar um capítulo relativo à metodologia aplicada (capítulo 4), com o planeamento necessário de forma a levar a cabo as etapas seguintes, as ferramentas utilizadas, as técnicas e métricas que se pretenderiam utilizar para o desenrolar da aplicação do presente caso.

No capítulo 5 deu-se início à aplicação de toda a parte prática, tendo sido analisadas as séries temporais referentes a três produtos distintos, com dados relativos a periodicidades diferentes. A presente dissertação culmina com a síntese e a discussão dos resultados obtidos, assim como apresentadas as contribuições para trabalhos futuros (capítulo 6).

2. Quadro teórico de referência

No decorrer dos anos, tem-se vindo a recolher e armazenar diversos conjuntos de dados nas mais diversas áreas, sendo que com a entrada de tecnologias de informação e sistemas de informação informatizados, essa recolha tem vindo a assumir proporções de tal modo elevadas que se tornou necessário desenvolver novas teorias e ferramentas com o intuito de apoiar os cidadãos na extracção de informação útil (conhecimento). A informação recolhida é tão vasta e complexa que existe o risco de não se conseguir extrair todo o conhecimento presente, ou por falta de tempo, ou devido à complexidade de interligar os diversos conjuntos que isolados poderão tornar-se pouco úteis ou até mesmo inúteis.

Durante anos se tem utilizado o chavão de vivermos na era da informação, no entanto com a massificação dos suportes de armazenamento de dados e com os actuais sistemas de bases de dados e *Data Warehouses* é possível que se possa afirmar que actualmente se vive na era dos dados. Dados esses que é necessário analisar e modelar no sentido de obter a informação certa na hora certa. É neste sentido que surge a área da descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases - KDD*) e o *Data Mining*. Neste capítulo esclarecem-se conceitos básicos, como dados, informação, conhecimento, inteligência e sabedoria, o que é o *Business Intelligence* e a necessidade da Descoberta de Conhecimento em Bases de Dados e qual a sua relação com *Data Mining*. O *Data Mining* associado à farmácia e a necessidade de gestão de medicamentos numa farmácia comunitária.

2.1 Dados, Informação, Conhecimento, Inteligência e Sabedoria

2.1.1 Conceitos e Definições

Existe actualmente, um conjunto de concepções de grande diversidade do conceito Informação. Tal facto comprova-se aquando da análise desta definição em diversos dicionários, sendo provavelmente através destes a grande causa desta multiplicidade, a qual é apenas encontrada e justificada através dos seus fins (Davis, 1974).

Entre várias definições de “Informação”, poderemos dizer que esta não é mais do que um conjunto de dados, ligados entre si, devidamente organizados e interpretados, sendo que ao mesmo tempo existe uma preocupação de os filtrar e analisar. É no fundo, uma sumarização, possivelmente formatada, para que seja apresentada com o objectivo do utilizador dai retirar proveito útil (Gordon & Gordon, 1999).

“Informação” é também entendida como qualquer tipo de relacionamento dos dados, por observação e análise para que esta se traduza em comunicação útil que se possa trocar entre os diversos utilizadores.

Todavia, não se pode falar de “Informação” sem se distinguir primeiro a sua diferença com a dos vocábulos “dados” e “conhecimento”. Deste modo, “dados” são factos isolados, sem qualquer tipo de ligação entre si, representações não estruturadas que poderão ou não ter utilidade numa determinada situação. Resumindo, “dados”, são conceitos, números, símbolos ou qualquer tipo de representação gráfica, sem qualquer tipo de utilidade ou benefícios.

O conceito “conhecimento” entende-se por grandes estruturas duradouras de factos com significado, com o intuito de resolver problemas, inovar e aprender, tendo como base experiências prévias. Todavia, esta capacidade só é possível através de um conjunto de directrizes, normas, regras e procedimentos usados para seleccionar e analisar, manipulando ao mesmo tempo os dados para que possam ser úteis num futuro próximo.

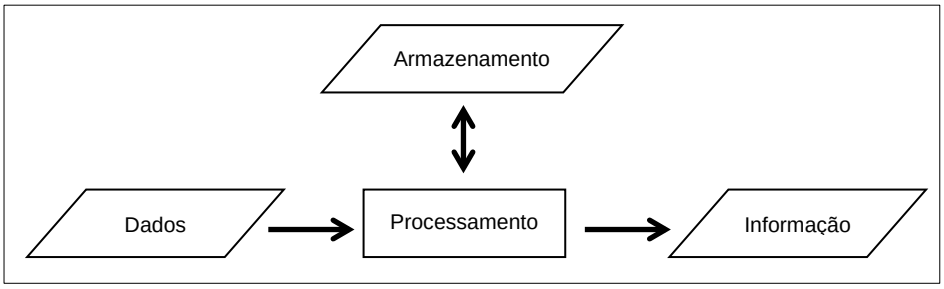


Figura 1 - Transformação de dados em informação, adaptado de (Davis, 1974)

Com significados diferentes, “dados” e “informação” estão directamente relacionados (Figura 1). Esta ligação verifica-se pela necessidade constante, nomeadamente nas organizações, em captar, identificar e analisar os dados, para que se possa obter algo de útil. No entanto, este conjunto de dados necessita de ser arquivado para possíveis tomadas de decisão.

Pode-se constatar, no entanto, que existe uma utilidade diferente no conceito de “informação”, na medida em que, o que é para um utilizador poderá ser diferente para outro, tal como um produto acabado de uma secção de fabrico poderá ser matéria-prima para a secção seguinte (Davis, 1974).

(Boisot & Canals, 2004) defendem que a informação é uma extracção de dados que, modificando as distribuições de probabilidade relevantes, tem capacidade para realizar um trabalho útil na base de um agente do conhecimento.

Na Figura 2 é possível verificar que os agentes operam dois tipos de filtros ao converterem estímulos recebidos em informação. Apenas os estímulos que passam pelos filtros perceptivos são registados como dados. Por sua vez, os filtros conceptuais extraem informação com base nos dados registados. Ambos os filtros são "sintonizado" pelas expectativas dos agentes cognitivos e afectivos, moldados de acordo com os conhecimentos ao longo da vida, no sentido de actuar selectivamente tanto nos estímulos como nos dados (Boisot & Canals, 2004).

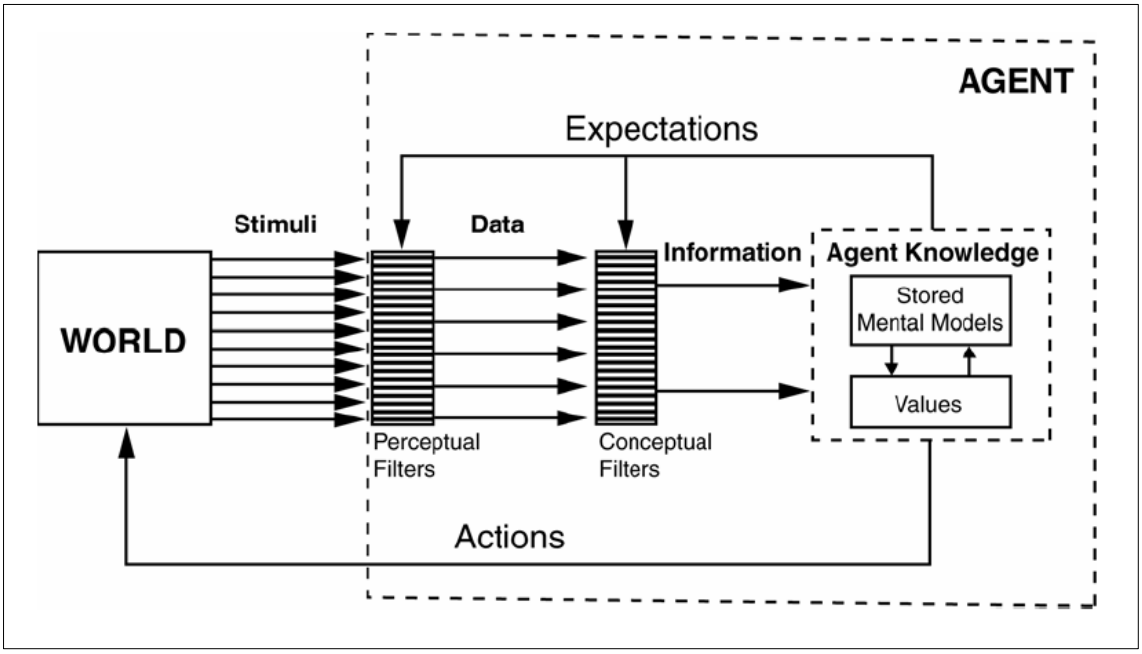


Figura 2 – Relacionamento entre dados, informação e conhecimento (Boisot & Canals, 2004)

Em aditamento aos conceitos acima descritos, Tuomi acrescenta mais dois (2) tipos de conhecimento, nomeadamente a inteligência e a sabedoria. Como é possível verificar através da Figura 3, os dados são descritos como simples factos isolados, que em determinado contexto e combinados com uma estrutura úncia, estes dão lugar a informação. À informação quando quando lhe é dado determinado significado, através da sua interpretação, esta transforma-se em conhecimento. Nesta fase os factos já existem com uma estrutura mental de tal forma complexa que a consciência humana pode processar no sentido de prever consequências futuras ou fazer inferências. A inteligência surge assim, na fase em que a mente humana usa este conhecimento para escolher entre alternativas. Por fim, quando os valores e os comportamentos culturais são como directrizes no comportamento humano, pode-se dizer que este comportamento baseia-se na sabedoria (Tuomi, 1999).

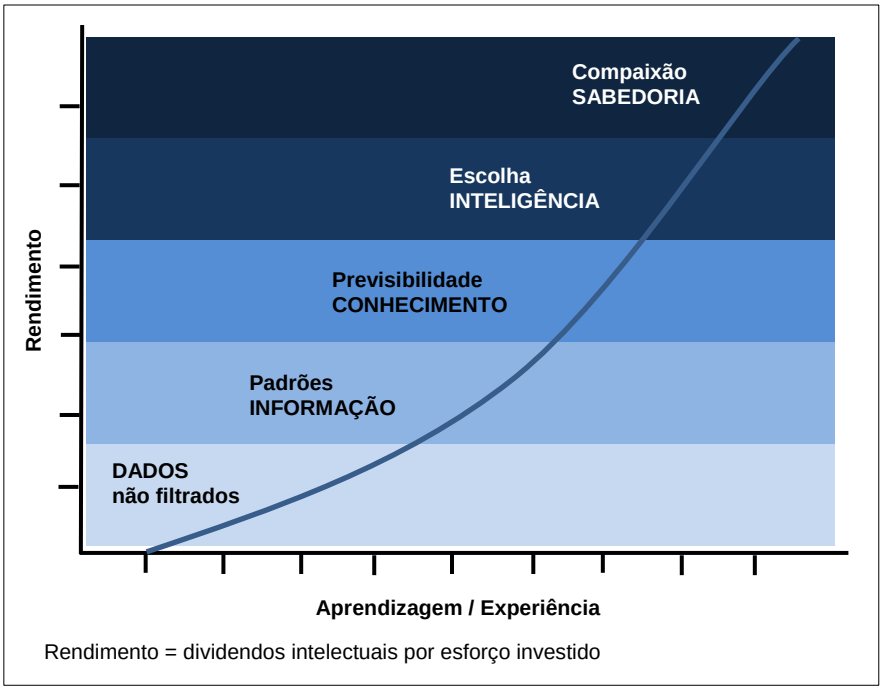


Figura 3 - A visão convencional da hierarquia do conhecimento, adaptado de (Tuomi, 1999)

2.1.2 Valor e Custo da Informação

Dos vários recursos da organização, tais como financeiros, humanos ou logísticos, a informação é provavelmente o mais valioso de todos, porque interliga, e descreve, os recursos físicos e o meio onde se encontram. Podendo assim, maximizar os recursos físicos, da melhor forma a obter a informação pretendida.

O valor da informação só é importante quando considerados os objectivos da organização, uma vez que esta é determinada pelo utilizador, nas suas acções e decisões, dependendo do contexto que é utilizada na tomada das decisões finais. A informação só é valorizada com base nas decisões eficazes, não tendo qualquer valor se esta não tiver qualquer utilidade para a tomada de decisões, no presente ou no futuro.

Segundo (Ein-Dor, 1985), o valor da informação é determinado pelo uso que daí provém. O uso da informação, como parte das operações diárias nas organizações, é o factor mais importante, na medida em que este factor constitui o valor real.

O valor da informação não deriva da frequência a que se tem acesso a este, mas sim do valor psicológico atribuído pelo utilizador ao ter acesso a dados disponíveis. Poderá, eventualmente, acontecer a modelação sofrida pelos Sistemas de Informação que poderão disponibilizar, segundo os objectivos pretendidos, modelos pré-definidos, de forma a facilitar o uso e a transmissão da informação pretendida (Davis & Olson, 1986).

A informação nas organizações está intimamente relacionada com o seu custo, isto acontece pelo facto da informação assumir um custo inferior ou superior, consoante as tecnologias e metodologias aplicadas, bem como a dimensão do factor humano envolvido.

Continuando com (Davis & Olson, 1986), estes afirmam que através de um conjunto de decisões possíveis, tendo em conta o valor da informação perante o custo, uma será seleccionada com base nesse conjunto, sendo que o retorno obtido na decisão tomada consistirá na diferença de decisões anteriores com o feedback da nova decisão. Isto é, se porventura a nova informação não provocar uma decisão diferente, o valor será nulo, apesar da sua não utilização útil, a informação continuará a ter um custo.

2.1.3 Importância da Informação na Organização

A importância da informação para as organizações é hoje, universalmente aceite, em alguns casos, entendida como o recurso mais importante, ou pelo menos, um dos recursos cuja gestão e aproveitamento detêm maior influência no sucesso empresarial.

Assim, a importância da informação nas organizações assume três vertentes, recurso, activo e mercadoria, onde (Gordon & Gordon, 1999):

- a informação é entendida como recurso, quando serve como forma de recolha de dados, respectivo tratamento, de modo a dar satisfação às exigências pretendidas;
- a informação como activo, verifica-se quando a organização consegue rentabilizar os recursos existentes de modo tornar-se mais competitiva;
- a informação como mercadoria, encontra-se quando as organizações podem vendê-la, sob forma de jornais, revistas e outras publicações.

2.2 *Business Intelligence*

Business Intelligence é um termo abrangente que agrega arquitecturas, ferramentas, bases de dados, aplicações e metodologias (Raisinghani, 2004). Um dos maiores objectivos de *Business Intelligence* é o facto de disponibilizar acesso e manipulação de forma interactiva (por vezes em tempo real) a dados, e facultar a gestores e analistas o acesso a análises mais controladas e detalhadas. O processo de *Business Intelligence,* em traços gerais, baseia-se na transformação de dados em informação, posteriormente em decisões e por fim em acções (Turban, et al., 2007).

Cada vez mais os gestores de topo têm consciência de que a informação é essencial para vingar no mundo empresarial, introduzindo novos produtos, ganhando cota de mercado,

angariando novos clientes, colmatando falhas de mercado, entre outras. É nesta sequência que se torna imprescindível a obtenção de informação correcta, no momento certo, permitindo analisar um conjunto de dados resultantes de meses, anos ou décadas de armazenamento e extrair conhecimento estratificado que permita alcançar o sucesso de uma organização.

O termo *Business Intelligence* foi originalmente divulgado através do *Gartner Group* em meados de 1990. Desde então, a utilização dos sistemas de BI massificou-se, sendo que actualmente é comum que toda a informação que os gestores de topo necessitam seja facultada a partir de um sistema de BI. De referir ainda que o BI agrega diversas tecnologias, em particular a Descoberta de Conhecimento em Bases de Dados/*Data Mining*.

2.3 KDD

2.3.1 Conceitos e Definições

Numa abordagem generalista, a Descoberta de Conhecimento em Bases de Dados (KDD[1]) tem como objectivo primordial o desenvolvimento de métodos e técnicas com o intuito de relacionar conjuntos de dados e estabelecer significados entre eles, com vista à extração de conhecimento útil. O processo consiste no mapeamento de grandes volumes de dados, armazenados em bases de dados e/ou *Data Warehouses*, transformando-os em formas mais compactas, mais abstractas, tornando-se assim mais úteis e de compreensão mais simples e rápida (Fayyad, et al., 1996).

A pertinência da utilização de KDD torna-se mais evidente quando comparado com métodos mais tradicionais, em que a transformação de dados para conhecimento incidem numa análise e interpretação efectuada de forma manual. Esta transformação, dependendo das diversas áreas, poderá ser algo desde um conjunto de relatórios detalhados tornando-se assim na base para futuras tomadas de decisão, planeamento e gestão das diversas actividades inerentes à organização, até à previsão de determinados acontecimentos, prevenção de crimes, acidentes, anomalias, entre muitas outras.

Para as aplicações acima identificadas, a análise manual de um conjunto de dados tona-se demasiado lenta, dispendiosa e altamente subjectiva. Como se pode verificar em (Fayyad, et al., 1996), os autores acreditam mesmo que esta tarefa não é, certamente, uma tarefa para seres humanos, sendo que a componente de análise tem de ser automatizada, pelo menos de forma parcial.

[1] Do inglês, *Knowledge Discovery from Databases*.

Perante toda esta panóplia de actividades disponíveis, o KDD é utilizado num vasto conjunto de áreas, nomeadamente no marketing, finanças, banca, ciência, seguradoras, prevenção, medicina, área farmacêutica, entre muitas outras áreas.

Em suma, KDD é um processo, não trivial, de identificação de padrões a partir de dados, sendo estes válidos, novos (preferencialmente para o utilizador), potencialmente úteis (que seja possível obter mais valias para o utilizador ou tarefa) e fundamentalmente que sejam compreendidos.

2.3.2 Processo de KDD

O princípio de KDD é um processo de identificação de padrões válidos perceptíveis a partir dos dados, com o propósito de criar conhecimento válido, capazes de serem compreendidos e interpretados pelos humanos. O processo de KDD depende de um conjunto de ferramentas e técnicas de análise de dados, que envolve diversas etapas, tendo como foco o processo global de descoberta de conhecimento a partir de dados, desde o processo em que os dados são armazenados e acedidos até ao modo em como os resultados são interpretados e apresentados.

O processo de KDD pode, igualmente, ser visto como uma actividade multidisciplinar que engloba técnicas bastante mais abrangentes do que apenas as técnicas disponíveis em área específicas como a aprendizagem máquina (*machine learning*).

Para (Fayyad, et al., 1996) o processo de KDD ocorre de forma interactiva e iterativa e desenrola-se ao longo de nove (9) etapas. A etapa de *Data Mining* é considerada como uma das mais relevantes, formando assim o núcleo do processo, e que muitas vezes se confunde com o processo em si. As nove (9) etapas são as seguintes:

1. Definir o domínio de aplicação e pré-conhecimento relevante, bem como o objectivo do ponto de vista do cliente;
2. Criar um conjunto de dados alvo;
3. Limpeza e pré-processamento de dados;
4. Redução e projecção de dados;
5. Escolha do método de *Data Mining* adequado;
6. Escolha da técnica de *Data Mining* mais apropriada;
7. Processo de *Data Mining*, ou seja procura de padrões relevantes;
8. Interpretação dos padrões identificados;
9. Utilização e/ou documentação do conhecimento obtido.

Para uma melhor compreensão de todo o processo de KDD e a forma de como se interligam as diversas fases encontra-se sistematizado na Figura 4.

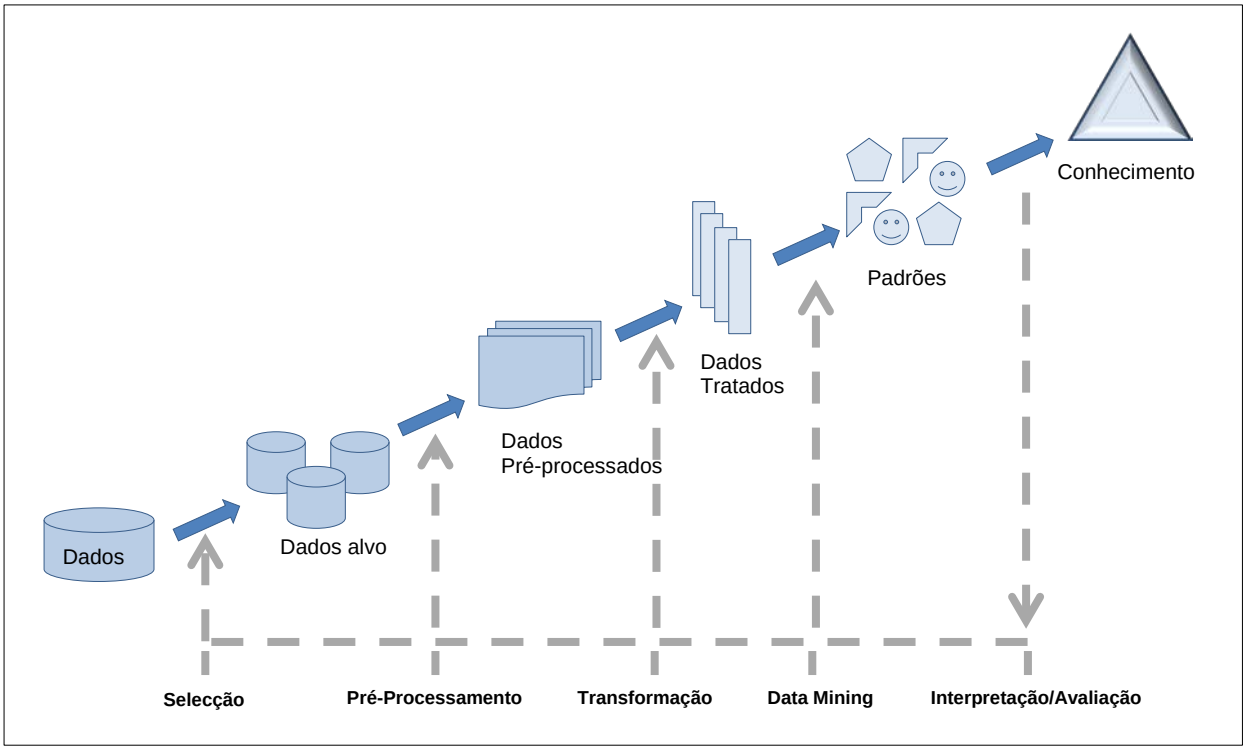


Figura 4 – Fases do processo de KDD, adaptado de (Fayyad, et al., 1996)

Por sua vez, (Adriaans & Zantinge, 1996), defende que a extracção de conhecimento de bases de dados é efectuada através de seis (6) etapas, nomeadamente:

1. **Selecção dos dados** – etapa onde é efectuada a recolha e a selecção frequentemente em bases de dados ou *Data Warehouses*;
2. **Limpeza** - os dados são analisados minuciosamente e são aplicadas técnicas específicas baseadas em critérios previamente definidos, de forma a remover dados duplicados, dados incompletos, incorrectos, tratamento de dados irrelevantes, entre outras;
3. **Enriquecimento** – este estádio prende-se com a possibilidade de, em determinadas situações, ser possível ter acesso a bases de dados suplementares e como tal poder complementar os dados anteriores com um conjunto de informação adicional. Este processo pode ser mais complexo, dada a necessidade de haver a inscrição de dados relacionados em diferentes bases de dados dependendo da finalidade de cada uma (Ex.: Agregar elementos referente aos cidadãos constantes na base de dados das páginas brancas, base de dados de determinado ISP, base de dados de determinada empresa de marketing);

4. **Codificação** – prende-se com a obtenção de um conjunto de informação, mediante a criação de filtros, por exemplo, através de instruções de SQL (Structured Query Language)[2] e obter informação pertinente, dependendo da natureza da informação que se pretende obter e qual a análise que se pretende efectuar;

5. ***Data Mining*** – nesta etapa são abordadas uma das áreas mais importantes no processo de descoberta de conhecimento, como a aprendizagem de máquina (*machine learning*) e algoritmos de reconhecimento de padrões. Os autores defendem que existe mais conhecimento escondido nos dados do que aquilo que aparentam numa análise superficial;

6. **Documentação/divulgação** – este estádio do processo combina duas funções distintas: a análise dos resultados dos algoritmos de reconhecimento de padrões e a aplicação dos resultados obtidos em novos dados. A documentação/divulgação pode ser efectuada através de diversas formas com maior ou menor grau de interactividade, isto é utilizando ferramentas que possibilitem a escolha de determinados dados através de gráficos, diagramas ou outros, seja a 2D ou 3D.

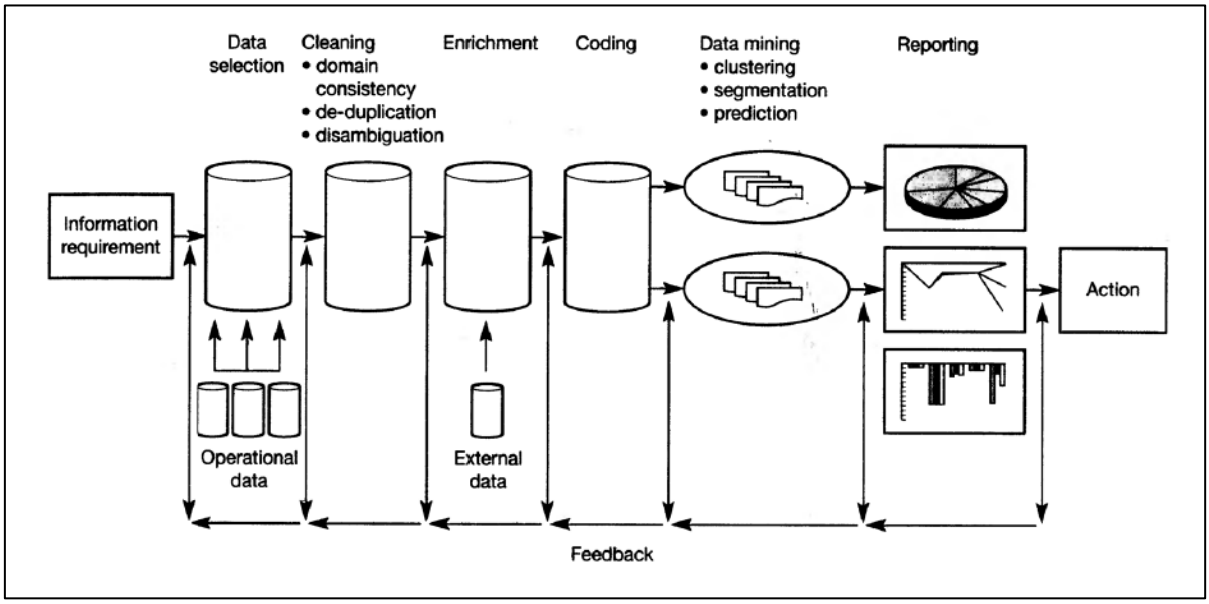


Figura 5 – O processo de KDD (Adriaans & Zantinge, 1996)

Ambas as vertentes de KDD têm pontos em comum como se pode facilmente verificar nas seis (6) etapas da Figura 5 que se encontram-se englobadas nas nove (9) etapas identificas por (Fayyad, et al., 1996). Salienta-se o facto dos autores da primeira vertente subdividirem a etapa de *Data Mining* em três, nomeadamente: a escolha do método, a escolha do algoritmo e o processo de *Data Mining*. Aliás, quando analisada a Figura 4 e a Figura 5 verifica-se que estes autores resumem o processo nas seguintes etapas:

[2] Poderá saber mais sobre este assunto em: Groff, James, Weinberg, Paul e Oppel, Andrew J., *SQL The Complete Reference*, 3rd Edition, McGraw-Hill Osborne Media, 2009.

1. **Selecção** – etapa destinada à identificação e selecção dos dados considerados relevantes para determinado estudo;
2. **Pré-Processamento** – aqui é efectuado um tratamento dos dados previamente seleccionados. É efectuado o tratamento ao nível de erros ocorridos durante o período de inserção, incoerências, valores duplicados, inclusão ou exclusão de dados omissos, distribuição de dados não uniformes, entre outros. Este tratamento deverá ser minucioso, cuidado e consciente pois poderá levar a resultados diferentes dos pretendidos[3];
3. **Transformação** – nesta etapa ocorrem as adaptações consideradas necessárias para que se possam aplicar as técnicas de *Data Mining*;
4. ***Data Mining*** – aplicação das diversas técnicas de análise tendo em vista a identificação de padrões relevantes;
5. **Interpretação/Avaliação** – esta fase encerra o processo de KDD, tendo em conta a análise e interpretação dos resultados alcançados, englobando o estudo e a avaliação dos resultados alcançados na fase anterior. Esta poderá, de forma mais abrangente, ser o fim do estudo em causa, alcançando o conhecimento e aplicando para determinados fins ou dando origem a novos dados e consequente novo estudos sobre a mesma ou nova matéria.

Em suma, o KDD é um processo dinâmico, interactivo e iterativo envolvendo inúmeras etapas e com muitas decisões a tomar por parte do utilizador.

Os mesmos autores acima referidos relativamente às duas (2) abordagens de KDD afirmam que o *Data Mining* assume uma grande importância em todo o processo e é de facto a fase da grande descoberta. A descoberta que se inicia em dados devidamente inseridos e armazenados em bases de dados e *Data Warehouses* e que através de um conjunto de técnicas minuciosas e complexas levam ao conhecimento e posteriormente a acções diversas.

No seguimento das abordagens acima mencionadas torna-se relevante mencionar a metodologia CRISP-DM. Esta metodologia assenta num ciclo interactivo e flexível, baseado em seis (6) fases: Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelação, Avaliação e Implementação. As fases do CRISP-DM apresentam semelhanças observáveis face às etapas de KDD, na medida em que é uma metodologia para desenvolver projectos de descoberta de conhecimento em bases de dados. Considerando que esta é aplicada a projectos de *Data Mining*, verifica-se, igualmente, um conjunto de pressupostos similares aos do processo de KDD como um todo. Esta metodologia será objecto de análise mais detalhada na Subsecção 2.4.4, no entanto considera-se oportuna a presente referência face às semelhanças nas etapas das vertentes em causa.

[3] Sobre esta matéria sugere-se a leitura de: Rahm, Erhard e Hai Do, Hong, *Data Cleaning: Problems and Current Approaches*, IEEE Techn. Bulletin on Data Engineering, 2000.

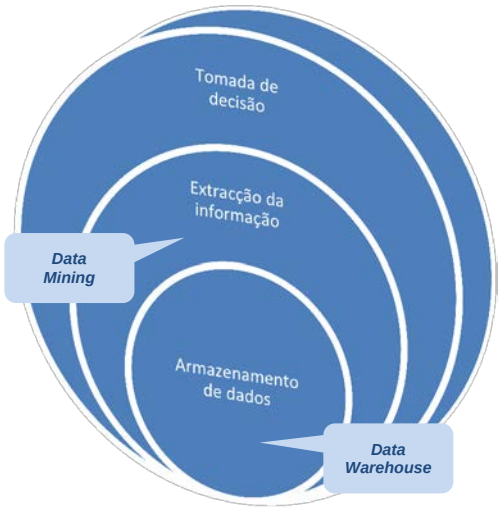


Figura 6 – Data Mining como um passo na descoberta de conhecimento.

2.4 Data Mining

2.4.1 Conceitos e definições

Conforme já mencionado no início deste capítulo, no decorrer dos anos e com a grande evolução das novas tecnologias temos vindo a recolher grandes volumes de dados. Pode-se enumerar alguns exemplos presentes no mundo dos negócios que só por si geram volumes de dados gigantescos, nomeadamente: transacções comerciais, registos contabilísticos, descrições de produtos, promoções de vendas, perfis e desempenho de empresas, *feedback* de clientes, entre outros. Deste explosivo crescimento surge a necessidade de criar ferramentas poderosas e versáteis, com a finalidade de descobrir informação útil, baseando-se nestes grandes volumes de dados e transformando-os em conhecimento organizado.

Muitos autores, investigadores e profissionais da área consideram o *Data Mining* como sendo sinónimo do processo do termo KDD, enquanto outros consideram-no como sendo apenas um passo essencial no processo de descoberta de conhecimento, conforme pode ser visível na Figura 4. Nesta dissertação, será em geral adotado o uso mais global para o termo *Data Mining*, ou seja, como sinónimo do KDD.

Segundo (Turban, et al., 2007), *Data Mining* é o termo utilizado para descrever a descoberta de conhecimento em bases de dados. Este não é mais do que um processo que utiliza um conjunto de técnicas estatísticas, matemáticas, inteligência artificial e aprendizagem máquina (*machine learning*) de forma a extrair e identificar informação útil, em grandes bases de dados, transformando-a posteriormente em conhecimento. O mesmo autor define *Data Mining* como sendo um processo de identificação matemática de padrões, normalmente através de grandes quantidades de dados.

Para além de identificação de padrões, o *Data Mining* pretende, também, extrair e identificar tendências de comportamentos. Essa extracção de dados é realizada, normalmente, num *Data Warehouse* (Brito, et al., 2006).

O Grupo Gartner define *Data Mining* como sendo um processo de descoberta de correlações significativas, padrões e tendências através de grandes quantidades de dados, armazenados em repositórios, utilizando tecnologias de reconhecimento de padrões, bem como técnicas estatísticas e matemáticas.

(Han, et al., 2011) preferem manter uma visão mais ampla, e definem *Data Mining* como sendo um processo de descoberta de padrões e conhecimento relevantes com base em grandes quantidades de dados. As fontes destes dados incluem bases de dados, *Data Warehouses*, Internet e outros repositórios de informação ou dados que são transmitidos para o sistema de forma dinâmica.

Este novo processo de análise de dados oferece às organizações uma melhoria no processo de decisão, cada vez mais indispensável, de forma a ser possível explorar novas oportunidades, novos mercados, novos produtos, transformando os dados recolhidos numa arma estratégica. Neste sentido, tem-se vindo a verificar a aplicação do *Data Mining* num conjunto de áreas cada vez mais abrangentes, destacando-se assim alguns exemplos da sua aplicação com maior pormenor (Turban, et al., 2007):

- **Marketing** através da gestão de carteira de clientes, previsão/estimativa da compra de determinados produtos e segmentação demográfica de clientes;
- **Sector bancário** através da análise de modelos de concessão de crédito habitação ou crédito pessoal;
- **Retalho e vendas** com a previsão de vendas, maior precisão na determinação de níveis de inventário, melhor distribuição de *stocks* entre lojas;
- **Fabrico e produção** com aplicação na previsão da avaria de máquinas e possível identificação de factores relevantes que impliquem melhor controlo e optimização;
- **Bolsa de valores** através da análise da flutuação de mercado;
- **Medicina** com particular melhoria na identificação de terapias de sucesso para diferentes tipos de tratamentos, detecção de padrões de sintomas de patologias;
- **Farmácias** com a análise de vendas, dispensa de produtos específicos fora de época, possível aposta em determinados produtos e/ou marcas;
- **Seguradoras** através da concessão baseada na análise de possíveis pacientes com doenças diversas;
- **Informática** com a previsão de avarias de *hardware* e possível violação da rede informática;
- **Governo e defesa nacional** mediante análise dos custos de equipamento militar, testes de estratégias militares e previsão do consumo de recursos;

- **Companhias aéreas** através da verificação dos destinos preferenciais dos clientes, análise de rotas com e sem escala de viagem, inclusão e exclusão de rotas;
- **Rádio e Televisão** com especial foco na previsão de programas que melhor possam ser rentáveis em horário nobre e maximizar os lucros através da introdução apropriada dos diferentes tipos de publicidade.

2.4.2 Evolução do *Data Mining*

O conceito de *Data Mining* tem vindo a evoluir ao longo dos vários anos. Em meados da década de 70 surgiram os primeiros sistemas de base de dados que rapidamente evoluíram para sistemas de base de dados relacionais.

No entanto, a evolução de mais uma década deu lugar ao aparecimento de sistemas de Gestão de Bases de Dados, *Data Warehousing* e *Data Mining* utilizando análises de dados avançadas e bases de dados baseadas em ambiente *web*. Foi a meio do ano de 1980 que começaram as surgir as bases de dados cada vez mais complexas. Estes sistemas começaram a incorporar modelos de dados mais recentes e poderosos tal como, orientadas a objectos, objecto-relacionais e modelos dedutivos.

Posteriormente surgiram novas arquitecturas de repositórios de dados com múltiplas fontes de dados com características heterogéneas organizadas através de um único sistema de forma a facilitar a tomada de decisão. Esta nova tecnologia inclui a limpeza e integração de dados e sistemas OLAP (*online analytical processing*) que não são mais do que a capacidade para manipular e analisar um grande volume de dados sob múltiplas perspectivas. A partir desta fase começaram-se a armazenar grandes quantidades de dados em grandes bases de dados e *Data Warehouses*, sendo que a partir de 1990 com o aparecimento da Internet e as bases de dados, alteraram-se os paradigmas na indústria da informação.

Com a massificação do armazenamento de dados rapidamente os tempos que se seguiram ficaram marcados como a década da informação. Actualmente e com a explosão das unidades de armazenamento cada vez maiores e mais eficientes há condições para acreditar que nos encontramos na era dos dados, dados esses que carecem de ferramentas poderosas e versáteis para que possam extrair o máximo de informação possível e a transformem em conhecimento estruturado (Han, et al., 2011).

Actualmente, existem novas abordagens com novos desafios como o objectivo de analisar e descobrir padrões num conjunto bastante diversificado de dados, destacam-se apenas alguns exemplos (Han, et al., 2011):

- *Text Mining* – Esta é uma área de actuação bastante interdisciplinar e é definido como um processo de descoberta de informação de qualidade a partir de texto. Assim, é

possível manipular mais facilmente informações não estruturadas como notícias, páginas da internet, *blogs* e outros tipos de documentos. Uma das partes mais importantes do processo de *Text Mining* é a preparação textual, cujo objectivo é armazenar um texto não estruturado numa base de dados com uma estrutura definida.

- *Web Mining* – Aqui verifica-se nada mais do que a aplicação de técnicas de descoberta de padrões, estruturas e conhecimento a partir da Internet. É possível subdividir este em três (3) categorias: Análise de Conteúdos da Internet (texto, multimédia, páginas e/ou ligações entre páginas); Análise da Estrutura da Internet (analisa as ligações e a estrutura de um site/portal completo) e a Análise do Histórico de navegação (analisa o histórico dos utilizadores como o intuito de descobrir padrões de navegação que possam ajudar a melhorar a navegabilidade das páginas de Internet. Deste modo, é possível prever o que os utilizadores preferem e melhorar a eficácia e eficiência de motores de pesquisa e/ou melhoria da aplicação de publicidade direccionada de acordo com um conjunto de preferências específicas).

- *Ubiquitous Data Mining* – Com a introdução dos computadores portáteis, PDA (*personal digital assistant*), telefones inteligentes e *tablets*, tem sido cada vez mais fácil o acesso ubíquo a grandes quantidades de dados, independentemente do lugar físico. Este é o processo de extracção utilizando um conjunto vasto de técnicas de forma a analisar dados com emissão contínua, nomeadamente com recurso a dispositivos móveis. Um dos objectivos prende-se com o facto dos dados estarem a ser transmitidos para um dispositivo móvel e o paradigma da necessidade de analisar dados em qualquer lugar e em qualquer altura.

- *Multimedia Data Mining* – Prende-se com o processo de descoberta de padrões em bases de dados com elementos multimédia, nomeadamente: imagens, vídeo, áudio, hiperligações e marcações de texto.

- *Spatial Data Mining* – Este refere-se à descoberta de padrões e conhecimento em dados espaciais. Para este efeito, têm vindo a ser criados grandes *Data Warehouses* com cubos de dados espaciais com dimensões e medidas espaciais e suporte OLAP para análises deste género.

- *Spatiotemporal Data Mining* – Este envolve dados relacionados com o tempo e o espaço, nomeadamente através da descoberta de padrões na evolução histórica de cidades e terrenos, mediante a descoberta de padrões de condições meteorológicas, previsão de terramotos, furacões e a determinação de tendências globais sobre o aquecimento global.

Tabela 1 - Tipos de dados analisados entre Junho de 2010 e Junho de 2011[4]

Dados em tabelas (n.º de colunas fixo) (143)	69.4%
Séries temporais (86)	41.7%
Conjuntos de itens / transações (67)	32.5%
Texto (formato livre) (53)	25.7%
Dados anónimos (45)	21.8%
Dados de localização/geográficos/móveis (40)	19.4%
Outros (29)	14.1%
Redes sociais (26)	12.6%
E-mail (22)	10.7%
Conteúdos da Internet (21)	10.2%
clickstream web (18)	8.7%
Imagens / video (14)	6.8%
Dados XML (10)	4.9%
Música / Áudio (7)	3.4%

Todos os dias, grandes volumes de dados, com as mais vaiadas características, são gerados a partir de todas as áreas, como no mundo empresarial, educação, comunidade científica, a Internet e muitas outras áreas. Como resultado surgem enormes desafios para a área de *Data Mining* e KDD. Através da Tabela 1 é possível ver um inquérito *online* sobre um conjunto de dados analisados entre Junho 2010 e Junho 2011, sendo de destacar as séries temporais (em segundo lugar).

2.4.3 Métodos de *Data Mining*

Existem vários tipos de métodos de *Data Mining* usados para diferentes propósitos e objectivos, através da Figura 7 é possível compreender a variedade de alguns dos seus métodos e a sua relação. Um dos objectivos, a *verificação* testa a hipótese do utilizador enquanto a *descoberta* procura novos padrões. O objectivo de *descoberta* por sua vez subdivide-se em dois (2), *previsão* e *descrição*. A *previsão* assume uma procura constante de padrões que permitam prever situações/comportamentos futuros e a *descrição* assenta numa procura de padrões que apresentem o conhecimento de forma compreensível.

Para o objectivo de *descrição* há vários métodos, nomeadamente: segmentação, sumarização, dependência, detecção de desvios, visualização e síntese linguística. Por outro lado, no objectivo de *previsão*, a problemática assenta em dois pontos essenciais: a *regressão* onde se pretende encontrar uma função desconhecida cuja variável dependente tem um domínio de

[4] Inquérito *online* realizado em Junho de 2011 (Piatetsky-Shapiro, 2011).

valores reais e a *classificação* onde se pretende encontrar uma função que faça o mapeamento dos dados em classes pré-definidas (Maimon & Rokach, 2010) e (Fayyad, et al., 1996).

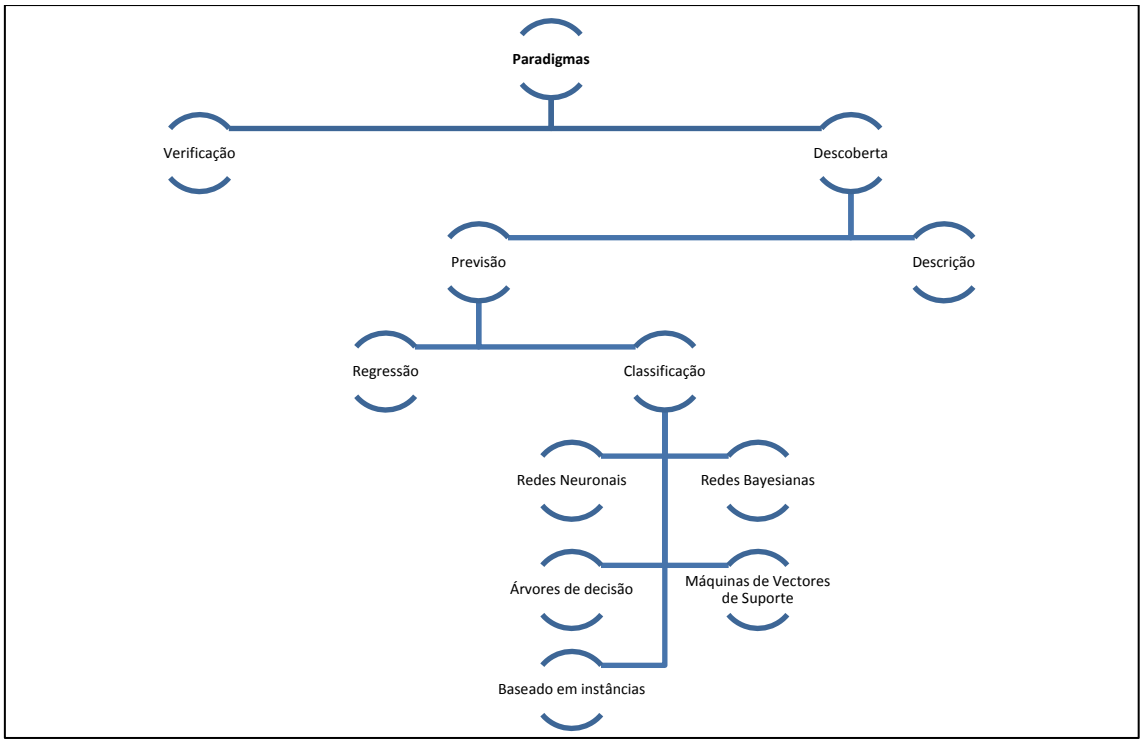


Figura 7 – Taxonomia do *Data Mining*, adaptado de (Maimon & Rokach, 2010)

Os algoritmos de *Data Mining* podem dividir-se em quatro (4) grandes categorias: classificação, segmentação, associação e descoberta de sequências. No entanto, poderão existir outras ferramentas de análise de dados, como a visualização, regressão e análise de séries temporais (Turban, et al., 2007):

Classificação – Esta é considerada como uma das actividades de *Data Mining* mais utilizadas e consiste na análise do histórico de dados, que se encontram em bases de dados, e gerar de forma automática um modelo que possa prever comportamentos futuros. Este modelo induzido consiste na generalização dos dados de "treino" que irão apoiar a distinguir classes redefinidas. O conceito assenta na esperança de que o modelo possa ser utilizado para prever classes de registos não classificados. Algumas das ferramentas mais utilizadas são as redes neuronais, árvores de decisão e outras regras sem uma estrutura definida. As redes neuronais envolvem o desenvolvimento de estruturas matemáticas com a capacidade de aprender. Por sua vez, as árvores de decisão classificam os dados num número finito de classes com base em valores das variáveis de entrada.

Segmentação – Este processo consiste divisão em segmentos que partilhem características similares. Contudo, ao contrário da classificação, na segmentação, os segmentos são

desconhecidos aquando da aplicação do algoritmo. Após a identificação de um conjunto de segmentos significativo ser identificado, estes podem ser utilizados para classificar novos dados, sendo que o objectivo passa pela criação de grupos com membros que partilhem características o mais semelhante possíveis entre si, e que os membros desses mesmos grupos apresentem características pouco significativas com membros de outros grupos.

Associação – As associações estabelecem relacionamentos entre registos que se encontram no mesmo registo.

Descoberta de sequências – Este consiste na identificação de associações ao longo do tempo. Quando uma determinada informação é facultada, pode realizar-se uma análise temporal, no sentido de verificar determinados comportamentos ao longo do tempo. Este processo fornece um conjunto de informação considerável e que pode ser utilizado, a título de exemplo, para o aumento de vendas ou detecção de fraudes.

Visualização – As percepções obtidas através deste processo não podem ser subestimadas. Dado o elevado volume de dados nas bases de dados a considerar, o processo de visualização apresenta, na generalidade, um esforço difícil.

Regressão – É uma técnica estatística utilizada para mapear dados tendo como fim a previsão. Para o efeito são utilizadas técnicas de regressão linear e de regressão não linear.

Forecasting – Trata-se de um caso particular da regressão, onde se tenta prever valores futuros tendo com base em padrões obtidos em grandes conjuntos de dados do passado, mediante aplicação de métodos estatísticos de séries temporais.

2.4.4 Metodologias de *Data Mining*

Como na maioria das iniciativas organizacionais, os projectos de *Data Mining* deverão igualmente seguir um processo de gestão estruturado. Segundo (Azevedo & Santos, 2005) quando o processo de *Data Mining* é enquadrado no contexto de uma metodologia, este torna-se mais fácil de compreender, implementar e desenvolver. Destacam-se três (3) metodologias principais para este efeito: CRISP-DM, DMAIC e SEMMA.

A metodologia **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM) será adotada neste trabalho. Foi concebida em 1996 por um consórcio de quatro (4) empresas, SPSS, NCR Corporation, Daimler-Benz e OHRA. A primeira versão foi lançada em 1999, versão que se descreve neste documento e que servirá de base para o desenrolar deste trabalho de investigação.

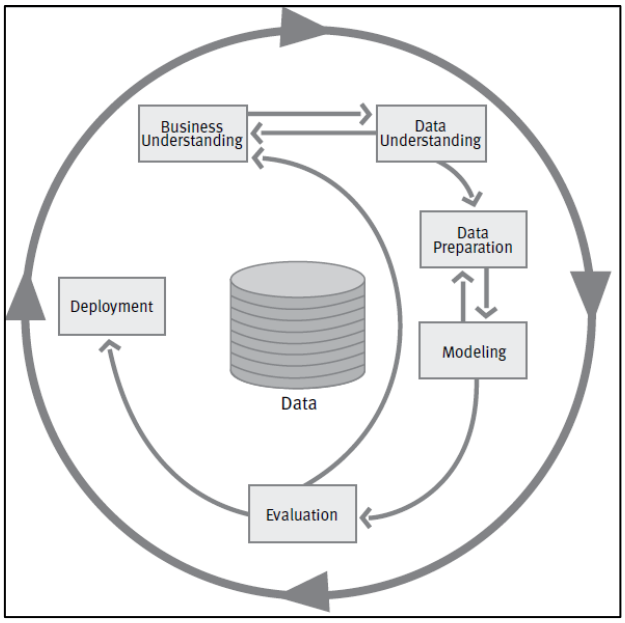


Figura 8 – Fases da metodologia CRISP-DM (Chapman, et al., 2000)

Conforme é visível na Figura 8 o processo é apresentado como um projecto com um ciclo de vida iterativo. O ciclo apresentado assenta em seis (6) fases cuja sequência não é rígida, mas dependente do resultado de cada fase. As fases mais interactivas incidem especialmente nos primeiros passos, de forma a haver uma maior compreensão das necessidades do negócio e da disponibilidade dos dados. É comum nos projectos de *Data Mining* que cerca de 60% do tempo estimado assente nas duas primeiras fases (Turban, et al., 2007).

As fases do ciclo desenvolvem-se da seguinte forma (Chapman, et al., 2000):

- **Compreensão do Negócio**
  Nesta fase será efectuada a análise dos objectivos e dos requisitos funcionais, técnicos e temporais na perspectiva do negócio. Será, ainda, efectuado o enquadramento desses objectivos e restrições na formulação do problema de *Data Mining*, bem como a elaboração de uma estratégia preliminar para alcançar os objectivos propostos.

- **Compreensão dos Dados**
  Nesta segunda fase é efectuada a recolha e análise dos dados.

- **Preparação dos Dados**
  Neste estágio é aplicado um conjunto de actividades com a finalidade de construção do conjunto de dados para que possam ser objecto de análise pelas ferramentas de modelação. É efectuada a selecção de casos e variáveis que se pretendem analisar. Poderá, caso necessário, efectuar-se algumas transformações.

- **Modelação**
  Posteriormente procede-se à selecção e aplicação de técnicas de *Data Mining* de acordo com os objectivos definidos, bem como à necessária calibração do modelo para que possam optimizar os resultados.

- **Avaliação**
  Nesta fase é efectuada a avaliação e revisão das actividades realizadas na construção do modelo e verificação da sua contribuição para o alcance dos objectivos de negócio.

- **Implementação**
  Actividades que conduzem à organização do conhecimento e à sua disponibilização, como relatórios e outros tipos de documentos. Por vezes os resultados obtidos poderão dar lugar a um novo projecto de *Data Mining*.

Uma outra metodologia, apesar de pouco referenciada e pouco presente na literatura desta área, talvez por não ser talhada para o *Data Mining* em particular, é a metodologia **DMAIC** - *Define, Measure, Analyze, Improve, Control*. Esta é baseada na metodologia Six Sigma, que é bastante estruturada e vocacionada para a eliminação de defeitos, desperdícios e para problemas de controlo de qualidade em diversas áreas de negócio (Turban, et al., 2007).

Figura 9 - Fases da metodologia DMAIC

Os passos descritos na Figura 9 demonstram o fluxo do procedimento, para cada uma das etapas, sendo definidos os objectivos, medidas de aplicação e mecanismos de *feedback*. Os processos desenvolvidos pela Six Sigma foram utilizados por indústrias de fabrico, um pouco por todo o mundo, considerando o incremento da qualidade e processos de controlo. Os seguidores desta metodologia defendem que com a mesma pode-se obter a mesma eficiência em projectos de *Data Mining*.

Uma terceira metodologia, bastante mais utilizada que a anterior, é a metodologia **SEMMA** - *Sample, Explore, Modify, Model, Assessment*, que foi desenvolvida pela empresa SAS, cuja área de negócio é a Estatística, Análise de Dados, *Business Intelligence*, *Data Mining* e Suporte à Decisão (Turban, et al., 2007).

Figura 10 - Fases da metodologia SEMMA

Conforme é possível ver na Figura 10 esta metodologia assenta em cinco (5) fases e é considerada um auxiliar na condução de um projecto em todas as suas etapas, desde a especificação do problema até à sua implementação, disponibilizando uma estrutura para a concepção, criação e evolução dos projectos de *Data Mining*, de forma a apresentar soluções para os problemas do negócio.

Tabela 2 – Comparação entre as metodologias SEMMA e CRISP-DM, adaptado de (Azevedo & Santos, 2008)

SEMMA	CRISP-DM
----------	Compreensão do Negócio
Amostra	Compreensão dos Dados
Exploração	
Modificação	Preparação dos Dados
Modelo	Modelação
Avaliação	Avaliação
----------	Implementação

Na Tabela 2 comparam-se apenas as metodologias SEMMA e CRISP-DM pelo facto de serem as mais usuais em projectos de *Data Mining*. Numa primeira análise pode-se depreender que em termos de processos para desenvolvimento de um projecto de *Data Mining* a metodologia CRISP-DM é mais completa que a SEMMA, pela incorporação das fases de *Compreensão do Negócio* e *Implementação*. Contudo, mediante uma análise mais pormenorizada é possível integrar o *Compreensão do Negócio* na fase *Amostra* da metodologia SEMMA considerando que não é possível constituir uma amostra coerente e sólida sem uma verdadeira compreensão de todos os aspectos apresentados. No que toca à fase de *Avaliação* da metodologia SEMMA se se considerar que o conhecimento obtido é aplicado assume-se assim que a fase de *Implementação* (presente na metodologia CRISP-DM) encontra-se também presente (Azevedo & Santos, 2008).

3. Aplicação de Técnicas de *Data Mining* em Farmácias Comunitárias

Através da revisão da literatura efectuada não foi possível encontrar estudos sobre a aplicação directa de técnicas de *Data Mining* aliadas ao tema a desenvolver, nomeadamente na tentativa de identificar padrões relevantes para apoio à gestão de uma farmácia comunitária. Os estudos que foram possível identificar reportam-se, em grande escala, à aplicação na indústria farmacêutica e ao seu âmbito, nomeadamente na criação, testes de desenvolvimento e previsão de reacções diversas relativamente a medicamentos. Estas técnicas são também aplicadas para analisar um conjunto de elevados volumes de dados no que toca a testes de ensaios clínicos.

Da pesquisa efectuada, verifica-se, também, um conjunto significativo de estudos na área de **farmacovigilância**. Existe, geralmente em todos os países, um sistema de farmacovigilância de medicamentos[5], com o objectivo de recolher e proceder à avaliação científica de informação relativa a suspeitas de reacções adversas no ser humano pela utilização de medicamentos, eventual comunicação a outros países, implementar medidas de segurança adequadas a minimizar os riscos, entre outras sobre esta matéria. Todavia, estes grandes volumes de dados provenientes deste assunto não são da competência das farmácias comunitárias e como tal também não constituem uma referência directa para o estudo em apreço. No entanto, poderão constituir uma mais-valia no que toca à compreensão da aplicação das diversas técnicas de *Data Mining*.

Este facto é bastante relevante uma vez que os objectivos destas organizações não são os mesmos. Em termos muitos generalistas, pode-se afirmar que a indústria farmacêutica fabrica e comercializa medicamentos após um longo processo, envolvendo meses ou anos de investigação, efectuando testes diversos e investimentos bastante elevados; os sistemas de farmacovigilância, acima descritos, têm atribuições muito específicas e que não se pretendem abordar neste trabalho.

O foco do presente trabalho de investigação irá incidir sobre o estudo em farmácias e estas assumem objectivos muito particulares, nomeadamente: dispensa de medicamentos nas condições legalmente previstas; na colaboração com a Autoridade Nacional do Medicamento e Produtos de Saúde, I. P. (INFARMED); e na identificação, quantificação, avaliação e prevenção dos riscos do uso de medicamentos, uma vez comercializados, permitindo o seguimento das suas possíveis reacções adversas.

É de realçar que, actualmente, as farmácias apresentam um portefólio de serviços bastante mais diversificado, nomeadamente: medicamentos, substâncias medicamentosas,

[5] Para o efeito, em Portugal, foi instituído o Sistema Nacional de Farmacovigilância de Medicamentos para Uso Humano e encontra-se regulamentado ao abrigo do Decreto-Lei n.º 176/2006, de 30 de Agosto.

medicamentos e produtos veterinários, medicamentos e produtos homeopáticos, produtos naturais, dispositivos médicos, suplementos alimentares e produtos de alimentação especial, produtos fitofarmacêuticos, produtos cosméticos e de higiene corporal, artigos de puericultura e produtos de conforto. Acresce, ainda, a prestação de serviços farmacêuticos de promoção da saúde e do bem-estar dos utentes, através da vacinação e da determinação de parâmetros bioquímicos. Alerta-se, ainda, para o facto de, actualmente, uma farmácia poder disponibilizar um conjunto restrito de serviços e venda de medicamentos de forma *online*.

Ainda que não tendo acesso ao documento integral, é relevante mencionar um trabalho elaborado como requisito parcial para obtenção do MBA – Engenharia de Software, da Universidade Federal do Rio de Janeiro, onde foi elaborado um projecto de sistema de apoio à decisão com o objectivo de gerar regras, a partir de uma base de dados, com a finalidade de identificar perfis de consumidores, aplicação de técnicas de marketing e auxiliar na estratégia de uma farmácia (Gonçalves, et al., 2008). O problema foi abordado com a aplicação de técnicas de associação e recurso à utilização do algoritmo APRIORI[6].

Tabela 3 - Relacionamento de regras e factores de suporte e confiança (Gonçalves, et al., 2008)

Nº. Regra	Regra Suporte		Confiança
1	[AGULHAS] ==> [SERINGA]	0,03	*0.87*
2	[ESPARADRAPO] +[LUVAS] ==> [COMPRESSA]	0,01	*0.73*
3	[COMPRESSA] + [HIDRATANTE] ==> [LUVAS]	0,01	*0.57*
4	[ATADURA] ==> [COMPRESSA]	0,01	*0.56*
5	[HIDRATANTE] +[LUVAS] ==> [COMPRESSA]	0,01	*0.56*
6	[ESPARADRAPO] ==> [COMPRESSA]	0,04	*0.51*
7	[ABSORVENTE-PARA-INCONTINENCIA] ==> [FRALDA-GERIATRICA]	0,02	*0.48*
8	[COMPRESSA] + [LUVAS] ==> ESPARADRAPO	0,01	0.44
9	[LUVAS] ==> [COMPRESSA]	0,03	0.42
10	[COMPRESSA] + [ESPARADRAPO] ==> [LUVAS]	0,01	0.40
11	[ALCOOL] ==> [COMPRESSA]	0,01	0.40
12	[COMPRESSA] + [LUVAS] ==> [HIDRATANTE]	0,01	0.37
13	[COMPRESSA] ==> [ESPARADRAPO]	0,04	0.34
14	[FRALDA-GERIATRICA] ==> [ABSORVENTE-PARA-INCONTINENCIA] 0,02		0.32
15	[COMPRESSA]==> [LUVAS]	0,03	0.32
16	[SHAMPOO] ==> [SABONETE]	0,01	0.31
17	CREME-DENTAL==> SABONETE	0,02	0.31
18	[DESODORANTE] ==> [SABONETE]	0,01	0.3
19	[SERINGA] ==> [AGULHAS]	0,03	0.29
20	[ALGODAO-HIDROFILO] ==> [COMPRESSA]	0,02	0.29

Como resultado destaca-se o facto de ser possível verificar, através da Tabela 3 que existe uma tendência de se vender produtos relacionados entre si, sendo dado alguns exemplos, como: luvas, adesivos e compressas; agulhas e seringas; e shampoo, sabonete e desodorizante. No entanto, os autores ressalvam que as percentagens de confiança revelam, na sua maioria, valores baixos. Através desta constatação, foi sugerida à Direcção Técnica a criação de promoções de produtos da mesma área.

Como caso de sucesso na utilização de um sistema de *Data Mining* efectivo surge a cadeia japonesa de farmácias *Pharma*, traduzindo-se esse sucesso no aumento de vendas e lucros. A cadeia em apreço encontra-se organizada de forma a melhor poder rentabilizar a informação e o seu armazenamento ao longo das suas vendas. Assim, a sede tem como objectivos: recolher as notas de encomenda e remeter aos distribuidores; armazenar informação dos clientes e das vendas; e criar relatórios de comportamentos de consumo remetendo-os para os fabricantes. É

[6] APRIORI é um algoritmo vocacionado para regras de associação (Agrawal & Srikant, 1994).

possível, assim, verificar que a descoberta de conhecimento é um dos factores chave do negócio da *Pharma*.

Como resultado, numa das aplicações de *Data Mining,* foi, através das regras de associação, a correlação na compra de artigos de higiene, e através de uma análise mais pormenorizada foi adicionada a localização das farmácias com indicação de *stock* traduzindo-se num aumento das vendas de analgésicos, em 1,5 vezes, sem necessidade de realizar qualquer tipo de desconto ou promoção como era habitual neste tipo de produtos (Hamuro, et al., 1998).

Continuando com o mesmo caso, verificou-se, ainda, que no verão a venda de dispositivos com a finalidade de transmitir calor para as mãos era suspensa. No entanto, numa das farmácias as vendas não só se mantiveram como levaram à liquidação deste mesmo artigo. Tal facto deveu-se à identificação de uma regra de associação nas vendas que permitiu compreender uma forte correlação com a medicina para o reumatismo. Após esta análise, verificou-se que afinal as outras farmácias continuavam a realizar vendas deste artigo, mesmo no verão, devido ao facto dos utentes que efectuavam esta compra trabalharem em escritórios equipados com ar condicionado. A mesma informação foi transmitida aos fornecedores que rapidamente voltaram a comercializar o produto independentemente da altura do ano.

Mais recentemente, (Sabhnani, et al., 2005) descrevem um sistema de vigilância que monitoriza, nos Estados Unidos da América, múltiplas fontes de dados, nomeadamente: vendas das farmácias não sujeitas a receita médica, visitas do departamento de emergência, indicadores climáticos, informação proveniente de censos, entre outras. Essa diversidade de fontes é utilizada tanto na identificação de surtos de natureza natural como surtos resultantes de ataques bioterroristas. As técnicas aplicadas neste estudo foram de Classificação, em que se pretendeu discriminar entre segmentos de surtos provenientes de doenças e surtos provenientes de outras causas. Assim, são apresentados alguns resultados como a venda de determinados produtos exactamente antes de ocorrer mau tempo (como tempestades de neve ou furacões) com o receio de rotura de *stock*. Uma segunda conclusão surgiu com o aumento de vendas no dia seguinte a um feriado nacional. Outro fenómeno que despertou interesse foi o aumento de vendas em destinos turísticos durante fins-de-semana prolongados.

Como exemplo de uma aplicação prática das aplicações de *Data Mining* numa farmácia comunitária, surge o estudo de (Silva, et al., 2007) que analisam regras que possam gerar possíveis localizações estratégicas para a criação de uma nova farmácia de maior ou menor dimensões. Para o efeito foi analisado um conjunto de informações dos clientes, como: local de residência; tipo de compra; modalidade de pagamento; altura do dia com maior afluência; entre outras, por forma a efectuar um levantamento socioeconómico dos clientes e conhecer as regras de classificação que possibilitem a descoberta de uma localização estratégica.

Para o estudo acima enunciado, os autores utilizaram regras de classificação capazes de prever valores de destino a partir de um conjunto de valores de entrada. Através da ferramenta WEKA[7] o algoritmo PRISM[8] mostrou ser eficiente e apto para o trabalho proposto.

Tabela 4 - Classificador induzido pelo algoritmo PRISM com selecção "Local de Compra" (Silva, et al., 2007)

1.	If FORM_PGTO = A VISTA and DS_BAIRRO_RES = CENTRO and TURN_COMP = NOITE then LOCAL
2.	If FORM_PGTO = A VISTA and DS_BAIRRO_RES = TELEGRAFO and TURN_COMP = TARDE then LOCAL
3.	If TURN_COMP = MANHA and DS_BAIRRO_RES = FATIMA and FORM_PGTO = CHEQUE then DOMICILIO
4.	If DS_BAIRRO_RES = COMERCIO and FORM_PGTO = A_VISTA and TURN_COMP = TARDE then DOMICILIO

Através da análise da Tabela 4 foi possível constatar que nos locais de "Fátima" e "Comércio" (Brasil) realizam, habitualmente, compras com recurso a entregas ao domicílio. Estes tornaram-se assim locais menos viáveis à abertura de uma nova farmácia de grandes dimensões. Por outro lado, verificou-se que nos locais "Telégrafo" e "Centro" apresentaram um número de consumidores bastante elevado e a modalidade de pagamento mais significativa foi a modalidade de numerário. Assim, a abertura de uma farmácia de grandes dimensões nestes últimos dois locais apresentam um investimento mais rentável e com um retorno mais rápido do investimento realizado. Por outro lado, os locais de "Fátima" e "Comércio" apresentaram características que poderiam, eventualmente, levar à abertura de uma farmácia de pequenas dimensões com o objectivo de servir como um espaço de distribuição da actual farmácia (Silva, et al., 2007).

[7] O WEKA (*Waikato Environment for Knowledge Analysis*) foi concebido em 1993, na Universidade de Waikato, Nova Zelândia sendo adquirido posteriormente por uma empresa no final de 2006. Este agrega algoritmos de diferentes abordagens/paradigmas que permitam a um computador "aprender" (no sentido de obter novo conhecimento) quer indutiva quer dedutivamente (Silva, et al., 2007).

[8] Poderá saber mais sobre este algoritmo em: Cendrowska, J. *PRISM: An algorithm for inducing modular rules,* International Journal of Man-Machine Studies, Elsevier, 1987, 27, 349-370

4. Metodologia

Pretende-se com este capítulo realizar uma descrição do caso em estudo, as suas características e particularidades, tendo em considerando tratar-se uma área de negócio muito particular e um mercado que tem sofrido alterações constantes.

Pretende-se, ainda, definir e descrever o planeamento que se propõe aquando do início dos trabalhos, tendo em consideração as ferramentas e técnicas de *Data Mining* utilizadas de forma a dar resposta à problemática levantada.

Para o prosseguimento de todo o trabalho científico é utilizada a metodologia CRISP-DM conforme anteriormente abordado e descrito na Subsecção 2.4.4.

4.1 Contextualização

O mercado de medicamentos tem vindo a sofrer constantes alterações legislativas na última década, havendo a necessidade de desenvolver novos modelos de organização administrativa, financeira e comercial, adaptando este tipo de organizações às novas realidades do mercado actual e, igualmente, de acordo com as regras vigentes no mercado europeu.

Assim, torna-se necessário investir cada vez mais, procurando rentabilizar novos recursos e tentando tirar o melhor partido da posição que as farmácias têm e da função que assumem perante a sociedade. É neste sentido e na necessidade de centrar o foco cada vez mais na gestão, partindo de uma filosofia de que todos os produtos têm de existir na farmácia, e que se deve assumir como princípio que a gestão não se resume somente à vertente comercial, como também a todo um aconselhamento técnico e especializado por profissionais da área da saúde.

É neste sentido, que surge cada vez mais a problemática da necessidade das farmácias se especializarem em determinadas áreas em particular, desde que tenham capacidade técnica e comercial para o fazer, criando assim elementos diferenciadores e potenciarem a criação de valor para o negócio.

Toda esta problemática acima referida, de capacidade comercial, assume hoje uma relevância acrescida, se considerarmos que as áreas de actuação das farmácias não se esgotam com a dispensa de medicamentos, mas sim na aposta da comercialização em novos mercados e em novas áreas, tais como a puericultura, cosmética, produtos ortopédicos, perfumaria/aromaterapia, veterinária, dietética, fitoterapia, dispositivos médicos e homeopatia.

É neste último ponto, que assenta a problemática que se pretende abordar e apoiar na gestão da farmácia tanto a nível de *stocks* como fazer a aposta em determinados produtos em

detrimento de outros, nomeadamente em produtos cuja venda seja frequente e que ao mesmo tempo apresentem particularidades que justifiquem a necessidade de estudar modelos de previsão tendo em vista o impacto no negócio.

Os produtos e serviços comercializados por uma farmácia são diversos tornando o estudo de todos uma tarefa morosa e com custos elevados, tendo em consideração que alguns desses produtos podem apresentar valor de vendas tão reduzidos que não importe qualquer necessidade de modelo de previsão, nem apresentam qualquer problema de investimento significativo. Assim, pretende-se que sejam objecto de estudo da aplicação de técnicas de *Data Mining*, e consequente necessidade de criar modelos de previsão apenas em produtos e/ou serviços que possam apresentar particularidades ao nível de algumas variáveis, como por exemplo processos especiais de armazenamento, prazos de validade, maior dificuldade na distribuição por parte dos fornecedores e a necessidade de um investimento significativo que permita uma maior ou menor negociação comercial com todas as vantagens e consequências associadas a essa variável.

No presente caso de estudo a gestão de compras é uma preocupação constante, no entanto é também algo que não segue uma linha de orientação única por parte da organização, isto é, a gestão das diferentes categorias são asseguradas por diversos farmacêuticos. Tal medida poderá ser compensatória no que toca à delegação de competências, maior motivação e poderá eventualmente levar a uma gestão com maior ou menor grau de personalização, dependendo do farmacêutico em causa. Assim, verifica-se uma gestão de categorias variada onde cada responsável negoceia directamente com os diferentes agentes comerciais existentes e posteriormente submete à consideração da Directora Técnica que por sua vez procederá ao deferimento ou indeferimento da aquisição dos artigos em causa e das quantidades em causa.

Contudo, a gestão acima referida denota uma gestão com incidência muito específica e com critérios muito pessoais não tendo em consideração toda uma estratégia global para a farmácia em causa no que toca à aposta de novos produtos, necessidade de eliminação ou substituição de categorias, possível existência de *lobbys* (sem descurar a competência e o profissionalismos dos colaboradores em apreço) para manutenção dessas mesmas categorias, entre outras variáveis a ter em consideração numa análise macro.

É igualmente sabido que aquando da negociação de determinados produtos é possível obter valores mais atractivos quando se efectuam volumes consideráveis, contudo essas aquisições deverão sempre ser muito ponderadas tendo em linha de conta factores como: elevado *stock* de produtos muito dispendiosos, artigos que que exijam processos especiais de armazenamento e artigos que tenham prazo de validade curto.

É neste sentido, que se pretende estudar determinados artigos mediante aplicação de técnicas de *Data Mining* e qual a sua previsão de venda para o futuro, de forma a melhor gerir o *stock*

dos mesmos e levar a uma melhor negociação junto dos actores de mediação de produtos farmacêuticos. Todavia, e dada a grande dimensão de produtos que fazem parte do portfólio de produtos e serviços disponíveis para venda, mais de 21 500 artigos, e tendo em consideração ainda a calendarização definida para este estudo foi necessário analisar apenas alguns desses artigos, restringindo-se assim apenas para 3 produtos que mais impacto pudessem ter para a farmácia na sua gestão diária. Os produtos seleccionados serão objecto de estudo aprofundado e devidamente justificados, no capítulo 5.3, referente à preparação dos dados.

4.2 Planeamento

Após a caracterização e uma breve descrição do caso que se pretende estudar, torna-se necessário definir um planeamento e uma metodologia adequada de forma a poder-se atingir os objectivos propostos. Assim, e tal como foi referido na Subsecção 2.4.4, a metodologia adoptada para esta dissertação foi a metodologia CRISP-DM (Chapman, et al., 2000) pelo facto de ser ajustável e dinâmica adaptando-se ao projecto em causa e igualmente por ser uma das metodologias mais usadas neste tipo de trabalhos de investigação.

Tendo em consideração que a metodologia CRISP-DM é simultaneamente iterativa e flexível, esta permite uma análise e avaliação constante, na medida em que possibilita retroceder nas diversas fases, constituindo assim uma mais-valia e permitindo igualmente, rever cada uma das fases sempre que se considere pertinente, reorganizando o trabalho em causa e redireccionando o trabalho de investigação. Neste sentido, e considerando o tempo disponível para a prossecução da dissertação a presente metodologia pretende proporcionar orientação no processo de descoberta de conhecimento válido.

Uma vez que o trabalho em causa é um caso de estudo foi necessário solicitar os dados necessários relativos às vendas de todos os produtos comercializados pela farmácia em causa. Os mesmos foram devidamente disponibilizados e autorizados única e exclusivamente para fins de estudo da presente dissertação. Para o efeito, foram salvaguardados todos os dados considerados confidenciais e de carácter identificador de clientes ou outros.

A presente dissertação teve início com objectivos muito generalistas, descritos na secção 1.2, como a demostração da aplicação de técnicas de *Data Mining* numa empresa desta natureza, com fins tão específicos, bem como identificar padrões relevantes para apoio à gestão e como tal efectuar uma melhor gestão de *stocks*. Todavia, e com o decorrer da aplicação da metodologia CRISP-DM os objectivos foram-se ajustando e tornando-se mais concretos. A descrição das diversas fases, bem como todo o tratamento efectuado dos dados será objecto de análise e descrição na secção 5.3.

4.3 Ferramentas Utilizadas

Para a elaboração da presente dissertação foi necessário proceder a diversas actividades de carácter prático e como tal foi igualmente necessário recorrer a ferramentas tecnológicas capazes de levar a cabo essas actividades.

Considerando que o pretendido era a aplicação de técnicas de *Data Mining* foi necessário equacionar várias hipóteses desde aplicações pagas, aplicações *open source*[9], aplicações já conhecidas ou explorar novas ferramentas. Para o efeito, e considerando os meios disponíveis, assim como o facto de ser um estudo de académico entendeu-se optar pela ferramenta de programação estatística **R**[10]. Esta é uma ferramenta *open source,* multiplataforma (disponível para Windows, Linux e Mac OS).

Apesar da ferramenta R não ser orientada especificamente para *Data Mining,* esta inclui uma grande variedade de algoritmos nesta área e é utilizada por um elevado número de especialistas nesta matéria, tendo havido um aumento na sua utilização de 23,3% em 2011 para 30,7% em Maio de 2012, segundo um inquérito *online* relativo à escolha de ferramentas nesta área para projectos reais (Piatetsky-Shapiro, 2012). Segundo um inquérito elaborado em 2011 a especialistas de *Data Mining* de cerca de 60 países, foi possível verificar que a ferramenta R é utilizada por 47% dos inquiridos, alegando ser uma ferramenta gratuita, *open source* e possuir uma grande variedade de algoritmos, seguindo-se as ferramentas SAS, SPSS, Weka e STATISTICA (Rexer, et al., 2011).

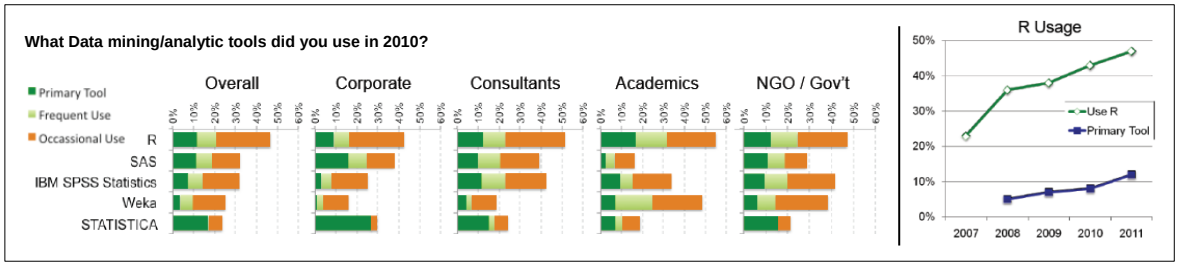


Figura 11 – Utilização da ferramenta R, adaptado de (Rexer, et al., 2011)

Outra potencialidade da ferramenta R reside no facto de possuir uma comunidade muito activa e dinâmica, estando novos pacotes a ser periodicamente criados e disponibilizados na sua página de Internet, havendo neste momento mais de 4000 pacotes disponíveis.

Em particular, recorreu-se à biblioteca *rminer*[11]. Esta é instalada em ambiente R, sendo igualmente *open source* e multiplataforma (Windows e Mac OS) e visa facilitar o uso de

[9] *Open Source* é tradicionalmente referenciado para aplicações livres e gratuitas, sendo possível ter acesso ao código, ou seja, qualquer pessoa tem acesso ao código podendo igualmente criar diferentes vertentes do mesmo programa.

[10] R Core Team (RCT), 2012. R: A Language and Environment for Statistical Computing. Vienna, Austria, http://www.R-project.org.

[11] Biblioteca disponível em http://cran.r-project.org/web/packages/rminer/index.html

algoritmos de *Data Mining* nas tarefas de regressão e classificação. Esta biblioteca é adaptada especialmente para Redes Neuronais e Máquina de Vectores de Suporte (Cortez, 2010). Neste trabalho em particular, esta ferramenta foi um recurso importante principalmente para a determinação das métricas de desempenho descritas na Secção 4.5.

Recorreu-se, igualmente, ao recurso da biblioteca *forecast*[12], para aplicação de funções de previsão de séries temporais univariadas e modelos lineares. Esta biblioteca da mesma forma que a anterior é *open source* e multiplataforma (Windows e Mac OS) (Hyndman & Khandakar, 2008).

Foi, igualmente, utilizado o programa EViews[13], que é um programa de estatística usado geralmente para análise Econométrica, sendo igualmente utilizado para análises de séries temporais. Este programa foi utlizado essencialmente como apoio para análise de testes de raiz unitária e análise de função de autocorrelação e função de autocorrelação parcial, tendo em consideração utilizar um ambiente gráfico, facilitando assim a utilização deste tipo de ferramentas por parte dos utilizadores e tendo, igualmente, em consideração o facto de mestrando não ter um conhecimento especializado na linguagem R.

Na fase inicial de recolha e tratamento dos dados foi utilizada a ferramenta *Microsoft Excel*[14]*.* Na disponibilização dos dados por parte da farmácia em estudo, surgiram alguns condicionantes sendo estes relativos aos dados passíveis de serem extraídos, bem como à forma de como esses dados seriam exportados e para que formato final. Todavia, foi possível exportar em formato xls[15], apresentando relatórios com os dados referentes às vendas diárias de todos os produtos e como tal foi necessário proceder à uniformização desses dados e transformá-los de forma a ser possível aplicar as diversas técnicas de *Data Mining*. Nesta ferramenta foram utilizadas fórmulas diversas de forma a agrupar os dados em vendas mensais e semanais, eliminar cabeçalhos e rodapés (elementos identificadores e usuais aquando da elaboração de relatórios de vendas), entre outras. Não se pretende nesta fase descriminar em detalhe todos os procedimentos levados a cabo sendo os mesmos objecto de maior análise na Secção 5.3.

[12] Biblioteca disponível em http://http://cran.r-project.org/web/packages/forecast/index.html

[13] IHS Inc., 2012. EViews 7: A statistical package for Windows. Irvine, Califórnia, Estados Unidos da América, http://www.eviews.com.

[14] Ferramenta de folhas de cálculo, parte integrante do pacote de produtividade *Microsoft Office,* comercializado pela *Microsoft*.

[15] xls é a extensão dos ficheiros Excel (folhas de cálculo). A partir de 2007 o Excel passou a utilizar a extensão xlsx, devido às novas funcionalidades introduzidas.

4.4 Técnicas de *Data Mining*

Para a etapa de modelação dos dados foi escolhida a técnica de análise de séries temporais, conforme já descrita em 0. As técnicas de séries temporais envolvem a análise estatística tendo como base padrões obtidos em grandes conjuntos de dados do passado com o intuito de se prever valores futuros. Esta opção foi equacionada tendo em consideração a metodologia CRISP-DM, relativamente à redução da quantidade dos dados de entrada e uma melhoria considerável da sua qualidade à medida que progride em cada processo iterativo. Foram escolhidos métodos estatísticos e populares na área de previsão: Alisamento Exponencial e *Box-Jenkins*.

Nos diversos modelos testados, a divisão entre os dados de teste e de treino foi efectuada dividindo o conjunto em 90% para treino (ajuste dos modelos de previsão) e 10% para teste (avaliação da capacidade de previsão dos modelos). Esta divisão foi efectuada de acordo com a sequência temporal e consequentemente os dados de teste dizem respeito às observações mais recentes.

4.4.1 Método de Alisamento Exponencial

O método de Alisamento Exponencial aplica uma média ponderada nas observações de uma série temporal, sendo atribuídos diferentes pesos aos dados, isto é, os dados mais antigos têm pesos menores em detrimento dos dados mais recentes. Este tipo de método é muito utilizado na análise de controlo de *stocks*, sendo um método de previsão bastante rápido, com alguma simplicidade e baixo custo (Hyndman & Athanasopoulos, 2012). Os métodos de previsão que aplicam alisamento exponencial dividem-se em 3 (Makridakis, et al., 1998):

- Alisamento Exponencial Simples

  A forma geral do método de Alisamento Exponencial Simples encontra-se através da seguinte equação:

  $$F_{t+1} = \alpha Y_t + (1 - \alpha) F_t \qquad (1)$$

  Na medida em que F_{t+1} é a previsão para t+1; F_t é a previsão para t; Y_t a procura realizada no período t; n como sendo o tamanho da série temporal e α como sendo a constante de alisamento com valor entre 0 e 1.

  O valor de α quanto mais próximo for de 1, maior será o ajuste do erro na previsão anterior, isto é, o modelo dá mais peso a observações recentes e torna-se mais sensível a mudanças. Por outro lado, quanto mais próximo de 0 for o valor α, menor será o ajuste e como tal o modelo irá dar mais peso a observações antigas, sendo o tratamento mais uniforme o que leva a previsões mais estáveis.

- Alisamento Exponencial Linear de Holt

  Este método expande o método anterior para previsões com dados que apresentam tendência linear, mas não sazonalidade. Desta forma, a previsão com Alisamento Exponencial Linear de Holt é efectuada com recurso a duas constantes de alisamento, nomeadamente α e β, com valores entre 0 e 1, não relacionados entre si, e descreve-se através das seguintes equações:

Previsão	$F_{t+m} = L_t + b_t m$	(2)
Nível	$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + b_{t-1})$	(3)
Tendência	$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$	(4)

  Sendo que F_{t+m} é a previsão para $t+m$; m como sendo o horizonte de previsão; L_t a estimativa do nível da série temporal no período t; b_t é a estimativa de tendência da série temporal para o período t e α e β são as constantes de alisamento.

  Tal como no primeiro método os valores de α e β podem ser determinados através da minimização do erro de previsão em dados de treino.

- Método de Holt-Winters

  Este terceiro método é utilizado para situações em que as séries temporais apresentam padrões com tendência linear e sazonalidade. O Método de Holt-Winters aplica suavizações para estimar o nível, tendência e sazonalidade da série em estudo no processo de previsão.

  O método apresenta duas abordagens distintas que incidem essencialmente na forma de como a sazonalidade é abordada, forma multiplicativa ou forma aditiva. A primeira é indicada sempre que a amplitude da sazonalidade varia com o nível e a forma aditiva é mais apropriada para séries cuja amplitude de sazonalidade é independente do nível.

  O método multiplicativo de Holt-Winters apresenta as seguintes equações:

  Previsão

  $$F_{t+m} = (L_t + b_t m) S_{t-s+m} \quad (5)$$

  Nível

  $$L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1}) \quad (6)$$

  Tendência

  $$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \quad (7)$$

  Sazonalidade

  $$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma)S_{t-s} \quad (8)$$

Onde S é o número de períodos por ciclo sazonal; S_t é a estimativa do componente sazonal da série no período t e α, β e γ são as constantes de alisamento (com valores entre 0 e 1, não relacionados entre si).

Por outro lado o método aditivo de Holt-Winters é menos comum que o multiplicativo, na medida em que trata o componente sazonal de forma aditiva. Este apresenta as seguintes equações:

Previsão

$$F_{t+m} = (L_t + b_t m)S_{t-s+m} \quad (9)$$

Nível

$$L_t = \alpha(Y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1}) \quad (10)$$

Tendência

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \quad (11)$$

Sazonalidade

$$S_t = \gamma(Y_t - L_t) + (1-\gamma)S_{t-s} \quad (12)$$

As diferenças entre as formas aditiva e multiplicativa deste método são que os índices sazonais e de nível são somados ou subtraídos ao contrário de serem multiplicados ou divididos.

De igual forma como os 2 primeiros métodos, os parâmetros α, β e γ podem ser determinados de forma a minimizar o erro de previsão.

4.4.2 Método de *Box-Jenkins* - ARIMA

A aplicação do método de *Box-Jenkins* é efectuada através da utilização de um algoritmo matemático, com termos autoregressivos e de média móvel, de forma a criar um modelo adaptado e adequado para uma série temporal. Este método modela a função de autocorrelação de uma série temporal estacionária com o mínimo de parâmetros possíveis, utilizando uma combinação de termos de autoregressão (AR – valor p), integração (I – valor d) e média móvel (MA – valor q) – ARIMA[16] (p,d,q) O objectivo fundamental da modelação ARIMA é a definição de um modelo que apresenta boas propriedades estatísticas e descreva a série em estudo. Para o efeito poderá ser seguida a metodologia de *Box-Jenkins*, esquematizada na Figura 12, onde se propõem três etapas conforme se encontram abaixo e o modelo geral desta metodologia é apresentado na seguinte equação:

$$\emptyset(B)(1-B)^d Y_t = \emptyset(B)e_t \quad (13)$$

[16] ARIMA - *AutoRegressive Integrated Moving Average*

- Identificação
  - Estacionaridade da série

    De forma a aplicar o este modelo é necessário verificar se a série em análise é estacionária ou não. Uma série é considerada estacionária quando não se verificam nos dados sinais de tendência e sazonalidade, sendo que os mesmos flutuam sobre uma média independentemente do tempo e uma variância constante. Com o intuito de aferir se uma série é estacionária é necessário analisar os coeficientes de autocorrelação (FAC) e de autocorrelação parcial (FACP) (Box, et al., 1994).

    Para séries não estacionárias, as mesmas deverão ser transformadas (remoção de padrões, tal como: tendência, sazonalidade, e outros) de forma a tornarem-se estacionárias em relação à média e variância. A estacionaridade na média é efectuada através da diferenciação, e deste modo obtém-se o valor d (parcela de integração).

    A necessidade de atingir a estacionaridade é igualmente relevante aquando da estimativa de previsão, uma vez que com séries não estacionárias torna-se mais difícil de prever a longo prazo, considerando que a amplitude dos intervalos de previsão aumenta com o horizonte de previsão.

  - Identificação dos parâmetros p e q através da FAC e FACP

    Aquando da análise de séries temporais estacionárias, os coeficientes de FAC e FACP tendem a ser valores próximos de zero, da mesma forma que para séries não estacionários estes coeficientes tendem a ser não zero para diversos períodos de tempo (Box, et al., 1994).

    Assim, p será a ordem máxima dos parâmetros de autoregressão simples (AR) e q será a ordem máxima dos parâmetros de média móvel simples (MA).

- Estimação

  Os modelos podem ser apenas de ordem AR ou MA, ou então híbridos, isto é modelos ARMA (modelos estacionários e sem diferenciação) ou modelos ARIMA (processos que foram objecto de diferenciação). Para este método, a série temporal é ajustada a um modelo que apresenta o menor erro em relação a outros possíveis modelos. Assim, os parâmetros são estimados através de procedimentos de optimização de forma a minimizar a Soma dos Erros Quadráticos (Makridakis, et al., 1998).

- Avaliação

  Após a identificação e estimação dos modelos efectua-se a avaliação dos mesmos, nomeadamente através da verificação dos resíduos e se os mesmos são aleatórios. Analisa-se igualmente os coeficientes de FAC FACP dos resíduos e espera-se que

para os erros aleatórios nenhum coeficiente seja ele de autocorrelação ou autocorrelação parcial seja significativo (Makridakis, et al., 1998).

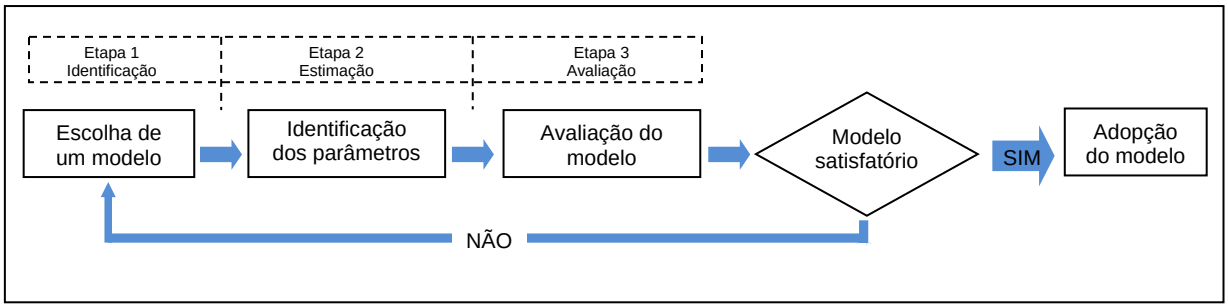


Figura 12 – Metodologia de *Box-Jenkins* (Box, et al., 1994)

Uma vez definido o modelo, as correlações históricas entre os dados são capturadas e extrapoladas para períodos futuros, obtendo-se assim previsões de valores futuros (Makridakis, et al., 1998).

4.5 Medidas de Desempenho da Previsão

De forma a aferir a qualidade dos modelos testados neste trabalho de investigação foram utilizadas as seguintes métricas:

- Erro Médio (ME – Mean Error)

$$MAE = \frac{1}{n}\sum_{t=1}^{n} e_t \quad (14)$$

- Erro Absoluto Médio (MAE - *Mean Absolute Error*)

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t| \quad (15)$$

- Raíz do Erro Quadrático Médio (RMSE - Root Mean Squared Error)

$$RMSE = \sqrt{\frac{SSE}{L}} \quad (16)$$

- Erro Quadrático Médio (MSE - *Mean Squared Error*)

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \quad (17)$$

Para (Makridakis, et al., 1998) o erro é determinado pela diferença entre uma observação no tempo t e a previsão para o mesmo tempo t, e representa-se através da seguinte equação: $e_t = Y_t - F_t$, sendo F_t os valores de previsão e Y_t os valores conhecidos ou valores reais.

O MAE calcula os valores absolutos do erro médio (erros positivos e erros negativos). O MSE parte do mesmo princípio que o erro absoluto médio, no entanto os valores dos erros são multiplicados por si mesmo e posteriormente é realizada a média dos seus resultados. Tal significa que o MSE penaliza erros individuais (para um dado valor) maiores, quando comparado com o MAE, sendo muito utilizado em diversos métodos estatísticos.

Estas métricas foram escolhidas tendo em consideração a literatura existente na previsão da procura onde estas geralmente estão sob a forma de percentagens de erro absoluto ou erro quadrático (Robert, 1996). As medidas Erro Percentual Absoluto Médio (MAPE - *Mean Absolute Percentage Error*) e Erro Médio Percentual (MPE - *Mean Percentage Error*) poderiam fornecer outras informações relevantes, no entanto, não será possível utilizar neste trabalho uma vez que surgem valores iguais a zero nas observações, conforme descrito na Secçao 5.3.

5. Gestão de Produtos Farmacêuticos via Técnicas de *Data Mining*

Tal como referido anteriormente a metodologia adoptada para dar seguimento à presente dissertação foi a metodologia CRISP-DM dado ser bastante flexível e cíclica. Para o efeito irá descrever em torno deste capítulo todas as actividades desenvolvidas e que se consideram pertinentes tendo em consideração as seis fases essenciais que caracterizam o CRISP-DM, nomeadamente: Compreensão do Negócio; Compreensão dos Dados; Preparação dos Dados; Modelação; Avaliação e Implementação (Chapman, et al., 2000).

Contudo, e considerando que o CRISP-DM é uma metodologia bastante adaptável, algumas das iterações poderão não ser executas nem descritas em todas as fases.

5.1 Compreensão do Negócio

De acordo com a metodologia CRISP-DM, a compreensão do negócio é a primeira fase a ser analisada, pois será necessário compreender do ponto de vista do negócio o que realmente se pretende e que objectivos se querem atingir. Desta forma será possível direccionar todos os esforços no sentido de dar resposta a todas as questões essenciais evitando assim eventuais esforços ao responder a falsas questões.

Do ponto de vista dos objectivos de negócio pretendia-se algo bastante rudimentar, (sem prejuízo de uma análise de produto e/ou de marketing) tal como efectuar uma melhor gestão de *stocks*, sem afectar a disponibilidade dos produtos aos clientes. Esta gestão pode ser tão simples como complexa, pois dependerá da visão e objectivos estratégicos que a Direcção Técnica possa definir para a farmácia, ou seja, se pretende ter em *stock* todos os produtos tornando-se numa referência para os clientes que sabem que podem sempre contar com aquela farmácia, ou por outro lado se pretende especializar-me em determinados produtos em detrimento de outros e deste modo incidir em nichos de mercado, com um público-alvo específico e bem definido. Poderá, ainda, apostar em *stocks* reduzidos ou nulos em produtos específicos, colmatando essas falhas com pedidos periódicos junto das empresas de distribuição farmacêutica. Sendo que neste último caso poderá não ser possível negociar o valor dos produtos, tendo em consideração não se efectuar compras em larga escala, e não ter acesso a bonificações diversas e contractos *rappel*[17].

Neste sentido, foi necessário proceder ao levantamento de todos os produtos vendidos na farmácia em apreço, e dentro de um período temporal definido. Assim, foram identificados mais

[17] Desconto concedido sobre a totalidade das compras efectuadas quando se ultrapassa determinado montante previamente negociado, durante um período de tempo estabelecido.

de 21 500 produtos e serviços, pelo que foi necessário restringir apenas para três produtos tendo em consideração factores específicos, nomeadamente: maior ou menor impacto, necessidade de processos especiais de armazenamento, prazos de validade, distribuição por parte dos fornecedores, bem como a necessidade de aposta em determinado produto incrementando assim o *stock* e ter em consideração os diferentes factores relativos à negociação comercial.

Do ponto de vista do negócio a possibilidade de prever a venda destes três produtos será uma forma de descobrir tendências e sazonalidades nas venda, identificar e compreender comportamentos por parte dos clientes e consequentemente possibilitar uma melhor e mais adaptada compra desses produtos, reduzindo investimento descontrolado. Será possível, ainda, reorganizar as compras e efectuar uma melhor gestão seja esta anual, semestral ou outra considerada por parte da Direcção Técnica.

Para o efeito, a análise dos dados será efectuada através de uma análise de séries temporais, ou seja, serão analisadas as vendas de um determinado período de tempo, que irá incidir basicamente deste a abertura da farmácia nas actuais instalações, desde o final de 2003 até início de 2012.

Sem prejuízo da análise incidir apenas em três artigos, este estudo nesta fase, assenta também, na verificação e viabilidade da aplicação de técnicas de *Data Mining* numa empresa desta natureza, sendo que posteriormente as mesmas técnicas poderão ser aplicadas e/ou revistas de forma a serem utilizadas a um conjunto de produtos mais alargado.

5.2 Compreensão dos Dados

Conforme já referido na Secção 4.3 os dados analisados neste trabalho, relativos às vendas efectuadas, foram cedidos e exportados através do sistema informático utilizado, nomeadamente o sistema WinPhar[18], para o formato xls, em formato de relatórios. Estes dados apenas foram possíveis de exportar em dois níveis: resumo de produtos vendidos e detalhe de vendas.

O primeiro apresentava informação muito sucinta, com uma periodicidade mensal, e com os seguintes campos: código do produto, produto, quantidade e data da última venda. Este tipo de relatório não foi utilizado para a aplicação de técnicas de *Data Mining* uma vez que se pretendia analisar as vendas diárias de um conjunto de produtos e o presente tipo de relatório apenas apresentava informação relativa à data da última venda de cada mês.

[18] Winphar é um Sistema de Informação para Farmácias desenvolvido pela empresa SimPhar.

O segundo tipo de relatório passível de extrair (relatório mensal com detalhe de vendas) apresenta informação mais completa, nomeadamente a discriminação de todos os produtos transaccionados por cada operação de venda, ou seja, todos os artigos vendidos a cada cliente, com detalhe da quantidade vendida e por dia. Os relatórios apresentavam, ainda, informação sobre o número do movimento (número sequencial relativo a cada venda diária), o número do documento (número associado à transacção), iva aplicado, denominação do sistema ou subsistema de saúde e valor da comparticipação, desconto associado quando aplicável, informação de venda anulada e outras com codificação não identificada. Devido ao detalhe de informação ser elevado, foram gerados pelo sistema informático WinPhar, por cada mês, uma média de 800 relatórios, ou seja 9 600 relatórios por cada ano e um total de mais de 76 800 relatórios referentes ao período de Outubro de 2003 e Abril de 2012. Este segundo tipo de relatório foi o escolhido e utilizado para a aplicação de técnicas de *Data Mining*, contudo houve a necessidade de tratar os dados existentes removendo assim toda a informação adicional, desde cabeçalhos e rodapés (característico de relatórios), colunas com informação adicional e não necessária e o detalhe associado ao cliente aquando da venda. Este tipo de tratamento será analisado com maior detalhe no capítulo seguinte, relativamente à preparação dos dados.

Numa fase inicial tornou-se impossível efectuar qualquer tipo de análise ou tentar formular hipóteses uma vez que o número de relatórios era tão elevado e com informação diversa em cada um que não era possível ter noção do panorama real apenas com uma simples visualização e sem qualquer tipo de uniformização e sistematização dos dados.

Contudo, foi possível verificar que apesar dos dados serem mensais e terem início em Setembro de 2003, constatou-se que a data da primeira venda foi a 20 do mês em questão e não iniciou no dia 1. Tal deveu-se ao facto da farmácia ter iniciado a sua actividade nas actuais instalações nessa mesma data. Por este motivo, e pelo facto dos dados não se reportarem na sua totalidade ao mês Setembro não sendo assim significativos, nem corresponder com a frequência que se pretenderia adoptar, entendeu-se utilizar apenas os dados referentes ao período de Novembro de 2003 a Abril de 2012.

No que diz respeito à verificação da qualidade dos dados, constatou-se problemas na verificação dos dados do mês de Dezembro de 2004, onde surgiam vendas efectuadas sem qualquer tipo de quantidade. Tal facto seria impossível, uma vez que a concretizar-se uma venda a mesma teria de corresponder a pelo menos uma unidade de saída no *stock* da farmácia. Ao confrontar com os dados do sistema de informação da farmácia verificou-se que estava a surgir um problema na exportação dos dados para o ficheiro xls. Tal obstáculo foi ultrapassado com a exportação dos dados para um ficheiro pdf[19] e posteriormente para xls. Nesta dissertação tem-se vindo a fazer referência apenas à extensão do tipo de ficheiros xls e

[19] pdf (*Portable Document Format*) é uma extensão de ficheiros desenvolvido pela *Adobe Systems* com o intuito de disponibilizar ficheiros independentes do *software* de origem e da sua resolução.

não necessariamente a uma ferramenta para visualização e edição desses mesmos tipos de ficheiros, pelo facto da farmácia ter adoptado e utilizado como ferramenta de trabalho e demonstrado os ficheiros aquando da sua cedência através do *software* OpenOffice Calc[20], todavia na presente dissertação a visualização, edição e todo o pré-processamento dos dados foi efectuado utilizando o *software* Microsoft Excel. O facto de os dados terem sido sempre exportados para o formato xls foi pela necessidade de se efectuar algum pré-processamento e adaptar os dados para a aplicação de técnicas de *Data Mining*.

5.3 Preparação dos Dados

Conforme já enunciado os dados apresentavam informação referente a Outubro de 2003 e Abril de 2012, sendo que a primeira venda foi efectuada a 20 de Outubro de 2003. Neste sentido, optou-se por não considerar os dados relativos ao mês de Outubro e utilizar uma série com dados desde Novembro de 2003 a Abril de 2012, forma a uniformizar a série considerando a frequência adoptada.

Foi igualmente referido na Secção 5.1 que a farmácia no período em análise comercializou mais de 21 500 produtos e serviços e como tal foi necessário proceder a uma selecção de artigos no sentido de se analisarem apenas esses e não todos, tendo em consideração o tempo disponível para levar a cabo este trabalho de investigação. Para além das razões já elencadas entendeu-se, igualmente, estudar artigos referentes as duas categorias que implicam uma maior ou menor rentabilidade aquando da gestão de uma farmácia. Ou seja, existem produtos vendidos, actualmente, com IVA a 6% (produtos com receita médica) e outros vendidos com IVA a 23% (produtos não sujeitos a receita médica). Não sendo o objectivo deste trabalho descrever esta gestão nem aprofundar esta temática, pretende-se apenas salientar que os valores estipulados para os produtos com IVA a 23% são da exclusiva competência da Direcção Técnica de cada farmácia e sem qualquer restrição por parte de qualquer entidade governamental ou entidade reguladora. Daí o entendimento neste trabalho em analisar dois produtos onde os preços são estipulados com algum condicionamento e será necessário gerir com maior rigor, bem como analisar dois produtos de venda livre e que poderão trazer uma maior rendibilidade à farmácia.

Para aplicação das técnicas de *Data Mining,* nomeadamente análise de séries temporais entendeu-se criar duas séries para o estudo de cada produto: uma série com periodicidade mensal e outra com periodicidade semanal.

[20] Ferramenta de folhas de cálculo, parte integrante do pacote de produtividade *Apache OpenOffice,* sendo este um *software* aberto e gratuito.

Tabela 5 – Descrição das séries temporais

Produtos	Quantidade		Descrição
	Série mensal	Série semanal	
Produto A	102	443	Vacina indicada para crianças desde as 6 semanas até aos 5 anos de idade.
Produto B	102	443	Medicamento utilizado em adultos com hipertensão essencial (pressão arterial elevada).
Produto C	102	443	Creme reparador antibacteriano indicado para peles irritadas e agredidas.
	Treino: 92 Teste: 10	Treino: 399 Teste: 44	

Conforme é possível verificar na tabela acima, os dados foram divididos em dados de teste e em dados de treino. Esta divisão foi efectuada dividindo o conjunto em 90% para treino (ajuste dos modelos de previsão) e 10% para teste (avaliação da capacidade de previsão dos modelos). Esta divisão foi efectuada de acordo com a sequência temporal pelo que os dados de teste dizem respeito às observações mais recentes.

Conforme referido no capítulo anterior os relatórios cedidos pela farmácia apresentavam muita informação suplementar e que não se enquadrava nos objectivos deste trabalho, como o número do movimento efectuado, número do documento, IVA, sistema ou subsistema de saúde, comparticipação, descontos, vendas anuladas e outras com codificação não identificada. Todas estas informações adicionais foram eliminadas, bem como foram removidas todas as referências aos rodapés e cabeçalhos de cada relatório, criando assim um ficheiro único com todos os produtos e as vendas realizadas por cada operação em cada dia.

Seleccionados os produtos e considerando que se dispunha apenas de relatórios com as vendas diárias foi necessário proceder a um pré-processamento no sentido de transformar esses dados em somatórios de vendas mensais e somatórios de vendas semanais, para posterior utilização das séries temporais com a frequência desejada. Para efeitos desse pré-processamento foi utilizada a ferramenta Excel com recurso às mais variadas fórmulas disponíveis para concatenar valores, eliminar duplicados, criar somatórios, devolver o dia da semana, somar produtos e outras manipulações de forma simples, rápida e intuitiva. Neste sentido, o campo referente à quantidade vendida em cada operação foi rapidamente substituído pelo somatório de vendas ou mensal ou semanal, isto é foram somadas as vendas diárias de forma a criar vendas mensais e semanais, dependendo da série utilizada, referentes a cada produto. Durante este processo de manipulação de datas e somatório de vendas foram devolvidos alguns erros ao longo da criação de cada série. Tais erros foram rapidamente identificados pelo facto de alguns dados, aquando do processo de exportação, terem adicionado um caracter espaço em branco. Uma vez identificado foi necessário proceder à remoção de todos os caracteres espaço em branco de todos dados e de todos os campos.

Tendo em consideração que se pretende fazer uma análise de séries temporais e entendeu-se criar duas séries de dados (mensal e semanal) para cada produto verificou-se rapidamente que

existem valores omissos, ou seja existem alguns produtos que não foram objecto de venda nem todos os meses nem todas as semanas. Assim, foi necessário preencher essas lacunas e completar as respectivas séries com valor 0, isto é na semana ou mês X foram vendidas 0 unidades.

Verificou-se, ainda, aquando da preparação dos dados para a criação de séries semanais, que existiam anos com um total de 52 semanas e outros com um total de 53 semanas e aconteceu devido ao facto de se tratar de anos bissextos. Esta diferença do número de semanas acontece apenas para os anos de 2004 e 2009 e Tal situação apesar de parecer apenas um detalhe levantou uma questão interessante na frequência a aplicar aquando da introdução dos dados nas ferramentas de *Data Mining*, onde teria de se indicar um valor para essa frequência, nomeadamente 52 ou 53. De forma a uniformizar anos dados e minimizar a questão entendeu-se adoptar uma frequência de 52 semanas, devido ao facto de ser a frequência dominante e os bissextos serem apenas 2 no conjunto em estudo.

Os produtos seleccionados e apresentados na Tabela 5 encontram-se mascarados de forma a manter a confidencialidade dos nomes dos produtos e evitar qualquer tipo de promoção de determinadas marcas. Os produtos A e B correspondem a produtos que carecem de receita médica para serem comercializados e o produto C é de venda livre ou seja os utentes não carecem de receita médica para a sua aquisição.

5.4 Modelação

Nesta fase, segue-se a aplicação das diversas técnicas de *Data Mining* implementadas pela biblioteca *forecast* e *rminer*. Numa primeira fase serão descritas as técnicas implementadas através da ferramenta *forecast*, nomeadamente: Alisamento Exponencial e *Box-Jenkins.*

Apenas nesta fase se deu início aos primeiros ensaios com a ferramenta R, uma vez que até aqui todo o tratamento tinha vindo a ser efectuado com recurso a outras ferramentas. Para este efeito, num primeiro momento irá descrever-se a análise efectuada e posteriormente a avaliação dos resultados obtidos nas diferentes séries analisadas para cada um dos produtos.

5.4.1 Aplicação do Método de Alisamento Exponencial

Aquando da análise de um modelo de previsão através de técnicas de *Data Mining* pretende-se sempre que o modelo criado possa apresentar o melhor desempenho possível e consequentemente possa ter o menor valor de erro de acordo com as métricas de desempenho existentes. Será tomado como valor de maior referência o MSE por ser frequentemente utilizado em estudos desta natureza.

Tabela 6 – Resultados do método de Alisamento Exponencial para o produto A

	Alisamento Exponencial Simples		Alisamento Exponencial Linear de Holt		Holt-Winters	
Série →	**Mensal**	**Semanal**	**Mensal**	**Semanal**	**Mensal**	**Semanal**
alpha	0,28	0,09	0,43	0,39	0,29	0,08
beta			0,18	0,22	0,05	0,01
gamma					0,52	0,35
MSE	59,55	**8,85**	**46,41**	179,60	78,37	11,96
RMSE	7,72	2,98	6,81	13,40	8,85	3,46
MAE	6,40	2,44	5,71	11,80	6,46	2,82

Desta forma, foram aplicados os três métodos de Alisamento Exponencial de forma a aferir qual apresentava melhores resultados, nomeadamente: Alisamento Exponencial Simples, Alisamento Exponencial Linear de Holt e Método de Holt-Winters. Assim, e como é possível verificar através da Tabela 6, o método de Alisamento Exponencial Linear de Holt é aquele que apresenta erros mais reduzidos para a série mensal, tal como é possível aferir aquando da análise do erro quadrático médio (MSE) com um valor de 46,41, em detrimento de 59,55 para o método Simples e 78,37 para o método de Holt-Winters, bem como através da análise dos erros SSE, RMSE e MAE que apresentam no segundo método valores diminutos face aos restantes. Por outro lado, na série semanal o método Simples apresenta-se como um melhor modelo de previsão, quando analisado os erros. Salienta-se que nesta série, os erros do método Linear de Holt são mais elevados que os outros dois métodos. Constata-se assim que apenas na série mensal se verifica tendência, apesar do valor beta (0,18) ser reduzido e como tal esta componente sofre pouca actualização ao longo da série.

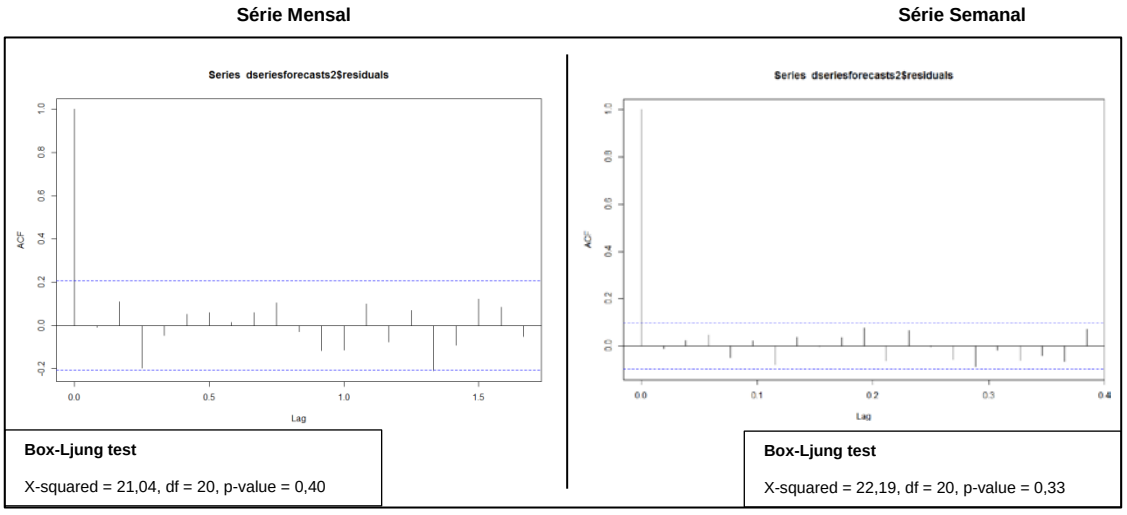


Figura 13 – Correlograma dos resíduos para o Produto A

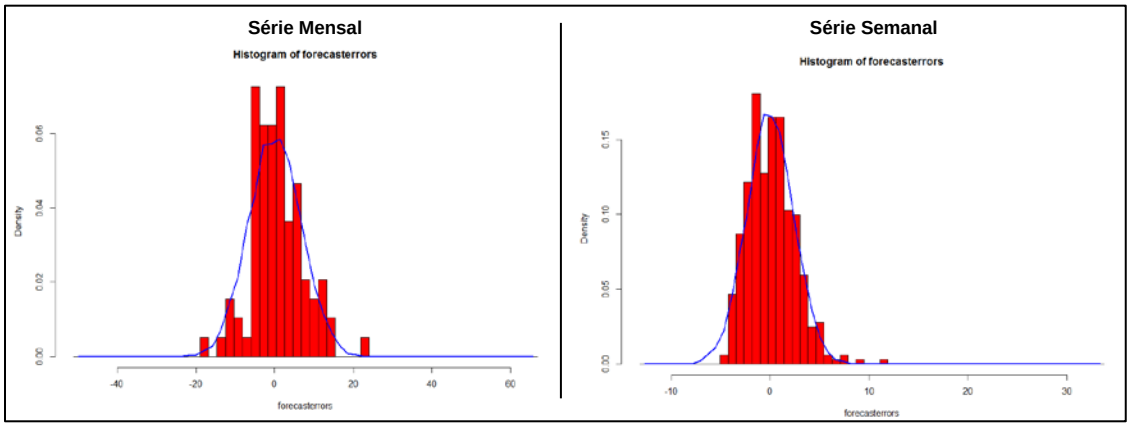


Figura 14 – Histograma dos erros de previsão para o produto A

Através da Figura 13, observa-se que os coeficientes de correlação para os erros de previsão da amostra, para a série mensal, não excedem os limites de significância para os valores de desfasamento de 1-20. Contudo, é de realçar que apenas o valor de desfasamento 16 (-0.210) ultrapassa os limites de significância de 95% e o valor 3 quase atinge este valor, pelo que foi necessário averiguar estes resultados através do teste Ljung-Box. Verifica-se, ainda, através da Figura 14, que o histograma apresenta uma distribuição dos erros de previsão aproximadamente centrados em zero, com distribuição mais ou menos normal, com uma curva ligeiramente inclinada para a direita em comparação com uma distribuição normal de média zero. Mediante execução do teste Ljung-Box não se rejeita a nulidade das autocorrelações para o nível de significância (0,05) estabelecido, e pode-se afirmar que a distribuição dos erros de previsão encontra-se distribuída de forma normal com média zero.

Relativamente à série semanal, aquando da observação do correlograma dos resíduos acima apresentado, verifica-se que a autocorrelação para os erros de previsão da amostra não excedem os limites de significância para os valores de desfasamento de 1-20. De forma a verificar se existe evidência significativa de correlações não zero para os valores de desfasamento 1-20 foi efectuado o teste estatístico Ljung-Box que retornou um valor de 0,33, indicando que existe pouca evidência de autocorrelações não zero nos erros de previsão. Verifica-se, ainda, na Figura 14, que o histograma apresenta uma distribuição dos erros de previsão aproximadamente centrados em zero, com distribuição mais ou menos normal, apresentando uma assimetria positiva. Todavia, esta assimetria positiva é relativamente pequena pelo que se pode afirmar que os erros de previsão encontram-se normalmente distribuídos com média zero.

Face ao exposto, verifica-se que o método de Alisamento Exponencial Linear de Holt fornece um modelo de previsão adequado para a série mensal, onde destaca a componente tendência. Para os dados semanais deste produto, a tendência não se evidência e o método de Alisamento Exponencial Simples apresenta melhores resultados.

Tabela 7 – Resultados do método de Alisamento Exponencial para o produto B

	Alisamento Exponencial Simples		Alisamento Exponencial Linear de Holt		Holt-Winters	
Série →	**Mensal**	**Semanal**	**Mensal**	**Semanal**	**Mensal**	**Semanal**
alpha	0,19	0,05	0,21	0,03	0,02	0,006
beta			0,14	0,01	0	0
gamma					0,07	0,13
MSE	45,76	7,95	117,08	**7,02**	**35,44**	7,94
RMSE	6,76	2,82	10,82	2,65	5,95	2,82
MAE	6,02	2,40	9,75	2,26	4,34	2,29

Em relação ao produto B, quando analisadas as séries temporais através dos diferentes métodos de alisamento exponencial verifica-se que para a série mensal o método de Holt-Winters é aquele que apresenta resultados mais satisfatórios, sendo o valor de alpha estimado de 0,02, um valor bastante reduzido que significa que o ajuste do modelo é mínimo e irá ser dado mais peso a observações antigas, sendo o tratamento mais uniforme e como tal leva a previsões mais estáveis. Também, ao nível da análise dos erros de previsão se verifica que este método é aquele que apresenta valores mais reduzidos, e como tal um melhor desempenho do modelo, conforme se constata com a análise do MSE (35,44), em detrimento de 45,76 e 117,08 para os métodos Simples e Linear de Holt, respectivamente. Os restantes erros RMSE e MAE apresentam igualmente valores mais reduzidos reforçando assim o desempenho e optimização do modelo. É possível, ainda, verificar que se encontra com componente de tendência e componente sazonal, com valores beta e gamma de 0 e 0,13, respectivamente. A componente tendência assume valor 0, o que indica que a estimativa desta componente não é actualizada ao longo da série temporal, sendo igual ao seu valor inicial, significando um bom sentido de intuição na medida em que o nível é ligeiramente alterado ao longo da série temporal, mas esta componente permanece aproximadamente a mesma. Por sua vez a componente sazonal é ligeiramente reduzida indicando que é baseada nas observações mais antigas.

Quando analisada a série semanal, constata-se um melhor desempenho através do método Linear de Holt, sendo que se continua a evidenciar a componente de tendência, mas os elementos sazonais não têm grande expressão ao contrário dos dados mensais. Assim, os valores de alpha e beta são 0,03 e 0,01, respectivamente, sendo alpha um valor bastante reduzido indicando que a estimativa é baseada essencialmente nas observações mais antigas e o valor beta, igualmente reduzido, indicando que a estimativa da componente tendência é muito pouco actualizada ao longo da série temporal. Igualmente, ao nível da análise dos erros verificam-se valores menores neste segundo método, como se comprova através do MSE (7,02) e demais erros.

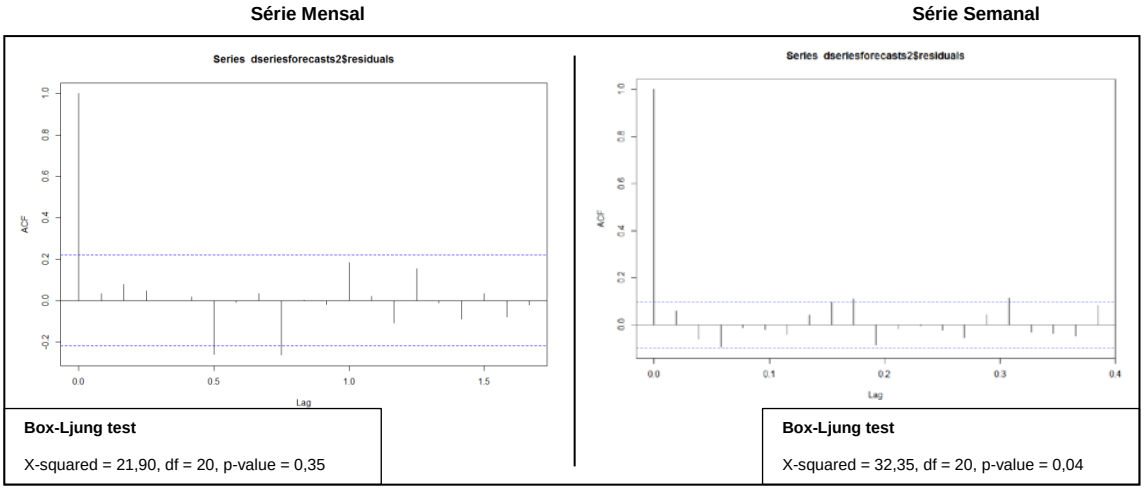


Figura 15 – Correlograma dos resíduos, Produto B

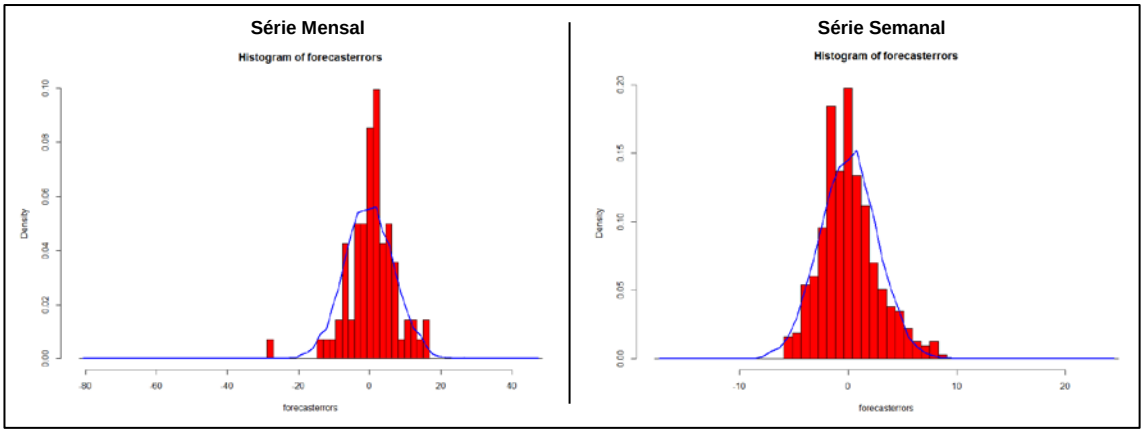


Figura 16 – Histograma dos erros de previsão para o produto B

Na série mensal, através do correlograma dos resíduos, presente na Figura 15, verifica-se que os coeficientes de correlação para os erros de previsão nos valores de desfasamento 6 (-0,26) e 9 (-0,26) excedem os limites de significância. Assim, é de esperar que 2 em 20 autocorrelações para os primeiros 20 valores de desfasamento excedam 95% dos limites. Verifica-se, ainda, através da Figura 16, que o histograma apresenta uma distribuição dos erros de previsão aproximadamente centrados em zero, com distribuição mais ou menos normal, sendo ligeiramente inclinada para a esquerda quando comparada com uma curva normal. Todavia, a inclinação é relativamente reduzida, pelo que é plausível afirmar que os erros de previsão estão normalmente distribuídos com média zero. O teste Ljung-Box comprova, igualmente, que não se rejeita a nulidade das autocorrelações para o nível de significância (0,05) estabelecido.

Ao analisar o produto B, na sua serie semanal, constata-se que o correlograma dos resíduos demonstra que as autocorrelações para os erros de previsão nos valores de desfasamento 9 e 16 excedem os limites de significância. Também o valor 8 quase atinge os limites de

significância pelo que deverá ser um dado a ter em consideração. Para verificar se existe ou não evidência de autocorrelações não zero procedeu-se à análise do teste Ljung-Box que retornou um p-value de 0,04, indicando que existe pouca evidência de autocorrelações não zero para os valores de desfasamento 1-20. Ao contrário do que acontece na série mensal, o histograma presente na Figura 16 apresenta uma distribuição dos erros de previsão aproximadamente centrados em zero, com distribuição mais ou menos normal, sendo ligeiramente inclinada para a direita, contudo é plausível afirmar que os erros de previsão estão normalmente distribuídos com média zero.

Verifica-se assim, que o método de alisamento exponencial Holt-Winters fornece um modelo de previsão adequado para a série mensal e o método de alisamento exponencial Linear de Holt com melhores resultados para a série semanal, referente ao produto B.

Tabela 8 – Resultados do método de Alisamento Exponencial para o produto C

	Alisamento Exponencial Simples		Alisamento Exponencial Linear de Holt		Holt-Winters	
Série →	**Mensal**	**Semanal**	**Mensal**	**Semanal**	**Mensal**	**Semanal**
alpha	0,13	0,06	0,32	0,06	0,002	0,004
beta			0,09	0	0,99	0,09
gamma					0,25	0,28
MSE	**11,24**	1,64	14,05	1,64	25,71	**1,62**
RMSE	3,35	1,28	3,75	1,28	5,07	1,27
MAE	2,71	1,03	3,08	1,03	3,90	1,04

Quando analisado o produto C, na Tabela 8 verifica-se que para a série mensal o método de alisamento exponencial simples é aquele apresenta resultados que minimizam os erros de previsão. O valor de alpha estimado é de 0,13, sendo bastante reduzido, o que significa que o ajuste do modelo é mínimo e irá ser dado mais peso a observações antigas, sendo o tratamento mais uniforme levando a previsões mais estáveis. Realça-se que esta série temporal, aquando da estimação do modelo de Holt-Winters, apresentou um erro na ferramenta adoptada, pelo que foi necessário diferenciar a série de forma a estabilizar e prosseguir a análise.

Ao nível da análise dos erros de previsão verifica-se, de forma clara, que o método de alisamento exponencial simples é aquele que apresenta valores mais reduzidos, e um melhor desempenho do modelo, conforme se constata com a análise do MSE (11,24), em detrimento de 14,05 e 25,71 para os métodos Linear de Holt e Holt-Winters, respectivamente. Os restantes erros RMSE e MAE apresentam igualmente valores mais reduzidos.

Ainda quanto à análise do mesmo produto C, mas na série semanal, num total de 52 semanas anuais, verifica-se que tanto o método de alisamento exponencial simples como o método Linear de Holt apresentam os mesmos valores de alpha (0,06) e os mesmos valores quanto aos erros MSE, RMSE e MAE. Todavia, entende-se que existe melhor desempenho através do método Holt-Winters. Para esta análise, recorreu-se aos erros MSE e RMSE, com valores de 1,62 e 1,27 respectivamente. Apenas o MAE é ligeiramente superior aos outros dois métodos,

com apenas com uma centésima de valor de diferença. Ao analisar-se o método de Holt-Winters, verifica-se que existe componente de tendência e sazonal, pelo que se terá de ter em consideração estas características, com valores beta e gamma de 0,09 e 0,28, respectivamente. A componente tendência apresenta um valor bastante reduzido indicando fraca actualização da estimativa ao longo da série temporal. Por sua vez a componente sazonal é ligeiramente reduzida indicando que é baseada nas observações mais antigas.

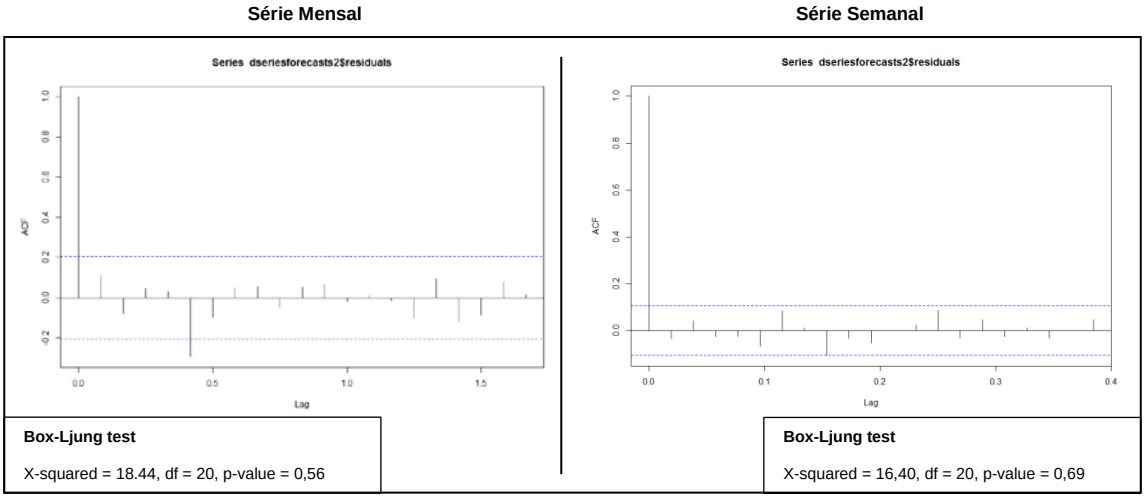


Figura 17 – Correlograma dos resíduos para o produto C

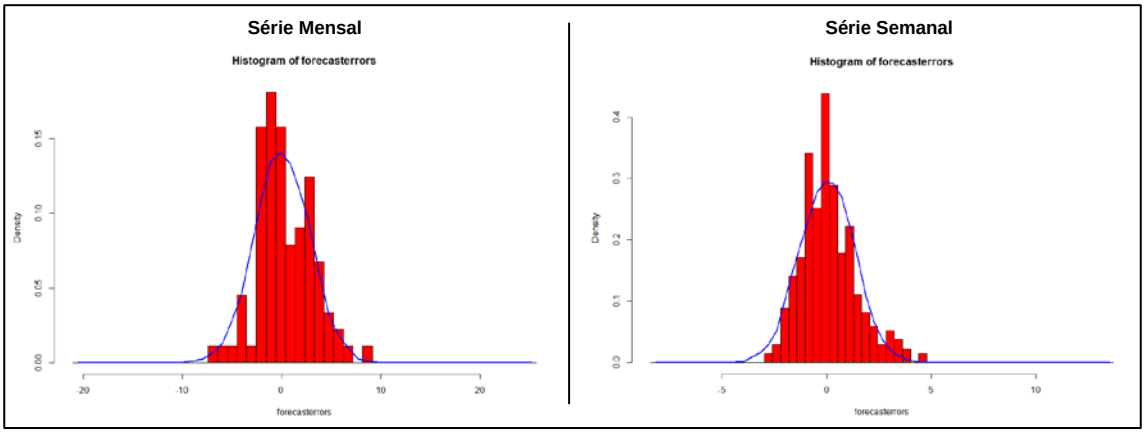


Figura 18 – Histograma dos erros de previsão para o produto C

Através da análise do correlograma dos resíduos, presente na Figura 17, da série mensal, a autocorrelação dos erros de previsão no valor de desfasamento 5 excede os limites de significância, apresentando um valor de -0,294. Desta forma, é de esperar que 1 em 20 autocorrelações excedam os limites de 95%. De seguida procedeu-se à análise do teste Ljung-Box de forma a apurar se existe ou não evidência de autocorrelações não zero. Uma vez efectuado o teste, não se rejeita a nulidade das autocorrelações para o nível de significância (0,05) estabelecido. É possível ainda constatar, através da Figura 18, que o histograma apresenta uma distribuição dos erros de previsão tem uma variação aproximadamente

constante ao longo do tempo. O histograma mostra que é plausível afirmar que os erros de previsão são normalmente distribuídos com média zero.

Também a série semanal, à semelhança dos dados mensais, no correlograma dos resíduos, se constata que o valor de desfasamento 8 excede os limites de significância com um valor negativo de -0,11, sugerindo assim que 1 em cada 20 autocorrelações excedam os limites de 95%. Este valor que excede os limites de significância sugeriu numa primeira fase a possibilidade de existir evidência de autocorrelações não zero, pelo que procedeu-se à execução do teste Ljung-Box que retornou um valor p-value de 0,69 comprovando assim que existe pouca evidência de autocorrelações não zero nos erros de previsão da amostra. Quanto à distribuição dos erros de previsão, Figura 18, e ao contrário dos dados mensais, o histograma apresenta uma distribuição aproximadamente centrados em zero, com distribuição mais ou menos normal, sendo ligeiramente inclinada para a direita. Todavia, a inclinação é bastante reduzida, pelo que os erros de previsão estão normalmente distribuídos com média zero.

Constata-se que para o produto C, a série mensal não apresenta uma tendência ou sazonalidade significativa que careçam de análise mais cuidada, pelo que o método de alisamento exponencial simples apresenta resultados bastantes satisfatórios. Quando aos dados semanais, verifica-se tendência e sazonalidade pelo que o modelo de previsão Holt-Winters obtém melhores resultados de previsão face à série mensal onde se obtém erros maiores, quando comparado através do MAE.

5.4.2 Aplicação do Método Box-Jenkins

Através da modelação ARIMA pretende-se definir um modelo em termos de parâmetros que apresente boas propriedades estatísticas e descreva a série em estudo. Para se atingir tal objectivo irá utilizar-se a metodologia de *Box-Jenkins*. Assim, e conforme anteriormente descrito na Figura 12, são propostas três etapas: identificação, estimação e avaliação. Estas etapas serão de seguida descritas para cada uma das séries, de cada um dos produtos.

Série Mensal

Null Hypothesis: PRODUTOA_MES has a unit root
Exogenous: None
Lag Length: 0 (Automatic - based on SIC, maxlag=0)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.779853	0.0714
Test critical values:	1% level	-2.588059	
	5% level	-1.944039	
	10% level	-1.614637	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(PRODUTOA_MES)
Method: Least Squares
Date: 21/10/12 Time: 17:48
Sample (adjusted): 2003M12 2012M04
Included observations: 101 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PRODUTOA_MES(-1)	-0.061663	0.034645	-1.779853	0.0781

R-squared	0.030704	Mean dependent var	0.009901
Adjusted R-squared	0.030704	S.D. dependent var	7.188178
S.E. of regression	7.076964	Akaike info criterion	6.761419
Sum squared resid	5008.342	Schwarz criterion	6.787311
Log likelihood	-340.4516	Hannan-Quinn criter.	6.771901
Durbin-Watson stat	2.920508		

Série Semanal

Null Hypothesis: PRODUTOA_SEMANA has a unit root
Exogenous: None
Lag Length: 0 (Automatic - based on SIC, maxlag=0)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-6.965263	0.0000
Test critical values:	1% level	-2.570216	
	5% level	-1.941543	
	10% level	-1.616217	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(PRODUTOA_SEMANA)
Method: Least Squares
Date: 21/10/12 Time: 17:49
Sample (adjusted): 17/11/2003 23/04/2012
Included observations: 441 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PRODUTOA_SEMANA(-1)	-0.201354	0.028908	-6.965263	0.0000

R-squared	0.099289	Mean dependent var	0.015873
Adjusted R-squared	0.099289	S.D. dependent var	3.248387
S.E. of regression	3.082907	Akaike info criterion	5.091888
Sum squared resid	4181.899	Schwarz criterion	5.101160
Log likelihood	-1121.761	Hannan-Quinn criter.	5.095546
Durbin-Watson stat	2.682707		

Figura 19 – Teste aumentado de *Dickey-Fuller* para o produto A

Para identificar qual o modelo a adoptar é necessário numa primeira fase verificar que a série temporal é estacionária e que não possua problemas de raiz unitária. Sempre que uma série apresente raiz unitária os pressupostos estatísticos de que a média e a variância devem ser constantes ao longo do tempo são violados pondo em causa os resultados obtidos por modelos econométricos. Dos testes de raiz unitária disponíveis destaca-se o teste aumentado de *Dickey-Fuller* (ADF). Este teste devolve um valor de estatística em comparação com os valores críticos de *MacKinnon* para os níveis de significância 1%, 5% e 10%. Assim, duas hipóteses são formuladas. A hipótese nula é rejeitada em determinado nível de significância quando o valor calculado da estatística ADF for menor que o valor crítico de *MacKinnon* correspondente, esta informação sugere que a série é estacionária. Quando o valor da estatística ADF é superior aos valores críticos de *MacKinnon* não é possível rejeitar a hipótese nula, o que sugere tratar-se de uma série não estacionária.

Como é possível verificar na Figura 19, para a série mensal, o teste ADF apresenta um valor de -1,78, sendo superior aos valores críticos de -2,59 e -1,94 para os valores críticos de 1% e 5%. O teste ADF apenas apresenta valor inferior para com o valor crítico de 10%. Assim, não se pode rejeitar a hipótese nula, o que significa que a série mensal do produto A tem um problema de raiz unitária e estamos perante uma série não estacionária. Por outro lado, os dados semanais no teste ADF apresentam um valor inferior (-6,97) para com os valores críticos de 1%, 5% e 10%, nomeadamente -2,57, -1,94 e -1,62, respectivamente, pelo que tais valores dão lugar a um teste ARMA.

Para o efeito, nos dados mensais, procedeu-se à diferenciação de um nível através do seguinte código: `ddiff1 <- diff(d, differences=1)` e posteriormente à análise da FAC e da FACP de forma a verificar se estes indicadores por si seriam suficientes para definir quais os melhores modelos ARIMA a levar a cabo. Da análise efectuada, entendeu-se realizar um conjunto de modelos ARIMA, não ficando apenas restrito à determinação por via FAC e FACP. Assim, para ambas as séries, relativas ao produto A, foram estimados diversos modelos, sendo que no presente documento, na Tabela 9, constam apenas aqueles que apresentam valores mais reduzidos, de forma a facilitar a comparação e leitura através dos critérios de ajustamento *Akaike* (AIC) e *Schwartz* (BIC), com o intuito de aferir qual o modelo que minimiza estes mesmos critérios estatísticos.

Tabela 9 – Modelos ARIMA identificados para o produto A

Série Mensal						Série Semanal		
Modelos	**AIC**	**BIC**	**Modelos**	**AIC**	**BIC**	**Modelos**	**AIC**	**BIC**
0,1,0	713,01	715,51	1,1,3	591,82	604,31	0,0,1	1849,77	1861,73
0,1,1	620,77	625,77	1,1,4	589,90	604,90	0,0,2	1844,83	1860,78
0,1,2	**588,97**	**596,47**	2,1,0	633,75	641,25	0,0,3	1838,12	1858,05
0,1,3	590,97	600,97	2,1,1	595,91	605,91	1,0,0	1848,03	1859,99
0,1,4	588,32	600,82	2,1,2	590,59	603,09	**1,0,1**	**1823,59**	**1839,54**
1,1,0	644,19	649,19	2,1,3	591,04	606,04	1,0,2	1825,27	1845,20
1,1,1	594,49	601,99	2,1,4	588,45	605,95	1,0,3	1827,17	1851,08
1,1,2	590,97	600,97				2,0,0	1840,56	1856,50

Verifica-se assim, que para a série mensal referente ao produto A, o melhor modelo ARIMA é o modelo p=0, d=1, q=2, apresentando valores de 588,97 e 596,47 para as estatísticas AIC e BIC respectivamente. Realça-se, ainda, o modelo ARIMA (0,1,4) apresenta um valor AIC inferior ao anterior (588,32), contudo o valor de BIC (600,82) é ligeiramente superior, pelo que entende-se que pela combinação dos dois factores o primeiro modelo é mais robusto.

Os dados presentes na série semanal destacam-se que pelo facto de não ter sido necessário realizar nenhuma diferenciação, o valor *d*=0 (integração), assim sendo foram efectuados modelos ARMA, em detrimento dos anteriores ARIMA. Neste sentido, o modelo que apresenta melhor desempenho é o modelo ARMA (1,0,1) com 1823,59 e 1839,54 para os valores de AIC e BIC respectivamente. Apesar de haver outros modelos com valores AIC algo semelhantes, a verdade é que aquando da combinação com BIC, o ajuste acaba por ter de ser maior.

O facto da Tabela 9 demonstrar mais modelos para a série mensal que para a série semanal, serve apenas para realçar o desempenho do modelo ARIMA (0,1,4), já referido anteriormente, e como tal entendeu-se pertinente essa inclusão no quadro. Nos dados semanais não se verificou nenhum outro modelo que se destacasse, sempre tendo em consideração os valores dos critérios de ajustamento *Akaike* e *Schwartz*.

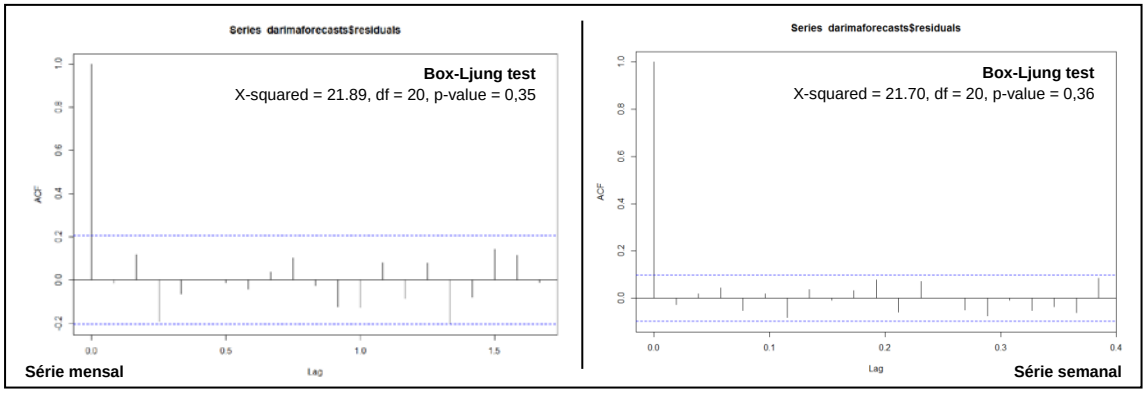
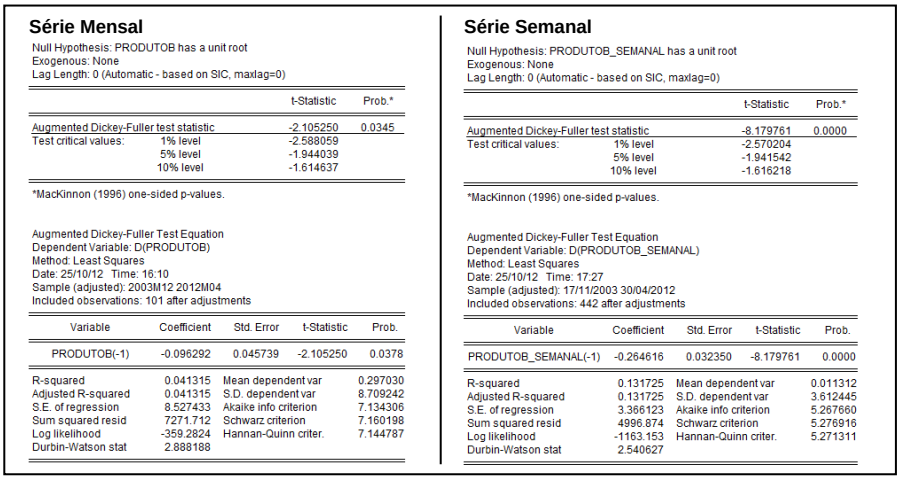


Figura 20 – Avaliação do modelo ARIMA, Produto A

Após a estimação do modelo a aplicar foi necessário proceder à avaliação do mesmo, nomeadamente através da verificação dos parâmetros de autocorrelação FAC e teste Ljung-Box, esta avaliação foi efectuada tanto para as séries mensais como semanais.

Após análise do correlograma constata-se que a série mensal e a série semanal apresentam comportamentos semelhantes, sendo que os valores de desfasamento das observações semanais não ultrapassam os limites de significância, em detrimento dos dados mensais onde o valor de desfasamento 16 (-0.207) é negativo e ultrapassa os limites, sendo por isso de esperar, para esta série (mensal), 1 em cada 20 valores excedam os 95% dos limites de significância. Constata-se, ainda, que para as duas séries o teste Ljung-Box retorna um valor muito semelhante de 0,35 (dados mensais) e 0,36 (dados semanais) indicando assim que não se rejeita a nulidade das autocorrelações para o nível de significância (0,05) estabelecido, nos erros de previsão para os valores de desfasamento 1-20.



Série Mensal

Null Hypothesis: PRODUTOB has a unit root
Exogenous: None
Lag Length: 0 (Automatic - based on SIC, maxlag=0)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-2.105250	0.0345
Test critical values:	1% level	-2.588059	
	5% level	-1.944039	
	10% level	-1.614637	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(PRODUTOB)
Method: Least Squares
Date: 25/10/12 Time: 16:10
Sample (adjusted): 2003M12 2012M04
Included observations: 101 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PRODUTOB(-1)	-0.096292	0.045739	-2.105250	0.0378

R-squared	0.041315	Mean dependent var	0.297030
Adjusted R-squared	0.041315	S.D. dependent var	8.709242
S.E. of regression	8.527433	Akaike info criterion	7.134306
Sum squared resid	7271.712	Schwarz criterion	7.160198
Log likelihood	-359.2824	Hannan-Quinn criter.	7.144787
Durbin-Watson stat	2.888188		

Série Semanal

Null Hypothesis: PRODUTOB_SEMANAL has a unit root
Exogenous: None
Lag Length: 0 (Automatic - based on SIC, maxlag=0)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-8.179761	0.0000
Test critical values:	1% level	-2.570204	
	5% level	-1.941542	
	10% level	-1.616218	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(PRODUTOB_SEMANAL)
Method: Least Squares
Date: 25/10/12 Time: 17:27
Sample (adjusted): 17/11/2003 30/04/2012
Included observations: 442 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PRODUTOB_SEMANAL(-1)	-0.264616	0.032350	-8.179761	0.0000

R-squared	0.131725	Mean dependent var	0.011312
Adjusted R-squared	0.131725	S.D. dependent var	3.612445
S.E. of regression	3.366123	Akaike info criterion	5.267660
Sum squared resid	4996.874	Schwarz criterion	5.276916
Log likelihood	-1163.153	Hannan-Quinn criter.	5.271311
Durbin-Watson stat	2.540627		

Figura 21 - Teste aumentado de *Dickey-Fuller* para o produto B

Para analisar o produto B, começou-se por realizar o teste aumentado de *Dickey-Fuller*, como se verifica na Figura 21, sendo que para a série mensal o teste ADF apresenta um valor de -2,11, sendo inferior aos valores críticos de -1,94 e -1,61 para os valores críticos de 5% e 10%. O teste ADF apenas apresenta valor inferior para com o valor crítico de 1%. Assim, não se considera possível rejeitar a hipótese nula, tendo esta série, deste produto B, um problema de raiz unitária e estamos perante uma série não estacionária. Por outro lado, os dados semanais no teste ADF apresentam um valor inferior (-8,18) para com os valores críticos de 1%, 5% e 10%, nomeadamente -2,57, -1,94 e -1,62, respectivamente, contudo entendeu-se realizar igualmente um nível de diferenciação de forma a garantir a estacionariedade da série temporal quanto à média e permitir uma melhor comparação face aos restantes produtos em análise.

Tal como aconteceu no produto anterior, procedeu-se à diferenciação de um nível através do seguinte código: `ddiff1 <- diff(d, differences=1)` e posteriormente à análise da FAC e da FACP de forma a verificar estes indicadores e realizar um conjunto diverso de modelos ARIMA de forma apurar qual o melhor modelo, tendo como base de comparação os critérios de ajustamento *Akaike* (AIC) e *Schwartz* (BIC).

Tabela 10 - Modelos ARIMA identificados para o produto B

Série Mensal			Série Semanal		
Modelos	**AIC**	**BIC**	**Modelos**	**AIC**	**BIC**
0,1,0	751,09	753,59	0,1,0	2560,19	2564,18
0,1,1	656,77	661,77	0,1,1	2149,72	2157,69
0,1,2	**608,04**	**615,54**	0,1,2	1924,89	1936,84
1,1,0	695,11	700,11	1,1,0	2357,47	2365,44
1,1,1	629,78	637,27	1,1,1	2069,29	2081,24
1,1,2	609,89	619,89	**1,1,2**	**1921,09**	**1937,03**

Seguindo a mesma linha de orientação face à análise ao produto A, aquando da análise dos modelos ARIMA, verifica-se que para a série mensal o melhor modelo é o ARIMA(0,1,2), apresentando este valores de 608,04 e 615,54 para as estatísticas AIC e BIC, respectivamente.

No caso da série semanal, estas estatísticas são mais reduzidas no modelo ARIMA (1,1,2) com valor AIC de 1921,09 e BIC de 1937,03.

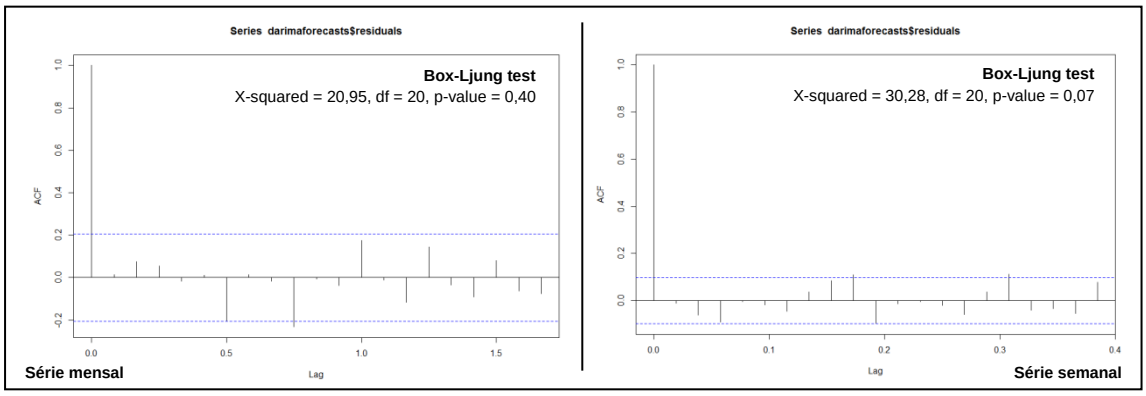


Figura 22 - Avaliação do modelo ARIMA para o produto B

Depois de estimados os modelos, conforme acima identificado, procedeu-se à análise do correlograma em que se constata que a série mensal nos valores de desfasamento 6 (-0,21) e 9 (-0,23) ultrapassam os limites de significância, sendo por isso de esperar, que 2 em cada 20 valores excedem os 95% dos limites de significância. O teste Ljung-Box (p-value: 0,40) esclarece que existe pouca evidência para autocorrelações não zero nos erros de previsão para os valores de desfasamento 1-20. Também a série semanal apresenta 2 valores de desfasamento que ultrapassam os limites de significância, nomeadamente o valor 9 (0,11) e 16 (0,11) e um p-value de 0,07, um valor bastante reduzido, no teste Ljung-Box, pelo que é possível afirmar que 2 em cada 20 valores excedem os 95% dos limites de significância, contudo não se rejeita a nulidade das autocorrelações para o nível de significância (0,05) estabelecido.

Série Mensal

Null Hypothesis: PRODUTOC_MENSAL has a unit root
Exogenous: None
Lag Length: 0 (Automatic - based on SIC, maxlag=0)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-2.872699	0.0044
Test critical values:	1% level	-2.588059	
	5% level	-1.944039	
	10% level	-1.614637	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(PRODUTOC_MENSAL)
Method: Least Squares
Date: 25/10/12 Time: 22:59
Sample (adjusted): 2003M12 2012M04
Included observations: 101 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PRODUTOC_MENSAL(-1)	-0.162500	0.056567	-2.872699	0.0050

R-squared	0.075873	Mean dependent var	0.069307
Adjusted R-squared	0.075873	S.D. dependent var	3.530602
S.E. of regression	3.394020	Akaike info criterion	5.291759
Sum squared resid	1151.938	Schwarz criterion	5.317651
Log likelihood	-266.2338	Hannan-Quinn criter.	5.302241
Durbin-Watson stat	2.601861		

Série Semanal

Null Hypothesis: PRODUTOC_SEMANAL has a unit root
Exogenous: None
Lag Length: 0 (Automatic - based on SIC, maxlag=0)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-11.64134	0.0000
Test critical values:	1% level	-2.570204	
	5% level	-1.941542	
	10% level	-1.616218	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(PRODUTOC_SEMANAL)
Method: Least Squares
Date: 25/10/12 Time: 23:03
Sample (adjusted): 17/11/2003 30/04/2012
Included observations: 442 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PRODUTOC_SEMANAL(-1)	-0.470501	0.040416	-11.64134	0.0000

R-squared	0.235065	Mean dependent var	0.002262
Adjusted R-squared	0.235065	S.D. dependent var	1.701672
S.E. of regression	1.488291	Akaike info criterion	3.635395
Sum squared resid	976.8201	Schwarz criterion	3.644651
Log likelihood	-802.4222	Hannan-Quinn criter.	3.639045
Durbin-Watson stat	2.433739		

Figura 23 - Teste aumentado de *Dickey-Fuller* para o produto C

No teste aumentado de *Dickey-Fuller* para a série mensal o valor de -2,87 é inferior aos valores críticos de 1%, 5% e 10%. Assim, considera-se possível rejeitar a hipótese nula não tendo um problema de raiz unitária e como tal estamos perante uma série estacionária. De igual forma,

os dados semanais no teste ADF apresentam um valor inferior (-11,64) para com os valores críticos de 1%, 5% e 10.

Assim sendo, considera-se não ser necessário realizar qualquer tipo de diferenciação, nem de remoção de tendência, com o intuito de proceder à realização dos modelos ARIMA. Foi posteriormente efectuada a análise da FAC e da FACP de forma a verificar qual o melhor modelo, tendo como base de comparação os critérios de ajustamento Akaike (AIC) e Schwartz (BIC).

Tabela 11 - Modelos ARIMA identificados para o produto C

Série Mensal		
Modelos	**AIC**	**BIC**
0,0,0	**443,06**	**448,11**
0,0,1	442,59	450,15
0,0,2	444,34	454,43
0,0,3	446,33	458,94
1,0,0	442,94	450,51
1,0,1	444,40	454,49
1,0,2	446,34	458,95
1,0,3	448,32	463,45

Série Semanal		
Modelos	**AIC**	**BIC**
0,0,0	**1276,53**	**1284,50**
0,0,1	1278,44	1290,41
0,0,2	1276,79	1292,74
0,0,3	1278,58	1298,53
1,0,0	1278,43	1290,39
1,0,1	1277,92	1293,88
1,0,2	1278,59	1298,54
1,0,3	1280,04	1303,97

Verifica-se através da Tabela 11 que para o produto C, na série mensal, o modelo que apresenta melhores resultados, ou seja critérios de ajustamento AIC e BIC é o modelo ARMA(0,0,0) com valores de 443,06 e 448,11, respectivamente. É possível, ainda, verificar que os modelos ARMA(0,0,1) e ARMA(1,0,0) apresentam também valores bastante reduzidos e chegam mesmo a ter AIC inferior, no entanto, quando conjugado com o BIC entende-se que são modelos que acabam por ser necessário maior ajustamento face ao modelo ARMA(0,0,0).

No que toca à série dos dados semanais, verifica-se algo semelhante, isto é, o modelo que apresenta critérios de ajustamento AIC e BIC é o ARMA(0,0,0) como se pode verificar na tabela acima. O modelo ARMA(0,0,2) apresenta um valor AIC muito semelhante, no entanto acaba por ser necessário um maior ajustamento através do critério BIC, não sendo assim o modelo mais adequado.

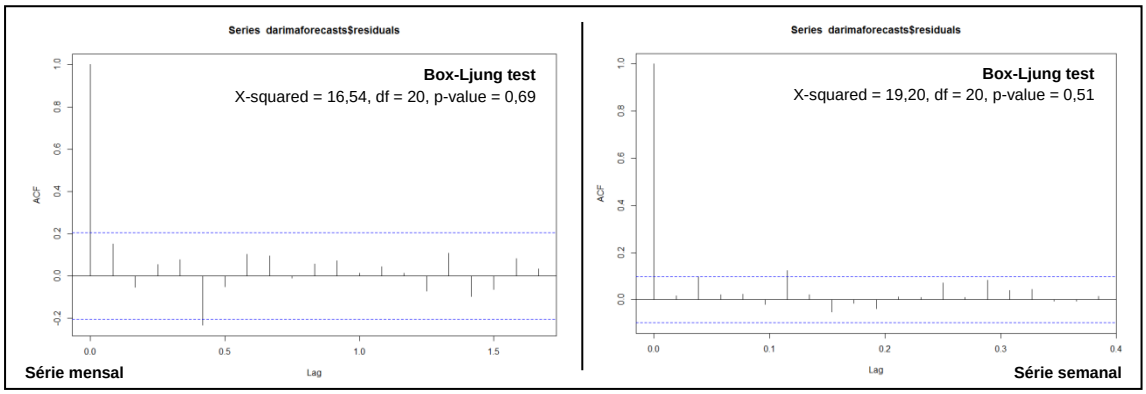


Figura 24 - Avaliação do modelo ARIMA para o produto C

Torna-se necessário após a definição dos modelos para cada uma das séries analisar os resíduos através do correlograma que se encontra na Figura 24. Ora, constata-se que para ambas as séries um dos valores de desfasamento ultrapassa os limites de significância, o valor 5 (-0,23), assumindo um valor negativo na série semanal e o valor 6 (0,12), sendo este positivo na série semanal. Realça-se, ainda, que para a segunda série (dados semanais) o valor de desfasamento 2 quase ultrapassa os limites, com um valor de 0,09. Assim é de esperar, para as duas séries, que 1 em cada 20 valores excedem os 95% dos limites de significância. Constata-se, ainda, que para as duas séries o teste Ljung-Box retorna um valor de 0,69 (dados mensais) e 0,51 (dados semanais) indicando assim que não se rejeita a nulidade das autocorrelações para o nível de significância (0,05) estabelecido nos erros de previsão para os valores de desfasamento 1-20.

5.5 Avaliação

De forma a avaliar os resultados obtidos das diversas modelagens tratadas no presente trabalho procurou-se comparar através das métricas de desempenho, nomeadamente os erros de previsão, e através da comparação dos resultados das previsões nos dois métodos.

Tabela 12 – Comparação dos métodos de previsão através de métricas de previsão

		Série mensal		Série semanal	
Prod.	**Métricas**	**Alisamento Exponencial**	**Box-Jenkins**	**Alisamento Exponencial**	**Box-Jenkins**
A	MAE	**5,71**	6,57	2,44	**2,30**
	MSE	**46,41**	75,94	8,85	**7,82**
	RMSE	**6,81**	8,71	2,98	**2,80**
B	MAE	**4,34**	6,07	**2,26**	3,35
	MSE	**35,44**	52,35	**7,02**	15,36
	RMSE	**5,95**	7,24	**2,65**	3,92
C	MAE	**2,71**	3,41	**1,04**	1,07
	MSE	**11,24**	17,17	**1,62**	2,07
	RMSE	**3,35**	4,14	**1,27**	1,44

Através da análise da Tabela 12, são comparadas as métricas definidas para esta dissertação de forma a comparar qual dos dois métodos, Alisamento Exponencial ou *Box-Jenkins*, consegue explicar melhor um modelo de previsão de vendas para produtos farmacêuticos tendo em consideração as vendas mensais e semanais, dos três produtos analisados. Foram, ainda, colocados a negrito e cor azul os valores dos erros que são mais diminutos e como tal evidenciam o método com melhores resultados e com melhor desempenho. A Figura 25 pretende, igualmente, demonstrar a comparação entre os erros de previsão, mas de forma visual e como tal de forma mais imediata.

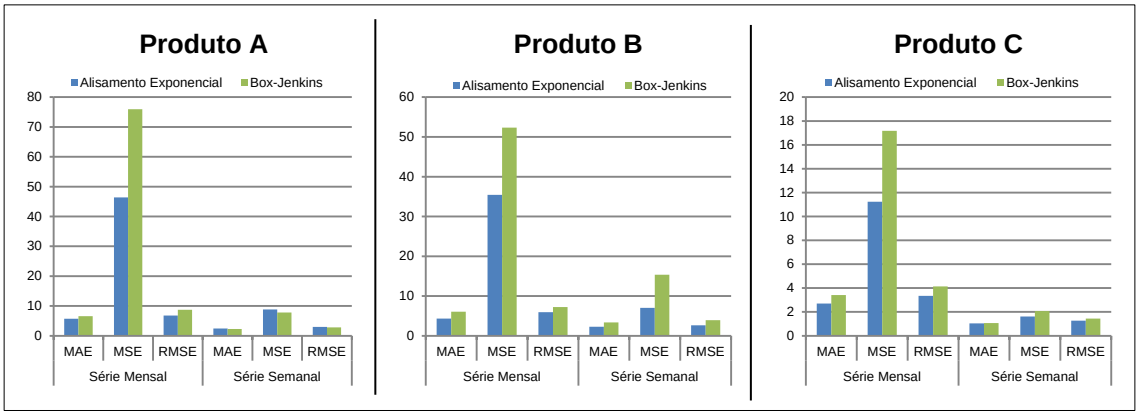


Figura 25 - Comparação e avaliação através de métricas de desempenho

Para o produto A, é possível verificar que não existe apenas um único modelo com bom desempenho na previsão de vendas para as duas escalas temporais analisadas. Ou seja, o modelo de alisamento exponencial apresenta melhores resultados quando analisada a série mensal e o modelo *Box-Jenkins* destaca-se para as vendas quando analisadas do ponto de vista semanal. Salienta-se, ainda, que ambos os modelos apresentam valores muito semelhantes na série semanal, contudo o MSE através de *Box-Jenkins* apresenta um valor de 7,82, mais reduzido que no outro método com 8,85. As métricas RMSE e MAE apresentam igualmente valores muito próximos. Nos dados mensais, os modelos destacam-se ligeiramente mais, sendo o método de alisamento exponencial o mais robusto, conforme se pode verificar através do valor MSE (46,41) destacando-se do segundo método para 75,94. Os valores MAE (5,71) e RMSE (6,81) não são tão evidentes, mas ainda assim evidenciam o método de alisamento exponencial, em detrimento do segundo com 6,57, e 8,71, respectivamente.

Para o produto B, o modelo que se evidencia como sendo mais interessante é o modelo de alisamento exponencial, tanto para os dados mensais como semanais. Conforme se pode verificar, o MSE (35,44), na série mensal, é claramente mais reduzido, e nos dados semanais este erro atinge valores de menos de metade com 7,02 quando comparado com o método de *Box-Jenkins*, com um erro de15,36.

Tal avaliação, também se verifica no que toca ao produto C, sendo que apesar do método de *Box-Jenkins* apresentar erros de previsão bastante reduzidos, o método de alisamento exponencial consegue ser ainda melhor. Assim, destacam-se os erros MAE, MSE e MSE, na série mensal, com 2,71, 11,24 e 3,35 face aos valores de 3,41, 17,17 e de 4,14 para o método em comparação. Também, para a série semanal, o método que apresenta melhores resultados, tem valores muito reduzidos com MSE de 1,62, o que denota um modelo de elevada qualidade.

É, ainda, de destacar o facto do erro MSE, para os dados mensais, ser sempre bastante mais reduzido no método de alisamento exponencial, conforme se pode ver através da Figura 25, bem como o facto dos erros nas observações semanais apresentar, para os três erros em análise, sempre valores bastante reduzidos, o que significa que é possível prever valores

futuros, com esta periodicidade, com menor probabilidade de erro. Realça-se, ainda, o facto do método de alisamento exponencial apresentar resultados mais satisfatórios em cinco das seis séries temporais em análise, sendo que na única série que apresenta erros mais elevados, esses mesmos erros de previsão são apenas ligeiramente superiores ao método em comparação, o método *Box-Jenkins*.

5.6 Implementação

O presente trabalho de investigação tem de entre um conjunto de princípios definidos, o de obter bons modelos em termos das suas capacidades de previsão, traduzindo-se em conhecimento e consequentemente que venham trazer mais-valias na gestão de medicamentos e produtos diversos, comercializados numa farmácia comunitária. Idealmente, estes modelos seriam integrados num Sistema de Apoio à Decisão, aplicado em ambiente real, de modo a apoiar na gestão de medicamentos (e.g. encomendas, *stocks*) destas farmáticas. Todavia, devido a restrições temporais, não foi possível tratar desta fase do CRISP-DM., ficando apenas referenciada aqui para trabalho futuro.

6. Conclusões

No presente capítulo pretende-se efectuar uma síntese de todo o trabalho realizado, bem como os impactos que este trabalho poderá trazer para este tipo de negócio. Pretende-se, igualmente, fazer levar à discussão a possibilidade e a pertinência da aplicação de técnicas de *Data Mining* em farmácias para a previsão de vendas e respectiva gestão de medicamentos, bem como os diversos objectivos que pretendem atingir.

6.1 Síntese

Portugal atravessa actualmente uma época de crise económica, não apenas por consequências de carácter nacional como também internacional, sofrendo mesmo uma intervenção do Fundo Monetário Internacional (FMI)[21], havendo por isso inúmeras dificuldades financeiras. Tais dificuldades têm levado a um conjunto de restrições orçamentos por parte do Ministério da Saúde e a uma regulação mais acentuada por parte da autoridade nacional do medicamento e produtos de saúde de forma a reduzir os custos de medicamentos para os cidadãos, levando assim a uma menor rendibilidade nesta área de negócio.

Devido a tais dificuldades torna-se necessário uma reavaliação do *modus operandi* das farmácias comunitárias pelo menos a dois níveis distintos, sendo o primeiro relativo à necessidade de reorganizar os actuais serviços e tomar acções proactivas junto dos clientes através de formas diversas, com o intuito de cativar e fidelizar os mesmos. O segundo prende-se essencialmente com a necessidade de uma reanálise periódica das necessidades da farmácia e dos seus clientes no sentido de reduzir a despesa. Essa redução poderá efectuar-se através da optimização dos níveis de *stocks*, num melhor controlo das compras junto dos diversos fornecedores, na aposta de volumes de compras de acordo com a previsão de vendas efectivas, evitando o investimento elevado e eventualmente redireccionar esses investimentos para outro tipo de medicamentos ou produtos de saúde, diversificando desta forma o seu portefólio de negócio, ou simplesmente reduzir o investimento e aumentar a liquidez de forma a fazer face aos actuais tempos conturbados.

Ora, é neste segundo nível que se pretende focar este trabalho de investigação, uma vez que pretende aplicar um conjunto de técnicas de *Data Mining* de forma a verificar a possibilidade de criar modelos capazes de apoiar na previsão de venda de medicamentos e outros produtos de saúde e como tal melhorar a gestão destes mesmos produtos, traduzindo em conhecimento útil e poder ser uma ferramenta de apoio à decisão numa altura em que todas as empresas necessitam de todos os mecanismos de forma a fazer face à conjuntura actual. Para auxiliar no

[21] Organização internacional que pretende assegurar o bom funcionamento do sistema financeiro mundial através de assistência técnica e financeira. Para mais informações poderá consultada a seguinte de Internet: http://www.imf.org

prosseguimento deste trabalho adoptou-se a metodologia CRISP-DM e optou-se pela ferramenta estatística R com recurso às bibliotecas *forecast* (Hyndman & Khandakar, 2008) e *RMiner* (Cortez, 2010).

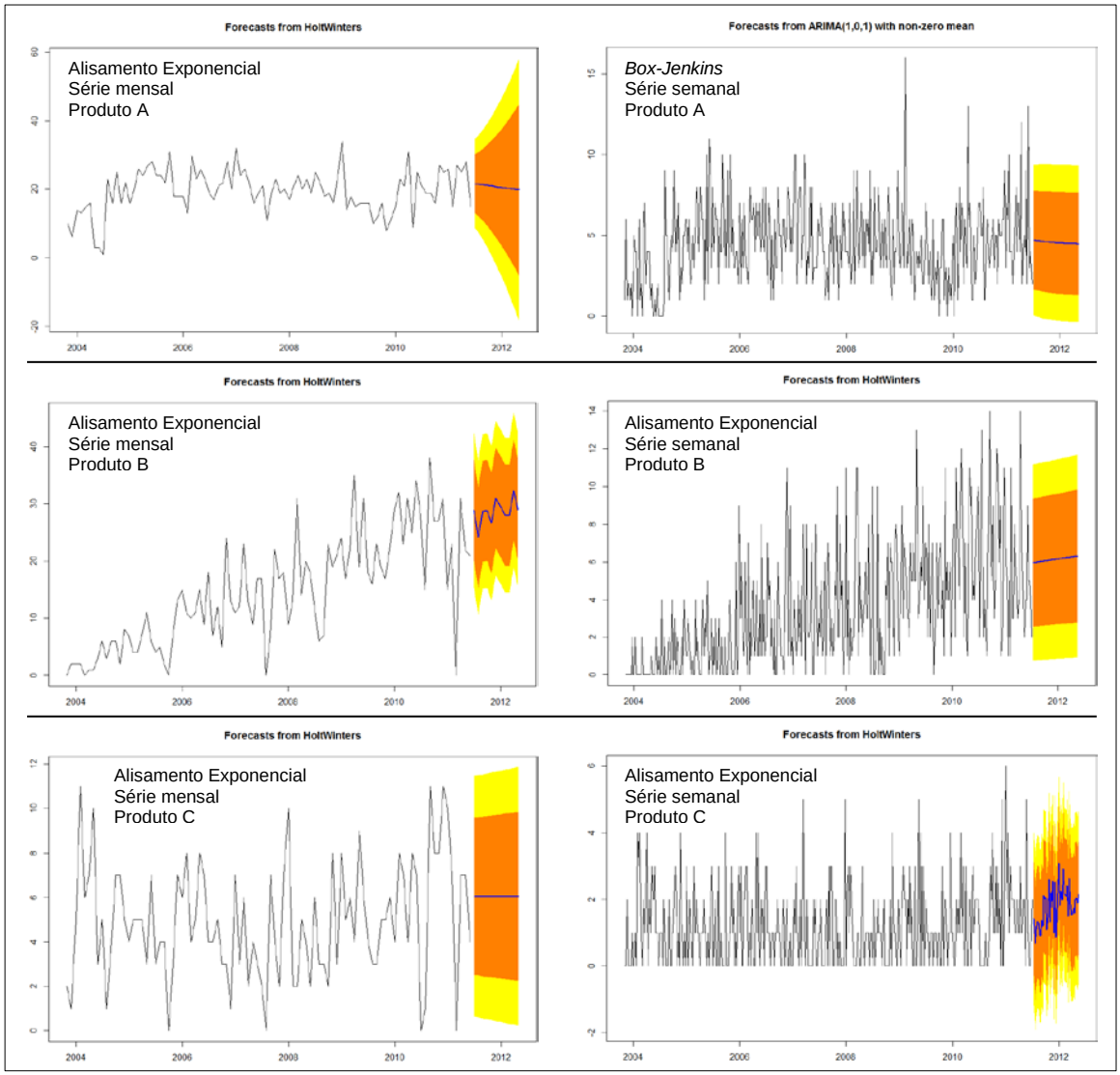


Figura 26 – Previsão das séries temporais

Através da Figura 26 é possível ter a percepção do tipo de previsão efectuado pelos diferentes modelos testados, para cada uma das 6 séries temporais, relativas aos 3 produtos em análise, com observações de vendas relativas a dados mensais e semanais. A mesma figura serve também como um exemplo daquilo que poderia surgir no sistema de apoio à decisão a ser implementado em farmácias.

É possível também averiguar que as séries semanais permitem a criação de modelos com resultados mais robustos e como tal de melhor qualidade. Tendo, também, em consideração os dados presentes na Tabela 12, os erros previstos nestas séries são relativamente reduzidos, sendo por isso expectável que a previsão de dados futuros possa ser realizada com maior precisão e exactidão.

Dá-se particular destaque à série semanal do produto C, em que o modelo adoptado apresenta erros de previsão muito diminutos e responde com bastante rigor e adaptação às componentes tendência e sazonal, nomeadamente através do modelo de alisamento exponencial Holt-Winters. Realça-se, uma vez mais, o facto do método de alisamento exponencial apresentar resultados mais satisfatórios em cinco das seis séries temporais em análise, sendo que na única série que apresenta erros mais elevados, esses mesmos erros de previsão são apenas ligeiramente superiores ao método em comparação, o método *Box-Jenkins*.

6.2 Discussão

No seguimento do já exposto e tendo em consideração o actual contexto económico e social, onde se verifica uma crise financeira a nível não apenas nacional como também internacional, torna-se imperativo a redefinição de objectivos, e o redesenho dos paradigmas de gestão das farmácias comunitárias, na medida em que actualmente se prevêem enormes dificuldades no sector, levando mesmo ao fecho de muitas destas entidades, e verificando-se já novas directrizes de gestão impostas e outras sugeridas pelas autoridades nacionais. É, ainda, de realçar o facto de se tratar de um sector muito regulamentado, com um mercado muito particular e com regras muito definidas e até mesmo com valores e uma cultura de trabalho muito solidificada e de difícil mudança.

É neste contexto que a necessidade de novas adaptações é necessária, e a aposta em soluções que permitam uma maior e melhor gestão, nomeadamente ao mais alto nível, se verifica como uma mais-valia no actual contexto. Assim, a aplicação de técnicas de *Data Mining* como elemento de diferenciação e de apoio à decisão poderão trazer uma melhoria da eficiência, nomeadamente através da gestão de medicamentos e produtos de saúde.

Ao terminar este trabalho considera-se pertinente confrontar, adoptando uma postura crítica, os resultados obtido com os objectivos inicialmente propostos. A base partiu da análise de um caso de estudo relativa à análise de vendas, com periodicidade mensal e semanal, de um conjunto de produtos comercializados numa farmácia comunitária. Os objectivos principais desta dissertação eram a verificação da aplicação de técnicas de *Data Mining* neste tipo de sector e possibilitar a melhoria de gestão de medicamentos e produtos de saúde, melhoria esta associada à gestão de *stocks*.

No final desta investigação pode-se concluir que as ferramentas utilizadas, nomeadamente a plataforma R, com recurso às bibliotecas *forecast* e *RMiner* são adequadas e robustas para aplicação em casos reais, de trabalhos desta natureza. As ferramentas utilizadas são *open source,* traduzindo-se assim numa vantagem, e a ferramenta R conta já com uma grande comunidade de apoio para o seu desenvolvimento e manutenção. Sem prejuízo do já

mencionado, foi possível ter acesso a uma versão paga da ferramenta *EViews* 7, tendo sido utilizada apenas para fins muito específicos de apoio à metodologia *Box-Jenkins*.

Verificaram-se também algumas desvantagens, nomeadamente no elevado volume de produtos comercializados pela farmácia analisada e como tal foi necessário definir um conjunto de critérios para melhor definir aqueles pudessem trazer mais-valias. Durante a realização deste trabalho constatou-se que um dos produtos (possível produto D) deixou de ser uma aposta para venda, por decisão da Direcção Técnica, em meados de Julho de 2011, e como tal não possível dar seguimento à análise do mesmo. Tendo em consideração as limitações temporais para a conclusão deste trabalho, não foi possível seleccionar e preparar dados relativos a um novo produto. De qualquer forma, pensa-se que tal não comprometeu os objectivos inicialmente propostos e resultados esperados, na medida em que foi possível estudar três produtos distintos, e um total de seis séries temporais. Uma outra desvantagem a referir é o facto da ferramenta R não possuir um interface gráfico, o que torna a curva de aprendizagem maior quando comparado com outras aplicações gráficas.

6.3 Limitações e Trabalho Futuro

Chegada a conclusão do presente trabalho importa referir as limitações que condicionaram o mesmo, bem como indicar alguns caminhos que permitam nova investigação nesta matéria.

Infelizmente, o tempo foi um dos principais factores condicionantes para que se pudesse alargar a investigação a outras técnicas, nomeadamente mediante análise outro tipo de métodos de previsão, tais como Redes Neuronais Artificiais ou Máquinas de Vetores de Suporte, tirando assim um maior partido da biblioteca *RMiner*. Consequentemente, não foi possível comparar estes métodos com os estudados (Alisamento Exponencial e *Box-Jenkins*).

Relativamente a trabalho futuro, descrevem-se aqui três sugestões:

1. Implementar os modelos propostos em ambiente real, integrando-os num sistema de apoio à decisão amigável, e como tal verificar o impacto e as mais-valias no negócio;

2. Alargar as análises efectuadas a um conjunto de produtos mais variado;

3. Explorar outros métodos de previsão, tais como Redes Neuronais e Máquinas de Vectores de Suporte.

Bibliografia

Adriaans, P. & Zantinge, D., 1996. *Data Mining.* Pearson Education: Addison-Wesley.

Agrawal, R. & Srikant, R., 1994. Fast Algorithms for Mining Association Rules. *Proc 20th Int Conf Very Large Data Bases VLDB*, Volume 1215, pp. 487-499.

Anon., 2007. *Decreto-Lei n.º 307/2007 de 31 de Agosto. Diário da República, 1.ª série — N.º 168.* Lisboa: Ministério da Saúde.

Azevedo, A. & Santos, M. F., 2008. KDD, SEMMA and CRISP-DM: A parallel overview. pp. 182-185.

Azevedo, C. S. & Santos, M. F., 2005. *Data Mining: Descoberta de Conhecimento em Bases de Dados.* Lisboa: FCA - Editora Informática.

Boisot, M. & Canals, A., 2004. Data, Information and Knowledge: Have We Got It Right?. *Journal of Evolutionary Economics*, Volume 14, pp. 43-67.

Box, G., Jenkins, G. M., Reinsel, G. & Reinsel, G., 1994. *Time Series Analysis: Forecasting and Control.* 3ª ed. Englewood Cliffs, New Jersey: Prentice Hall.

Brito, C. M., Ramos, C. & Carvalho, P., 2006. Parcerias no Negócio Electrónico. In: *Parcerias no Negócio Electrónico.* Porto: Sociedade Portuguesa de Inovação, pp. 64-65.

Brockwell, P. J. & Davis, R. A., 2002. *Introduction to Time Series and Forecasting.* 2ª ed. New York, USA: Springer.

Campos, C. V. C., 2009. *Previsão da Arrecadação de Receitas Federais: Aplicações de Modelos de Séries Temporais para o Estado de São Paulo.* Ribeirão Preto, Brasil. Dissertação de Mestrado, Universidade de São Paulo.

Chapman, P. et al., 2000. *CRISP-DM 1.0 Step-by-step data mining guide.* s.l.:The CRISP-DM consortium.

Cortez, P., 2010. *Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. In P. Perner (Ed.), Advances in Data Mining - Applications and Theoretical Aspects.* Berlin, Germany, Springer, pp. 572-583.

Davis, B. D., 1974. *Management Information Systems: Conceptual foundations, Structure and Development.* New York: McGraw-Hill Book Company.

Davis, G. B. & Olson, M., 1986. *Information Systems Concepts for Management.* 3ª ed. s.l.:McGraw-Hill International Editions.

Dias, R. M. G. M., 2009. *Implementação de modelos de Previsão da Procura em combustíveis petrolíferos: O caso da Companhia Logística de Combustíveis, S.A..* Lisboa. Dissertação de Mestrado, Instituto Superior Técnico, Universidade Técnica de Lisboa.

Ein-Dor, P., 1985. *Information Systems Management.* Amsterdam: Elsevier Science Publishing Co. Inc..

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases. *AI MAGAZINE*, pp. 37-54.

Gonçalves, A., Pereira, D., Pinheiro, E. & Aguiar, P., 2008. Inteligência Artificial & Data Mining: Projeto de Sistema de Apoio à Decisão. In: Rio de Janeiro: s.n.

Gordon, S. R. & Gordon, J. R., 1999. *Information Systems – A Management Approach.* 2ª ed. Orlando FL: Fort Worth - The Dryden Press.

Hamuro, Y., Katoh, N., Matsuda, Y. & Yada, K., 1998. Mining Pharmacy Data Helps to Make Profits. *Data Mining and Knowledge Discovery*, Volume 2, pp. 391-398.

Han, J., Kamber, M. & Pei, J., 2011. *Data Mining: Concepts and Techniques.* 3ª ed. Waltham: Morgan Kaufmann.

Hyndman, R. & Athanasopoulos, G., 2012. *Forecasting: principles and practice.* [Online]
Disponível em: http://otexts.com/fpp/
[Acedido em 1 Outubro 2012].

Hyndman, R. J. & Khandakar, Y., 2008. Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software,* Volume 27.

Lakatos, E. M. & Marconi, M. d. A., 1992. *Metodologia do trabalho científico.* 4ª ed. São Paulo: Editora Atlas SA.

Lemos, F. d. O., 2006. *Metodologia para seleção de métodos de previsão de demanda.* Porto Alegre, Brasil. Dissertação de Mestrado, Universidade Federal do Rio Grande do Sul.

Maimon, O. & Rokach, L., 2010. *Data Mining and Knowledge Discovery Handbook.* 2ª ed. London: Springer.

Makridakis, S. G., Wheelwright, S. C. & Hyndman, R. J., 1998. *Forecasting: Methods and Applications.* 3rd Ed. ed. New York, USA: Wiley.

Olson, D. L. & Delen, D., 2008. *Advanced Data Mining Techniques.* s.l.:Springer.

Piatetsky-Shapiro, G., 2011. *What data types you analyzed/mined in the past 12 months?.* [Online]
Disponível em: http://www.kdnuggets.com/polls/2011/data-types-analyzed-mined.html [Acedido em 22 Janeiro 2012].

Piatetsky-Shapiro, G., 2012. *Inquérito online: What Analytics, Data mining, Big Data software you used in the past 12 months for a real project (not just evaluation).* [Online] Disponível em: http://www.kdnuggets.com/2012/05/top-analytics-data-mining-big-data-software.html
[Acedido em 7 Setembro 2012].

Quivy, R. & Campenhoudt, L. V., 1998. *Manual de Investigação em Ciências Sociais.* Lisboa: Gradiva.

Raisinghani, M., 2004. *Business Intelligence in the Digital Economy: Opportunities, Limitations and Risks.* s.l.:Idea Group Publishing.

Rexer, K., Allen, H. & Gearan, P., 2011. *5th Annual Data Miner Survey. 2011 Survey Summary Report,* Winchester, USA: s.n.

Robert, J. T., 1996. Estimating demand for services: issues in combining sales forecasts. *Journal of Retailing and Consumer Services*, Outubro, Volume Volume 3.

Sabhnani, M., Neill, D. & Moore, A., 2005. Detecting Anomalous Patterns in Pharmacy Retail Data. *Data Mining Methods for Anomaly Detection*, p. 58.

Shumway, R. H. & Stoffer, D. S., 2011. *Time Series Analysis and Its Applications: With R Examples.* 3ª ed. New York, USA: Springer.

Silva, L. S., Lima, R. S., Moreira, F. C. & Monteiro, O. L., 2007. Descoberta de Padrões Relevantes na Base de Dados da Farmácia Homeobel Center.

Tuomi, I., 1999. Data is more than knowledge: Implications of the reversed knowledge hierarchy for knowledge management and organizational memory. Volume 16, pp. 107-121.

Turban, E., Aronson, J. E., Sharda, R. & King, D., 2007. *Business Intelligence: a Managerial Approach.* s.l.:Prentice Hall.

Zuur, A., Ieno, E. N. & Meesters, E., 2009. *A Beginner's Guide to R (Use R!).* 1ª ed. New York, USA: Springer.