Instituto Universitário de Lisboa

Departamento de Ciências e Tecnologias da Informação

# Previsão de falhas em infraestruturas e equipamentos com recurso a técnicas de *Data Mining*

José Rafael Lopes Silva Charrua

**Dissertação submetida ao ISCTE/IUL para a obtenção do grau de Mestre em Gestão de Sistemas de Informação, na área de conhecimento de *Data Mining*/*Business Intelligence***

Orientador:
**Professor Doutor Paulo Cortez, Universidade do Minho**

Co-orientador:
**Professor Doutor Pedro Ramos, ISCTE-IUL**

Outubro de 2013

ISCTE IUL Instituto Universitário de Lisboa
Previsão de falhas em infraestruturas e equipamentos com recurso a técnicas de *Data Mining*
José Rafael Charrua
Outubro 2013

# Agradecimentos

O longo caminho que foi a elaboração desta dissertação só poderia ter sido percorrido com a colaboração, apoio e motivação de inúmeras pessoas. Apresento assim os meus mais sinceros agradecimentos a todos os que, direta ou indiretamente, colaboraram para a sua elaboração.

Ao orientador deste trabalho, o Prof. Doutor Paulo Cortez, sempre disponível, apesar da distância física, e respondendo sempre a todas as solicitações, de forma ativa e preocupada. Sem a sua orientação, esclarecimentos e palavras de apoio e motivação este trabalho não teria sido concluído.

Ao Prof. Doutor Pedro Ramos, pela forma como me motivou e apoiou para o desenvolvimento deste trabalho, desde o primeiro instante.

Ao colega da ANA – Aeroportos de Portugal, SA, Eng.º Nuno Duarte, pelo interesse que sempre demonstrou na realização deste trabalho e pela disponibilização dos dados e condições necessários à sua realização.

Aos meus amigos, pelo incentivo e motivação.

Aos meus Pais e irmão Bruno, pelo incondicional apoio em tudo, ao longo da vida.

Aos meus filhos, Matilde e Manuel, por terem prescindido do seu pai muitas vezes, para que este trabalho se pudesse desenvolver, e pela inspiração que são todos os dias.

E a ti, Susana, minha mulher de sempre e para sempre, que muito me ajudaste e incentivaste. Por isso, e por tudo, a ti este trabalho eu dedico.

# Resumo

Um Aeroporto internacional reúne um conjunto complexo de instalações, equipamentos e serviços, de difícil gestão e que funciona permanentemente, sendo premente a sua disponibilidade em condições de operação e utilização. A previsão de falhas e avarias reveste-se de uma grande importância para o aumento da disponibilidade da infraestrutura, bem como para uma gestão eficiente dos recursos humanos alocados à manutenção. Com a previsão de indisponibilidade de equipamentos ou infraestruturas poder-se-ão ainda tomar decisões atempadas e eficazes para a realização de manutenção preventiva, bem como para a gestão dos fluxos de passageiros ou até mesmo de aeronaves.

Pretende este trabalho demonstrar a fiabilidade de um modelo de previsão de avarias numa infraestrutura aeroportuária através de técnicas de *Data Mining* aplicadas à base de dados do Sistema de Gestão de Manutenção do Aeroporto de Lisboa. Para além das questões de ordem técnica e operacional, existe outro problema interessante de responder, que é o de demonstrar aos gestores a importância das previsões/modelos criados para a organização em que se insere o estudo de caso. Se os modelos tiverem qualidade suficiente e se a organização se decidir pela sua incorporação, poder-se-á posteriormente medir o impacto da sua aplicação.

**Palavras Chave:** Gestão de Manutenção, *Business Intelligence, Data Mining,* previsão, Aeroportos.

# Abstract

An international airport encloses a complex set of facilities, equipments and services which are difficult to manage and have to operate permanently. The accessibility of operating and usage conditions is of utmost importance. Forecasting failures and malfunctions, to increase the infrastructure's performance, as well as managing the human resources allocated to maintenance efficiently is a matter of major significance. By forecasting the equipment or infrastructure unavailability it is also possible to make timely and efficient decisions to perform a preventive maintenance, and also to manage the passengers and even aircrafts' flow.

This work demonstrates the dependability of a malfunction forecast model for an airport infrastructure, through *Data Mining* techniques, applied to the Lisbon Airport Maintenance Management System data base. Besides the issues of technical and operational nature, there is another interesting problem which is to demonstrate the managers the importance of forecasts/models, created for the company used in the case study. If the models have sufficient quality and if the company decides on its incorporation, then it will be possible to measure the impact of its application.



**Key-Words:** Maintenance Management, Business Intelligence, *Data Mining*, forecast, Airports.

# Índice

# Índice de Figuras

# Índice de Tabelas

# Lista de Abreviaturas

**ANA:** empresa ANA – Aeroportos de Portugal, SA (ANA)

**ALS:** Aeroporto de Lisboa (segundo a nomenclatura em vigor na ANA)

**BI:** *Business Intelligence*

**CMT:** Centro de Monitorização Técnica

**CRISP-DM:** *Cross Industry Standard Process for Data Mining*

**DM**: *Data Mining*

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

**KDD:** *Knowledge Discovery in Databases*

**MAE:** *Mean Absolute Error*

**MC:** Manutenção Corretiva

**MLPE:** *Multilayer Perceptron*

**MP:** Manutenção Preventiva

**MR:** *Multiple Regression*

**MSE:** *Mean Squared Error*

**NGM:** Núcleo de Gestão de Manutenção

**NN:** *Neural Networks*

**OT:** Ordem de Trabalho

**REC:** *Regression Error Characteristic*

**RMSE:** *Root Mean Squared Error*

**SEMMA:** *Sample, Explore, Modify, Model, Assess*

**SGM:** Sistema de Gestão de Manutenção

**SVM:** *Support Vector Machines*

# 1. Introdução

A previsão de falhas e avarias é de grande importância para aumentar a disponibilidade das infraestruturas e equipamentos aeroportuários, bem como para uma gestão eficiente dos recursos humanos alocados à manutenção. Com a previsão de indisponibilidade de equipamentos ou infraestruturas poder-se-ão tomar decisões atempadas para a realização de manutenção preventiva, bem como para uma melhor gestão dos fluxos de passageiros e de aeronaves. Pretende-se com este trabalho desenvolver uma ferramenta que auxilie na previsão de falhas em infraestruturas e equipamentos importantes através de técnicas de *Data Mining* aplicados à base de dados do Sistema de Gestão de Manutenção do Aeroporto de Lisboa. Para auxiliar na condução deste projeto, foi adotada a metodologia CRISP-DM, adiante detalhada.

## 1.1 Enquadramento

A empresa ANA – Aeroportos de Portugal, SA (ANA) é uma empresa que até Setembro de 2013 pertencia ao Setor Empresarial do Estado. Esta empresa, à qual foi atribuída a exploração do serviço público aeroportuário de apoio à aviação civil, foi entretanto privatizada, estando neste momento com 100% de capital privado. A ANA tem então a exploração dos Aeroportos de Lisboa, Porto e Faro, bem como do terminal civil de Beja, e ainda os Aeroportos de Ponta Delgada, Santa Maria, Horta e Flores na Região Autónoma dos Açores e os Aeroportos da Madeira e Porto Santo, na Região Autónoma da Madeira. O portfólio de negócios da ANA compreende as áreas de negócios de Aviação e Não Aviação, sendo que o negócio Aviação é regulado e consiste na gestão de infraestruturas para o tráfego de aeronaves, passageiros e carga. O negócio Não Aviação, tendencialmente não regulado, respeita à exploração de espaços comerciais e publicitários nos Aeroportos e na oferta de imóveis, de parques de estacionamento e de espaços para estações de *rent-a-car* nas imediações dos Aeroportos (ANA – Aeroportos de Portugal, SA, 2011).

De modo a dar uma dimensão do tráfego do Aeroporto de Lisboa, apresenta-se a Tabela 1, que contém o resumo dos dados referentes ao número de passageiros e movimentos de aeronaves no ano de 2012, para a totalidade dos Aeroportos ANA e também, em particular, para o Aeroporto de Lisboa (ALS).

Tabela 1 - Número de passageiros e de movimentos de aeronaves em 2012 no Aeroporto de Lisboa

| ANO DE 2012 | Total Aeroportos do Grupo ANA | Aeroporto de Lisboa |
|---|---|---|
| **Número de passageiros** | 30.515.564 | 15.301.176 |
| **Número de movimentos de aeronaves** | 280.346 | 140.909 |

Fonte: adaptado do Relatório de Gestão e Contas '12 da ANA – Aeroportos de Portugal, SA (www.ana.pt)

O autor desta dissertação é colaborador da empresa supra referida onde desempenha funções de Técnico Superior no Núcleo de Eletrónica, parte integrante da Divisão de Gestão Técnica e Manutenção do ALS. Esta Divisão tem a responsabilidade de manter em elevados níveis de disponibilidade e operacionalidade as infraestruturas e equipamentos do ALS.

Sendo a manutenção uma das atribuições da referida Divisão do ALS, é interessante a realização de um estudo de caso que possa auxiliar no planeamento e gestão da manutenção. Todas as incidências e trabalhos no âmbito desta atividade são registados na Base de Dados (BD) do Sistema de Gestão de Manutenção (SGM) do ALS, atualmente a aplicação Máximo 7, de onde são retirados diversos dados e indicadores essenciais para a gestão de manutenção.

Foi apresentada a ideia para este trabalho aos responsáveis do Núcleo de Gestão de Manutenção (NGM) corporativo da ANA, tendo o retorno sido muito positivo e permitido o acesso aos dados do SGM para os fins previstos.

## 1.2 Motivação

O tema anteriormente descrito reveste-se de grande interesse para a organização em causa. Através de uma investigação sobre as técnicas de *Business Intelligence* mais adequadas, e em particular de *Data Mining*, pretende-se extrair conhecimento útil a partir dos dados que já estão disponíveis no SGM do ALS, conhecimento esse que poderá auxiliar e melhorar a tomada de decisão.

Assim, e como resultado deste trabalho, existe um potencial para melhorar a gestão de infraestruturas e equipamentos aeroportuários, levando em última análise a uma redução de custos e a uma maior eficiência, sendo por isso de interesse para os gestores da área de manutenção e, consequentemente, para a globalidade da ANA.

Para o autor desta dissertação, este trabalho corresponde também à forma ideal de materializar (em termos práticos) os conceitos mais teóricos aprendidos durante a parte letiva do mestrado, numa área de aplicação que corresponde aos seus interesses profissionais. Como expectativa, aponta-se o facto de poder efetivamente causar impacto positivo na organização, conduzindo a benefícios diretos, bem como a possibilidade de despertar interesse académico por parte de outros estudantes ou autores interessados em conhecer o trabalho realizado no âmbito desta dissertação.

## 1.3 Objetivos

Esta dissertação pretende, através de um estudo de caso, demonstrar a fiabilidade de um modelo de previsão de avarias numa infraestrutura aeroportuária, tentando daí retirar benefícios para a atividade em causa.

Em particular, foram exploradas diversas técnicas de aprendizagem a partir de dados, avaliando-se a capacidade preditiva das mesmas. Depois de escolhido e validado o modelo, será dada resposta a um problema concreto da problemática da gestão de manutenção, que é o de prever com menor ou

maior certeza o número de falhaa ou avarias (ocorrências) e identificar quais os fatores internos e externos que mais contribuem para a ocorrência de falhas em equipamentos e infraestruturas.

Se uma organização que depende quase exclusivamente das suas instalações físicas para a sua atividade, tal como ocorre com o Aeroporto de Lisboa, conseguir determinar com alguma precisão quais os períodos em que poderá ter maiores incidências que provoquem falhas ou avarias, poderá planear melhor os seus trabalhos e recursos, sobretudo aqueles que estão alocados à manutenção.

Para além das questões de ordem técnica e operacional, foi analisado o impacto que as previsões/modelos criados tiveram dentro da própria organização em que se insere o estudo de caso.

Em suma, e de forma sistematizada, de modo a servir o propósito específico deste estudo de caso, temos os seguintes objetivos para esta dissertação: escolher a técnica de *Data Mining*/modelo mais adequada para o problema a abordar e criar modelos de previsão para o número de ocorrências de manutenção.

## 1.4 Organização

Esta dissertação, para além da presente introdução que pretende apresentar e enquadrar o conceito do trabalho, bem como as motivações que levaram ao seu desenvolvimento, apresenta-se dividida em mais 4 capítulos.

Desta forma, e excluindo o capítulo introdutório, a presente dissertação possui 4 capítulos organizados da seguinte forma:

**Capítulo 2 – Quadro teórico de referência: *Business Intelligence* e *Data Mining*** – neste capítulo será apresentado o levantamento realizado do quadro teórico de referência sobre os temas *Business Intelligence* e *Data Mining*, bem como das metodologias utilizadas nestas áreas de conhecimento, e em que passos estas se dividem. Apresenta-se ainda um breve estudo sobre as diferentes ferramentas disponíveis para a realização de trabalhos desta natureza.

**Capítulo 3 - Manutenção de infraestruturas e equipamentos** – apresentam-se neste capítulo as definições e ideias principais da temática da manutenção de infraestruturas e equipamentos, bem como da gestão da manutenção. São ainda apresentados trabalhos realizados no âmbito da manutenção/gestão da manutenção e que utilizem de alguma forma o *Business Intelligence* e/ou *Data Mining*.

**Capítulo 4 - Previsão de falhas em infraestruturas e equipamentos com recurso a técnicas de *Data Mining*** – após a contextualização e descrição do estudo de caso, é formulado o problema, bem como a forma da sua abordagem. Apresenta-se então neste capítulo o trabalho prático realizado, evidenciando a metodologia e ferramentas utilizadas, exibindo e analisando os resultados práticos obtidos.

**Capítulo 5 – Conclusões** – neste capítulo é realizada a síntese e discussão dos resultados obtidos, apresentadas as conclusões e as contribuições para futuros trabalhos.

# 2. Quadro teórico de referência: *Business Intelligence e Data Mining*

*Business Intelligence* e *Data Mining* são dois termos que frequentemente aparecem associados, sendo até por vezes confundidos. Se bem que ambos os termos estão associados às tecnologias e sistemas de apoio à decisão, a verdade é que não representam necessariamente o mesmo, pelo que adiante serão diferenciados.

Num mundo complexo e em que todo o tipo de ações ocorrem de forma muito rápida e dispersa, as organizações precisam de auxílio na definição da sua estratégia e consequente tomada de decisão. Urge tomar decisões precisas e rápidas, alinhadas com a estratégia das organizações e respondendo a pressões e fatores externos difíceis de prever e muito mais de controlar. Neste sentido, os sistemas de apoio à decisão poderão ser auxiliares preciosos e determinantes para a atividade das organizações

## *2.1 Business Intelligence*

Existem várias definições do termo *Business Intelligence* (BI), mais ou menos coincidentes. Segundo Moss *et al.* (2003), BI não é um produto nem um sistema, mas uma arquitetura e uma coleção de aplicações e bases de dados integradas de apoio à decisão, providenciando um acesso fácil aos dados necessários para um determinado decisor ou organização.

Uma definição mais completa e precisa é a apresentada por Michalewicz *et al.* (2007), que definem BI como um conjunto alargado de aplicações e tecnologias para recolher, armazenar, analisar e providenciar o acesso aos dados, definindo ainda o termo como um "estado" (um relatório que contém conhecimento) ou um "processo" (*software* responsável pela conversão de dados em conhecimento). Estes autores acrescentam ainda que o objetivo de muitos dos sistemas de BI são:

a) Aceder aos dados de uma diferente variedade de fontes;
b) Transformação dos dados em conhecimento; e
c) Providenciar um interface gráfico fácil de utilizar para a apresentação do conhecimento conseguido.

Como complemento, os autores desenharam ainda um diagrama que ilustra o processo dos sistemas tradicionais de BI, que a seguir se apresenta.

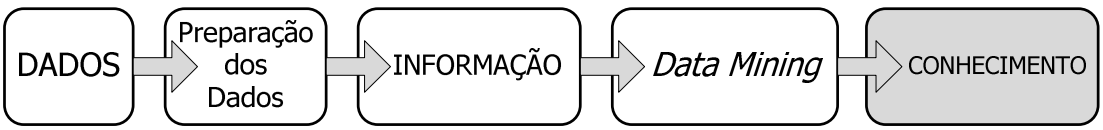


**Figura 1 – Processo de um sistema tradicional de *Business Intelligence*, segundo Michalewicz *et al.* (2007)**

Como se pode visualizar na Figura 1, para estes autores o *Data Mining* é referido como uma fase ou uma componente do processo de BI.

Esta definição é de certa forma confirmada e aumentada por Turban *et al.* (2008), segundo os quais o termo BI é agregador e sob o qual se enquadram arquitetura, ferramentas, bases de dados, aplicações e metodologias. Segundo estes autores, que citam Raisinghani (2004) e Zaman (2005), existe alguma confusão relacionada com o *Business Intelligence* e outras ferramentas, como o BPM – *Business Performance Management*. O objetivo do BI é o de permitir que exista um acesso interativo aos dados, fornecendo aos gestores e analistas formas de conduzir análises adequadas ao que desejam. Ao analisar dados actuais e históricos, indicadores e desempenhos, os decisores terão visões detalhadas sobre o seu negócio ou organização, levando a melhores decisões ou, pelo menos, a decisões com maior suporte de informação. Para terminar, estes autores referem que o processo de BI é baseado na transformação de dados em informação, seguida de decisão e, finalmente, ação.

## *2.2 Data Mining*

O termo *Data Mining* (DM) é originalmente descrito por Fayyad et al. (1996) e é referido como uma etapa no processo de descoberta de conhecimento em bases de dados (KDD)[1]. Segundo estes autores, DM consiste na aplicação de algoritmos específicos para extrair padrões a partir de grandes quantidades de dados.

Segundo Turban *et al.* (2008), DM é um termo utilizado para descrever a descoberta de conhecimento em bases de dados. É um processo que utiliza matemática e estatística, inteligência artificial e técnicas de aprendizagem de máquina, de modo a identificar e extrair informação útil para a produção de conhecimento. Anteriormente este termo descrevia o processo de identificação de padrões de dados mas, com o tempo, a definição original foi evoluindo para incluir múltiplos tipos de análise automática de dados. DM é, então, o processo de descoberta de padrões matemáticos em grandes conjuntos de dados. Estes padrões podem ser regras, afinidades, correlações, tendências ou modelos previsionais, estando algures entre as ciências da computação e a estatística, pois utilizam os avanços em ambas estas áreas da ciência para fazer progressos na extração de informação de grandes bases de dados.

Ainda segundo os mesmos autores, o DM é um campo que tem atraído muita atenção, pois é uma ferramenta muito útil para a extração de conhecimento, exploração de dados, processamento de dados, entre outras tarefas desta natureza, que permitem a descoberta de informação de forma rápida, mesmo para indivíduos não programadores.

A definição de uma disciplina científica é uma tarefa sempre controversa, tal como expresso por Hand *et al.* (2001), que apresentam o DM como a análise de conjuntos de dados observados, de modo a encontrar relações à partida insuspeitas, permitindo ainda o resumo dos dados em formas que sejam simultaneamente percetíveis e úteis para quem para eles olha. As relações e resumos que derivam

[1] Conhecido em inglês pelo termo: *Knowledge Discovery in Databases* (KDD).

de um exercício de DM são geralmente descritos como “modelos” ou “padrões”, sempre relativos a “dados observados”, por oposição a “dados experimentais”.

Outro ponto importante, apresentado por estes mesmos autores, é a noção de que o DM lida habitualmente com dados que foram previamente recolhidos com outro propósito que não o da própria análise de dados. Tal significa ique o *Data Mining* não representa um papel ativo na estratégia de recolha de dados.

Uma outra definição é a de Witten e Frank (2005), que afirmam que o DM é o processo de descoberta de padrões em dados. Este processo terá de ser automático ou, como mais habitualmente sucede, semiautomático. Os padrões descobertos deverão ter significado, levando a uma vantagem ou benefício.

Esta simples definição adiciona um dado importante às anteriores definições apresentadas, pois coloca como importante o “significado” dos resultados da aplicação do DM a um determinado conjunto de dados. Os padrões de dados deverão ser representados de forma a poderem ser examinados e utilizados para decisões futuras. Este tipo de padrões foram designados pelos autores como “estruturais”, pois capturam a estrutura de decisão de uma forma explícita, ou seja, contribuem para explicar algo sobre os dados.

Maimon e Rokach (2010) referem que o DM não é apenas um passo no processo de descoberta de conhecimento em bases de dados (KDD), mas sim a atividade principal deste processo, através de algoritmos que permitem explorar os dados, desenvolver modelos e descobrir padrões anteriormente desconhecidos. Ainda segundo estes autores, e dada a abundância de dados existente nos nossos dias, o DM é matéria de considerável importância e necessidade, apresentando o vasto leque de métodos e formas ao dispor dos investigadores e utilizadores.

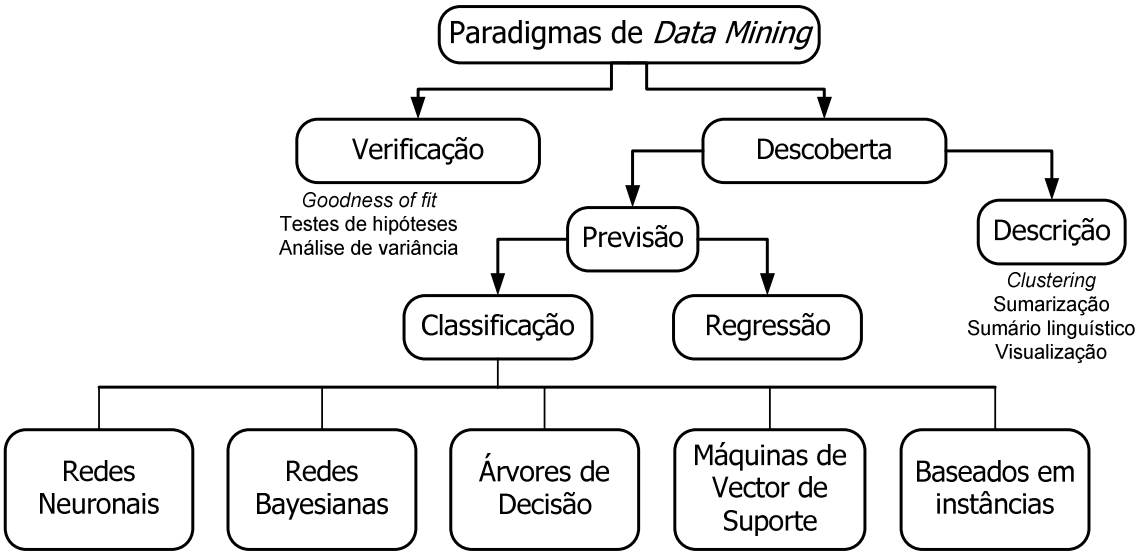


**Figura 2 – Taxonomia do *Data Mining*, segundo Maimon e Rokach (2010)**

Para alcançar diferentes objetivos, existem métodos variados de DM, pelo que se reveste de grande utilidade classificar ou definir uma taxonomia, distinguindo dois tipos principais de DM: orientados à verificação, em que o processo de DM verifica hipóteses formuladas, ou orientados à descoberta, em que o sistema encontra novos padrões e regras autonomamente. Na Figura 2 é apresentada a taxonomia proposta por estes autores.

Para a identificação automática de padrões nos dados, são utilizados métodos de descoberta. Este ramo dos paradigmas de DM é desenvolvido a partir de métodos de previsão ou de descrição. Os primeiros estão orientados para a interpretação dos dados, focando-se em perceber a forma como os dados se relacionam entre si. Os métodos de previsão procuram a construção automática de um modelo comportamental, que obtenha novas amostras. Desenvolvem também padrões, formando conhecimento descoberto, numa forma que seja percetível e utilizável. Alguns dos métodos de previsão permitem ainda ajudar à compreensão e interpretação dos dados.

Muitas técnicas de descoberta de DM, são baseadas em aprendizagem indutiva. Nestas técnicas, é construído um modelo, explícito ou implícito, através da generalização de um determinado número de exemplos de treino. O pressuposto para a utilização de técnicas indutivas, é o de que o modelo de treino é aplicável a exemplos futuros, desconhecidos do próprio modelo.

Para a validação de hipóteses propostas, por exemplo, por um perito, são utilizados métodos de verificação. Estes métodos incluem muitos dos métodos tradicionais de estatística, como os testes de hipóteses, análises de variância, entre outros. Estes métodos são menos utilizados no DM do que os métodos de descoberta, uma vez que muitos dos problemas de DM estão relacionados com a descoberta de uma hipótese, retirada de um número alargado de hipóteses, ao invés de testar uma nova hipótese.

Ainda segundo os mesmos autores, os métodos de previsão são descritos pela comunidade de aprendizagem por máquina como aprendizagem supervisionada, opondo-se à aprendizagem não supervisionada, que se refere à modelação de instâncias em que não há erro ou sinais que permitam avaliar uma solução potencial.

O conceito de aprendizagem não-supervisionada aplica-se sobretudo a técnicas que agrupam instâncias sem um atributo dependente, que não tenha sido previamente especificado. Este termo, aprendizagem não supervisionada, abrange apenas uma parte dos métodos descritivos da figura anterior, como por exemplo os métodos de *clustering,* mas não os métodos de visualização. Por outro lado, os métodos de aprendizagem supervisionada são aplicados na descoberta da relação entre vários atributos de entrada ou variáveis independentes e o atributo de saída por vezes designado como variável dependente. A relação descoberta através destes métodos é representada numa estrutura designada como um modelo. Normalmente, os modelos descrevem e explicam fenómenos que estão de certa forma escondidos no conjunto de dados, e podem ser utilizados para prever o valor do atributo de saída conhecendo os valores das variáveis independentes. Os métodos supervisionados podem ser implementados numa grande variedade de domínios, como no *marketing*, finanças ou produção industrial. Torna-se ainda útil distinguir entre dois grandes modelos de

aprendizagem supervisionada: modelos de classificação e de regressão. Os modelos de regressão mapeiam os atributos de entrada num domínio de valores reais. A título de exemplo, um modelo de regressão pode prever a procura de um determinado produto, dadas as suas características. Por outro lado, os modelos de classificação mapeiam as entradas em classes pré-definidas. Por exemplo, os modelos de classificação podem ser utilizados para classificar os clientes de créditos bancários como bons (pagam o empréstimo a tempo) ou maus (pagam com atraso), ou num outro qualquer número de classes necessárias. Exemplos de técnicas são máquinas de vetores de suporte ou redes neuronais. No ponto 2.5, são descritas em maior detalhe estas e outras técnicas.

Relativamente aos algoritmos de *Data Mining,* os mesmos poder-se-ão dividir em quatro grandes categorias: classificação, segmentação, associação e descoberta de sequências. Para além destas categorias, existem diversas ferramentas para se realizar a análise de dados, nomeadamente a visualização, regressão e análise de séries temporais (Turban, et al., 2010).

Segundo os mesmos autores, os algoritmos de classificação serão dos mais utilizados no *Data Mining e* consistem na análise de conjuntos de dados históricos, gerando automaticamente modelos que possam prever comportamentos futuros. Este modelo generaliza os dados de treino que irão apoiar na distinção de novas definições de classes. O conceito assenta na esperança de que o modelo possa ser utilizado para prever classes de registos não classificados. Alguns dos métodos de *Data Mining* mais utilizados para este tipo de dados são as redes neuronais, árvores de decisão e outros métodos com estruturas pouco definidas. Naturalmente que os algoritmos de classificação tratam a previsão de atributos nominais.

Existe por vezes a necessidade de agrupar elementos com características semelhantes, *i.e.*, em vez de se prever o valor de um atributo, agrupam-se elementos em que os valores dos seus atributos sejam aproximados o suficiente para que a técnica de *Data Mining* utilizada considere que esses elementos pertencem ao mesmo grupo. Neste caso, trata-se de um problema de *clustering* ou segmentação.

Os algoritmos de associação são um método para descobrir relações interessantes entre diferentes variáveis num determinado conjunto de dados, habitualmente de grandes dimensões. Têm o propósito de identificar regras fortes descobertas nos dados através de diferentes métricas. Os algoritmos de associação tipicamente não consideram a ordem dos atributos, ao contrário dos algoritmos de descoberta de sequências, que procuram padrões estatisticamente relevantes entre vários exemplos que são sequenciais. Tipicamente assume-se que os valores dos atributos são valores discretos.

Quando tratamos de prever um atributo numérico, temos um problema de regressão, em que o atributo a prever terá um valor numérico que o algoritmo de regressão considere o mais provável face ao que aprendeu com os dados iniciais.

Um caso particular da regressão é o *forecasting,* onde se tenta prever valores futuros tendo como base padrões obtidos em grandes conjuntos de dados do passado, mediante aplicação de métodos estatísticos de séries temporais.

## *2.3*Metodologias de *Data Mining*

Esforços têm sido realizados no sentido de serem estabelecidos *standards* no que concerne às metodologias de *Data Mining,* sem no entanto se chegar a um consenso generalizado. No entanto, segundo Azevedo *et al.* (2008), o SEMMA e o CRISP-DM são considerados os mais populares, pelo que serão os descritos. É ainda apresentada uma breve descrição do DMAIC.

### *2.3.1* Metodologias de *Data Mining:* SEMMA

O SAS Institute Inc. criou o processo de DM SEMMA, (*Sample, Explore, Modify, Model, Assess*) que descreve um processo de descoberta de conhecimento em bases de dados. Segundo Azevedo *et al.* (2008), o SEMMA deve o seu nome às 5 fases deste processo que constituem o nome do acrónimo. Através de uma amostra representativa dos dados que se pretende estudar, o SEMMA procura facilitar o processo de DM, criando modelos de previsão e respetiva confirmação.

Segundo o documento "*Data Mining 101: How to Reveal New Insights in Existing Data*", publicado pelo SAS Institute Inc., o SEMMA contribui genericamente para a melhoria do desempenho, acrescentando ainda que o SEMMA não é uma metodologia de DM, mas um processo lógico de desenvolvimento de um modelo, uma vez que consegue incluir qualquer técnica iterativa de DM que seja adotada. O mesmo documento descreve então as 5 etapas deste processo:

1. *Sample:* criar amostras de dados através de uma ou mais tabelas de dados que representem o objeto de estudo. Uma vez que o DM apenas consegue descobrir padrões já existentes nos dados, a amostra deverá ser suficientemente grande para conter toda a informação significativa, em quantidades passíveis de serem processadas. Os dados a estudar são divididos habitualmente em dois conjuntos: treino e teste. Os dados de treino são utilizados para treinar os algoritmos de DM, enquanto os dados de teste são usados para verificar a precisão dos algoritmos criados;

2. *Explore:* exploração preliminar dos dados, que poderá utilizar técnicas variadas, procurando relações, tendências e anomalias, de modo a poder-se retirar algum conhecimento desta exploração;

3. *Modify:* modificação dos dados através da criação, seleção e transformação de variáveis para tornar o modelo mais concentrado no objetivo;

4. *Model*: modelação dos dados através de ferramentas analíticas, de modo a poder-se encontrar a combinação de dados que preveja com fiabilidade o pretendido;

5. *Assess*: verificação dos dados e modelos criados através da avaliação da utilidade e fiabilidade dos resultados do processo de DM. Nem todos os resultados retirados dos algoritmos de DM serão válidos.

### *2.3.2* Metodologias de *Data Mining:* DMAIC

A metodologia DMAIC – *Define, Measure, Analyze, Improve, Control* baseia-se na metodologia *Six Sigma*, eminentemente vocacionada para a melhoria dos processos e redução de defeitos e desperdícios. Esta é uma metodologia generalizada sobretudo no fabrico e na indústria, uma vez que tem como objectivo fundamental um aumento da eficiência dos processos. Os defensores do DMAIC consideram que com a sua utilização se consegue obter maior qualidade e controlo nos projetos de DM. (Turban *et al.,* 2008)

### 2.3.3 Metodologias de *Data Mining*: CRISP-DM

Ao pesquisarmos sobre as metodologias de DM, tendo como o objetivo a extração de conhecimento a partir de conjuntos de dados, verificamos que o *standard* dominante nesta matéria é o designado como CRISP-DM (*Cross Industry Standard Process for Data Mining*). Este é um *standard* europeu criado em 1999 e que já teve tentativas de evolução em 2006, mas oficialmente existe ainda apenas a versão 1.0.

Chapman *et al*., (2000) descrevem de forma muito prática e objetiva o CRISP-DM como um modelo de processo de DM, composto por várias fases. A principal vantagem de se utilizar esta metodologia prende-se com o facto de ser independente da indústria ou do projeto específico, i.e., esta metodologia poderá ser utilizada para a análise de dados de setores diversificados como comércio, ciência ou finanças, pois trata-se de uma metodologia "aberta" que pode ser utilizada por qualquer pessoa.

Para além do anteriormente exposto, esta metodologia não está dependente de nenhum *software* ou ferramentas específicas, tornando-a compatível com qualquer ferramenta de DM que se queira utilizar, uma vez que existem inúmeras ferramentas disponíveis no mercado, inclusivamente ferramentas gratuitas *open source*.

O CRISP-DM tem-se tornado uma referência no DM, permitindo estabelecer comparações entre trabalhos realizados e técnicas, uma vez que a base metodológica é a mesma, que se baseia nas seguintes seis fases (descritas por Chapman *et al*.):

- *Business Understanding* (estudo/compreensão do negócio);
- *Data Understanding* (estudo dos dados);
- *Data Preparation* (preparação dos dados);
- *Modeling* (modelação);
- *Evaluation* (avaliação);
- *Deployment* (implementação).

Estas fases do CRISP-DM estão organizadas entre si da forma descrita na Figura 3.

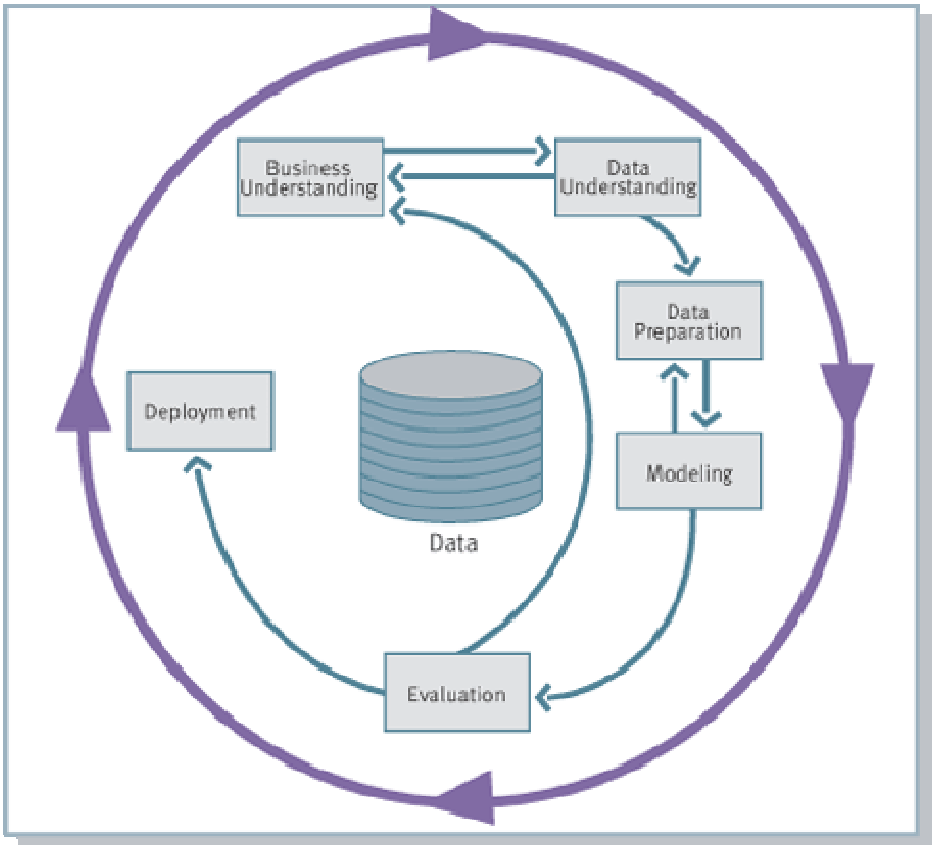


**Figura 3 – Fases do modelo de referência do CRISP-DM (Chapman et al., 2000)**

O círculo exterior do diagrama anterior demonstra a natureza cíclica e iterativa do DM, uma vez que um processo de DM pode continuar após a sua implementação. As lições aprendidas durante o processo poderão despoletar novas questões e frequentemente levantam novos problemas de granularidade mais fina para a compreensão do negócio.

As "major phases" do CRISP-DM, e tal como expresso pelos mesmos autores, são descritas nos pontos 2.3.3.1a 2.3.3.6.

Comparando o CRISP-DM com o SEMMA, poderemos afirmar que o CRISP-DM é uma metodologia mais completa, pois inclui a fase de *Business Understanding*, adiante mais detalhada. Para além disso, poderão ser utilizadas diversas ferramentas para o desenvolvimento de projetos de DM utilizando esta metodologia, tornando-a independente da ferramenta e da indústria. Para além disso, várias empresas como a Daimler-Chrisler, a NCR Corporation e a seguradora OHRA adotam esta metodologia para os seus projetos de DM. Por estes motivos, foi adotada esta metodologia para o estudos dos dados objeto desta dissertação.

O SEMMA não prevê o *Business Understanding*, tal como o CRISP-DM, em que se procura entender a forma e a importância de determinado conjunto de dados para os objetivos do estudo. No entanto, na fase *Sample*, terá de haver sempre algum conhecimento sobre os dados que se estão a trabalhar, pelo que, segundo os mesmos autores, as diferenças entre o CRISP-DM e o SEMMA não são tão

grandes como à partida poderia parecer. Por sua vez, a metodologia DMAIC não foi considerada, dado não ser específica da área temática deste trabalho, o DM.

#### *2.3.3.1 Business Understanding*

Esta fase, que tem como objetivo enquadrar e compreender o negócio, bem como a importância dos dados/previsão que o *Data Mining* se vai encarregar de tratar/prever, subdivide-se em 4 subfases onde são definidas tarefas (*tasks*) e *outputs*: *Determine business objectives, Assess situation, Determine Data Mining goals e Produce Project plan.*

Nas subfases *Determine business objectives* e *Assess situation* a principal tarefa para a análise de dados é efetivamente a compreensão exaustiva daquilo que se precisa obter, na perspetiva do seu negócio. É importante definir com o cliente ou similar todos os fatores que possam influenciar o projeto, de modo a que não haja um grande esforço para a produção de "*right answers to the wrong questions*". Para a execução destas tarefas, ter-se-á de ter em mente os principais outputs que o CRISP-DM enuncia: *Background*, Objetivos de Negócio, Critérios de Sucesso, Inventário de Recursos, Requisitos e constrangimentos, Riscos, entre outros.

Seguidamente, na subfase *Determine Data Mining goals,* a tarefa é transpor os objetivos de negócio enunciados anteriormente para a análise de dados. Por exemplo, enquanto um objetivo de negócio poderá ser: "aumentar as vendas para os clientes existentes", um objetivo do DM será: "prever quais as compras que os clientes existentes farão, tendo em conta os seus dados biográficos e as compras realizadas nos últimos 5 anos". Na subfase seguinte, *Produce Project Plan*, é elaborado um plano descrevendo todos os passos necessários para atingir os objetivos de análise de dados e, consequentemente, os objetivos de negócio ou do problema que se quer resolver.

#### *2.3.3.2 Data Understanding*

O CRISP-DM refere que esta fase se subdivide novamente em subfases. Na subfase *Collect initial data*, o objetivo será a aquisição dos dados ou do acesso a estes dados enunciados no plano de projeto. Nas subfases *Describe data*, *Explore data* e *Verify data quality*, irão ser descritos os dados no que se refere ao seu tipo, formato, quantidade, identificação dos campos, etc., proceder-se-á à sua exploração, tentando compreender relações simples entre alguns dos dados, refinando um pouco a perceção do analista sobre os mesmos e será executada uma verificação da sua qualidade (se existem dados omissos, incompletos, etc.).

#### *2.3.3.3 Data Preparation*

Depois de compreendidos os dados, os mesmos terão de ser selecionados (subfase *Select data*), onde se decide quais devem ser usados e quais os que não terão interesse para o atingir dos objectivos. Posteriormente, os dados serão "limpos" (subfase *Clean data*), fase em que se procederá a um aumento da qualidade dos mesmos, usando modelos de substituição ou estimativa de dados omissos. Nas subfases *Construct data* e *Integrate data*, a tarefa principal é a de criar novos atributos que são construídos a partir de outros atributos (exemplo: área = comprimento x largura). Os dados

deverão ainda ser formatados (subfase *Format data*), fase onde a alteração é de sintaxe dos mesmom dados, sem mudança do seu conteúdo ou significado.

#### *2.3.3.4 Modeling*

Na fase de *modeling* ir-se-á proceder à seleção do modelo ou técnicas de modelação de DM a usar para atingir os objetivos. Depois de selecionada a técnica, dever-se-á fazer um desenho de teste, de modo a validar a qualidade do modelo escolhido. Depois de validado, poder-se-á então construir o modelo final onde serão afinados todos os parâmetros já testados. Nesta fase serão verificados e interpretados os resultados da implementação deste modelo, tentando o analista apresentá-lo a outros analistas de dados e de negócio de modo a poder discutir e a interpretar melhor os dados. Caso esteja a ser utilizado mais que um modelo, serão verificados quais os mais adequados para a análise/previsão em causa.

#### *2.3.3.5 Evaluation*

Tal como evidenciado no nome, esta fase tem como objetivo a avaliação dos resultados obtidos, qual a sua precisão e qual a possibilidade de generalização do modelo construído. É nesta fase que se irá comprovar se o modelo responde aos requisitos de negócio, tentando compreender quais as razões para este modelo ser de maior ou menor qualidade. Todo o processo é revisto nesta fase, mesmo que os resultados sejam satisfatórios, pois novas ideias ou erros poderão ser detetados neste momento. É também o momento de determinar os passos seguintes, sejam eles o de terminar o projeto, proceder à sua implementação ou realizar mais iterações.

#### *2.3.3.6 Deployment*

Caso se tenha decidido pela implementação, esta fase prossegue a partir dos resultados da etapa anterior (*Evaluation)* e determina uma estratégia para a implementação, detalhando todos os passos, incluindo técnicas, monitorização e manutenção. Com o fecho desta fase será feita uma revisão a todo o projeto e a elaboração de um relatório final.

#### 2.3.3.7 Aplicações práticas do CRISP-DM

Existem inúmeras aplicações para o CRISP-DM, enquanto metodologia de extração de conhecimento a partir de grandes quantidades de dados, pois este é bastante utilizado tanto pela indústria como pelos meios académicos. A título de exemplo apresentamos alguns trabalhos realizados neste âmbito.

Moro *et al.* (2011) apresentam um estudo de caso de uma instituição bancária onde se utilizam técnicas de *Data Mining*, inseridas em várias iterações do CRISP-DM, com o objetivo de otimizar campanhas de subscrição de depósitos, nomeadamente no que respeita à capacidade de previsão do sucesso dos contactos realizados para os clientes.

Uma outra aplicação, no âmbito da tecnologia militar, é a proposta por Erskine *et al.* (2010), onde é utilizado e validado o CRISP-DM na redução de falsos alarmes nos sistemas de deteção de intrusão

em sistemas. Este trabalho demonstrou a melhoria de deteção de atividade maliciosa, reduzindo simultaneamente a incidência de falsos positivos em cerca de 91%.

Esta foi a metodologia escolhida para o desenvolvimento deste trabalho prático, uma vez que é uma metodologia flexível, adaptável a diferentes cenários de estudo e oferece uma estrutura completa para a conceção e desenvolvimento de trabalhos de DM.

## 2.4 Ferramentas de *Data Mining*

Existem no mercado diversas ferramentas de *software* para DM, contendo inúmeras formas e técnicas, que vão de ferramentas comerciais até *open source*[2]. Uma sondagem (*poll*) em que participaram 1880 utilizadores, realizada pelo site www.kdnuggets.com em Junho de 2013 e com a questão: "*What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project*?"[3], obteve resultados conforme a Tabela 2 que de seguida se apresenta.

**Tabela 2 – Resultados de uma sondagem sobre utilização de ferramentas de DM**

| ***What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real projet?*** [1880 votantes] | |
|---|---|
| Legenda: Vermelho: Ferramentas gratuitas/*open-source*<br>Verde: Ferramentas | % utilizadores |
| Rapid-I RapidMiner/RapidAnalytics free edition (737) | 39.2% |
| R (704) | 37.4% |
| Excel (527) | 28.0% |
| Weka / Pentaho (269) | 14.3% |
| Python with any of numpy/scipy/pandas/iPython... packages(250) | 13.3% |
| Rapid-I RapidAnalytics/RapidMiner Commercial Edition (225) | 12.0% |
| SAS (202) | 10.7% |
| MATLAB (186) | 9.9% |
| StatSoft Statistica (170) | 9.0% |
| IBM SPSS Statistics (164) | 8.7% |
| Microsoft SQL Server (131) | 7.0% |
| Tableau (118) | 6.3% |
| IBM SPSS Modeler (114) | 6.1% |
| KNIME free edition (110) | 5.9% |
| SAS Enterprise Miner (110) | 5.9% |
| Rattle (84) | 4.5% |
| JMP (77), | 4.1% |
| OUTROS | 3.6% |

Fonte: http://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html

[2] *Open-source:* software de código aberto e de utilização gratuita
[3] Tradução livre: "que ferramentas de análise, *Big Data, Data Mining, Data Science* utilizou nos últimos 12 meses para um projeto real (não apenas para avaliação)?"

Como podemos ver por esta tabela, existem inúmeras ferramentas disponíveis para a análise de dados, de utilização comercial e *open-source*, sendo que as mais utilizadas são o *Rapid-I RapidMiner/RapidAnalytics*, o *R* e ainda o *Microsoft Excel*. De notar que muitos dos votantes não usam apenas uma ferramenta para o desenvolvimento dos seus projetos, existindo muitas vezes a necessidade de se recorrer a várias destas ferramentas. Os números desta sondagem são a confirmação e atualização do que já havia sido constatado por Rexer *et al.* (2011), aquando de um inquérito direcionado a especialistas de *Data Mining,* e levado a cabo em cerca de 60 países, apresentou o resultado em que se evidenciava que a ferramenta *R* era utilizada por 47% dos inquiridos, precisamente por ser gratuita, *open source* e possuir uma grande variedade de algoritmos.

Na Tabela 3, que a seguir se apresenta, encontramos a distribuição dos resultados da mesma sondagem conforme o local e tipo de ferramenta, e podemos verificar que a maioria dos utilizadores de ferramentas para a realização de projetos de *Data Mining* utilizam ferramentas gratuitas, sendo que uma elevada percentagem destes mesmos utilizadores utiliza em exclusivo ferramentas desta natureza.

No caso da Europa Ocidental, 35% dos utilizadores responderam que apenas utilizam ferramentas gratuitas e 40% dos utilizadores utilizam ferramentas comerciais e gratuitas. Analisando estes dados, podemos verificar que 75% dos utilizadores de ferramentas de *Data Mining* utilizam ferramentas gratuitas.

**Tabela 3 – Distribuição dos resultados da mesma sondagem conforme o local e tipo de ferramenta**

| Região | % utilizadores que utilizam apenas ferramentas comerciais<br>% utilizadores que utilizam apenas ferramentas gratuitas<br>% utilizadores que utilizam ambas |
|---|---|
| EUA/Canada (33%) | 34% 17% 48% |
| Europa Ocidental (28%) | 24% 35% 40% |
| Europa de Leste (13%) | 34% 29% 36% |
| Asia (11%) | 25% 34% 41% |
| América Latina (8.8%) | 24% 45% 31% |
| África/Médio Oriente 4.0%) | 34% 29% 36% |
| Austrália/NZ (2.3%) | 25% 32% 43% |

Fonte: http://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html

## 2.5 Técnicas de *Data Mining*

Existem diversas técnicas de DM para análise de dados, conforme a categoria em que se insere o conjunto de dados a estudar. Para este trabalho em particular, e tratando-se de um problema de regressão, como adiante se verá, foram utilizados três algoritmos distintos: Redes Neuronais (*Neural Networks* – NN*)*, Máquinas de Vetores de Suporte (*Support Vector Machines* – SVM*)* e Regressão Múltipla (*Multiple Regression* – MR*)*.

As Redes Neuronais (NN – *Neural Networks)*, ou Redes Neuronais Artificiais são métodos que realizam análises de dados multi-dimensionais, sendo particularmente eficazes com atributos de natureza numérica. Segundo Dayhoff and Deleo (2001), as NN são inspiradas na Biologia, em particular nas redes de neurónios que contêm várias camadas de nodos computacionais simples que operam como dispositivos de soma não-lineares. Estes nodos estão interligados por linhas de conexão, cujos pesos são ajustados durante um processo de treino, sempre que novos dados são apresentados à rede neuronal. Através do treino, podemos ter redes neuronais que executam tarefas como classificar um atributo, prever um valor de saída, reconhecer um padrão em dados multi-factoriais, entre outras. As NN são atualmente utilizadas em muitas áreas de estudo, por serem robustas, terem capacidade de ser utilizadas para diferentes propósitos e serem computacionalmente viáveis. A mais popular das arquiteturas das *NN* é o *Multilayer Perceptron* (MLPE).

Uma metodologia mais recente que a NN é a das Máquinas de Vetores de Suporte (*Support Vector Machines* – SVM*)*. O modelo SVM baseia-se em teoria estatística da aprendizagem, transformando o espaço original das entradas num espaço imaginário. Esta transformação é dependente de uma função de *kernel*. Para além disso, o algoritmo de SVM deteta automaticamente quais os exemplos relevantes (*support vectors*) no espaço imaginário para uma classificação ou regressão ótima.

A Regressão Múltipla (*Multiple Regression* – MR*)* é um método estatístico utilizado para executar a análise das relações entre uma única variável de saída e diversas variáveis independentes. A regressão múltipla serve igualmente para verificar quais as variáveis independentes que mais influenciam a variável de saída. Dentro de um modelo de *MR,* todas as variáveis terão de ser numéricas, tal como expresso por Hastie *et al.* (2008). A relação entre a variável dependente e uma outra considerada como independente, pode ser representada através de um diagrama de dispersão. Os valores da variável dependente são representados no eixo das ordenadas, enquanto os valores da variável independente são representados no eixo das abcissas. Cada par de valores das duas variáveis, fornecerá um ponto, e através de métodos como o do desvio mínimo quadrado poder-se-á calcular a equação de um reta.

Existem casos particulares dentro dos problemas de regressão, onde se analisa a evolução de uma determinada variável de saída ao longo do tempo. Para estes casos poder-se-ão utilizar técnicas de análise de séries temporais, em particular os métodos de alisamento exponencial. Este método aplica uma média ponderada nas observações de uma série temporal, sendo atribuídos diferentes pesos aos dados, *i.e.*, os dados mais antigos têm peso menor, em detrimento dos dados mais recentes.

Este método é muito utilizado na análise de séries temporais, uma vez que é um método de previsão bastante rápido, simples e de baixo custo. (Hyndman e Athanasopoulos, 2012)

Existem três métodos de alisamento exponencial: Alisamento Exponencial Simples, Alisamento Exponencial de Holt e ainda o método de Holt-Winters (Makridakis et al., 1998). O Alisamento Exponencial Simples atribui aos dados pesos exponenciais decrescentes ao longo do tempo, sendo similar às técnicas de média móvel, e a forma geral deste método encontra-se descrita através da equação (1):

$$F_{t+1} = \alpha X_t + (1 - \alpha) F_t \quad (1)$$

$F_{t+1}$ é a previsão para t+1, $F_t$ é a previsão para t e $X_t$ a procura realizada no período t. O valor de alfa (α) está dentro do intervalo de 0 a 1. Quanto mais alto o seu valor, maior será o peso alocado às observações mais recentes e tornando o modelo mais sensível a mudanças. Quanto mais próximo o valor de alfa (α) for de 0, menor será o ajuste, ou seja, o modelo dará mais peso às observações mais antigas, levando a previsões mais estáveis.

O Alisamento Exponencial Linear de Holt é uma expansão do método anterior para previsões com dados que apresentam uma tendência linear, mas não sazonalidade. Assim este método utiliza duas contantes de alisamento, alfa (α) e beta (β), igualmente com valores entre 0 e 1, e sem relação entre si. A componente de tendência é definida através da equação (2):

$$b_t = \beta(L_t - L_{t-1}) + (1 - \beta) b_{t-1} \quad (2)$$

Um valor mais elevado de beta (β) indica que uma porção maior da tendência mais recente será adicionada à previsão para o período seguinte. O valor de beta (β) também poderá ser iterativamente modificado como o alfa (α) de modo a minimizar valores de erros. $b_t$ é a estimativa de tendência da série temporal para o período t.

Para além da componente de tendência, este método tem ainda uma equação para a previsão (3) e nível (4):

$$F_{t+m} = L_t + b_t m \quad (3)$$

$$L_t = \alpha X_t + (1 - \alpha)(L_{t-1} + b_{t-1}) \quad (4)$$

Para situações em que as séries temporais apresentam padrões com tendência linear e sazonalidade pode ser aplicado o Método de Holt-Winters, que utiliza suavizações para estimar o nível, tendência e sazonalidade da série em estudo no processo de previsão. O método recorre a duas abordagens diferenciadas que incidem essencialmente na forma de como se aborda a sazonalidade: aditiva ou multiplicativa. A forma aditiva é mais apropriada para séries cuja amplitude de sazonalidade é independente do nível, e a forma multiplicativa é indicada sempre que a amplitude da sazonalidade varie com o nível.

## 2.6 Medidas de Desempenho da Previsão

Várias medidas alternativas poderão ser utilizadas para avaliar o sucesso de previsões numéricas (Witten & Frank, 2010). O Erro Absoluto Médio (*Mean Absolute Error* – MAE) é um desses métodos que executa a média da magnitude dos erros individuais, ou seja, ignorando os seus sinais. Este método tem a vantagem de ser pouco influenciado por *outliers* (valores atípicos), uma vez que todas as dimensões de erro são tratadas de forma proporcional à sua magnitude. O cálculo do MAE poderá ser efetuado através da fórmula expressa na equação (5). Os valores previstos nas instâncias de teste são $p_1$, $p_2$,…, $p_n$; Os valores observados são $a_1$, $a_2$,…, $a_n$.

$$MAE = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n} \tag{5}$$

Uma das medidas mais habitualmente utilizadas é o Erro Quadrático Médio (*Mean Squared Error* – MSE). Muitas técnicas de regressão utilizam o MSE como métrica, sendo também mais sensível aos efeitos de *outliers.* O cálculo do MSE poderá ser efetuado através da fórmula expressa na equação (6).

$$MSE = \frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n} \tag{6}$$

A Raiz quadrada do Erro Quadrático Médio (*Root Mean Squared Error* – RMSE) executa a raiz quadrada do MSE, passando habitualmente para valores semelhantes aos do atributo que se tenta prever, tal como expresso na equação (6):

$$RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}} \tag{6}$$

Quanto mais baixos forem os valores de MSE e RMSE, menor será o erro medido, ou seja, mais preciso será o modelo de previsão utilizado.

Uma outra medida habitualmente utilizada, é o designado coeficiente de determinação ou $R^2$, uma vez que é um bom comparador de modelos de regressão em que se utilizem as mesmas variáveis dependentes e os mesmos períodos de estimativa. Este coeficiente indica de que forma os pontos se distribuem ao longo de uma linha ou curva. Este método fornece uma métrica de como são replicados os valores observados pelo modelo de previsão. O valor de $R^2$ varia entre -∞ e 1, e quanto maior for o seu valor, melhor será o modelo de previsão utilizado.

# 3. Aplicação de Técnicas de *Data Mining* na manutenção de infraestruturas e equipamentos

## 3.1 Introdução

O meio para desenvolvimento deste trabalho são as técnicas de *Data Mining*, enquadradas num desenvolvimento de BI, mas o seu fim são questões relacionadas com a temática da manutenção e da gestão de manutenção. De notar que esta temática, por si só, constitui uma área de conhecimento muito desenvolvida na Engenharia e na Gestão Industrial, estando também sob regulamentação própria para algumas das suas atividades. Não é objetivo desta dissertação fazer um trabalho exaustivo sobre esta matéria, mas será importante apresentar e descrever alguns dos objetos da manutenção. Toda a terminologia relacionada com estes termos está descrita em normas e *standards* internacionais, existindo inclusivamente uma norma europeia, a EN 13306:2010 – *Maintenance terminology,* que trata apenas da definição destes termos.

Cabral (2009) descreve a manutenção como o conjunto das ações destinadas a assegurar o bom funcionamento das máquinas e instalações, de acordo com as boas práticas técnicas e exigências legais, de forma a evitar a perda de função ou redução no rendimento destas e, em caso de necessidade, repor, com a maior brevidade possível, as suas condições de operacionalidade.

Outra definição que se encontra na literatura é a de Márquez (2007), que refere que a manutenção é uma combinação de ações tais que conservem ou restaurem um determinado bem num estado em que possa realizar a função para o qual foi desenhado ou implantado. “Conservação” e “restauro” são denominações para tipos de ações que depois se convertem em “preventivo” ou “corretivo” em termos de vocabulário de manutenção. Esta definição encaixa na descrita na norma EN 13306:2010 – *Maintenance terminology*, que define a manutenção como a combinação de todas as ações técnicas, administrativas e de gestão, durante o ciclo de vida de um item, com o propósito de o conservar ou restaurar num estado em que possa realizar a função requerida.

Conforme o propósito e o momento da manutenção, tal como anteriormente referido, poderemos ter Manutenção Preventiva (MP) ou Manutenção Corretiva (MC), cujas definições se podem encontrar na supra referida norma europeia. Manutenção Preventiva é a manutenção realizada em intervalos pré-determinados ou de acordo com critérios especificados no intuito de reduzir a probabilidade de falha ou degradação no funcionamento de um item. Por outro lado, Manutenção Corretiva é a manutenção realizada após o reconhecimento de uma falha, com o intuito de colocar um item num estado em que possa realizar a função requerida.

Existem outros tipos de manutenção e indicadores para a aferição dos níveis de desempenho de manutenção de um determinado grupo ou organização, mas que não serão, pelo menos nesta fase, incluídos no âmbito desta dissertação.

## 3.2 Gestão de manutenção e planeamento

Márquez (2007), refere que gestão caracteriza o processo de liderança e direção de parte ou totalidade de uma organização, através da criação e da manipulação de recursos. A gestão de manutenção caracteriza assim o processo de liderança e direção de uma organização de manutenção. Este autor define gestão de manutenção como o conjunto de todas as atividades de gestão que determinam os objetivos ou as prioridades de manutenção, estratégias e responsabilidades, de uma forma que possam ser implementadas através de meios como o planeamento da manutenção, o seu controlo e supervisão. Esta definição é praticamente igual à constante na norma EN 13306:2010, que define gestão da manutenção como "*todas as atividades da gestão que determinam os objetivos, a estratégia e as responsabilidades respeitantes à manutenção e que os implementam por meios, tais como o planeamento, o controlo e supervisão da manutenção e a melhoria de métodos na organização, incluindo os aspetos económicos*".

## 3.3 Técnicas de *Data Mining* aplicadas à área de manutenção

Não sendo o objetivo deste trabalho analisar o desempenho de uma organização de manutenção, mas sim o de fazer previsões de modo a auxiliar a gestão da manutenção, procurou-se encontrar na literatura trabalhos científicos que pudessem apresentar algo na integração destas áreas.

Létourneau *et al.* (1999) realizaram um trabalho no qual se utilizou o DM para previsão da substituição de componentes de aeronaves.

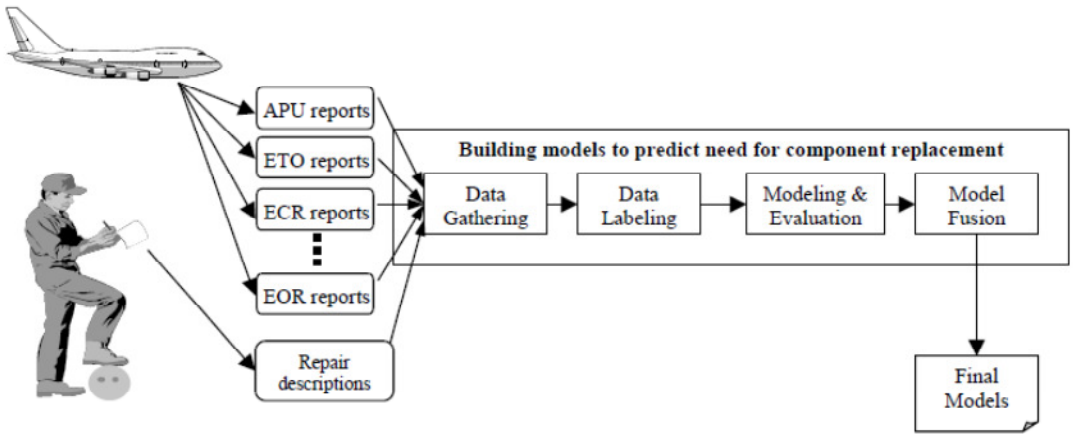


**Figura 4 – Os 4 passos para a construção de modelos de substituição de componentes de aviões, segundo Létourneau *et al.* (1999)**

A Figura 4 apresenta a forma de construir a previsão proposta por estes autores, que apesar de não estar sob nenhuma metodologia específica, apresenta os passos fundamentais para a preparação e análise dos dados e posterior modelação e avaliação dos resultados obtidos. Este trabalho é extremamente interessante, pois faz a previsão para 16 componentes diferentes da aeronave Airbus A320, criando uma escala para pontuação dos vários modelos utilizados, de modo a podermos facilmente verificar qual o modelo mais adequado para cada previsão.

Depois da preparação e da escolha dos dados relevantes, estes autores utilizaram 3 técnicas diferentes de DM para 16 componentes do avião em causa. Estas técnicas foram: árvore de decisão com C4.5, "instance-based" e Naive Bayes, recorrendo à biblioteca MLC++ (*Machine Learning* C++). Para todos os componentes foi utilizada uma função para classificar cada uma das técnicas para cada uma das 16 componentes. Os resultados obtidos permitiram verificar que, para cada um dos componentes do avião, uma técnica poderá ser mais adequada, concluindo os autores que deverão ser sempre testadas várias técnicas para cada situação.

De seguida, faz-se um breve apanhado de trabalhos realizados na área da manutenção, que de alguma forma tenham explorado os conceitos de BI e/ou DM. Existem inúmeros trabalhos publicados no âmbito da manutenção industrial ou de infraestruturas, mas quase sempre no âmbito da análise da fiabilidade e desempenho, na determinação do custo dos ciclos de vida e/ou na avaliação do desempenho da gestão de equipamentos. Esta é uma área extremamente desenvolvida, contando com diversas metodologias e inclusivamente *standards* e regulamentação própria.

Outro trabalho interessante é o de Reffat *et al.* (2004), no qual é utilizado o DM como um meio para tornar úteis enormes quantidades de dados gerados através dos processos de manutenção. Este trabalho foi realizado utilizando algoritmos de classificação, de modo a descobrir padrões e correlações em grandes volumes de dados. O trabalho destes autores investigou o potencial de se aplicar técnicas de DM a dados provenientes da manutenção de edifícios. Neste trabalho, os autores não aplicaram nenhuma das metodologias de *Data Mining* anteriormente expostas, mas referem a importância de analisar e preparar os dados para aplicação de técnicas de DM, que são diferentes das técnicas utilizadas na estatística, pois não utilizam à partida nenhuma distribuição probabilística (*v.g.* distribuição de Gauss), tal como exposto na Figura 5.

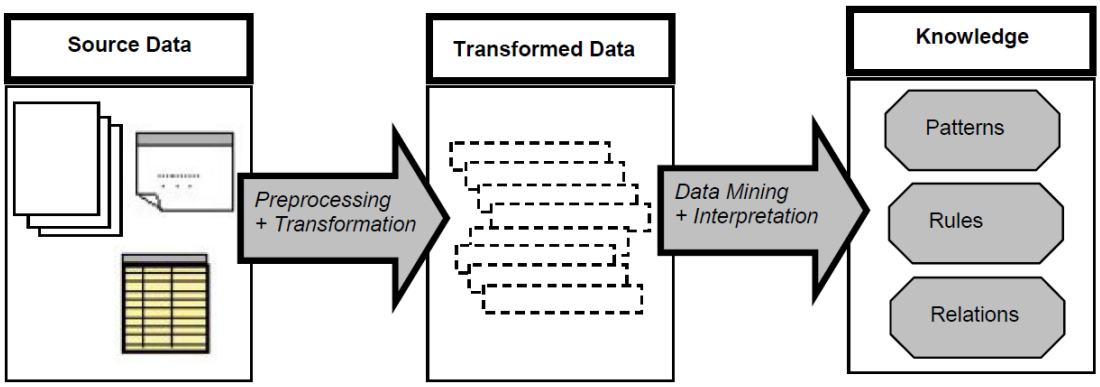


**Figura 5 – Pré-processamento e Transformação dos dados para utilização em DM, segundo Reffat *et al.* (2004)**

Estes autores comparam diversas ferramentas para o fim em questão e apresentam como escolha a ferramenta Weka[4], por apresentar diversos métodos de classificação, ser de distribuição gratuita e ser *open-source* baseado em Java, permitindo a integração futura num sistema de apoio à decisão.

[4] Weka: *software* de DM em Java que apresenta uma coleção de algoritmos para tarefas de DM e que contém diversas ferramentas para pré-processamento, classificação, regressão, entre outras.

No trabalho destes autores, várias técnicas de *Data Mining* foram utilizadas e foram demonstrados resultados práticos. A título de exemplo, a Tabela 4, retirada do artigo destes autores, apresenta resultados da aplicação.

**Tabela 4 – Exemplo de resultados da aplicação de técnicas de DM, conforme Reffat *et al.* (2004)**

| Técnica de DM | Resultados de DM | Potenciais benefícios para gestores de manutenção de edifícios |
|---|---|---|
| Árvore de decisão (C4.5) | Todos os trabalhos mensais de prioridade mais elevada são realizados na fase mais tardia do ano (Julho a Novembro) | Distribuição da prioridade dos trabalhos de manutenção é importante no planeamento e agendamento das atividades e recursos |
| Regra de associação | Existe uma relação crescente entre a prioridade do trabalho, o tempo estimado para a sua execução e o *budget associado* | Um melhor planeamento e agendamento poderá contribuir para o reforço deste relacionamento |

Como conclusão, Reffat *et al.* (2004) reconhecem a utilidade do DM para descobrir relações antes desconhecidas, através da análise de dados retirados da gestão da manutenção. Algo muito interessante, e que se enquadra no objetivo desta dissertação é o que estes autores referem na conclusão do seu trabalho: "*knowing when a building or system component will break before it breaks, in plenty of time for repairs, to be conveniently and cost-effectively scheduled and executed, is an exciting application of this technology that can add dollars to the bottom line"*. Para além disso, é referido ainda que o DM, ao ajudar na melhoria dos processos de gestão da manutenção, pode aumentar a disponibilidade dos equipamentos, reduzindo também desta forma os custos de manutenção.

Existem também diversos trabalhos publicados em que, embora não abordando a temática da manutenção, abordam a gestão de Aeroportos sobre diversas vertentes em que o BI e o DM estão presentes. A título de exemplo, Loan *et al* (2007) usam várias técnicas de DM para ilustrar, por exemplo, o uso ineficiente da capacidade dos *runways*[5] do Aeroporto de LaGuardia em Nova Iorque, concluindo que poderão ser utilizados de forma mais eficiente.

[5] *Runway*: designação do local onde aterram e descolam aviões. Em português, o termo mais utilizado é "pista".

# 4. Previsão de falhas em infraestruturas e equipamentos com recurso a técnicas de *Data Mining*

De um modo muito genérico, se conseguíssemos prever o nosso futuro, teríamos a oportunidade de tomar decisões de forma atempada, contribuído para a tomada de melhores decisões. Numa organização atual, sujeita a uma conjuntura económica e social particularmente difícil, qualquer melhoria nos seus processos técnicos ou de gestão, poderá significar ganhos operacionais e eventualmente financeiros, trazendo melhorias significativas para a sua atividade.

Este trabalho tem um objetivo concreto, o estudo de modelos de previsão para determinar o número de ocorrências de manutenção no Aeroporto, e que se enquadra num objetivo mais geral, que é o do aumento da eficiência. Sendo este um trabalho predominantemente exploratório, as próprias iterações de DM realizadas ajudarão a definir os modelos mais adequados, apontando em direções mais concretas. Neste capítulo é contextualizado e descrito o estudo de caso, nomeadamente na apresentação de vários conceitos utilizados no Sistema de Gestão de Manutenção (SGM) do Aeroporto de Lisboa (ALS), e que representam o conjunto de dados utilizados como base deste trabalho. É também exposto o trabalho prático realizado, evidenciando a metodologia e ferramentas utilizadas, apresentando também e analisando os resultados práticos obtidos.

O SGM é a plataforma em cuja base de dados são registadas todas as incidências e ordens de trabalho de manutenção do Aeroporto, tal como anteriormente descrito. É este sistema que os elementos da divisão de gestão técnica e manutenção do Aeroporto usam como suporte à atividade do dia-a-dia, de modo a poder documentar o trabalho realizado e poder retirar informação e indicadores de manutenção e gestão tidos como importantes. Se para além dos indicadores mais comuns de manutenção, fosse possível adicionar informação que auxiliasse os gestores das diversas áreas de manutenção a planear trabalhos ou futuras intervenções, a utilização deste sistema poderia ser potenciada, tornando-o num sistema mais completo.

## 4.1 Enquadramento

Dando seguimento ao descrito no ponto anterior sugere-se que o problema geral das organizações é habitualmente o mesmo: fazer mais e melhor com menos recursos, otimizando resultados financeiros e operacionais. Desta forma, o problema que se pretende abordar nesta dissertação está enquadrado neste problema mais geral, e é o de especificamente prever, com a melhor qualidade possível, o número de incidências ou ordens de trabalho, que traduz o número de avarias e falhas em equipamentos e infraestruturas do Aeroporto. Com este conjunto de previsões pretende-se ter informação atempada com o maior grau possível de confiança, sobre qual a forma, setor e frequência em que a infraestrutura aeroportuária será afetada, naquilo que diz respeito à manutenção.

Se uma organização que depende muito das suas instalações físicas para a sua atividade, tal como ocorre com o Aeroporto de Lisboa, conseguir determinar com alguma precisão quais os períodos em que haverá um maior número de incidências que provoquem falhas ou avarias, poderá planear

melhor os seus trabalhos e recursos alocados à manutenção. Adicionalmente, poderão ser planeados trabalhos mais aprofundados ou de natureza preventiva em períodos que se preveem como de menor número de incidências de manutenção.

Mais do que um problema, esta é uma oportunidade de dar início a um processo que poderá mudar a forma de encarar e planear a manutenção e a utilização dos seus recursos. Nesse sentido, este trabalho pretende ser uma contribuição efetiva para a escolha, aplicação de técnicas e análise dos modelos de previsão obtidos para uma organização aeroportuária e no que diz respeito à manutenção de infraestruturas e equipamentos.

Não foram encontrados na revisão de literatura efetuada casos de utilização direta de técnicas de *Data Mining* tendo como objetivo a identificação de padrões relevantes para apoio à gestão de manutenção. Apesar de se encontrar muita literatura sobre DM e sobre manutenção e gestão de manutenção, poucos trabalhos existem onde ambos os temas sejam abordados. O DM surge como uma ferramenta de apoio em diversas áreas, com estudos e publicações em diferentes contextos e problemas, sendo possível identificar inúmeras aplicações práticas destas técnicas, tal como anteriormente evidenciado.

O presente trabalho incide sobre o estudo de uma realidade muito concreta, com uma organização e objetivos muito particulares, que vão ao encontro da resolução de problemas concretos de estimativa de quais os recursos a alocar em cada momento para a área de manutenção.

## 4.2 Organização e Planeamento

A Divisão de Gestão Técnica e Manutenção do Aeroporto de Lisboa (ALSGTMAN) inclui um centro que recebe pedidos de assistência técnica de variadas formas. Estes pedidos têm geralmente origem num telefonema de um colaborador de um qualquer serviço ou área do Aeroporto, reportando incidências descrevendo anomalias e avarias num espaço físico ou equipamento do Aeroporto. Por outro lado, este centro, designado como Centro de Monitorização Técnica (CMT), tem também sistemas que permitem aos colaboradores que ali se encontram efetuar a monitorização de infraestruturas e equipamentos. Havendo a deteção de uma falha num qualquer sistema sob supervisão do CMT, poder-se-á também abrir um novo pedido de assistência técnica.

Os pedidos de assistência explicitados no parágrafo anterior dão origem ao que se designa de incidências, que são registadas no Sistema de Gestão de Manutenção (SGM) do ALS. Em suma, estas incidências têm origens diversas, e poderão ser abertas de diversas formas: telefonema ou *e-mail* de um passageiro ou colaborador do Aeroporto; pedido para alteração de um determinado parâmetro num sistema (v.g. alteração da temperatura do sistema de climatização numa determinada sala); deteção de uma falha num determinado sistema através de sistemas de supervisão; anomalia detetada no âmbito de uma ação de manutenção preventiva.

De notar que o âmbito deste trabalho é sobretudo o de prever o número de incidências que deem origem a ações de natureza corretiva. Incluímos neste lote de ações de manutenção todas aquelas

que tenham natureza não preventiva e que procurem colocar ou alterar um local ou equipamento para que possam realizar a sua função normal ou a nova função para que foram designadas.

Em termos de processo organizacional podemos resumir que as ações de Manutenção Corretiva podem ter as seguintes origens:

- Registo de incidências pelo Centro de Monitorização Técnica (CMT), através dos meios anteriormente descritos. Exemplo prático: colaborador do Aeroporto informa telefonicamente o CMT que não existe energia elétrica num balcão de *check-in*;
- Deteção de falha ou insuficiência de um equipamento ou sistema por um técnico, através de inspeção visual ou sistema de monitorização. Exemplo: o técnico, através de um sistema de supervisão, verificou que uma câmara do sistema de videovigilância está inoperativa;
- Deteção de falhas resultantes de ações de Manutenção Preventiva. Exemplo: o técnico, ao realizar a Manutenção Preventiva a um equipamento de rastreio de bagagem por Raio-X, verificou que as cortinas de chumbo do referido equipamento apresentam um desgaste grande, tendo necessidade de ser substituídas;
- Situações decorrentes de auditorias ou alterações da forma de funcionamento ou local de um determinado equipamento. Exemplo: por alteração de uma norma de segurança, será necessário ajustar ou calibrar um equipamento de rastreio de passageiros de modo a poder cumprir com a referida norma.

As incidências podem ser resolvidas rapidamente por uma equipa ou um técnico designado como de primeira linha, por se tratar de uma incidência simples, sem necessidade de intervenção de um técnico mais especializado, designado habitualmente como de segunda linha. Quando a incidência é resolvida por um técnico de primeira linha, pode ser completado o seu registo no SGM e fechada. A complexidade destas incidências é muito variável, podendo até por vezes ser resolvidas rapidamente ao telefone com a pessoa que coloca o problema encontrado. Se a incidência registada tiver de ser passada a um técnico de segunda linha, pela sua especificidade ou complexidade, é aberta uma ordem de trabalho (OT) para a resolução da incidência anteriormente. A título de exemplo, a uma incidência reportada para um equipamento de rastreio de bagagem por Raio-X, deverá responder um técnico de segunda linha, especializado neste tipo de equipamentos, pois não só estes são equipamentos com uma especificidade de funcionamento muito grande, como utilizam radiações ionizantes. A OT corresponde a um documento que no SGM contém toda a informação relativa à intervenção realizada.

Em termos de dados que queremos prever, ou seja variáveis de saída, temos as incidências, importantes para se poder aferir as necessidades para o CMT e equipas de manutenção de primeira linha, e ainda as OT, importantes para se poder atempadamente planear os trabalhos das equipas de segunda linha. Dos dados retirados da base de dados do SGM, temos o número de incidências e OT, através dos quais podemos ter uma ideia de qual a ordem de grandeza para os meses e anos seguintes, mas não muito mais do que isso, pois apenas poderemos fazer uma análise estatística destes valores e tentar projetar, sem grande suporte, o número futuro de incidências.

Se para além destas variáveis, juntarmos outros fatores, como por exemplo o número de passageiros, o número de movimentos de aeronaves, as condições climatéricas, etc., poderemos tentar aferir qual a contribuição destes para a ocorrência de mais ou menos incidências ou OT. O número de aeronaves e o número de passageiros são fatores que se conseguem obter com alguma antecedência e bastante precisão, sendo que, caso estes tenham influência no número de incidências, poderemos construir um modelo baseado nestes valores. Se esse modelo for preciso, teremos dados para uma melhor distribuição de recursos e um melhor planeamento das intervenções de manutenção preventiva, trabalhos de conservação, etc.

A Tabela 5 apresenta um excerto do número de incidências e OT, e ainda o número de passageiros e de movimentos de aeronaves. O período em estudo será de Janeiro de 2007 a Dezembro de 2012 e os dados têm uma distribuição mensal. A apresentação desta tabela tem apenas o propósito de dar ao leitor uma ideia sobre os valores das variáveis em estudo. Dando seguimento ao exposto anteriormente, poderemos incluir muitas mais variáveis, mesmo algumas que aparentemente possam ter pouca influência no aumento do número de incidências de manutenção, de modo a podermos estudá-las e comparar os resultados com ou sem a sua presença.

**Tabela 5– Número de incidências e ordens de trabalho, número de passageiros e movimentos de aeronaves (excerto do período em estudo)**

| Mês | Incidências | Ordens de Trabalho (OT) | Número de Passageiros | Movimentos de Aeronaves |
|---|---|---|---|---|
| **Jan-11** | 1600 | 1795 | 972570 | 11215 |
| **Fev-11** | 1393 | 1534 | 845683 | 9638 |
| **Mar-11** | 1393 | 1631 | 1072166 | 11174 |
| **Abr-11** | 1452 | 1694 | 1296414 | 11847 |
| **Mai-11** | 1567 | 1843 | 1280050 | 12297 |
| **Jun-11** | 1347 | 1601 | 1338406 | 12711 |
| **Jul-11** | 1454 | 1635 | 1568301 | 13823 |
| **Ago-11** | 1498 | 1711 | 1621971 | 13917 |
| **Set-11** | 1511 | 1588 | 1458718 | 13004 |
| **Out-11** | 1533 | 1574 | 1335767 | 12533 |
| **Nov-11** | 1467 | 1622 | 976217 | 10114 |
| **Dez-11** | 1448 | 1528 | 1039338 | 11058 |

As variáveis a incluir serão dados meteorológicos, tais como dados de temperatura, humidade, velocidade do vento, etc., e serão utilizados modelos que possam incorporar estas variáveis, e comparados com outros que não as incorporem, de modo a podermos comparar e escolher o melhor dos modelos.

Para além do número de variáveis que contribuem para o número de incidências, poderemos subdividir as incidências e as OT por setor de manutenção (eletricidade, civil, pavimentos, mecânica,

eletrónica, etc.), pois cada um destes serviços possui equipas alocadas para os diversos níveis e tipos de manutenção.

Estando caracterizado o estudo de caso a analisar através desta dissertação, é então necessário definir uma metodologia e planeamento adequados ao cumprimento dos objetivos propostos. Tal como referido anteriormente, o CRISP-DM (Chapman et al., 2000), dada a sua flexibilidade de adaptação a diferentes contextos, foi a metodologia escolhida para a realização deste trabalho. Esta metodologia apresenta uma grande vantagem, que é a de ser iterativa e cíclica, permitindo enriquecer os modelos obtidos ou rever completamente as condições que implicaram um determinado resultado. Vários autores referem que os projetos estruturados em CRISP-DM começam por vezes com um objetivo amplo. No entanto, a metodologia CRISP-DM foi encarregue de dirigir os esforços de pesquisa para um objetivo mais concreto, como veremos adiante. Desta forma, e dada a natureza exploratória deste trabalho, assumiu-se que o planeamento inicial serviria apenas como uma base do trabalho a realizar.

Uma outra mais-valia desta metodologia, é a de permitir uma avaliação constante dos resultados produzidos, permitindo inclusivamente a revisão e retrocesso em cada uma das fases sempre que exista essa necessidade, podendo organizar ou orientar o trabalho para uma persecução mais efetiva dos objetivos, que são os de produzir conhecimento dos dados estudados.

Esta dissertação tem o objetivo geral de prever incidências e ordens de trabalho de manutenção num Aeroporto, que são efetivamente o número de ocorrências de manutenção. Com recurso ao *Data Mining* e à metodologia CRISP-DM, procura-se este fim específico, que poderá identificar padrões úteis para a gestão de manutenção e, consequentemente, para uma melhor gestão dos recursos humanos e infraestruturas aeroportuárias. Ao longo do desenvolvimento deste trabalho, e com o decorrer da aplicação do CRISP-DM, os objetivos foram sendo afinados, tornando-se mais específicos.

## 4.3 Ferramentas utilizadas no trabalho

Para o desenvolvimento deste trabalho de dissertação foi necessário recorrer a várias ferramentas, uma mais relacionadas com o *Data Mining* em si, outras relacionadas com o caso de estudo, de forma a se poder extrair os dados em causa.

A primeira ferramenta utilizada, embora não especificamente na persecução dos objetivos de *Data Mining*, foi o Sistema de Gestão de Manutenção atualmente utilizado no Aeroporto de Lisboa, que é a aplicação Maximo 7, da IBM. Esta aplicação permite aos utilizadores o registo das ocorrências de manutenção (incidências e ordens de trabalho), a sua validação pelos responsáveis de cada setor e ainda a produção de indicadores de manutenção, como a percentagem de conclusão de manutenções preventivas, disponibilidade de equipamentos, etc. Por outro lado, contém os planos de manutenção preventiva, abrindo automaticamente, com a periodicidade definida para cada caso, pedidos de manutenção de carácter preventivo direcionados a cada um dos sectores. Esta ferramenta permite ainda aos seus utilizadores, fazer o registo da mão-de-obra utilizada para cada

incidência ou ordem de trabalho, descrever a intervenção realizada, fazer requisições de materiais ou peças, etc. Foi através desta ferramenta, utilizada diariamente pelo autor desta dissertação, que se fez uma exploração prévia dos dados, anda numa fase embrionária deste trabalho.

Para a utilização de técnicas de *Data Mining*, e tendo em conta as ferramentas disponíveis no mercado, tal como apresentado no ponto 2.4, foi utilizado o ambiente de programação *R*. Esta ferramenta foi utilizada não só pelas suas características de desempenho mas também porque o autor tem um gosto particular pela utilização de ferramentas *open-source*, e ainda porque são de utilização livre e gratuita.

O *R* é um projeto aberto, e que funciona em diversos sistemas operativos (Windows, Unix, e MacOS). O *R* não é mais que uma linguagem de programação desenhada para a estatística e análise de dados. Embora não tenha sido especificamente desenvolvida para o *Data Mining*, esta ferramenta inclui uma elevada variedade de algoritmos como a regressão múltipla, redes neuronais, máquinas de vetores de suporte, entre outros, sendo que na atualidade é utilizada por um elevado número de analistas. Tal como exposto no quadro teórico de referência, esta é uma das ferramentas que muitos dos analistas de *Data Mining* utilizam nos seus projetos. O *R* tem a vantagem de ser muito flexível e extensível, permitindo a instalação de diversos pacotes que são desenvolvidos pela comunidade de utilizadores para fins específicos. Esta é uma comunidade particularmente ativa, levando a que novos pacotes/bibliotecas com novos modelos sejam codificados e atualizados muito frequentemente, respondendo à evolução e necessidade dos muitos utilizadores desta ferramenta. Segundo o site oficial do *R* (http://cran.r-project.org/web/packages), à data de escrita deste trabalho, existiam disponíveis 4938 bibliotecas *R.*

Para a realização deste trabalho foram instalados várias bibliotecas no *R*, de onde se destacam o *rattle, rminer* e o *forecast*.

A biblioteca *rattle* (*R Analytical Tool To Learn Easily*) é um interface gráfico para utilização do *R,* e *foi* utilizada numa fase exploratória dos dados, apresenta sumários visuais e estatísticos dos dados, permite algumas operações para a sua transformação, e ainda a criação e avaliação de alguns modelos de DM. Apesar da sua ação ser algo limitada, o *rattle*, devido ao seu interface gráfico simples e intuitivo, permite a um utilizador sem conhecimento do *R*, a realização de trabalhos de *Data Mining* (http://rattle.togaware.com/)

Foi igualmente utilizada a biblioteca *rminer* (Cortez, 2010), que é uma biblioteca de DM para o *R*. Esta é uma biblioteca que facilita o uso de algoritmos de classificação e regressão, apresentado um conjunto de funções simples e coerentes. Embora possam ser utilizados diversos algoritmos de DM, esta biblioteca está particularmente adaptada para redes neuronais e máquinas de vetores de suporte (Cortez, 2012). Contém funções várias que permitem, por exemplo, o treino e o teste de um determinado modelo, a computação de métricas de validação desse modelo, bem como a criação de gráficos dos modelos criados, entre outras funções. Esta biblioteca pode ser encontrada em http://www3.dsi.uminho.pt/pcortez/rminer.html e foi utilizada neste trabalho, pois permite de uma

forma expedita criar modelos de regressão, métricas de avaliação e ainda o desenho de gráficos para os modelos criados.

Recorreu-se ainda à biblioteca *forecast (*Hyndman & Khandakar, 2013). Esta biblioteca fornece métodos e ferramentas para a análise univariada de séries temporais, incluindo alisamento exponencial, entre outros. Foi utilizada neste trabalho pois permite realizar previsões baseadas numa série temporal. Esta biblioteca está disponível em http://robjhyndman.com/software/forecast/

Foi ainda utilizada a ferramenta *Microsoft Excel*, para a criação de tabelas contendo todos os dados alvo de estudo. Para além das variáveis de saída, extraídas do SGM do Aeroporto, foram recolhidos dados meteorológicos de várias fontes, bem como os dados relativos ao número de passageiros processados e movimentos de aeronaves. Esta ferramenta é de simples utilização, construindo facilmente conjuntos de dados coerentes, com proveniências diferenciadas.

## 4.4 Técnicas de *Data Mining*

As técnicas de *Data Mining* utilizadas para a análise dos dados deste trabalho são as já anteriormente descritas: Redes Neuronais (*NN)*, Máquinas de Vetores de Suport (SVM) e ainda a Regressão Múltipla (*MR*), uma vez que estamos perante um problema de regressão, com vários atributos de entrada e uma variável de saída.

Como adiante veremos, e uma vez que os dados em estudo são de carácter sequencial e periódico, optou-se também por estudar a previsão de incidências e ordens de trabalho através análise de séries temporais, ou seja, criando modelos de previsão com base apenas na evolução temporal da variável de saída.

## 4.5 Medidas de Desempenho da Previsão

Em termos de métricas para o desempenho das previsões, e uma vez que estamos perante um problema de regressão, foram utilizadas três formas diferentes para a avaliação de cada um dos modelos: $R^2$, MAE e RMSE. A explicação da forma como cada uma destas métricas é utilizada está expressa no ponto 2.6 desta dissertação. São utilizadas três métricas em cada um dos modelos precisamente para existir uma validação sucessiva dos seus valores.

## 4.6 Trabalho Prático

A limitação humana para a execução de modelos de previsão complexos, levou à necessidade da utilização do *Data Mining* para analisar os dados em estado bruto e extrair informações de alto nível que possam ser úteis para os decisores.

O trabalho prático desenvolve-se de forma sequencial, seguindo o CRISP-DM, dado o carácter cíclico, flexível e iterativo, e que levaram á escolha desta metodologia para a realização desta dissertação. Algumas das iterações poderão não ser descritas ou até mesmo executadas nalgum ponto da aplicação da metodologia. Nos próximos pontos, descrevem-se de forma sistematizada as atividades e tarefas realizadas tendo em conta as seis fases do CRISP-DM: *Business Understanding*

(estudo/compreensão do negócio); *Data Understanding* (compreensão dos dados); *Data Preparation* (preparação dos dados); *Modeling* (modelação); *Evaluation* (avaliação) e; *Deployment* (implementação).

### 4.6.1 Compreensão do negócio [CRISP-DM: *Business Understanding*]

Nesta fase, o CRISP-DM não só descreve os passos para a compreensão de negócio, como auxilia na verificação na viabilidade do projeto. Dentro desta fase temos vário passos que devem ser executados. O primeiro é a determinação dos objetivos de negócio. Tal como descrito anteriormente, o objetivo principal de negócio será o de conhecer melhor o volume e a distribuição das incidências e ordens de trabalho de manutenção do Aeroporto de Lisboa. Através deste conhecimento, poder-se-ão gerir melhor os recursos disponíveis e planear necessidades futuras. O critério de sucesso deste trabalho será a verificação de qual a melhor técnica ou modelo de DM para este propósito. Futuramente, e caso exista uma aplicação prática deste trabalho, poderemos compreender qual o impacto ou utilidade no negócio em questão.

Um resultado de negócio que advém da persecução do objetivo de negócio, não vem exclusivamente da qualidade das previsões e que permitirão dimensionar os recursos alocados a estas tarefas. A previsão de um maior ou menor número de incidências, permite dimensionar o número de recursos humanos adequados para o Centro de Monitorização Técnica (CMT), e que monitorizam os sistemas em tempo real e respondem às chamadas e pedidos de assistência técnica. Permite ainda dimensionar o número de recursos humanos alocados à manutenção de primeira linha, vista como incidência. Por outro lado, a previsão do número de ordens de trabalho, permitirá igualmente ao gestor dimensionar o número e tipo de recursos humanos alocados a tarefas de manutenção corretiva não resolvida no âmbito das incidências/primeira linha. Igualmente importante, para além destes resultados, será o de se conseguir determinar quais os melhores períodos para trabalho de manutenção preventiva, reparações simples e, inclusivamente, trabalhos de conservação profunda. Os técnicos de manutenção têm um conjunto de trabalhos planeados, designado como Manutenção Preventiva, e que visa intervir em espaços físicos, equipamentos ou sistemas em momentos que não causem transtorno operacional. Este tipo de trabalho representa uma boa parte do trabalho dos técnicos, sendo planeado de forma atempada e em coordenação com a área operacional. Desta forma, saber com antecedência, e com alguma precisão, quantas incidências de manutenção irão ocorrer, poderá ser muito importante para o planeamento do trabalho e de turnos dos vários sectores.

A título de exemplo, se conseguirmos determinar com qualidade quais os meses de menores ocorrências de manutenção, poderemos não só planear uma melhor distribuição e alocação dos recursos humanos o mesmo período, como planear os trabalhos de manutenção preventiva ou de conservação para esses meses, levando a um impacto menor sobre o equipamento ou local em causa.

Recorde-se que as incidências de manutenção são eventos que ocorrem de forma extemporânea, tendo a sua resolução de ser o mais rápida possível. Com esta noção poder-se-ão tomar medidas

para organizar turnos e equipas, reforçando ou reduzindo determinados períodos de tempo, de modo a minimizar o impacto e o custo decorrente das incidências de manutenção.

O facto de se conseguir prever com exatidão o objetivo, não só poderá aumentar o desempenho das equipas de manutenção, como reduzir o risco de problemas operacionais graves.

A título de exemplo podemos referir que o CMT e os vários sectores asseguram, entre outros, a manutenção dos equipamentos. Em momentos de grande fluxo, a não existência de um técnico disponível para rapidamente reparar um equipamento de Raio-X poderá levar a um constrangimento no processamento dos passageiros. Se houver simultaneidade neste tipo de incidências, o constrangimento poderá passar a graves transtornos operacionais que poderão levar a custos de indemnização de clientes e companhias aéreas.

O recurso principal para a realização deste trabalho é a base de dados do Sistema de Gestão de Manutenção do Aeroporto de Lisboa. Este sistema contém toda a informação relacionada com todas as intervenções de manutenção no Aeroporto de Lisboa. Foi garantida a autorização para o acesso e tratamento destes dados. Sendo este um trabalho académico, não existe uma relação custo-benefício que se possa determinar de forma objetiva. Se o resultado deste trabalho for de interesse para os gestores da organização e, simultaneamente, provocar alguma melhoria, será sempre positivo o saldo em termos de retorno do investimento, uma vez que o mesmo é nulo por parte da organização.

Em termos de riscos do presente trabalho, os mesmos também são reduzidos, pois este é um estudo de caso, não levando necessariamente a uma implementação dos seus resultados práticos, independentemente da sua qualidade ou interesse para a organização.

### 4.6.2 Compreensão dos dados [CRISP-DM: *Data Understanding*]

Apresentados que estão os objetivos de negócio, e avaliados os eventuais riscos, passemos aos dados em estudo. Foram então retirados dados de diversas fontes, nomeadamente o número de incidências e ordens de trabalho de manutenção ocorridas no Aeroporto de Lisboa, mensalmente, e para um período de seis anos, bem como dados meteorológicos históricos, relativos aos meses em questão, retirados de vários sítios na internet, e que foram validados, por comparação, com dados oficiais do Instituto Português do Mar e da Atmosfera. Os dados relativos à pluviosidade foram obtidos através do site do SNIRH – Sistema Nacional de Informação de Recursos Hídricos.

Para se alcançarem os objetivos de negócio, ir-se-ão comparar diversos modelos, para o período de tempo em questão, analisando a evolução das incidências e ordens de trabalho de manutenção ao longo do tempo e, dependendo da técnica utilizada para a criação do modelo, utilizando os outros dados compilados (dados meteorológicos, passageiros, movimentos de aeronaves) caso se revelem de interesse.

Na Tabela 6 apresenta-se o nome de cada um dos atributos, bem como uma descrição sucinta do seu significado.

**Tabela 6 – Resumo dos atributos dos dados utilizados no trabalho**

| Atributo | Descrição | Tipo |
|---|---|---|
| Mes | Mês do ano [AAAA-MM] | Nominal |
| Tmax | Temperatura máxima registada no mês [ºC] | Numérico |
| Tmed | Temperatura média registada no mês [ºC] | Numérico |
| Tmin | Temperatura mínima registada no mês [ºC] | Numérico |
| Hmax | Humidade relativa máxima registada no mês [%] | Numérico |
| Hmed | Humidade relativa média mensal [%] | Numérico |
| Hmin | Humidade relativa mínima registada no mês [%] | Numérico |
| Chuva | Valor máximo de precipitação [mm] | Numérico |
| WindMax | Velocidade máxima do vento registado no mês [km/h | Numérico |
| WindMed | Velocidade média do vento registado no mês [km/h] | Numérico |
| RajadaMax | Valor máximo da rajada no mês [km/h] | Numérico |
| Mov | Número de movimentos de aeronaves por mês | Numérico |
| Pax | Número de passageiros por mês | Numérico |
| Inc | Número total de incidências por mês | Numérico |
| OT | Número total de ordens de trabalho por mês | Numérico |

Descrevendo mais em detalhe alguns dos atributos da tabela anterior:

**Mov – Movimentos:** número total de movimentos (aterragens e descolagens) de aeronaves por mês.

**Pax – Passageiros:** número total de passageiros processados por mês. Este valor inclui passageiros que partem e chegam ao Aeroporto, e ainda os que realizam ligações de transferência.

**Inc - Incidência:** É o documento que contém toda a informação relacionada com uma ação de manutenção, resultante de um pedido direcionado ao Centro de Monitorização Técnica (CMT) ou registado a partir de uma anomalia verificada através de um sistema de supervisão presente neste local. Este é um dos atributos que se pretende prever.

**OT - Ordem de trabalho:** Formulário que contém toda a informação de uma determinada ação de manutenção, resultante de uma incidência não resolvida através da primeira linha, ou de um trabalho de manutenção corretiva que não passou pelo CMT. O número de ordens de trabalho corresponde ao número de intervenções realizadas num determinado período que caiam dentro desta descrição. Este atributo é uma variável de saída, ou seja o atributo que se pretende prever.

Todos os atributos são relativos a dados meteorológicos ou operacionais, pelo que é fácil de compreender a sua natureza. Ao explorarmos os dados, verificamos que um atributo como Hmax, provavelmente, será pouco relevante para se prever o número de incidência ou ordens de trabalho, uma vez o valor máximo de humidade em quase todos os meses do período em estudo é de 100%. Apenas em sete meses foi diferente deste valor.

Relativamente à qualidade dos dados, dados omissos e *outliers* (valores atípicos) verificou-se que existem dados omissos para alguns meses dos atributos Chuva e Tmin. Estes dados são obtidos a partir de estações meteorológicas, e por vezes devido a problemas técnicos ou falhas de

comunicação não é possível obter todos os dados. Por exemplo, os dados relativos ao atributo Chuva, que são obtidos através de uma estação do Sistema Nacional de Informação de Recursos Hídricos, situada em São Julião do Tojal (a mais próxima do Aeroporto de Lisboa), apresentam por vezes dados omissos, e inclusivamente nesse campo apresentam uma descrição da avaria.

Relativamente à relação entre atributos, uma maneira de ver a relação de forma isolada, entre cada par de atributos é através da correlação linear. Através do R, e de um simples comando aplicado à tabela contendo todos os atributos para o período em estudo, conseguimos ver a relação entre cada atributo de forma isolada. A Tabela 7 apresenta a correlação entre os atributos Inc e OT e o peso individual de cada um dos outros atributos. Nesta tabela, foi também colocado o valor dos atributos Mov e Pax, de modo a poder-se confirmar algo que está intimamente ligado: o número de movimentos de aeronaves e o número de passageiros processados no Aeroporto. Podemos desta forma verificar que, como seria de esperar, existe uma correlação forte de +0,93 entre o número de passageiros (Pax) e o número de movimentos de aeronaves (Mov). No caso da variável de saída Inc que pretendemos prever, a correlação é fraca com o atributo Mov, sendo que, de forma isolada, o que mais contribui para este valor é também o atributo Pax e ainda o atributo Chuva. Esta relação mais forte faz sentido, pois sabe-se de antemão que quando há maiores valores de precipitação, existem mais incidências de manutenção, muito devido a inundações ou outros danos provocados pela água.

**Tabela 7 – Correlação entre OT, Inc e os demais atributos**

| | **Inc** | **OT** | **Chuva** | **Tmax** | **Tmed** | **Tmin** | **Hmax** | **Hmed** | **Hmin** | **WMax** | **WMed** | **RMax** | **Mov** | **Pax** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Inc** | 1,00 | **0,52** | **0,27** | 0,06 | 0,10 | 0,13 | 0,03 | 0,00 | 0,03 | -0,06 | -0,09 | 0,11 | 0,09 | **0,23** |
| **OT** | **0,52** | 1,00 | 0,05 | 0,08 | 0,02 | 0,10 | 0,05 | -0,04 | **-0,32** | -0,09 | -0,08 | -0,04 | **0,26** | **0,36** |
| **Mov** | 0,09 | 0,26 | -0,51 | 0,76 | 0,78 | 0,73 | -0,21 | -0,57 | -0,61 | 0,17 | 0,60 | -0,20 | 1,00 | 0,93 |
| **Pax** | 0,23 | 0,36 | -0,48 | 0,84 | 0,85 | 0,79 | -0,28 | -0,68 | -0,64 | 0,13 | 0,56 | -0,21 | 0,93 | 1,00 |

Por outro lado, um maior número de passageiros no Aeroporto, faz com que o número de incidências aumente, pois existe uma maior utilização da infraestrutura aeroportuária e o número de solicitações aos serviços de manutenção aumentam. A análise destes dados tendo como objetivo a previsão do número de ordens de trabalho e incidências por setor de manutenção, também poderá ser importante, para se poder ter informação para a gestão de cada um dos setores individualmente.

### 4.6.3 Preparação dos dados [CRISP-DM: *Data Preparation*]

Para dar mais coerência aos dados, a opção foi substituir os valores omissos utilizando o comando "imputation" existente no package *rminer* do *R*. Este comando permite substituir os valores omissos por outros valores. Neste caso foi utilizado o método "hotdeck", que procura primeiro nos dados o exemplo mais parecido. O "imputation hotdeck" substitui os dados omissos por dados comparáveis retirados da mesma tabela, ou seja através da observação dos dados existentes no próprio ficheiro de dados. Para o caso, por exemplo, dos dados de pluviosidade, o algoritmo vai olhar para os meses com características semelhantes de modo a preencher os valores omissos com o valor mais próximo da realidade. De notar que temos de tratar com algum cuidados os resultados do "hotdeck" pois estes dados são retirados da mesma tabela/ficheiro. Apesar de válidos, estes valores não são

necessariamente consistentes pois baseiam-se apenas numa coleção de dados específica, mas serão, com alguma probabilidade, dados próximos da realidade. Como se tratam apenas de alguns exemplos num universo de 72 ocorrências, considera-se que este método é adequado.

Nas primeiras iterações em que se utilizem modelos de *Data* Mining, irão ser utilizados todos os atributos, e feitos testes e comparações com modelos em que se retirem os atributos com menor correlação com OT e Inc. Um modelo com menos atributos é um modelo que poderá ser mais facilmente explicado.

Eventualmente nas iterações seguintes serão utilizados menos atributos, tentando obter um modelo mais simples e que tenham maior qualidade que o das iterações anteriores.

Os dados a utilizar, e tal como descrito no ponto anterior, têm várias proveniências pelo que foi necessário proceder à sua compilação e integração, nomeadamente através do *Microsoft Excel*, onde facilmente se pode reunir várias origens e tipos de dados numa só tabela e depois exportá-los para ficheiros de texto para serem utilizados no *R*.

Para além dos modelos de *Data Mining* de regressão, foi também utilizada a análise de séries temporais univariadas, o que levou à necessidade de se criar uma série para cada um das variáveis em estudo: Inc e OT, com periodicidade mensal, que foi a que se conseguiu obter através da extração dos dados.

Sendo este um trabalho de natureza exploratória, e dada a flexibilidade do CRISP-DM, optou-se por avançar para as primeiras iterações sem especial preparação dos dados, de modo a poder-se comparar previsões várias com diferentes modelos e atributos.

### 4.6.4 Modelação [CRISP-DM: *Modeling*]

Para o desenvolvimento desta fase, que é a fase em que efetivamente se vão realizar testes para a criação de modelos de previsão, foram então utilizadas as já anteriormente referidas bibliotecas do *R*, o *rminer* e o *forecast*. Em termos mais concretos, e tendo em conta os objetivos de negócio, teremos de criar modelos para estudar o número total de incidências por mês, o número total de OT por mês, e ainda o valor total para incidências e OT para um setor de manutenção selecionado, que servirá de exemplo para trabalhos futuros. O setor selecionado foi o setor de Manutenção Eletrónica (ELN).

#### 4.6.4.1 Modelação: Número total de Incidências

Foram selecionados três modelos de análise de regressão de tipologia diferente: Redes Neuronais (*NN)*, Máquinas de Vetores de Suport (SVM) e ainda a Regressão Múltipla (*MR)* com "holdout" ordenado no tempo. Desta forma usa-se como treino do modelo uma percentagem dos dados que correspondem aos dados mais antigos, ficando os restantes dados para teste. Em termos de métrica foram utilizadas três formas diferentes para cada um dos modelos: $R^2$, MAE e RMSE. O modelo de *NN* selecionado, foi o MLPE (*multilayer perceptron ensemble*).

Neste trabalho, foi fixado um holdout ordenado com 80% dos dados mais antigos para treino e os restantes 20% mais recentes para teste dos três modelos de previsão.

Para tal foi utilizado o comando *mining* da biblioteca *rminer* para modelar os dados, que permite de uma forma muito expedita construir o modelar os dados. A título de exemplo, apresentamos a linha de código *R* utilizada para criar um modelo de *Data Mining* que faça a previsão da variável de saída Inc, constante no objeto otinc (objecto que contém todos os dados apresentados anteriormente):

*m1mlpe=mining(Inc~.,otinc,model="mlpe",method=c("holdoutorder",0.8),Runs=1,search="heuristic10")*

Foram feitos diversos testes, com recurso à mesma tabela de dados mas ignorando alguns dos atributos, nomeadamente aqueles que menos valores de correlação com a variável de saída Inc apresentavam. Na Tabela 8 é apresentado o resumo dos resultados.

**Tabela 8 – Comparação dos modelos de regressão na previsão do número total de Incidências**

| Objeto R | Modelo | $R^2$ | MAE | RMSE |
|---|---|---|---|---|
| **m1mlpe** | NN (MLPE) | -1,13 | 167,47 | 195,96 |
| **m2svm** | SVM | -0,60 | 140,91 | 169,30 |
| **m2mr** | MR | -2,11 | 192,87 | 237,32 |

O resultado da aplicação desta função é guardado no objeto "m1mlpe" e que poderá facilmente ser avaliado através do comando *mmetric* também da biblioteca *rminer.* Este comando permite obter valores de várias métricas de desempenho apenas através de uma linha de código.

Para este conjunto de dados e aplicando métodos de regressão, o melhor modelo é o que utiliza o algoritmo de Máquinas de Vector de Suporte (*Support Vector Machines* – SVM*)*, pois apresenta o valor de $R^2$ mais próximo de zero. Este desempenho é também confirmado pelas métricas MAE e RMSE.

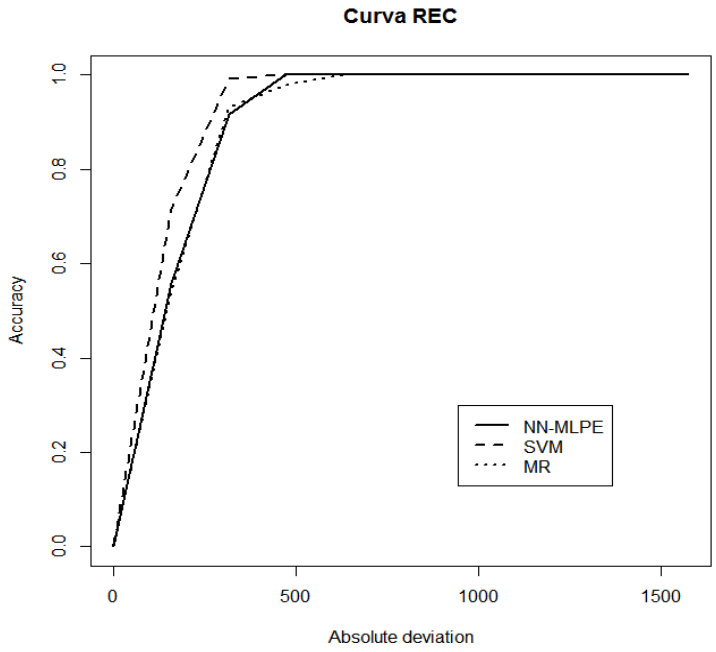


**Figura 6 – Curvas REC para os 3 modelos de *Data Mining* utilizados, total de Incidências**

Através da função *mgraph*, também existente na biblioteca *rminer*, foi possível desenhar um gráfico contendo as curvas REC (*Regression Error Characteristic)* para cada um dos modelos testados, e em que se pode verificar que o melhor dos 3 modelos estudados é o SVM. A curva REC apresenta no eixo das abcissas a tolerância de erro absoluto que se pode admitir e no eixo das coordenadas a percentagem de exemplos corretamente previstos dentro dessa tolerância. Quando mais elevada for a área da curva REC, melhor o modelo de previsão.

Sendo os dados em estudo de carácter temporal e periódico, optou-se também por estudar a previsão de incidências através da análise de séries temporais, ou seja, criando um modelo de previsão com base apenas na evolução temporal da variável de saída, neste caso, o número de incidências, Inc.

Desta forma, e da mesma forma que para os modelos de regressão, criou-se então uma tabela contendo todos os valores de incidências por mês, para o período histórico de 72 meses. No *R,* esta tabela terá de ser transformada num objeto de série temporal, utilizando-se para isso o comando *ts*, identificando também que o valor é mensal. A biblioteca *forecast* do *R*, apresenta diversas funções para a análise temporal univariada.

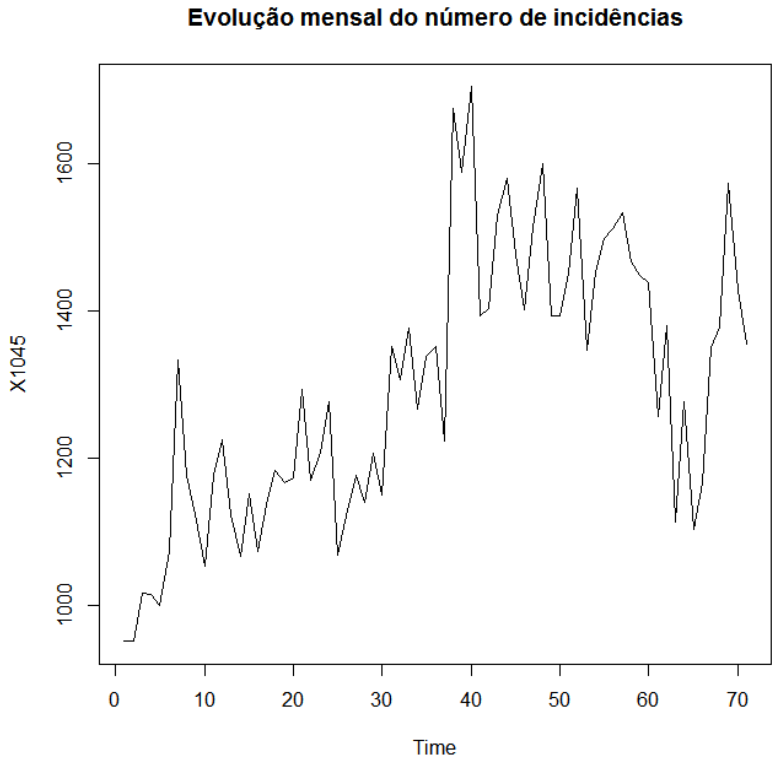


**Figura 7 – Evolução mensal do número de incidências**

Na Figura 7 podemos ver a evolução do número de incidências ao longo dos vários meses da amostra em estudo. O mês 0 corresponde a Janeiro de 2007, que é o início do período em estudo. No eixo das ordenadas temos o número de incidências. Podemos ver que não é um valor constante, apesar de ter ocorrido alguma tendência para aumentar. Através desta série temporal podemos utilizar a função "HoltWinters" no R para criar modelos de alisamento exponencial simples e

alisamento linear de Holt, através da mudança de parâmetros nesta função. O modelo sazonal Holt-Winters não foi executado porque de acordo com a Figura 7, a série temporal é de curta dimensão e não aparenta ter um padrão sazonal."

Para o caso do alisamento exponencial simples, e tal como anteriormente descrito, consegue-se obter um gráfico em que são apresentados os valores observados e os valores previstos, com a série temporal original em preto e os valores previstos por este método a vermelho.

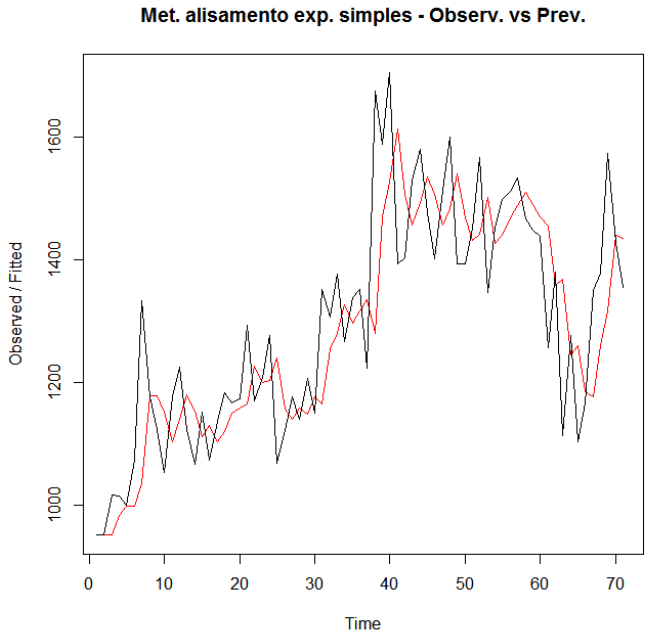


**Figura 8 – Valores Observados vs Previstos – Método de Alisamento Exponencial Simples**

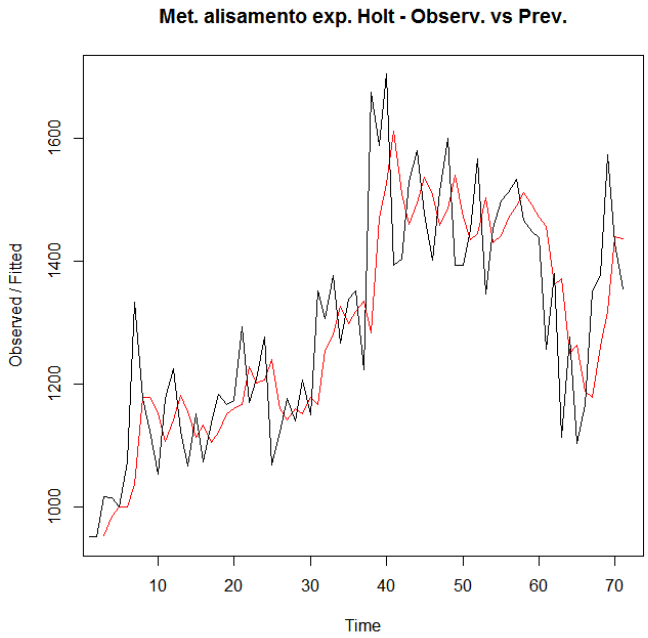


**Figura 9 – Valores Observados vs Previstos – Método de Alisamento Exponencial de Holt**

Podemos obter um gráfico semelhante para o método de alisamento exponencial de Holt, em que os dados são apresentados da mesma forma (Figura 9).

Com a função do *rminer*, *mmetric,* podemos, tal como anteriormente determinar as métricas de erro para estes dois modelos de análise de séries temporais, de modo a podermos compará-los com os modelos de regressão utilizados.

**Tabela 9 – Comparação dos modelos de regressão e séries temporais na previsão do número total de Incidências**

| Objeto R | Modelo | $R^2$ | MAE | RMSE |
|---|---|---|---|---|
| **m1mlpe** | NN (MLPE) | -1,13 | 167,47 | 195,96 |
| **m2svm** | SVM | -0,60 | 140,91 | 169,30 |
| **m2mr** | MR | -2,11 | 192,87 | 237,32 |
| **holt_ts_inc** | Alis. Exp. Simples | **0,82** | **49,61** | **76,16** |
| **holt_ts_inc2** | Alis. Exp. Holt | 0,62 | 73,52 | 107,91 |

Através destas métricas, podemos verificar que o método de Alisamento Exponencial Simples se apresenta como o melhor método, com um valor de $R^2$ de 0,82, que representa uma relação forte entre os resultados previstos e observados.

#### 4.6.4.2 Modelação: Número total de OT (Ordens de Trabalho)

Para a previsão do número total de OT, foram selecionados os mesmos três modelos de análise de regressão: redes neuronais, Redes Neuronais (*NN)*, Máquinas de Vetores de Suport (SVM) e ainda a Regressão Múltipla (*MR)*, com um "holdout" ordenado no tempo com 80% dos dados mais antigos para treino e os restantes 20% (os mais recentes) para teste dos três modelos de previsão.

O modelo com melhor desempenho para este caso é também o *SVM* (Máquinas de Vetores de Suporte), tal como se pode verificar na Tabela 9, que apresenta a comparação dos vários modelos de regressão testados para a previsão da variável OT.

**Tabela 10 – Comparação dos modelos de regressão na previsão do número total de OT**

| Objeto R | Modelo | $R^2$ | MAE | RMSE |
|---|---|---|---|---|
| **m3mlpeot** | NN (MLPE) | -0,56 | 146,35 | 183,34 |
| **m3svmot** | SVM | **-0,11** | **131,78** | **154,46** |
| **m1mrot** | MR | -2,12 | 192,87 | 237,32 |

De notar que, à semelhança do ponto anterior, foram realizados diversos testes com menos atributos de entrada, ignorando aqueles cuja análise de correlação era a mais baixa, entre outros testes, sendo seguro que o melhor modelo, apesar de ser mais complexa a sua interpretação, foi o modelo que continha todos os atributos constantes do conjunto de dados analisados.

Sendo os dados em estudo de carácter temporal e periódico, optou-se, à semelhança da previsão anterior, por estudar a previsão de OT, através da análise de séries temporais, ou seja, criando um

modelo de previsão com base apenas na evolução temporal da variável de saída, neste caso, o número de Ordens de Trabalho, OT.

**Tabela 11 – Comparação dos modelos de regressão e séries temporais na previsão do número total de OT**

| Objeto R | Modelo | $R^2$ | MAE | RMSE |
|---|---|---|---|---|
| **m3mlpeot** | NN (MLPE) | -0,56 | 146,35 | 183,34 |
| **m3svmot** | SVM | -0,11 | 131,78 | 154,46 |
| **m1mrot** | MR | -2,12 | 192,87 | 237,32 |
| **holt_ts_ot** | Alis. Exp. Simples | **0,88** | **38,18** | **214,27** |
| **holt_ts_ot2** | Alis. Exp. Holt | 0,69 | 153,95 | 344,74 |

Através destas métricas, podemos verificar que o método de Alisamento Exponencial Simples se apresenta como o melhor método, com um valor de $R^2$ de 0,88 que representa uma relação forte entre os resultados previstos e observados. Este valor é também confirmado pelo melhor erro absoluto médio (MAE) dos vários testes realizados.

#### 4.6.4.3 Modelação: Número de Incidências de um setor de Manutenção

De modo a poder-se demonstrar o interesse e a pertinência do desenvolvimento este trabalho por setor de Manutenção, foi selecionado um setor em particular para ser estudado: o setor de Manutenção Eletrónica (ELN). Caso se revele importante, poder-se-ão fazer, como trabalho futuro, os modelos de previsão aplicados aos dados de outros setores de manutenção. Os modelos utilizados foram os mesmos três modelos de regressão e ainda os modelos de análise de séries temporais.

Para o estudo deste setor, as variáveis de entrada serão as mesmas, mas como os valores mensais da variável de saída são diferentes, fez-se nova análise de correlação, no sentido de se verificar a relação individual entre cada um dos atributos. A Tabela 12 apresenta esta correlação.

**Tabela 12 – Correlação entre OT e demais atributos para o setor de Manutenção Eletrónica**

| | Inc | OT | Chuva | Tmax | Tmed | Tmin | Hmax | Hmed | Hmin | WMax | WMed | RMax | Mov | Pax |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Inc** | 1,00 | 0,52 | -0,10 | -0,02 | 0,00 | 0,00 | 0,05 | 0,11 | 0,22 | 0,00 | -0,03 | 0,02 | -0,06 | -0,13 |
| **OT** | 0,52 | 1,00 | -0,04 | 0,01 | 0,03 | 0,01 | 0,03 | -0,01 | 0,01 | -0,05 | -0,14 | -0,04 | -0,21 | -0,19 |

O número de incidências será menor uma vez que estamos a analisar apenas um setor de manutenção e não a totalidade dos dados. De modo a evidenciar a ordem de grandeza do número de incidências para o setor em estudo, é apresentado na tabela seguinte um excerto dos dados mensais contendo o número de incidências que foram resolvidas pelos técnicos de primeira linha alocados a este setor de manutenção.

**Tabela 13– Número de ordens de trabalho executadas pelo setor de Manutenção Eletrónica (excerto do período em estudo)**

| Mês | Incidências |
|---|---|
| **Jul-11** | 77 |
| **Ago-11** | 53 |
| **Set-11** | 52 |
| **Out-11** | 61 |
| **Nov-11** | 73 |
| **Dez-11** | 70 |
| **Jan-12** | 51 |
| **Fev-12** | 69 |
| **Mar-12** | 53 |
| **Abr-12** | 39 |
| **Mai-12** | 59 |
| **Jun-12** | 40 |

Através das métricas cujos valores se encontram resumidos na Tabela 14, podemos verificar que o método de Alisamento Exponencial Simples se apresenta como o melhor método, com um valor de $R^2$ de 0,53 que representa uma relação relativamente forte entre os resultados previstos e observados. Este valor é também confirmado pelos valores mais baixos para as medidas de erro MAE e RMSE.

**Tabela 14 – Comparação dos modelos de regressão e séries temporais na previsão do número de OT para o setor de Manutenção Eletrónica**

| Objeto R | Modelo | $R^2$ | MAE | RMSE |
|---|---|---|---|---|
| **m1mlpeelninc** | NN (MLPE) | 0,10 | 11,64 | 13,24 |
| **m1svmelninc** | SVM | -0,14 | 12,97 | 14,98 |
| **m1mrelninc** | MR | 0,10 | 11,65 | 13,24 |
| **holt_eln_inc** | Alis. Exp. Simples | **0,53** | **6,83** | **8,62** |
| **holt_eln_inc2** | Alis. Exp. Holt | 0,06 | 10,07 | 12,27 |

#### 4.6.4.4 Modelação: Número de OT (Ordens de Trabalho) de um setor de Manutenção

Para a previsão do número de OT para um setor de manutenção foi também selecionado o setor de Manutenção Eletrónica (ELN), de modo a poder demonstrar o interesse e a pertinência de se fazer este tipo de modelação por setor de Manutenção. Caso se revele importante, poder-se-ão fazer, como trabalho futuro, os modelos de previsão para outros setores de manutenção. Os modelos utilizados foram os mesmos três modelos de regressão e ainda os modelos de análise de séries temporais.

Neste caso, o número de OT, tratando-se de apenas de um setor específico de manutenção, será menor, pois corresponde às ordens de trabalho realizadas apenas pelos técnicos deste setor, e têm a ordem de grandeza da tabela seguinte.

**Tabela 15– Número de ordens de trabalho executadas pelo setor de Manutenção Eletrónica (excerto do período em estudo)**

| Mês | Ordens de Trabalho (OT) |
|---|---|
| **Jul-11** | 55 |
| **Ago-11** | 52 |
| **Set-11** | 38 |
| **Out-11** | 54 |
| **Nov-11** | 60 |
| **Dez-11** | 60 |
| **Jan-12** | 79 |
| **Fev-12** | 79 |
| **Mar-12** | 68 |
| **Abr-12** | 39 |
| **Mai-12** | 46 |
| **Jun-12** | 45 |

À semelhança dos pontos anteriores, foram realizados diversos testes com menos atributos de entrada, ignorando aqueles cuja análise de correlação era a mais baixa. Os melhores modelos de regressão, apesar de ser mais complexa a sua interpretação, foram sempre os modelos que continham todos os atributos constantes do conjunto de dados analisados. Na Tabela 16 temos o resumo dos valores dos erros para cada um dos modelos.

**Tabela 16 – Comparação dos modelos de regressão e séries temporais na previsão do número de OT para o setor de Manutenção Eletrónica**

| Objeto R | Modelo | $R^2$ | MAE | RMSE |
|---|---|---|---|---|
| **m1mlpeeln** | NN (MLPE) | 0,22 | 10,59 | 13,41 |
| **m1svmeln** | SVM | -0,11 | 13,37 | 15,27 |
| **m1mreln** | MR | -0,29 | 14,59 | 17,21 |
| **holt_eln** | Alis. Exp. Simples | **0,83** | 38,90 | **8,20** |
| **holt_eln2** | Alis. Exp. Holt | 0,54 | **9,58** | 13,17 |

Através destas métricas, podemos verificar que o método de Alisamento Exponencial Simples se apresenta como o melhor método, com um valor de $R^2$ de 0,83 que representa uma relação forte entre os resultados previstos e observados. Este valor é também confirmado pelo melhor erro RMSE dos vários testes realizados.

### 4.6.5 Avaliação CRISP-DM: *Evaluation*]

Após a modelação para cada um dos objetivos, é importante sistematizar e avaliar os resultados, nomeadamente através da comparação dos resultados das previsões realizadas através dos vários métodos de regressão e de análise de séries temporais.

Para a previsão do número de incidências foram avaliados três modelos de regressão, com base nos dados relativos às OT e Incidências, aos quais foram acrescentados dados meteorológicos e de natureza operacional. Após a verificação de que, para este caso, estes modelos tinham fraca qualidade, optou-se também por realizar análise de séries temporais. Como poderemos facilmente constatar na figura seguinte, o modelo de regressão com erros mais baixos é o SVM – Máquinas de Vetores de Suporte, mas, quando comparado com qualquer dos modelos de análise de séries temporais, apresenta valores de erro muito superiores.

Os modelos de séries temporais apresentam melhores resultados neste caso, e ainda a vantagem de serem de mais simples visualização e análise. Para este caso concreto, o modelo realizado com o método de Alisamento Exponencial Simples, revelou-se como o mais adequado, por apresentar erros mais baixos e, simultaneamente, apresentar um $R^2$, ou seja, um coeficiente de determinação mais próximo de 1. Este é um modelo que apesar de ter um intervalo de erro relativamente elevado, apresenta uma grande proximidade entre os valores previstos e reais.

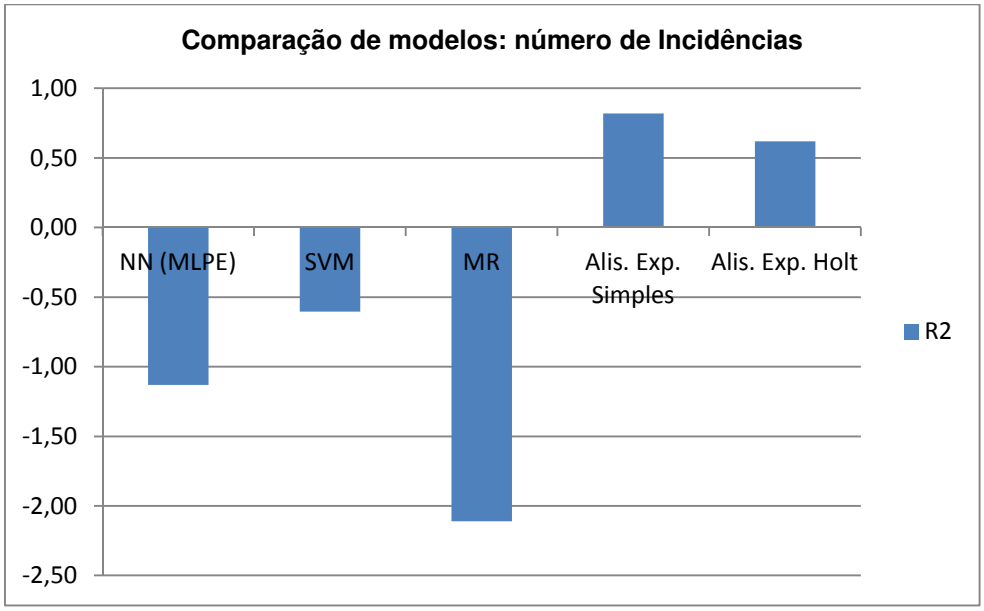


**Figura 10 – Gráfico do valor de $R^2$ para os diferentes modelos (número de Incidências)**

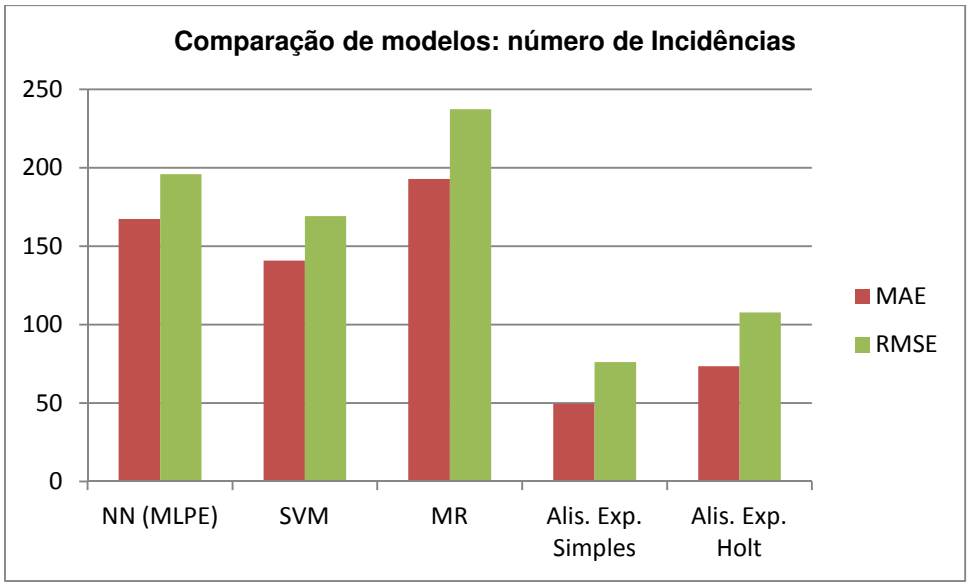


**Figura 11 – Gráfico dos valores de MAE e RMSE para os diferentes modelos**

No caso dos modelos utilizados para a previsão do número de ordens de trabalho, foram utilizados os modelos de regressão e ainda os dois métodos de análise de séries temporais. Como poderemos facilmente constatar na figura seguinte, o modelo de regressão com erros mais baixos é igualmente o SVM – Máquinas de Vetores de Suporte, mas, quando comparado com qualquer dos modelos de análise de séries temporais, apresenta valores de erro muito superiores, como poderemos facilmente constatar nos gráficos das figuras seguintes.

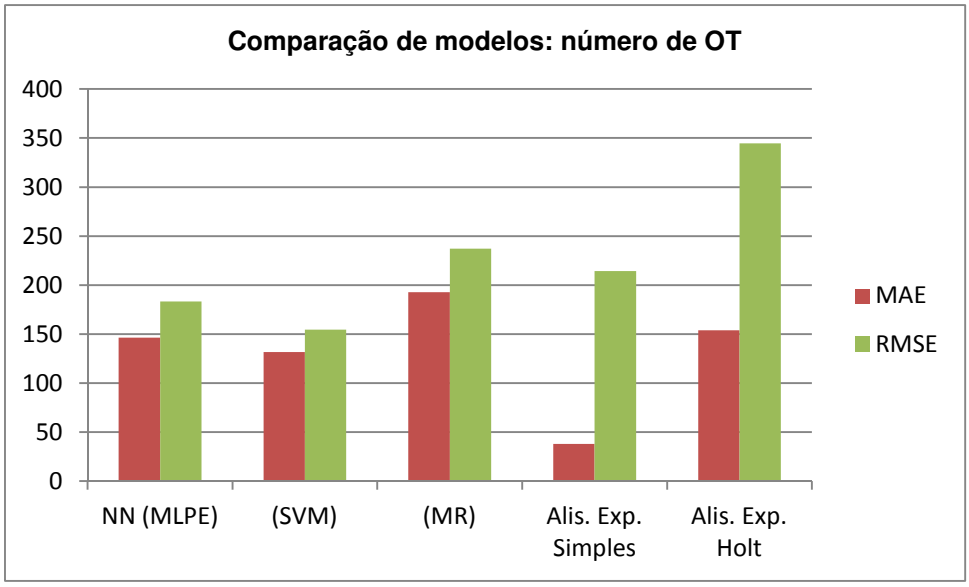


**Figura 12 – Gráfico do valor de MAE e RMSE para os diferentes modelos (número de OT)**

Igualmente para este caso, o da previsão do número total de OT, os modelos de séries temporais apresentam melhores resultados, e apresentando ainda a vantagem de serem de visualização e análise mais simples. Para este caso, o método de Alisamento Exponencial Simples revelou-se como o mais adequado, por apresentar erros mais baixos e um coeficiente de determinação ($R^2$) mais

próximo de 1, tal como expresso no gráfico da figura seguinte. Este é um modelo que, apesar de ter um intervalo de erro relativamente elevado, apresenta uma grande proximidade entre os valores previstos e reais.

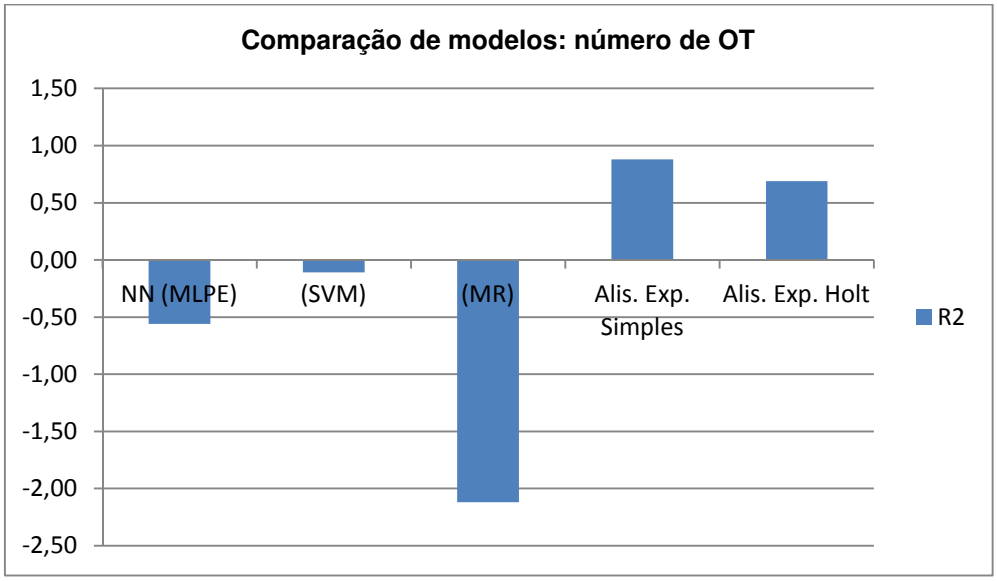


**Figura 13 – Gráfico do valor de MAE e RMSE para os diferentes modelos (número de OT)**

Para o caso das previsões realizadas para o setor específico de Manutenção Eletrónica, os resultados finais da comparação entre modelos acabam por ser semelhantes aos dos modelos criados para a previsão dos valores totais de Incidências e ordens de trabalho, em que os modelos de análise de séries temporais têm sempre um desempenho superior, quando comparados com os modelos de regressão.

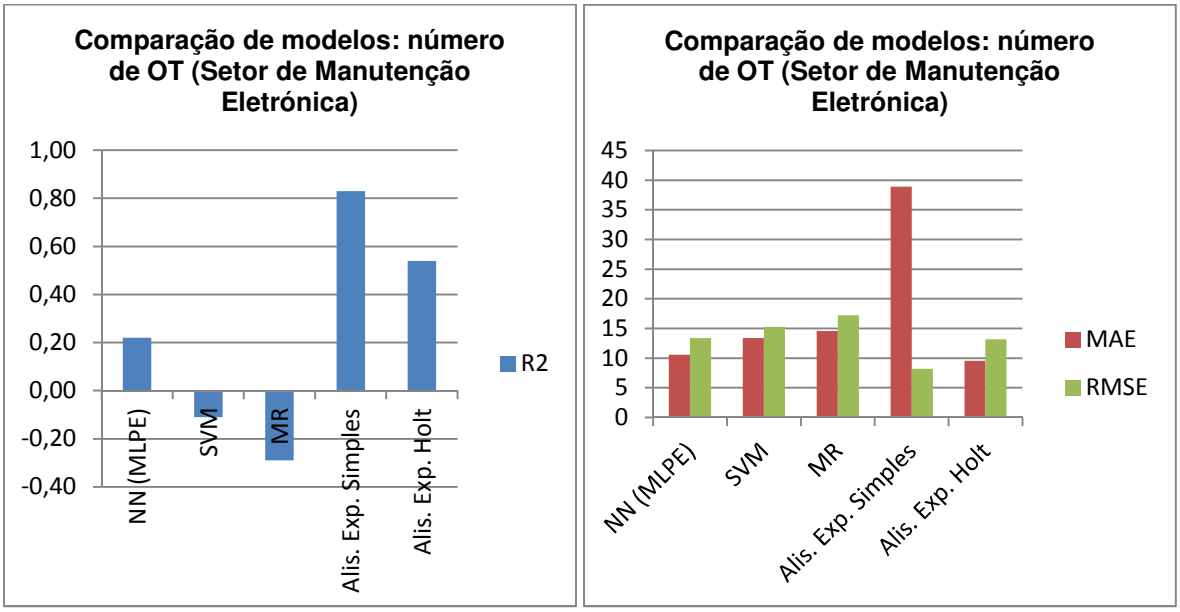


**Figura 14 – Gráficos do valor de $R^2$, MAE e RMSE para os diferentes modelos para o Setor de Manutenção Eletrónica (número de OT)**

Em particular para o caso da previsão do número de ordens de trabalho, o método de Alisamento Exponencial Simples revelou-se como de melhor desempenho, como poderemos verificar através das métricas inscritas nos gráficos da Figura 14.

No caso da previsão do número de incidências, e seguindo o desempenho dos modelos anteriormente avaliados, o método de Alisamento Exponencial Simples revelou-se como o que apresenta os valores menores para as medidas de erro MAE e RMSE e ainda o que apresenta um valor de $R^2$ acima de 0,8. Na figura 15 estão presentes os gráficos que apresentam a comparação das várias métricas de desempenho.

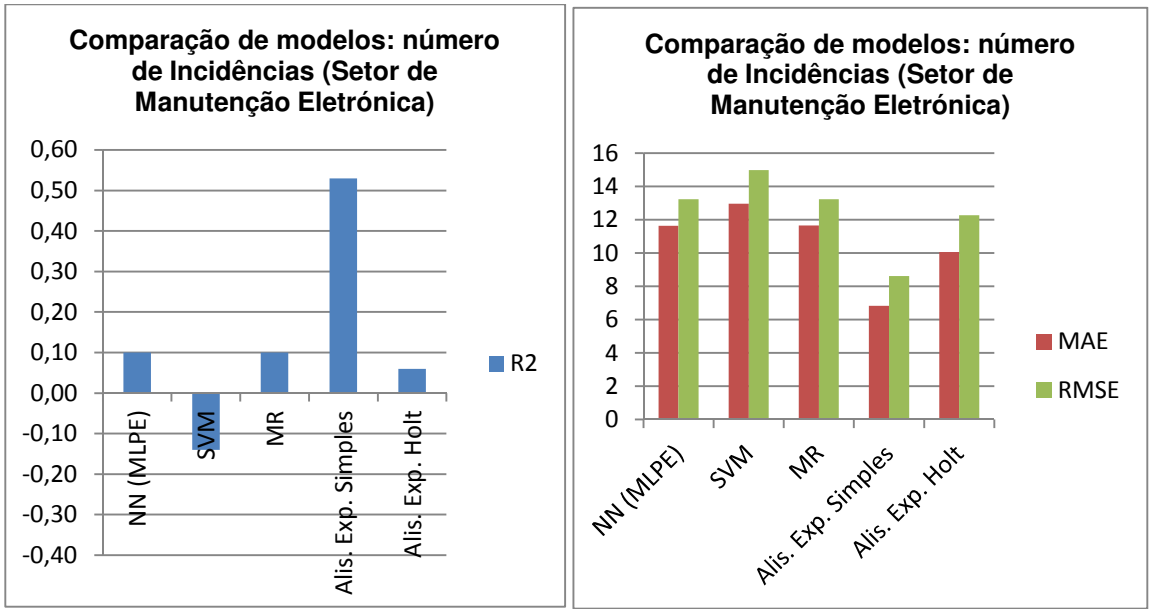


**Figura 15 – Gráficos do valor de $R^2$, MAE e RMSE para os diferentes modelos para o Setor de Manutenção Eletrónica (número de Incidências)**

# 5. Conclusões

No presente capítulo pretende-se efetuar uma síntese do trabalho realizado, bem como avaliar os impactos que poderão advir do desenvolvimento prático deste trabalho. Da mesma forma, pretende-se discutir a possibilidade e a pertinência da aplicação de técnicas de *Data Mining* para previsão de ocorrências de manutenção, que são o indicador da existência de falhas em infraestruturas e equipamentos.

## 5.1 Síntese do trabalho realizado

O momento atual, em que existe uma crise financeira global, e que afeta em particular Portugal, tem levado a que as organizações e os seus gestores procurem minimizar custos, procurando de forma permanente métodos e formas de tornar os processos organizativos mais eficientes.

Este trabalho apresenta uma contribuição nesse sentido, para uma realidade e um setor em particular: manutenção de um Aeroporto. Através da previsão do número de ocorrências de manutenção, os decisores poderão ter na sua mão informação que lhes permita planear melhor a alocação de recursos humanos, a indisponibilidade das infraestruturas e, consequentemente, aumentar a eficiência na gestão da manutenção do Aeroporto.

Este trabalho de dissertação focou-se na aplicação de uma metodologia de *Data Mining*, com o propósito de produzir conhecimento que pudesse responder ao objetivo preconizado. A metodologia adotada, o CRISP-DM revelou-se como muito importante para o desenvolvimento do trabalho prático, auxiliando na abordagem ao problema, e sendo capaz de incorporar qualquer técnica ou modelo de Data Minining mais adequado ao problema em causa.

Através desta metodologia e com recurso a técnicas de modelação e avaliação dos modelos criados, espera o autor ter traduzido em conhecimento útil o trabalho realizado. As técnicas de modelação utilizadas através da ferramenta estatística R e respetivas bibliotecas forecast e rminer, revelaram-se adequadas para o objetivo em questão, embora o problema pudesse ainda ser alvo de previsões com recurso a outras técnicas de *Data Mining*, de modo a poder estabelecer-se uma comparação ainda mais aprofundada.

Foram quatro as situações estudadas: número total de incidências, número total de ordens de trabalho, número de incidências para o setor de Manutenção Eletrónica e número de ordens de trabalho para este mesmo setor.

Através da Figura 16 é possível de uma forma gráfica e visual o tipo de previsão efetuado pelos melhores modelos testados para cada uma das situações alvo de estudo. Aplicando estes modelos na organização em causa, de forma frequente ou automática, poder-se-ia sempre ter previsões do número de ocorrências de manutenção para os meses seguintes.

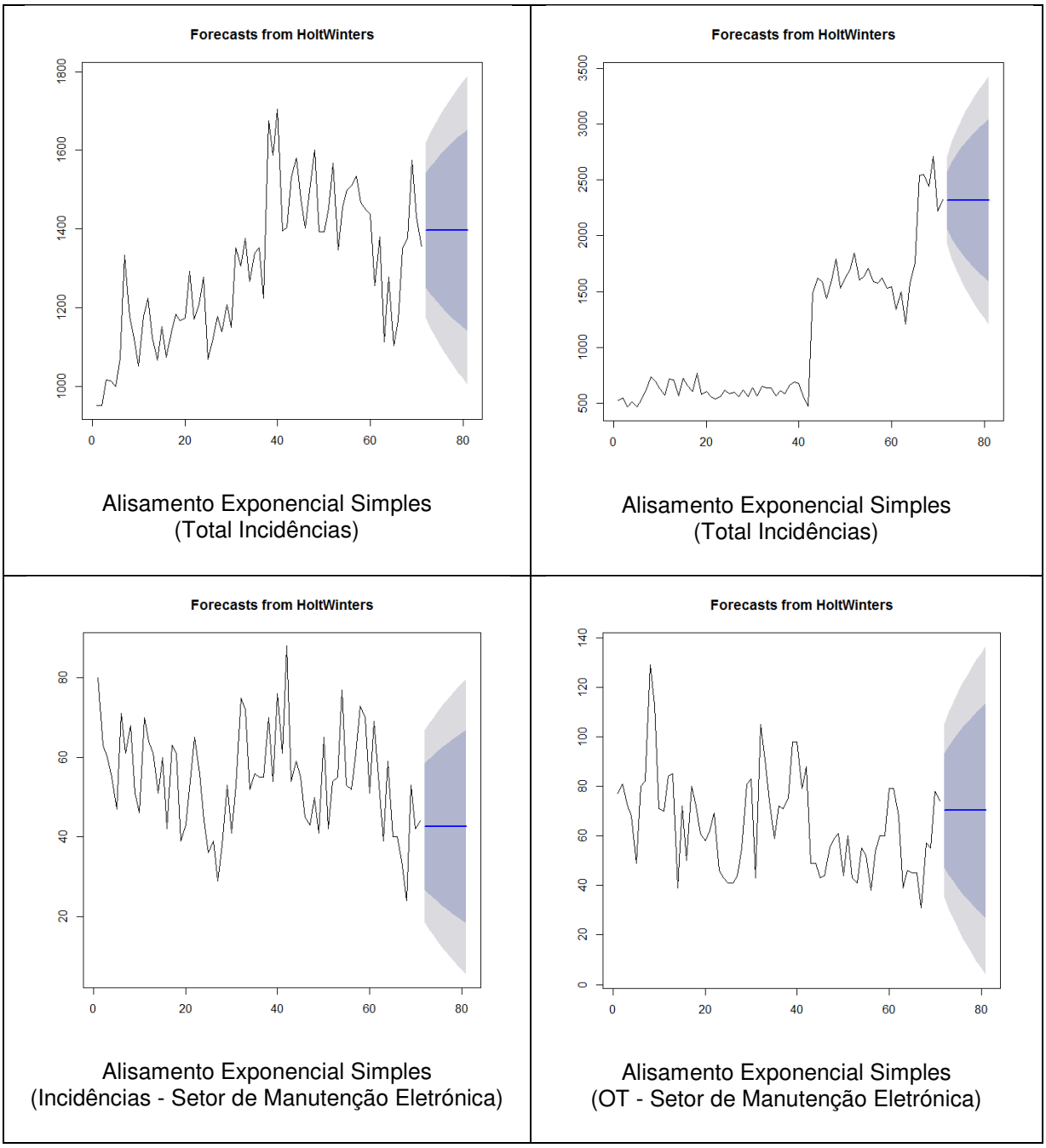


**Figura 16 – Previsão das séries temporais (melhor modelo para cada uma das quatro situações)**

Automatizando estas previsões, ou incorporando essa função no Sistema de Gestão de Manutenção, os gestores de cada setor poderiam antecipar problemas relacionados com picos em termos do número de incidências de manutenção, ou planear trabalhos de conservação ou de manutenção preventiva para momentos de vazio.

## 5.2 Discussão

De forma a dar seguimento ao exposto anteriormente e tendo em conta o momento em que vivemos, a necessidade de melhoria é uma realidade, passando a aposta muito por soluções que permitam uma gestão mais eficiente. Neste caso a gestão é de um processo, o de manutenção de infraestruturas e equipamentos, mas, como comporta custos diretos e indiretos elevados, a otimização destes processos poderá ser de grande valia para a organização.

A aplicação de técnicas de *Data Mining* poderá fazer a diferença como elemento de apoio à decisão, trazendo consequentemente uma melhoria da eficiência na gestão dos processos de manutenção. Com a realização deste trabalho, importa agora confrontar os objetivos propostos com os resultados obtidos.

O objetivo principal desta dissertação passava por escolher a técnica de *Data Mining*/modelo mais adequada para o problema a abordar. Por outro lado, e com base nesta técnicas, pretendia-se criar modelos de previsão para o número de ocorrências de manutenção, expressos em incidências e ordens de trabalho.

A aposta deste trabalho incidiu sobretudo numa abordagem de regressão, sendo que foi despendido um esforço considerável na coleta de dados externos à organização, em particular no que diz respeito aos dados meteorológicos. Dados estes, que se julgavam ter impacto nas variáveis a prever. Contudo, a experimentação, que incluiu modelos sofisticados de regressão, como o caso do SVM, não revelou resultados interessantes nesta abordagem, com valores de erro elevados (especialmente no que diz respeito ao coeficiente de determinação). Tal poderá dever-se a duas razões possíveis e que terão de ser estudadas com detalhe no futuro:

a) A quantidade reduzida de amostras, sendo que se utilizaram somente 72 exemplos, dos quais somente 58 foram utilizados no treino dos modelos de DM;

b) A não inclusão de outros atributos mais relevantes para a determinação de incidências ou ordens de trabalho. Este ponto será discutido na Secção 5.3.

Dada a falta de qualidade dos modelos de regressão, optou-se por uma alternativa, que envolveu a aplicação de uma abordagem pura de séries temporais. Dado os limites temporais de execução desde trabalho, foi possível somente explorar modelos simples de previsão univariada, como o alisamento exponencial simples. Também por limitações temporais, foi realizada a previsão apenas para um setor, embora este trabalho possa ser estendido a outros setores de manutenção.

Os resultados obtidos para os métodos de previsão univariada são interessantes, tendo-se obtido valores de erro que são substancialmente melhores que os obtidos pelos métodos da abordagem de regressão. E julgamos que mesmo este modelo simples pode ser útil parao domínio estudado, sendo fornecida informação do género da apresentada na Figura 16 e que poderá auxiliar a tomada de decisões dos gestores da manutenção aeroportuária.

Quanto à ferramenta *R*, esta apresenta um interface de linha de comando, ou seja, todas as operações realizadas são obtidas através de uma linha de código digitada pelo utilizador. Esta é à

partida uma desvantagem desta ferramenta, pois torna a aprendizagem mais lenta quando comparada com outras ferramentas com interface gráfico mais intuitivo. Por outro lado, estas desvantagens são ultrapassadas por já existirem bibliotecas contendo ambientes gráficos para o *R*, e pelo facto de, após alguma experiência e treino com o R, a linha de comandos ser extremamente flexível, permitindo parametrizar com mais detalhe todas as operações realizadas.

## 5.3 Limitações e trabalho futuro

Conforme já descrito na secção anterior, este trabalho apresentou algumas limitações que aqui se realçam. Descrevem-se ainda sugestões de trabalho futuro com vista à colmatar algumas destas limitações e melhoria do trabalho desenvolvido.

Importa primeiro realçar que este trabalho teve uma elevada natureza exploratória. Foi a primeira vez que foi executada uma abordagem de DM na secção de manutenção da organização aeroportuária estudada, sendo necessário efetuar uma morosa coleta de dados, bem como tomar algumas decisões iniciais e que só mais tarde puderam ser devidamente avaliadas. De realçar ainda que são escassos, a nível internacional, os estudos que utilizam DM para melhoria de atividades de manutenção em aeroportos, o que releva mais ainda a natureza exploratória deste trabalho.

Conforme já descrito, a aposta residiu na abordagem de regressão, com um foco no uso de atributos meteorológicos. Os resultados obtidos mostram que estes (aparentemente) não têm um impacto nas incidências e ocorrências a prever. Daí que há uma necessidade de explorar outros atributos. Como alternativa aos resultados não convincentes dos modelos de regressão, recorreu-se a métodos simples de previsão de séries temporais, sendo que por falta de tempo, não foi possível estudar modelos mais complexos de previsão univariada (ou multivariada), como o ARIMA ou mesmo de DM (e.g. NN, SVM).

Além disso, apostou-se uma análise ao nível mensal, que é útil para planeamento e tende a obter bons resultados em algumas áreas, como as vendas. Para esta periodicidade, o conjunto de dados analisado apresentava um número total de amostras que é bastante reduzido, quando comparado com aplicações normais de DM, onde existem milhares de exemplos de treino. Também por limitações temporais, estudou-se somente a previsão de valores totais e de um setor da manutenção.

Assim, como direções interessantes de trabalho futuro, de modo a melhorar o trabalho apresentado, sugere-se:

1. Coleta e análise de outros atributos que possam ter impacto na previsão de incidências e ocorrências, tais como número de dias desde que foi realizada a última manutenção, cumprimento ou não com os planos de manutenção preventiva previstos, etc.;

2. Coleta de mais amostras e posterior análise se há ou não uma melhoria dos modelos de previsão quando se treinam os modelos com mais amostras;

3. Alargamento deste trabalho a mais setores de manutenção;

4. Experimentação de modelos de previsão univariada e multivariada mais complexos, tais como a metodologia ARIMA ou mesmo modelos de DM como NN e SVM;

5. Análise de outros períodos temporais, como por exemplo uma periodicidade semanal ou diária;

6. Implementação dos modelos de previsão no Sistema de Gestão de Manutenção, de modo a avaliar o real impacto destes modelos (e que também contribuiria para melhorar e aumentar a coleta de dados).

# Bibliografia


ANA – Aeroportos de Portugal, SA. (2011); "Relatório de Gestão e Contas 2010"; Lisbon, Portugal: ANA – Aeroportos de Portugal, SA.

Azevedo A., Santos M. F.; "KDD, SEMMA and CRISP-DM: a parallel overview"; IADIS European conf. *Data Mining* (pp. 182–185), 2008

Cabral, J. P. S.; "Gestão de Manutenção de Equipamentos, Instalações e Edifícios"; LIDEL, 2009

Chapman, P. *et al.*; "CRISP-DM 1.0 Step-by-step *Data Mining* guide"; 2000

Cortez, P., "*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"; Springer, pp. 572-583, 2010

Dayhoff, J. and DeLeo, J.; "Artificial neural networks. CA A Cancer Journal for Clinicians", 91(S8), pp. 1615-1635, 2001

Erskine, J., Peterson, G., Mullins, B., Grimaila, M.; "Developing Cyberspace Data Understanding: Using CRISP-DM for Host-based IDS Feature Mining"; Air Force Institute of Technology, USA, 2010

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.; "From *Data Mining* to Knowledge Discovery in Databases"; AI Magazine Volume 17 Number 3, American Association for Artificial Intelligence, 1996

Hand, D., Mannila, H., Smyth P.; "Principles of *Data Mining*"; The MIT Press, 2001

Hastie, T., Tibshirani, R., and Friedman, J.; "The Elements of Statistical Learning: *Data Mining*, Inference, and Prediction"; Springer-Verlag, 2nd Edition, 2008

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

Hyndman, R. & Athanasopoulos, G., Forecasting: principles and practice. Disponível em: http://otexts.com/fpp/, 2012

Maimon, O., Rokach, L.; "*Data Mining* and Knowledge Discovery Handbook"; 2nd Edition, Springer, 2010

Makridakis, S. G., Wheelwright, S. C. & Hyndman, R. J.; "Forecasting: Methods and Applications"; 3rd Edition, Wiley, 1998

Márquez, Ad. C.; "The Maintenance Management Framework"; Springer, 2007

Moro, S., Cortez, P., Laureano, R.; "Optimização da Gestão de Contactos via Técnicas de *Business Intelligence*: aplicação na banca"; ISCTE-IUL, Lisbon, Portugal, 2011

Moss , L. T., Atre, S.; "Business Intelligence Roadmap**:** The Complete Project Lifecycle for Decision-Support Applications"; Addison Wesley, 2003

Patel, T., Thompson, W., Stephens, C.; "*Data Mining* 101: How to Reveal New Insights in Existing Data to Improve Performance"; SAS Institute Inc., 2011

SS-EN 13306:2010; "Maintenance–Maintenance terminology"; 2nd edition, Swedish Standards Institute, Stockholm, 2010

Turban, E., Sharda, R., Delen, D.; “Decision Support and Business Intelligence Systems”; 9th Edition, Prentice Hall, 2010

Turban, E. *et al.*; “Business Intelligence: A Managerial Approach”; 9th Edition, Prentice Hall, 2nd Edition, Prentice Hall, 2010

Witten, I. Frank, E.; “*Data Mining* – Practical Machine Learning Tools and Techniques”; 2nd Edition, Elsevier, 2005

Michalewicz, Z., Schmidt, M, Chiriac, C.; “Adaptive Business Intelligence”; Springer, 2007.