Universidade do Minho
Escola de Engenharia

Rui Pedro Brás Veloso

# Suporte Inteligente à Decisão sobre Terapias e Procedimentos através de Modelos para a Otimização de Custos e Outcome

outubro de 2014

Suporte Inteligente à Decisão sobre Terapias e Procedimentos através de Modelos para a Otimização de Custos e Outcome

Rui Pedro Brás Veloso

UMinho | 2014

Universidade do Minho
Escola de Engenharia

Rui Pedro Brás Veloso

# Suporte Inteligente à Decisão sobre Terapias e Procedimentos através de Modelos para a Otimização de Custos e Outcome

Dissertação de Mestrado

Mestrado Integrado em Engenharia e

Gestão de Sistemas de Informação

Trabalho efetuado sob a orientação do

**Professor Doutor Carlos Filipe Portela**

e do

**Professor Doutor Manuel Filipe Santos**

Outubro de 2014

DECLARAÇÃO

Nome: Rui Pedro Brás Veloso

Endereço electrónico: a51046@alunos.uminho.pt

Telefone: 969809533

Número do Bilhete de Identidade: 13398951

Título dissertação

Suporte Inteligente à Decisão sobre Terapias e Procedimentos através de Modelos para a Otimização de Custos e Outcome

Orientador(es):

Professor Doutor Carlos Filipe da Silva Portela

Professor Doutor Manuel Filipe Vieira Torres dos Santos

Ano de conclusão: 2014

Designação do Mestrado: Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação



Universidade do Minho, __/__/____

Assinatura: _____

# AGRADECIMENTOS

Aos meus país, Fernando Veloso e Maria Brás por todo o acompanhamento e incentivo nesta última caminhada, assim como ao longo de toda a minha vida e percurso académico.

Aos meus orientadores Filipe Portela e Manuel Filipe Santos por todo o empenho, dedicação e sabedoria que se revelaram fundamentais na realização desta dissertação. Trabalhar com pessoas assim torna qualquer trabalho mais fácil.

Por último um agradecimento à Catarina Fernandes, por toda a presença, paciência e apoio ao longo deste último ano.

# RESUMO

Com o aumento a cada ano das despesas com saúde e a detioração da situação económica mundial, a gestão de custos e recursos nos hospitais assume um papel cada vez mais importante. No caso das unidades de cuidados intensivos este problema atinge ainda uma dimensão maior, pois os custos com cada doente são bastante altos, devido a diversos fatores como o grande número de equipamentos que estas unidades possuem para efetuar uma constante monitorização dos doentes, o maior número de enfermeiros e médicos presentes nestas unidades e a grande diversidade de terapêuticas e fármacos que têm que ser administrados de modo a tentar reverter as situações em que os doentes se encontram. Todos estes fatores levam a um acréscimo do custo diário de um doente de uma unidade de cuidados intensivos.

Com o intuito otimizar custos de internamento e tratamento, assim como otimizar o *outcome* de doentes, pretendeu-se com este trabalho de investigação aferir se existe viabilidade na utilização de modelos em Unidades de Cuidados Intensivos, para suportar a decisão ao nível clínico e administrativo. Foram criados e otimizados modelos de Data Mining que usam técnicas de *clustering* e *classificação* e também desenvolvidos modelos de decisão. O trabalho assentou essencialmente em três grandes estudos que consistiram na utilização de *clustering* na previsão de readmissões em medicina intensiva, previsão do tempo de internamento de doentes em unidades de cuidados intensivos através de técnicas de classificação e modelos de suporte à decisão no combate a infeções bacteriológicas em unidades de cuidados intensivos baseados em heurísticas.

Ao nível dos modelos *clustering*, identificaram-se claramente dois potenciais grupos de doentes readmissíveis em unidades de cuidados intensivos. No caso da previsão de readmissões em medicina intensiva os melhores modelos apresentaram acuidades na ordem dos 80% e sensibilidades em torno dos 95%. Ao nível dos modelos de decisão para infeções bacteriológicas criou-se um algoritmo que utiliza heurísticas de pesquisa e que permite prestar aconselhamento aos intensivistas sobre o rol de soluções ao nível da administração de antibióticos.

Palavras-Chave: Unidades de Cuidados Intensivos, Data Mining, Gestão Hospitalar, Gestão de Custos

# ABSTRACT

With a constant increase in health expenses and an aggravation of the global economic situation, managing costs and resources in healthcare increasingly assumes a more important role. In the case of intensive care units this problem is even greater, because the costs with each patient are quite high due to several factors such as the large number of equipment that these units have to perform constant monitoring to patients. Also are important factors that affect costs in this units the bigger number of nurses and doctors present and the great variety of therapies and drugs that must be administered to try to reverse the situation in which patients are. All these factors lead to an increase in the daily cost of a patient in an intensive care unit.

In order to help optimize spending with inpatients and treatments, it was intended with this research work to assess if there is viability of using models in intensive care units, to support clinical and administrative decision. Were created and optimized data mining models that used techniques like clustering and classification, and developed decision models. The work done consists in three major studies: use of clustering on the readmission prediction in intensive medicine, prediction of inpatients length of stay in intensive care units and decision support models for bacteriological infections in intensive care units.

In the clustering models it's clearly possible to identify two greater groups of inpatients that are more likely to have an readmission on the unit. In the prediction of inpatients length of stay the best models achieved accuracies of about 80% and sensibilities rounding 95%. In the decision models to support bacteriological infections was created an algorithm that uses search heuristics that allowed counseling the clinical staff about the treatment options at the level of antibiotic prescription.

Keywords: Intensive Care Unit, Data Mining, Hospital Management, Resource Management

# ÍNDICE

# LISTA DE FIGURAS

# LISTA DE TABELAS

# LISTA DE ABREVIATURAS, SIGLAS E ACRÓNIMOS

AD – Árvores de Decisão

APACHE - Acute Physiology and Chronic Health Evaluation

AUC – Area Under Curve

CC – Correlation Coefficient

CM – Case Mix

CHP – Centro Hospitalar do Porto

CRISP-DM – Cross Industry Standard Process for Data Mining

DCBD – Descoberta Conhecimento em Bases de Dados

DM – Data Mining

DSR – Design Science Research

HSA – Hospital Santo António

MAE – Mean Absolute Error

MPM – Mortality Probability Model

MSE – Mean Squared Error

PPML –Predictive Model Markup Language

Psaer - Pseudomonas aeruginosa

RAE – Relative Absolute Error

RNA – Redes Neuronais Artificiais

ROC – Receiver operating characteristic

ROCCH – ROC Convex Hull

RSE – Relative Squared Error

SAPS – Simplified Acute Physiology Score

SEMMA – Sample, Explore, Modify, Model and Assess

SOFA - Sequential Organ Failure Assessment

SVM – Support Vector Machines

SWIFT - Stability and Workload Index for Transfer

TI – Tempo de Internamento

TOS –Taxa de Ocupação de Serviço

UCI – Unidades de Cuidados Intensivos

# 1. INTRODUÇÃO

Neste capítulo será introduzida a temática e motivação para a realização deste trabalho assim como apresentados os objetivos e questão de investigação elaborada. Serão também apresentadas a abordagem metodológica seguida durante todo o trabalho assim como a estrutura do trabalho.

## 1.1 Motivação

O objetivo dos sistemas de informação para a saúde centra-se, essencialmente, na garantia de qualidade e eficiência no tratamento dos doentes. Estão também, englobados, os tratamentos médicos e de enfermagem assim como as tarefas administrativas e de gestão associadas a esses mesmos tratamentos (Haux et al., 2004).

As unidades de cuidados intensivos (UCI) fornecem serviços constantes a doentes que se encontram em estado crítico. Estas unidades são locais onde está intensificado o suporte por parte de médicos e enfermeiros através da constante monitorização do estado dos doentes, fornecendo sempre que necessário um suporte vital à sobrevivência do doente, por exemplo, através de ventilação (Adhikari, 2010).

A grande quantidade de dados gerados nas unidades hospitalares e mais concretamente nas de cuidados intensivos, são bastante complexas e volumosas para serem analisadas com recurso aos métodos mais tradicionais. A aplicação de técnicas de Data Mining (DM) nestes tipos de dados fornece os processos e a tecnologia com vista à transformação desses dados em informação útil para os decisores clínicos, traduzindo-se em uma mais-valia no tratamento de doentes (Koh & Tan, 2005).

Os modelos de previsão de risco em medicina intensiva foram desenhados para prever o risco de mortalidade ou complicações para populações de doentes, mas não são precisos o suficiente para serem aplicados a doentes. Até ao surgimento do INTCare (Portela et al., 2014a) não existiam ferramentas para prever com fiabilidade a probabilidade de um doente desenvolver uma complicação como por exemplo a falha de um órgão (Ramon et al., 2007).

No âmbito dos sistemas de informação hospitalares e do apoio à decisão clínica, surgiu o projeto INTCare. Este projeto constituí uma mudança dos paradigmas de registo de monitorização de doentes em unidades de cuidados intensivos hospitalares (UCI), suportando os decisores clínicos através de:

- Prognóstico fornecido por modelos de DM, ao nível da previsão da falência de órgãos e na avaliação da mortalidade (Portela et al., 2013a);
- Desmaterialização de processos (Portela et al., 2013b);
- Automatização de tarefas (Portela et al., 2013b);
- Monitorização e análise de eventos críticos/sinais vitais (Portela et al.,2012a; Portela et al., 2013c);
- Aquisição e cálculo automáticos de scores clínicos em medicina intensiva (Portela, et al., 2012b; Portela et al., 2014).

Este projeto encontra-se em constante desenvolvimento e testes na unidade de cuidados intensivos do Centro Hospitalar do Porto (CHP) – Hospital Santo António (HSA)[1]. Com o intuito de ajudar a evoluir ainda mais o projeto INTCare, surgiu este projeto de dissertação que está enquadrado na fase II do mesmo, sendo que para demonstrar o contributo científico desta dissertação se formulou a seguinte questão de investigação:

**"Qual a viabilidade do desenvolvimento de modelos de otimização de custos e outcome em medicina intensiva?"**

A introdução de novas ferramentas de aquisição de informação na medicina intensiva possibilita o desenvolvimento de modelos DM que permitam suportar o processo de decisão clínico/administrativo. Com este trabalho de investigação pretendeu-se criar e avaliar um conjunto de modelos de previsão e decisão relacionados com o tempo de internamento de doentes em UCI, *outcome* de doentes e gestão de custos (com enfase na gestão de custos derivada da administração de medicamentos). Ao nível das

[1] http://www.chporto.pt/

técnicas de DM foram essencialmente abordados problemas de *clustering* e *classificação*. Estas técnicas foram utilizadas na criação de modelos de DM no âmbito do tempo de internamento e ocupação de camas. Estes modelos foram depois otimizados através da criação de diferentes abordagens. Foi também desenvolvido um algoritmo que utiliza heurísticas de pesquisa no âmbito da gestão de custos e suporte à decisão no combate a infeções bacteriológicas, permitindo assim ter um modelo de decisão associado a esta área.

Como forma de suporte a todo este trabalho de investigação foram utilizadas duas metodologias, uma de investigação e outra técnica, sendo elas respetivamente a *Design Science Research* (DSR) *e Cross Industry Standard Process for Data Mining* (CRISP-DM) que irão acompanhar todo o processo.

## 1.2 Objetivos e Resultados a Atingir

Este trabalho de investigação está centrado num grande objetivo, que visa:

- Obter modelos com boas capacidades de previsão, no apoio à gestão de Unidades de Cuidados Intensivos;

Estas previsões para além de ajudarem na gestão hospitalar vão também permitir uma redução de custos assim como suportar os intensivistas na decisão clínica. Este objetivo será atingido através de dois objetivos específicos obtidos pelo desdobramento do principal:

- Modelos de otimização de custos de internamento/tratamento;

- Modelos de otimização e *outcome* dos doentes do serviço de cuidados intensivos;

Caso exista viabilidade no desenvolvimento de modelos de Data Mining em medicina que permitam otimizar os custos e o *outcome* dos doentes, espera-se que os mesmos possam ser integrados num sistema de apoio à decisão em tempo real que esteja a operacionalizar num hospital, mais concretamente no INTCare.

Estes objetivos serão desenvolvidos e atingidos no serviço de cuidados intensivos do CHP. Irão ser utilizados dados reais, provenientes do CHP, que serão recolhidos de diversas fontes, nomeadamente sinais vitais, terapêuticas, gestão hospitalar, processos clínicos eletrónicos, laboratórios e outros. A utilização destas diferentes variáveis estará sempre dependente da importância que as mesmas demonstrem ter para os modelos de Data Mining que irão ser criados aquando da implementação prática do trabalho.

De referir que todo este trabalho de investigação foi realizado segundo condutas éticas relativas à utilização de informação humana e sigilo profissional associadas à medicina intensiva, pelo não existe qualquer referência que permita identificar um doente, não são divulgados dados na totalidade, e quando parcialmente, foi tido em conta se a conjugação dos mesmos permite de algum modo identificar o doente. Foi também feito um encapsulamento dos dados de modo a uma vez mais garantir todo o sigilo necessário.

## 1.3 Abordagem Metodológica

A elaboração deste trabalho de investigação esteve assente sobre duas metodologias que acompanharam todo o processo, a *Design Science Research* (DSR) e a *Cross Industry Standard Process for Data Mining* (CRISP-DM). A aplicação da metodologia de investigação DSR estará diretamente associada à metodologia CRISP-DM pelo que algumas das fases da DSR são representadas por fases da CRISP-DM.

A metodologia DSR é uma metodologia de investigação que se baseia na construção e avaliação de artefactos inovadores. Esta metodologia assenta no conhecimento que permite compreender e resolver problemas (Peffers et al., 2008). Este conhecimento é gerado através da criação de artefactos. Esta metodologia possuí seis fases sendo elas (figura 1, adaptado de Peffers, 2008):

1. Identificação do problema e motivação;
2. Especificação de objetivos para uma solução;
3. Desenho e desenvolvimento;
4. Demonstração;
5. Avaliação;
6. Comunicação.


**Figura 1 – Ciclo de vida DSR**

Com o crescimento expectável do mercado de *Data Mining* nos anos 90 surgiu a necessidade da existência de um *standard* para estes projetos, sendo esta a génese da metodologia CRISP-DM. A mesma consiste num processo hierárquico com um conjunto de tarefas definidas em quatro níveis de abstração. Estas tarefas assentam em seis fases sendo elas: Compreensão de Negócio, Compreensão dos dados, Preparação dos Dados, Modelação, Avaliação e Instalação (figura 2, adaptado de Chapman, 2000). Na fase de compreensão de negócio pretende-se entender os objetivos e os requisitos de negócio. A segunda fase consiste na compreensão dos dados onde se realizam um conjunto de passos para ganhar conhecimento dos dados e identificar problemas de qualidade nos mesmos, sendo que após esta fase vem a fase de preparação dos dados onde se vão executar um conjunto de tarefas para tratar os dados. Estas tarefas incluem seleção, transformação, limpeza, etc. Seguidamente vem as duas fases mais importantes desta metodologia, a modelação e a avaliação. Na modelação serão criados e calibrados os modelos e na avaliação, verifica-se se foram atingidos os resultados e objetivos especificados. Por último existe a fase de instalação que consiste na organização e apresentação de conhecimento. Esta última fase não será executada neste projeto de investigação (Chapman et al.,2000).

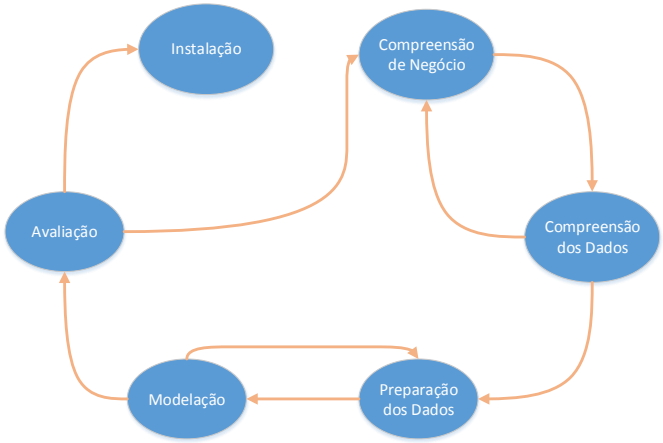


**Figura 2 - Ciclo de vida CRISP-DM**

Ao longo deste trabalho de investigação ambas as metodologias estiveram conjugadas de modo a garantir a qualidade do trabalho investigativo assim como de todo o trabalho realizado na área de DM. Sendo assim criou-se o seguinte mapeamento de modo a elucidar as sinergias entre as duas metodologias, apresentando também os resultados dessas mesmas fases. (Tabela 1).

**Tabela 1 – Mapeamento entre metodologia DSR e CRISP-DM**

| Metodologia | DSR | CRISP-DM | Resultados/Artefactos |
|---|---|---|---|
| **Fases** | -Identificação do Problema e Motivação | -Compreensão do Negócio | -Pré-Dissertação |
| | -Especificar Objetivos de uma Solução | -Compreensão do Negócio | -Pré-Dissertação;<br>-Relatório de Dissertação |
| | -Desenho e Desenvolvimento | -Compreensão dos Dados<br>-Preparação dos Dados<br>-Modelação | -Modelos de previsão do tempo de internamento;<br>-Modelos de *clustering* para caracterização de doentes readmissíveis;<br>-Algoritmo heurístico para suporte à decisão no combate a infeções |
| | -Demonstração | -Modelação | -Indução dos modelos usando dados em tempo real para prever hora em que o doente tem alta; |

| | | | |
|---|---|---|---|
| | | | -Indução dos modelos de *clustering* para a descoberta de dois grupos de doentes readmissíveis;<br><br>-Aplicação do algoritmo a um conjunto de tratamentos a infeções para aconselhamento de antibióticos |
| | -Avaliação | -Avaliação | - Avaliação dos modelos induzidos;<br><br>-Avaliação do algoritmo heurístico; |
| | -Comunicação | | - 4 Artigos Científicos (1 publicado, 2 a aguardar publicação e 1 submetido para revisão por pares);<br>-Relatório de Dissertação |

Nas fases de identificação do problema e motivação, e especificação dos objetivos de uma solução, que englobam a tarefa de compreensão do negócio da metodologia CRISP-DM, definiram-se os objetivos do trabalho assim como as principais linhas orientadoras, tendo sido produzido como *output* o relatório de pré-dissertação. Na fase de desenho e desenvolvimento, que coincidiu com as fases de compreensão dos dados, preparação dos dados e modelação da metodologia CRISP-DM, foram analisados os dados disponíveis, transformados e induzidos os modelos e criado o algoritmo heurístico. Os modelos induzidos e o algoritmo criado correspondem aos artefactos da metodologia DSR. Na fase de demonstração, que está alinhada com a fase de modelação do CRISP-DM foram aplicados os modelos e algoritmos a um problema de contexto real, utilizando dados do CHP e posteriormente foram analisados os resultados. Ao nível da fase de avaliação, que corresponde também à fase de avaliação do CRISP-DM foram avaliados os modelos induzidos e os seus resultados assim como os resultados da aplicação do algoritmo de pesquisa no combate às infeções bacteriológicas. Por último ao nível da fase de comunicação foi elaborado o relatório de dissertação que descreve todo o trabalho de investigação executado, assim como elaborados quatro artigos científicos, estando um deles já publicado, e três a aguardar publicação. O Anexo I apresenta uma visão geral sobre estes artigos.

## 1.4 Estrutura do Documento

Este relatório de pré-dissertação está organizado com a seguinte estrutura de capítulos:

- **Introdução** – este capítulo apresenta uma breve introdução sobre a temática assim como a motivação, apresentação da questão de investigação e objetivos do trabalho, a abordagem metodológica seguida e a estrutura do documento de dissertação.
- **Enquadramento Conceptual** – nesta secção é apresentada a revisão de literatura feita onde foram abordados temas como: hospitais, UCI, sistemas de informação hospitalares, registos eletrónicos de saúde e dados hospitalares, gestão de custos e hospital em UCI assim como gestão de medicamentos, tempos de internamento (TI) e taxas de ocupação de serviços. Seguidamente será apresentado o processo de Descoberta de Conhecimento em Bases de Dados, Data Mining (onde serão abordados as técnicas de classificação, regressão e *clustering*, as metodologias *standard* para este tipo de projetos, DM na área da medicina intensiva, o projeto INTCare e o DM na gestão de fármacos, tempo de internamento e taxa de ocupação de serviços hospitalares).
- **Trabalho Realizado** – Neste secção será apresentado todo o trabalho realizado. Este capítulo subdivide-se em três sendo eles: *clustering* na previsão de readmissões em medicina intensiva, previsão do tempo de internamento de doentes em UCI e suporte a decisão de infeções bacteriológicas em medicina intensiva. Cada um destes subcapítulos terá os seus subcapítulos de modo a ser mais fácil apresentar o trabalho efetuado.
- **Conclusões** – Será apresentada uma síntese do trabalho efetuado assim como as contribuições científicas do mesmo. Serão também lançadas as principais linhas orientadoras para o trabalho futuro.

# 2. ENQUADRAMENTO CONCEPTUAL

Para a elaboração do enquadramento conceptual foram consultados e utilizados diversos motores de pesquisa de publicações científicas. Entre os utilizados destacam-se:

- Web of Knowledge
- ScienceDirect
- Springer
- IEEE Xplore
- B-on
- RCAAP
- Scopus
- Google Scholar

A pesquisa efetuada baseou-se nas seguintes palavras-chaves/expressões:

- Data Mining
- Data Mining Methodologies
- Data Mining Healthcare
- Data Mining ICU
- Data Mining Drugs
- Data Mining in Quality of Hospitals
- Hospital
- Hospital Management
- Hospital Resources
- Quality in Hospitals
- Performance Indicators in Hospitals
- Bed Occupancy Rate
- Length of Stay
- Drugs
- INTCare

Para a revisão de literatura e artigos a utilizar foram considerados fatores como a relevância do autor na área, a reputação dos artigos publicados assim como o ano em que o artigo foi publicado.

## 2.1 Hospitais, Unidades de Cuidados Intensivos, Medicina Intensiva e Infeções Bacteriológicas

Os centros médicos e os hospitais são ambientes complexos que oferecem tecnologias e cuidados a pessoas que em alguns casos estão em pontos vulneráveis das suas vidas. Os primeiros hospitais tinham como objetivo alojar peregrinos, indigentes e vítimas de pragas. Mais tarde estas instituições começaram a ser responsáveis pelo diagnostico e recuperação de problemas nos doentes (Griffin, 2006).

Um hospital é constituído por diversos departamentos/serviços sendo que o número de departamentos varia de hospital para hospital e está dependente de fatores como economia, local, número de doentes e outros. Entre os mais comuns destacam-se os departamentos de emergência médica, unidades de cuidados intensivos (que podem incluir unidades pediátricas e neonatais e também unidades para adultos), neurologia, cardiologia, oncologia, obstetrícia e ginecologia. Estes departamentos são suportados por equipas clínicas que englobam médicos, enfermeiros e auxiliares e possuem serviços auxiliares como os serviços de análises laboratoriais, imagiologia, anestesia e farmácia. Existem depois departamentos de suporte à normal funcionalidade de um hospital e que incluem por exemplo a gestão de materiais, cantinas e toda a alimentação do hospital, finanças e contabilidade, entre outros (Griffin, 2006).

As pessoas que estão gravemente doentes são normalmente admitidos nas unidades de cuidados intensivos (UCI) para que consigam manter as suas funções fisiológicas através dos diversos dispositivos de suporte à vida. Nestas unidades as funções vitais dos doentes são monitorizadas continuamente, bem como o estado de cada um dos sistemas orgânicos: neurológico, respiratório, hepático, hematológico, cardiovascular e renal. De modo a precaver a vida e estado do doente estas funções podem ser suportadas através de medicamentos ou através de dispositivos mecânicos, até que as funções de um doente voltem a ser autónomas (Ramon et al., 2007).

Por si só a quantidade de dispositivos de monitorização e dados provenientes do doente em estado crítico não é suficiente para efetuar toda a avaliação do estado de um doente e posteriormente tomar uma decisão efetiva relativamente aos tratamentos. Há diversos anos foram desenvolvidos diferentes sistemas de classificação da severidade das doenças dos doentes assim como da mortalidade dos mesmos em

UCI. Entre esses sistemas destacam-se o Simplified Acute Physiology Score (SAPS), o Acute Physiology and Chronic Health Evaluation II (APACHE II), o Mortality Probability Model (MPM) e o Sequential Organ Failure Assessment Score (SOFA). Estes sistemas são tipicamente baseados em scores que permitam avaliar o estado do doente assim como o caminho a seguir no tratamento/recuperação do mesmo.

O SAPS é um sistema baseado em scores para calcular a probabilidade de mortalidade hospitalar. O SAPS II, a segunda versão deste modelo foi desenvolvido através de um estudo entre europeus e norte americanos envolvendo doentes de UCI médicas e cirúrgicas de dez hospitais europeus e dois norte-americanos. Este sistema considera como parâmetros de avaliação da probabilidade de mortalidade a idade, frequência cardíaca, pressão arterial sistólica, temperatura corporal, taxas de oxigenação e respiração, *output* urinário, quantidade de ureia, potássio e sódio, contagem de glóbulos brancos, nível de bilirrubina, escala de coma de *Glasgow*, tipo de admissão, portador de HIV, portador de patologia hematológica e se possuí algum cancro metastizado (Le Gall et al., 1993). Em 2005 surgiu uma nova versão deste modelo, o SAPS 3. Este novo sistema utiliza técnicas estatísticas para selecionar e pesar diferentes variáveis. Este modelo assume vinte variáveis divididas em três *subscores* que tem a ver com as características na admissão, após a admissão e o grau de perturbação fisiológica após a primeira hora na UCI (Moreno et. al, 2005).

O APACHE II foi desenvolvido para permitir medir a severidade de uma patologia em doentes adultos que estão em UCI. É um sistema baseado em *scores*, que são expressos em valores inteiros de 0 a 71 e avalia as medições de treze parâmetros clínicos, a saber, idade, temperatura corporal, pressão arterial, frequência cardíaca, taxa de respiração, oxigenação, pH arterial, quantidades de sódio, potássio e creatinina, hematócrito, contagem de glóbulos brancos e escala de coma de *Glasgow* (Knaus et al., 1985).

O MPM é um modelo de *scores* para aferir a probabilidade de morte de um doente na UCI. O primeiro modelo MPM foi publicado em 1985 e consistia num modelo de admissão com sete variáveis e por um modelo de vinte e quatro horas usando sete variáveis recolhidas durante as primeiras vinte e quatro horas do doente numa UCI. Em 1993 saiu a versão MPM II que consiste também em dois modelos de scores, um na admissão que considera quinze variáveis e outro após 24 horas de internamento que

considera cinco das variáveis utilizadas no modelo de admissão e oito variáveis adicionais recolhidas durante as primeiras vinte e quatro horas (Ronco et al., 2009).

O SOFA é um modelo de *scores* para avaliar o estado das funções dos órgãos dos doentes ou para prever a taxa de um desses sistemas falhar. É composto por seis scores para sistemas de órgãos, o respiratório, cardiovascular, hepático, hematológico, renal e neurológico. Os scores são atribuídos de 0 a 4 de acordo com o grau de disfunção ou falha (Janssens et al., 2001).

Embora o desenvolvimento dos sistemas de monitorização e das modalidades terapêuticas tenha levado a um aumento da probabilidade de sobrevivência de doentes nas UCI ainda restam imensos desafios nesta área. Estima-se que cerca de 70% dos doentes das UCI que necessitem de suporte vital e de monitorização durante alguns dias tenham uma taxa de sobrevivência bastante alta, no entanto quando estes suportes aumentam estendendo-se por semanas ou até meses a taxa de sobrevivência dos doentes é bastante baixa (cerca de 30% após três semanas na UCI). Quando o internamento dos doentes perdura por mais de três semanas começa a existir uma deterioração do sistema imunitário, o que deixa os doentes mais sujeitos a infeções severas ou a estados híper-inflamatórios, estados que ameaçam os sistemas dos órgãos vitais do doente como pulmões, coração, rins, fígado e cérebro (Ramon et al., 2007).

As infeções são a maior causa de mortalidade em UCI tanto na Europa (Angus et al., 2001) como nos Estados Unidos da América (Vicent, et al., 2006). De acordo com a comunidade médica este tipo de problema é mais comum em doentes que estão internados há mais de cinco dias. Vicente et al. (2009) conduziram um estudo sobre infeções em UCI e do conjunto de doentes avaliados 50% tinha uma infeção. Dos doentes infetados em cerca de 70% dos casos foram administradas terapêuticas associadas a antibióticos e culturas microbiológicas sendo que na generalidade dos casos se obteve resultados bastante positivos. A prescrição de antibióticos varia bastante de país para país sendo que não se sabe as razões para tal acontecer (Lindbaek, 2006). Em muitos dos casos o corpo do doente resiste a um antibiótico administrado sendo que é necessário testar e administrar outro antibiótico ao doente.

## 2.2 Sistemas de Informação Hospitalares e Registo de Dados Hospitalares

Em meados dos anos 50 assistiu-se ao começo do uso de computadores nos hospitais, sendo que atualmente este tipo de tecnologia está presente na grande maioria dos mesmos. Embora os sistemas de informação hospitalares cubram quase todas as áreas hospitalares e de suporte, o maior grau de sofisticação está presente na área financeira. A tendência atual é para que os sistemas interliguem os departamentos, verificando-se assim um aumento do número de hospitais que usam os sistemas de informação para ajudar a tomada de decisão clínica, investigação na área clínica, análises laboratoriais, atividades dos médicos e enfermeiros assim como as atividades de tratamento aos doentes (Griffin, 2006).

Atualmente nos grandes hospitais existem pelo menos vinte sistemas de informação hospitalares, de modo a facilitar toda a gestão de processos e informação que são necessárias pelos diferentes departamentos de uma unidade hospitalar (Chenhui & Xudong, 2008). Os utilizadores deste tipo de sistema necessitam de aceder a informação médica de modo a terem as melhores decisões clínicas e a fornecer um tratamento custo-eficiente aos diferentes doentes das unidades hospitalares. No entanto ainda existem alguns problemas a serem resolvidos neste tipo de sistemas, entre os quais, falta de consistência dos dados clínicos inter-departamento, o acesso a informação médica proprietária que está muitas vezes descentralizada em diversos sistemas de informação, e as grandes diferenças entre interfaces dos diferentes sistemas que torna muitas vezes a consulta ou inserção de dados um processo difícil (Chenhui & Xudong, 2008).

Um processo clínico eletrónico é segundo Iakovidis (1998) "informação sobre cuidados saúde guardada digitalmente, da vida de um individuo, com o propósito de suportar a continuação dos cuidados, educação ou pesquisa e assegurando sempre a confidencialidade".

O processo clínico eletrónico possuí alguns atributos importantes, entre os quais se destaca (Iakovidis,1998):

- **Acessibilidade e disponibilidade** – é fulcral poder-se aceder sempre que necessário aos registos;

- **Fiabilidade** – este tipo de registos tem que ser fiáveis pois poder-se-ão tomar decisões clinicas com base nos mesmos;
- **Confidencialidade** – estão-se a tratar dados que contém informações delicadas dos doentes pelo que a confidencialidade dos mesmos e das suas patologias e estado de saúde tem que que estar em total confidencialidade para que não sejam observadas por indivíduos que nada tenham a ver com o tratamento dos doentes.

Nas UCI devido à constante monitorização de doentes, o espetro dos dados recolhidos é bastante grande, sendo que Ramon et al. (2007) destacam um conjunto de dados chave do doente nas UCI:

- **Dados Demográficos** – este conjunto de dados considera todos os dados demográficos do doente, dados como a data de nascimento, sexo, morada, telefone, etc.;
- **Historial Clínico** – todo o historial de patologias e tratamentos do doente poderá ser relevante para o seu estado atual e para as decisões a tomar relativamente ao tratamento a executar;
- **Dados do internamento na UCI** – aqui estão englobados todos os dados registados do doente durante o sua internamento na UCI. Engloba diferentes tipos de dados como medições dos parâmetros clínicos (temperatura, frequência cardíaca), resultados das análises laboratoriais, dados sobre as infeções do doentes, registos sobre as observações subjetivas efetuados por médicos ou enfermeiras, dados sobre os tratamentos efetuados aos doentes e fármacos administrados, sobre alimentação dada, etc.;
- **Informação sobre medicamentos** – tipicamente existe informação sobre os efeitos secundários provocados pelos medicamentos e sobre as situações e as razões pelos quais são administrados;
- **Conhecimento especialista** – este tipo de dados considerada o conhecimento específico como interações conhecidas entre medicamentos ou tratamentos, modelos de funcionamento do corpo humano e descrição da interpretação dos sintomas dos doentes.

## 2.3 Gestão de Custos e Gestão Hospitalar em UCI

A avaliação da qualidade nas unidades de cuidados intensivos deve assumir um papel bastante importante tanto pelos tipos de patologias e sintomas dos doentes que são tratados, que muitas vezes se encontram em risco de vida e com sintomas bastante agravados, assim como pelos diversos serviços

prestados. É também necessário avaliar todos os custos associados a este tipo de unidades pois os mesmos são bastante altos e com o aumento dos custos em saúde toda a otimização e diminuição de custos é essencial.

Na figura 3 (adaptado de Portada,2013) podemos ver a despesa corrente em cuidados de saúde em % do PIB Português. Verifica-se que até aos anos de 2011, 2012, devido aos cortes orçamentais, na maioria dos anos a fatia da saúde teve tendência a aumentar.

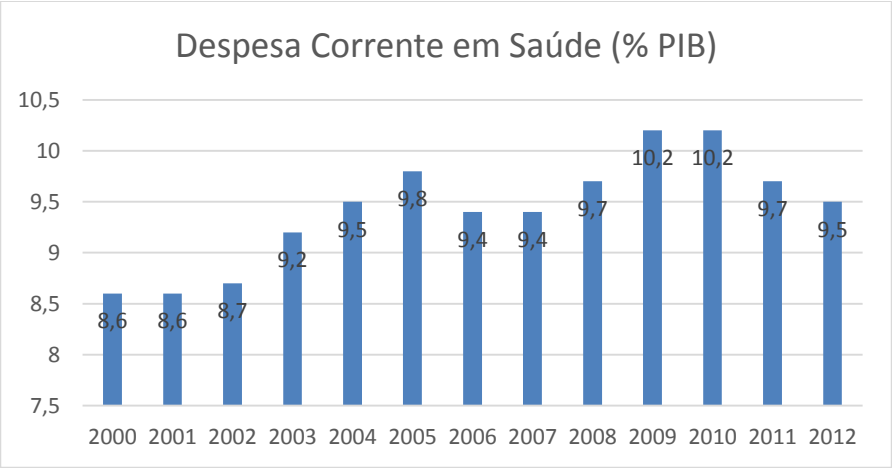


**Figura 3 - Despesa corrente em cuidados de sáude em Portugal em % do PIB**

Como se pode verificar pela figura 4 (adaptado de Portada, 2014a), nas duas últimas décadas têm-se assistido a um aumento bastante significativo nas despesas do estado em saúde *per capita* em Portugal, pelo que, como já referido, surge aqui um dos principais motivadores da tentativa de otimização de custos hospitalares e custos com a saúde em geral.

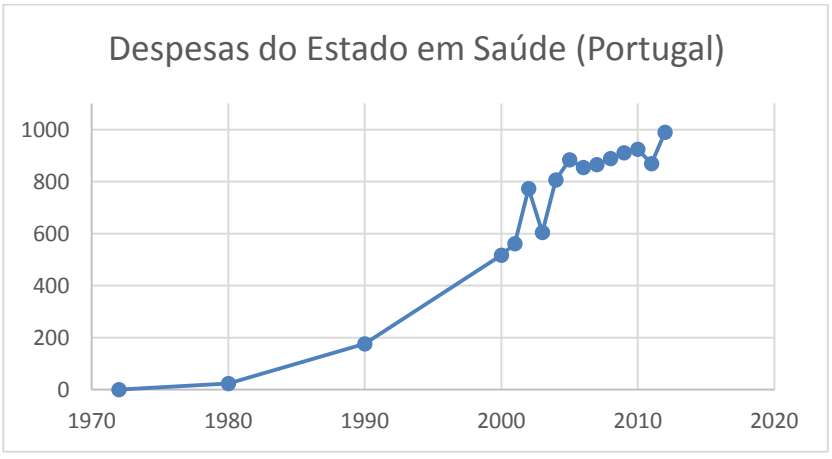


**Figura 4 - Despesas do estado em saúde - Execução orçamental per capita em Portugal**

De referir que a realidade europeia e mundial encontra-se com as mesmas alterações, pelo que a gestão de custos representa um desafio para a saúde a um nível global. A figura 5 (adaptado de Pordata, 2014b) permite observar a despesa com saúde em % do PIB de diversos países europeus assim como do Japão e Estados Unidos da América.

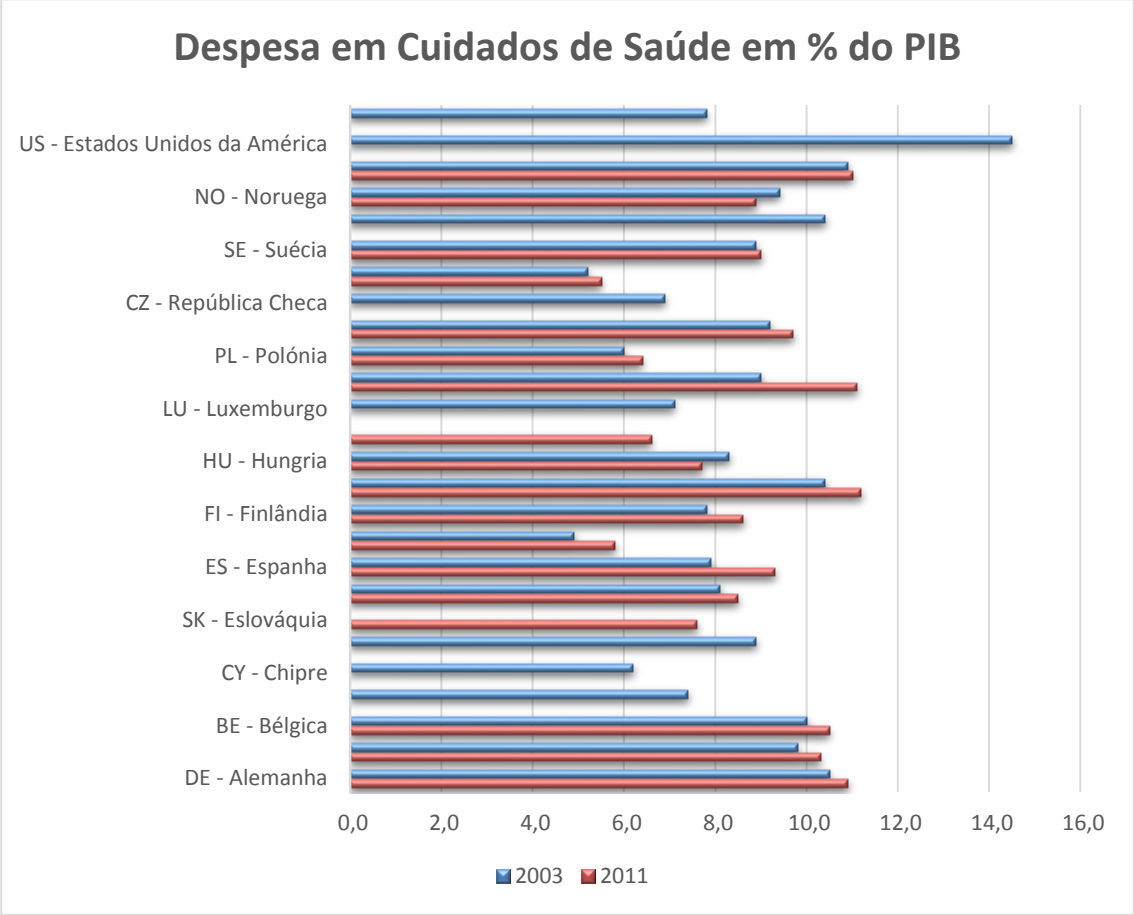


**Figura 5 - Despesa em saúde em % do PIB, em 2003 e 2011**

Existem um elevado número de indicadores de gestão de custos e hospitalar, no entanto, neste documento apenas serão abordados os que apresentem uma maior relevância na gestão de custos e de recursos humanos das UCI assim como os mais críticos no âmbito dos doentes.

A Direção Geral de Saúde (DGS) Portuguesa considera um conjunto de indicadores como sendo fulcrais para a avaliação contínua da qualidade nas UCI portuguesas sendo eles (Direção Geral de Saúde, 2003):

- Registo das escalas de gravidade dos doentes internados;
- Registo da carga de trabalho;
- Tempo médio de internamento;
- Mortalidade nas UCI;
- Taxa de reinternamento até às 48 horas;
- Tempo médio de ventilação;
- Taxa de reintubações ate às 48 horas;
- Incidência de complicações iatrogénicas;
- Avaliação dos custos.

Apesar de apenas serem destacados estes, existem outros indicadores hospitalares que permitem aferir a qualidade das UCI assim como o seu desempenho que também devem ser considerados. Existem diversas métricas interessantes para análise da gestão de custos e hospitalar, sendo que Nerenz e Neill (2001) identificaram algumas que agruparam por diferentes domínios, entre as quais se destacam:

- **Qualidade dos Cuidados** – neste domínio incluem-se indicadores como a mortalidade hospitalar, taxa de complicações e de infeções; taxas de retorno à cirurgia e de erros médicos específicos;

- **Utilização e Eficiência** – Taxa de ocupação da unidade (que consiste na análise do número de camas ocupadas e livres), tempo de internamento nas unidades hospitalares, admissões e dias de internamento por cada mil doentes, custos por alta, entre outros;

- **Satisfação** – *Reports* de satisfação dos doentes e dos cuidados prestados;

- **Financeiros** – Englobam despesas e margens operacionais, custos com fármacos.

Poderão também ser considerados outros indicadores relacionados por exemplo com questões culturais (capacidade para atender cidadãos que se expressam e possuem documentos clínicos em outras línguas) e benefícios para a comunidade, como cuidados prestados em programas públicos de vacinação ou rastreios.

Berenholtz et al. (2002) propôs uma taxionomia diferente para as métricas hospitalares subdividindo-as em três categorias, **estrutura**, **processo** e ***outcome*** dos doentes. A tabela 2 (adaptado de de Vos et. Al, 2007) propõe um conjunto de indicadores de avaliação e gestão em unidades de cuidados intensivos e utiliza a taxonomia de Berenholtz et al.

**Tabela 2 - Indicadores em UCI**

| Domínio | Indicador |
|---|---|
| **Estrutural** | Disponibilidade dos intensivistas |
| | Rácio de enfermeiros por doente |
| | Estratégia para prevenir erros de medicação |
| | Medição da satisfação doente/familiares |
| **Processual** | Tempo de Internamento na UCI |
| | Duração da ventilação mecânica |
| | Taxa de ocupação das unidades |
| **Outcome dos Doentes** | Avaliação da Mortalidade |
| | Número de extubações não planeadas |
| | Incidência de decúbito |

Um dos pontos bastante importantes na gestão de recursos no hospital é o seu planeamento e abastecimento ao nível de fármacos, aparelhos, alimentação e outros.

Uma das maneiras de efetuar esta gestão é recorrendo à análise de métricas e indicadores que permitam avaliar e otimizar as mesmas.

A gestão de medicamentos assume um papel bastante importante na medicina, tendo em conta duas grandes vertentes. Uma delas são os custos associadas à administração de fármacos e das próprias farmácias hospitalares. No caso das UCI estes custos assumem valores ainda mais elevados pelo que a otimização de custos ao nível dos fármacos desempenha um papel bastante importante ao nível de gestão hospitalar.

Ainda relativamente aos fármacos, outra das vertentes que é bastante interesse de se analisar são os efeitos adversos provocados pelos medicamentos. Estes efeitos são considerados atualmente um dos maiores problemas ao nível de saúde pública (Koutkias et al., 2009), pois acarretam um aumento de

custos com saúde assim como colocam em risco a vida dos doentes. Interessa portanto efetuar estudos na gestão destes efeitos assim como estudos nos efeitos adversos provocados pela combinação e administração de diferentes fármacos, levando assim a uma segurança acrescida no que toca aos doentes e a uma redução de custos com a saúde derivado a não serem necessários tratamentos ou medicamentos adicionais para contrariar estes efeitos adversos.

Outro dos indicadores é o tempo de internamento (TI) que consiste no número de dias de tratamento a que um doente é sujeito desde a sua data de admissão até ao momento em que tem alta. Este valor pode ser contabilizado num hospital ou em casa caso o doente estiver a receber cuidados médicos em casa. A utilização deste indicador permite além de saber quais os gastos que o doente ou grupo representam para o hospital, avaliar a qualidade dos serviços prestados ao doente, quando os tratamentos assim o permitam. O TI nos hospitais tem-se revelado uma boa métrica uma vez que permite medir o consumo de recursos hospitalares com um determinado doente, ou de uma determinada unidade hospitalar ou grupo de doentes (Marshall et al., 2005).

Uma vez que estas medições tem que ser corretas, de modo a maximizar a otimização de recursos, torna-se fulcral desenvolver modelos que permitam prever o TI dos doentes internados, essencialmente em unidades críticas como é o caso das Unidades de Cuidados Intensivos (UCI). Obtendo o TI passa a ser possível determinar a taxa de ocupação de um determinado serviço.

A taxa de ocupação de um serviço hospitalar é o rácio entre o número de doentes por dia e o número de leitos por dia num determinado período.

$$\text{Taxa de Ocupação do Serviço} = \frac{\text{Doentes/dia}}{\text{Leitos Operacionais/dia}} \times 100\%$$

A análise deste indicador é bastante importante pois permite contornar um problema que por vezes surge, que é o de a taxa de ocupação de camas ser superior a 100%, sendo que no caso das UCI, devido ao estado dos doentes, este problema apresenta um risco acrescido para os mesmos.

## 2.4 Descoberta de Conhecimento em Bases de Dados

O ritmo elevado e de constantes mudanças às quais as organizações estão sujeitas levam a uma cada vez maior pressão no modo como operam. É necessário ter um processo de decisão seguro, rápido e estratégico por parte dos gestores de modo a se obter uma vantagem competitiva no tipo de mercado onde a organização atua. Nos últimos anos o custo de armazenamento tem diminuído consideravelmente, levando as organizações a investir fortemente em tecnologias de base de dados. Hoje em dia a produção de informação já ultrapassou a capacidade de análise humana pelo que é necessário recorrer a outras formas de a analisar. Sendo assim a Descoberta de Conhecimento em Bases de Dados e o Data Mining começam a assumir um papel bastante importante para as organizações no apoio ao processo de tomada de decisão e ao ganho de vantagem competitiva junto dos seus pares (Maimon e Rokach, 2010).

A Descoberta de Conhecimento em Bases de Dados (DCBD), figura 6 (adaptado de Maimon & Rokac, 2010) é uma análise e modelação automática e exploratória de grandes repositórios de dados. É o processo organizado de identificar padrões válidos, inovadores, úteis e que se consigam entender em *datasets* grandes e complexos (Maimon e Rokach, 2010).

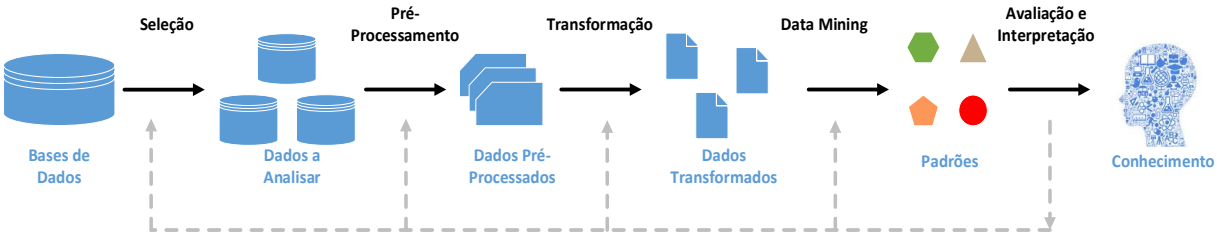


**Figura 6 - Processo de DCBD**

Este processo começa com a aquisição de conhecimento do domínio assim como a definição dos objetivos de DCBD de modo a iniciar o processo sendo que depois se vai prosseguir pelas seguintes fases:

- **Seleção** – Neste passo pretende-se selecionar ou criar um *data set* sobre o qual a descoberta de dados vai ser feita. Tendo em conta os objetivos definidos no início do processo deve-se então

determinar quais os dados que irão ser utilizados. Esta determinação incluí descobrir quais os dados que estão disponíveis, obtendo caso seja necessário, dados adicionais, e integrar todos os dados necessários ao processo de descoberta em um único *data set*, incluindo os atributos que serão considerados para o processo. Este processo assume um papel bastante importante pois o DM aprende e descobre através dos dados fornecidos. Esta é a evidência base para a construção dos modelos. Caso se falhe a identificação de alguns atributos importantes, todo o projeto poderá estar em risco. Para garantir o sucesso deve-se considerar o maior número de atributos possíveis nesta etapa. Mas por outro lado, colecionar, organizar e operacionalizar repositórios de dados complexos é bastante dispendioso mas ganha-se a oportunidade de melhor entender os fenómenos (Maimon e Rokach, 2010).

- **Pré-Processamento** – Na fase de pré-processamento pretende-se melhorar a fiabilidade dos dados. Estão englobadas tarefas como a limpeza dos dados, gestão de valores em falta, remoção de ruído e valores fora dos intervalos de normalidade nos dados (Maimon e Rokach, 2010).

- **Transformação** – Nesta etapa pretende-se fazer uma redução dos dados e uma projeção dos mesmos. Deve-se encontrar caraterísticas úteis que permitam representar os dados, sendo que através dos métodos de redução de dimensão ou transformação, o número efetivo de variáveis a considerar pode ser reduzido assim como também se podem encontrar representações dos dados invariáveis (Fayyad et al.,1996).

- **Data Mining** – O DM consiste na aplicação da análise de dados e de algoritmos de descoberta que produzem um conjunto de padrões ou modelos sobre os dados. Esta fase engloba a escolha da tarefa de DM a utilizar assim como o(s) algoritmo(s) e a implementação dos mesmos (Fayyad et al.,1996).

- **Avaliação e Interpretação** – Nesta fase efetua-se a avaliação e interpretação dos padrões resultantes do processo de DM, respeitando sempre os objetivos definidos no início do processo. Deve-se avaliar a compreensibilidade e utilidade dos modelos induzidos, assim como documentar a descoberta de conhecimento efetuado (Maimon e Rokach, 2010).

Concluídas com sucesso estas fases é descoberto novo conhecimento que poderá então ser visualizado ou integrado noutros locais, quer seja, por exemplo, através de outras tecnologias ou no apoio ao processo de tomada de decisão através relatórios.

## 2.5 Data Mining

Ao longo da história da computação tem sido atribuídas diversas denominações à descoberta de padrões nos dados sendo que entre estas atribuições se podem destacar Data Mining (DM), extração de conhecimento, descoberta de informação, colheita de informação, arqueologia de dados e processamento de padrões de dados. O termo de DM tem sido utilizado essencialmente pelas comunidades estatísticas, de analistas de dados e de gestão de sistemas de informação (Fayyad, 1996).

O DM consiste numa aplicação de técnicas e métodos em grandes bases de dados de modo a ser possível encontrar tendências ou padrões como forma de apoiar a descoberta de novo conhecimento (Santos e Azevedo, 2005). É o núcleo do processo de Descoberta de Conhecimento em Bases de Dados e envolve a inferência de algoritmos que exploram os dados, desenvolvem os modelos e que descobrem padrões anteriormente desconhecidos. O modelo é usado para entender os fenómenos relacionados com os dados, análise e predições (Maimon e Rokach, 2010). O DM está associado a diversas áreas podendo ser considerado uma combinação multidisciplinar (Turban, 2010). A figura 7 (adaptado de Turban, 2010) pretende explicitar todas as áreas associadas ao DM.

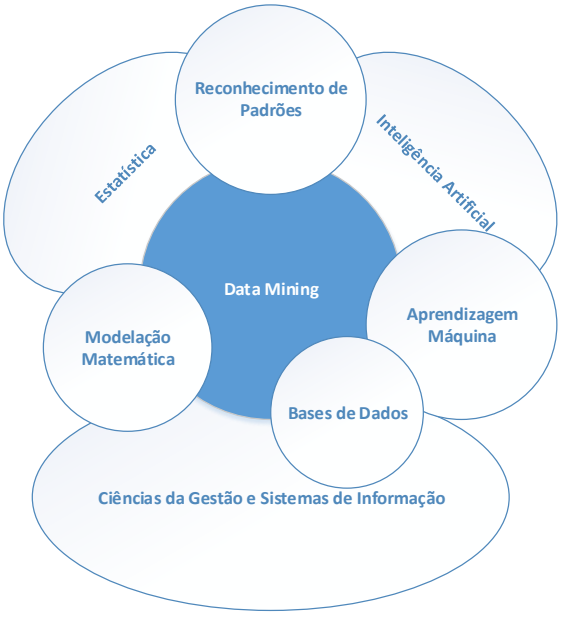


**Figura 7 - Data Mining e Áreas Associadas**

O DM tem-se revelado uma ferramenta bastante útil no apoio ao negócio e a diversas áreas. A utilização de DM é feita na maioria das vezes para que se possam identificar oportunidades ou ameaças ao negócio ou para tentar solucionar problemas. Entre as áreas de aplicação mais conhecidas encontram-se *Costumer Relationship Management* (CRM), banca, empresas de retalho e logística, indústria do entretenimento, desporto, seguradoras, indústria das viagens (companhias áreas, hotéis, aluguer de carros), saúde e medicina (Turban, 2010).

Entre as tarefas de *Data Mining* mais comuns destacam-se (Goebel & Gruenwald, 1999):

- **Processamento de Dados** – automatizar tarefas de processamento de dados de modo a melhorar a produtividade dos analistas;
- **Previsão** – consiste em prever um valor específico de um atributo;
- **Regressão** – análise da dependência de alguns atributos nos valores de outro atributo;
- **Classificação** – determinar a que classe específica a que um item pertence;
- ***Clustering*** – particionar um conjunto de dados em classes tendo em conta as suas características;
- **Associações** – dado um conjunto de dados, identificar relações entre os atributos e os itens;
- **Visualização** – o sistema olho-cérebro é o melhor dispositivo para descoberta de padrões, pelo que se recorre a gráficos, histogramas e outras técnicas visuais para melhorar a interpretação humana;
- **Análise de Dados Exploratória** – é uma análise dos dados interativa sem grandes dependências de modelos ou conhecimento dos dados.

Tipicamente as tarefas de DM podem ser divididas em duas categorias, as descritivas e as preditivas. As descritivas tem como objetivo caracterizar as propriedades gerais dos dados numa determinada base de dados e as preditivas pretendem inferir sobre os dados para se poder obter previsões dos mesmos (Han & Kamber, 2000). As tarefas de previsão mais comuns são a classificação, regressão, deteção de anomalias, enquanto as tarefas descritivas mais comuns são o *clustering*, regras de associação e padrões sequenciais (Gorunescu, 2011).

As tarefas de DM podem também ser classificadas em supervisionadas e não supervisionadas. Nas supervisionadas o processo de aprendizagem é direto através de um atributo ou objetivo conhecido. Este tipo de tarefas tenta explicar o comportamento do *target* (variável independente) como uma função de um grupo independente de atributos ou dependente. Tipicamente a aprendizagem supervisionada resulta em modelos de previsão, e envolve o treino que é o processo através do qual se analisa diversos casos onde o valor do *target* já é conhecido, ou seja, o modelo aprende a lógica que lhe vai permitir efetuar a previsão. Nos modelos não supervisionados a aprendizagem não é direta, pois não existe diferença entre atributos dependentes e independentes e não existem resultados prévios que permitam guiar os algoritmos na construção de modelos. Este tipo de tarefas é utilizado par construir modelos descritivos, mas convém salientar que também pode ser utilizado para fazer predições (Taylor, 2010).

### 2.5.1 Classificação

A classificação consiste na descoberta de uma função que vai associar um determinado caso a uma classe de entre as classes de classificação, para que depois seja possível classificar um novo objeto através do modelo de classificação (Santos & Azevedo, 2005). Ou seja os modelos de classificação mapeiam o espaço de *inputs* em classes pré-definidas. Por exemplo este tipo de modelos pode ser utilizado para classificar o historial bancário dos utilizadores como bom ou mau (Rokach & Maimon, 2010).

Entre as técnicas ou algoritmos mais comuns na tarefa de classificação encontram-se a análise de árvores de decisão, análise estatística, redes neuronais, raciocínio baseado em conhecimento, classificadores de *Bayes* e algoritmos genéticos (Turban, 2010). Existem também outras técnicas para representar modelos de classificação como é o caso das *support vector machines*, sumários probabilísticos e funções algébricas (Rokach & Maimon, 2010).

As árvores de decisão (AD) revelam-se um método eficiente na produção de classificadores dos dados. Um sistema de aprendizagem por AD adota normalmente uma estratégia *top-down* (de cima para baixo) que procura uma solução numa parte do espaço de pesquisa. Uma AD consiste em nós onde os atributos são testados, e em que os ramos de saída de um nó correspondem a todos os possíveis resultados desse mesmo nó. A figura 8 (adaptado de Kantardzic, 2011) pretende representar o conceito de árvore de

decisão. Entre os algoritmos de árvores de decisão mais utilizados encontram-se o *ID3* e o *C4.5* (Kantardzic, 2011).

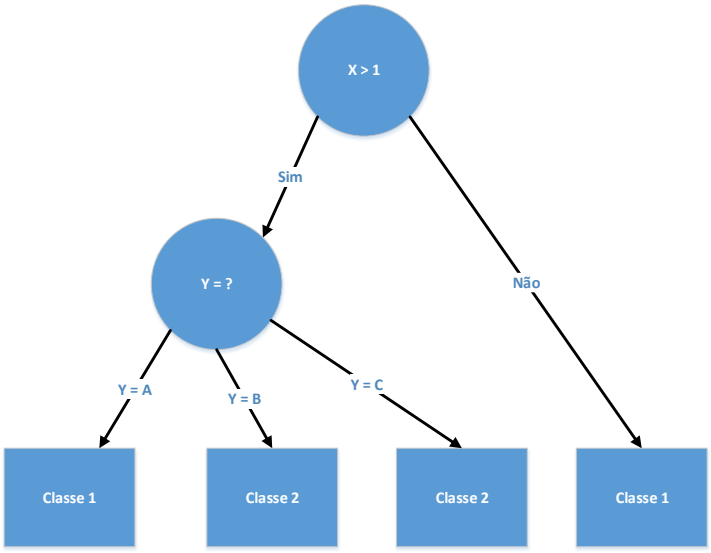


**Figura 8 – Árvore de decisão com testes às variáveis X e Y**

Uma **Rede Neuronal Artificial (RNA)** consiste num conjunto de elementos de processamento simples, que incluem nós, unidades ou neurónios artificiais que se caracterizam pelo grande número de ligações entre eles. As RNA pretendem imitar o comportamento do cérebro humano e a sua principal característica (capacidade de aprender e de se auto corrigir), sendo que a sua estrutura é bastante próxima do mesmo (Santos & Azevedo, 2005).

A **Classificação *Bayesiana*** é uma técnica para a classificação de padrões num conjunto de dados. Este tipo de técnica assume que a classificação de padrões é expressa em termos probabilísticos. A classificação baseada na teoria de *Bayes* pretende classificar objetos baseando-se na informação estatística sobre os mesmos de modo a minimizar a probabilidade de uma classificação ser mal efetuada (Cios et al. 2007). A regra de *Bayes* pode ser expressa pela seguinte expressão,

$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{p(x)},$$

Onde

$P(c_i|x)$ representa a probabilidade *a posteriori*,

$P(c_i)$ a probabilidade *a priori*,

$P(x|c_i)$ a função densidade de probabilidade (a probabilidade da classe $c_i$) e

$p(x)$ a função densidade de probabilidade incondicional.

Segundo Han & Kamber (2000) classificadores ***K Nearest Neighbors*** são baseados na aprendizagem através de analogia, onde os dados de treino são descritos como atributos número de dimensão *n*. Cada amostra representa um ponto no espaço dimensional, ou seja, todas as amostras de treino estão alojadas no espaço dimensional. Quando é dado ao classificador uma amostra desconhecida ele procura padrões no espaço para os espécimes de treino *k* que estão mais próximos da amostra desconhecida. Os espécimes *k* são os vizinhos mais próximos do desconhecido. A proximidade dos vizinhos é definida pela distância Euclidiana, que pode ser expressa pela forma:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$, onde X e Y são dois pontos

As **Support Vector Machines** (SVM) são um método para classificar dados lineares e não lineares. Tipicamente os algoritmos de SVM usam um mapeamento não linear para transformar os dados de treino originais em dimensões superiores. Dentro destas dimensões é procurado o hiperplano ótimo efetuando uma separação linear do mesmo, surgindo então uma fronteira de decisão que separa os tuplos das diferentes classes. Na figura 9 (adaptada de Han et al., 2009) podemos verificar um conjunto de dados que são linearmente separáveis e onde existe um número infinito de hiperplanos e fronteiras de decisão. A escolha do melhor hiperplano é feita através da procura do hiperplano com maior margem, onde a maior margem representa uma maior separação entre as classes (Han et al. 2011).

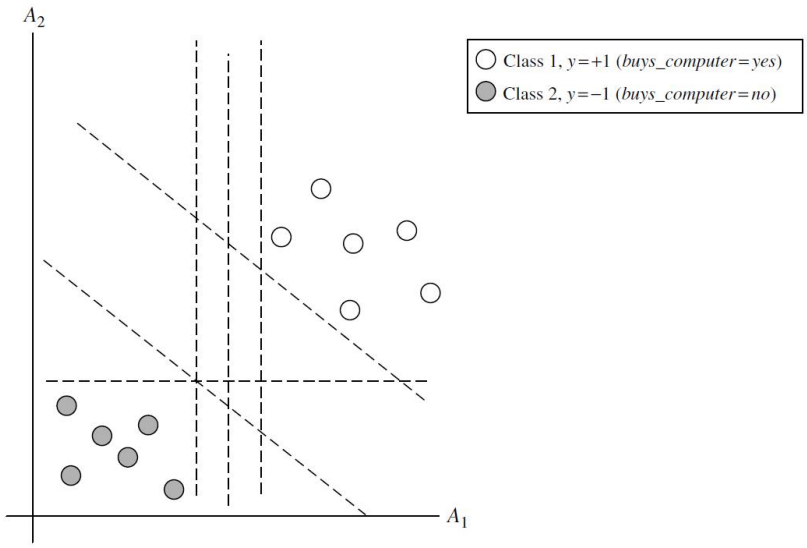


**Figura 9 – Seperação linear de duas classes e hiperplanos/fronteiras de decisão gerados**

Após a criação dos modelos de DM de classificação é necessário avaliar os mesmos. Turban (2010) considera que os fatores mais importantes na avaliação de um modelo de classificação são:

- **Acuidade da previsão** – a capacidade que o modelo tem de prever corretamente a classe é o fator mais comum na avaliação dos modelos de computação;
- **Velocidade** – avaliar os custos computacionais envolvidos na geração e utilização do modelo, em que quanto mais rápido melhor;
- **Robustez** – a capacidade do modelo realizar previsões com uma acuidade razoável, existindo ou não ruido nos dados, valores em falta ou valores incorretos;
- **Escalabilidade** – a habilidade de construir um modelo de previsão eficiente com grandes quantidades de dados;
- **Interpretabilidade** – o nível de conhecimento e foque dados pelo modelo.

Para se fazer a avaliação destes modelos recorrem-se a algumas técnicas que permitem depois analisar métricas de modo a medir a eficiência dos modelos. A mais utilizada é a **matriz de confusão** que permite depois calcular as taxas de erro e definir a curva *Receiver Operating Characteristic* (ROC).

A partir da matriz de confusão (tabela 3) podemos saber qual o número de classificações corretas (Real C1 e C2) versus as previsões efetuadas (Previsão C1 e C2) para cada classe de um determinado modelo.

**Tabela 3 - Matriz de confusão**

| Matriz de Confusão | Previsão C1 | Previsão C2 |
|---|---|---|
| **Real C1** | Verdadeiros Positivos (VP) | Falsos Negativos (FN) |
| **Real C2** | Falsos Positivos (FP) | Verdadeiros Negativos (VN) |

Essencialmente as possibilidades de acerto e erro resumem-se a quatro indicadores (Santos & Azevedo, 2005):

- **Verdadeiros Positivos** (VP) – correspondem ao número de exemplos positivos classificados como tal (corretamente);
- **Verdadeiros Negativos** (VN) – correspondem ao número de exemplos negativos classificados como tal (corretamente);
- **Falsos Positivos** (FP) – correspondem ao número de exemplos positivos classificados como negativos (incorretamente);
- **Falsos Negativos** (FN) – correspondem ao número de exemplos negativos classificados como positivos (incorretamente).

A partir da matriz de confusão podem ser retiradas diversas métricas mas convém salientar as seguintes (Santos & Azevedo, 2005):

- **Acuidade** – é a métrica mais simples, que consiste no cálculo da proporção das instâncias corretamente classificadas, é dada pela expressão

$$\text{Acuidade} = \frac{TP+TN}{n} \times 100(\%)$$

- **Sensibilidade** – a sensibilidade é a proporção de atuais positivos que foram corretamente identificados como positivos pelo modelo. A expressão que permite calcular esta métrica é

$$\text{Sensibilidade} = \frac{TP}{TP + FN} \times 100(\%)$$

- **Especificidade** – esta métrica relata a capacidade do modelo identificar resultados negativos. Considerando um exemplo clínico, a especificidade de um modelo é a proporção de doentes que não tem uma determinada doença e que vão ter um teste para a doença negativo. É dado pela expressão

$$\text{Especificidade} = \frac{TN}{TN+FP} \times 100(\%)$$

A nível clínico uma das métricas mais utilizadas é a sensibilidade, pois os clínicos preferem prever que algo de mal vai acontecer ao doente e agir em conformidade do que ter modelos mais equilibrados que nem sempre acertam nos verdadeiros positivos.

A curva ROC permite avaliar o desempenho de um modelo de classificação, e estabelece a relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos, variando num determinado *treshold* (limiar). Através dela é também possível visualizar o compromisso entre sensibilidade e especificidade. Na análise da curva ROC é também possível avaliar a métrica de *Area Under Curve* (AUC) e o *ROC Convex Hull* (ROCCH). A AUC é uma métrica que permite avaliar o desempenho dos modelos onde é calculada a área por baixo da curva ROC, enquanto a ROCCH permite declarar um subconjunto de classificadores como potencialmente ótimos (Santos & Azevedo, 2005).

### 2.5.2 Regressão

A regressão consiste em procurar uma função que represente de uma forma aproximada os comportamentos das variáveis e apenas pode ser utilizada quando a variável a prever é um número (Santos & Azevedo, 2005). Os modelos mapeiam o espaço que contém os *inputs* em domínios de valor real. Por exemplo um modelo de regressão pode prever a procura de um determinado produto tendo em conta um conjunto de características que foram dadas ao modelo (Benjamini & Leshno, 2010).

Entre as técnicas mais utilizadas na regressão destaca-se a regressão linear, regressão não linear e modelos lineares generalizados (Han & Kamber, 2000).

Na **regressão linear**, os dados são modelados usando uma linha reta. Este tipo de regressão é a mais simples. Normalmente os modelos de regressão linear são bivariáveis e modelam uma variável aleatória, *y* (variável de resposta), como uma função linear de outra variável aleatória, *x* (variável de previsão), sendo que a regressão linear pode então ser expressa pela fórmula (Han & Kamber, 2000):

$$Y= \alpha+ \beta X,$$

Onde $\alpha$ representa a interseção com Y e $\beta$ o declive da reta.

A **regressão não linear** é utilizada quando os dados dos modelos não têm uma dependência linear. Um exemplo de regressão não linear é quando uma variável de resposta e de previsão apresentam uma relação que pode ser modelada por uma função polinomial. Neste caso iriam-se adicionar termos polinomiais ao modelo linear básico, onde através da transformação das variáveis se consegue converter um modelo não linear em um modelo linear que pode ser resolvido com recurso a métodos estatísticos como o dos mínimos quadrados (Han & Kamber, 2000).

Os **modelos lineares generalizados** representam a aplicação de regressão linear em modelos para prever a resposta das variáveis categoricamente. Aqui a variância da variável de resposta *y* é uma função do valor médio de *y*, ao contrário da regressão linear onde o *y* é constante. Entre os modelos lineares generalizados mais comuns encontra-se a **regressão logística**. A regressão logística modela a probabilidade de um acontecimento ocorrer como uma função linear de um conjunto de variáveis de previsão.

Um dos grandes objetivos da aplicação do método de regressão é a conceção do "melhor" modelo tendo em conta uma medida da estimativa do erro. Quando estamos perante um problema de regressão o erro é medido pela expressão $e=d-d'$ , onde $d$ representa o valor desejado e $d'$ representa o valor estimado para o modelo (Santos & Azevedo, 2005).

Para quantificar este erro e por consequência avaliar o modelo de regressão através do sucesso da avaliação da previsão numérica utilizam-se diversas métricas entre as quais se encontram a *Mean Squared Error* (MSE), a *Mean Absolute Error* (MAE), *Relative Squared Error* (RSE), *Relative Absolute Error* (RAE) e *Correlation Coefficient* (CC).

A MSE é a métrica principal e a mais utilizada e é aplicada para dar as mesmas dimensões do que as do valor previsto em si. É bastante utilizada em regressão pois é uma métrica bastante fácil de manipular

matematicamente (Witten et al., 2011). É calculada através da seguinte expressão (*p* representa os valores previstos das instâncias de testes e *a* os valores atuais):

$$\text{MSE} = \frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$$

A MAE é uma alternativa à métrica de MSE e efetua uma centralização da magnitude dos erros individuais. Embora, como já referido, seja bastante utilizada ela tende em exagerar nos efeitos dos *outliers* (instâncias onde o erro de previsão é maior que em outras) (Witten et al., 2011). Esta métrica pode ser calculada pela forma (*p* representa os valores previstos das instâncias de testes e *a* os valores atuais):

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

A métrica RSE funciona de uma maneira um pouco diferente, uma vez que se assume que o erro é relativo em relação ao que deveria ter sido caso se tivesse utilizado um preditor simples. Um preditor simples é a média dos valores atuais dos dados de treino. Então, a RSE pega no *total squared error* (TSE) e normaliza-o dividindo-o pelo TSE do preditor de origem (Witten et al., 2011). É expresso pela fórmula (*p* representa os valores previstos das instâncias de testes, *a* os valores atuais e $\bar{a}$ representa o valor médio dos dados de treino):

$$\text{RSE} = \frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}$$

O RAE consiste no erro total absoluto, sendo normalizado da mesma forma que a métrica RSE (Witten et al., 2011). É calculado recorrendo à fórmula matemática (*p* representa os valores previstos das instâncias de testes, *a* os valores atuais e $\bar{a}$ representa o valor médio dos dados de treino):

$$\text{RAE} = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{|a_1 - \bar{a}| + \cdots + |a_n - \bar{a}|}$$

O CC consiste em efetuar uma medição da correlação estatística entre os valores previstos e os valores atuais (Witten et al., 2011). Calcula-se com recurso à seguinte expressão (*p* representa os valores previstos das instâncias de testes, *a* os valores atuais, $\bar{a}$ representa o valor médio dos dados de teste e $p_i$ representa a enésima instância de teste):

$$\text{CC}= \frac{S_{PA}}{\sqrt{S_P S_A}}, onde\ S_{PA} = \frac{\sum(p_i - \bar{p})(a_i - \bar{a})}{n-1}, S_P = \frac{\sum(p_i - \bar{p})^2}{n-1}, S_A = \frac{\sum(a_1 - \overline{a})^2}{n-1}$$

### 2.5.3 Clustering

A análise de *clusters* divide os dados em grupos (*clusters*) que têm sentido, que são úteis, ou ambos. Nos que possuem sentido os grupos são o objetivo, em que os *clusters* devem capturar a estrutura natural dos dados. Ao contrário de outras técnicas, como a classificação, os *clusters* não são definidos pelo analista, sendo descobertos ao longo do processo. Eles caracterizam-se por ter uma grande homogeneidade interna e uma heterogeneidade externa (Tufféry, 2011).

Em alguns casos, a análise de *clusters* só é um ponto de partida útil para outros propósitos, como a sumarização de dados. Quer seja para a compreensão ou para a utilidade, a análise de *clusters* desempenha um papel importante numa grande variedade de áreas: psicologia e ciências sociais, biologia, estatística, reconhecimento de padrões, recuperação de informação, aprendizagem máquina e Data Mining (Ta net al., 2005).

Existe um grande número de algoritmos de *clustering*, sendo que a escolha do método a utilizar depende tanto do tipo de dados sobre os quais se pretende trabalhar assim como o propósito e aplicação que se pretendem. A grande maioria dos métodos de *clustering* engloba-se nas seguintes categorias (Han & Kamber, 2000):

- **Métodos de partição –** Um método de partição constrói um conjunto de partições dos dados, em que cada partição representa um *cluster*, ou seja, classifica os dados num determinado número de grupos, que satisfazem dois requisitos. O primeiro é o de que cada grupo tem

obrigatoriamente que conter pelo menos um objeto. O segundo requerer que cada objeto só pode pertencer a um e um só grupo.

- **Métodos hierárquicos** – Estes métodos executam uma decomposição hierárquica de um conjunto de dados. Um método deste tipo pode ser aglomerativo ou divisivo, dependendo da forma como a decomposição é efetuada. O aglomerativo começa com cada objeto a formar um grupo isolado. Depois, sucessivamente, os objetos ou grupos são fundidos até que no fim todos os grupos são fundidos num só (topo da hierarquia). O divisivo funciona de forma contrária, começando todos os objetos por estar no mesmo *cluster*, sendo que ao longo de diversas iterações os *clusters* vão sendo divididos em *clusters* mais pequenos até cada um dos objetos representar um *cluster* ou até se atingir uma condição de término.

- **Métodos baseados em densidade -** A maioria dos métodos de partição é baseada na distância entre os objetos, sendo que só conseguem encontrar *clusters* com forma esférica e têm bastantes dificuldades em encontrar *clusters* com outro tipo de formas. Os métodos baseados em densidade são bastante usados para filtrar *outliers* ou para descobrir *clusters* que tem uma forma arbitrária. Têm como princípio, fazer crescer um *cluster* tanto que a densidade (número de objetos ou pontos) na vizinhança exceda um limiar, isto é, por cada ponto de um determinado *cluster*, a vizinhança num certo raio tem que conter pelo menos um número mínimo de pontos.

- **Métodos baseados na *Grid*** – Os métodos baseados na *Grid* restringem o espaço de objetos a um número finito de células que formam uma estrutura em grelha. Depois são feitas todas as operações de *clustering* na estrutura da grelha. A grande vantagem deste método é o rápido tempo de processamento que é tipicamente independente do número de objetos, e apenas dependente do número de células de cada dimensão do espaço restringido.

- **Métodos baseados em modelos** – Os métodos baseados em modelos formulam uma hipótese de modelo para cada um dos *clusters*, e encontram o melhor ajuste dos dados ao modelo. Estes algoritmos podem localizar *clusters* através da construção de funções de

densidade que reflitam a distribuição espacial dos pontos. Este tipo de método leva a uma forma automática de determinar o número de *clusters*, assim como identificar ruído ou *outliers*, sendo por isso um método de *clustering* bastante robusto.

Ao nível da avaliação dos resultados da tarefa de *clustering* podem-se considerar dois grandes fatores a *compactness* e a separabilidade. A *compactness* é uma propriedade que expressa o quanto os elementos de um *cluster* estão próximos. Ou seja quanto menor for o valor da variância maior será a *compactness* do *cluster*. Uma técnica útil para aferir esta propriedade é calcular a distância *intra-cluster*, que quanto menor melhor e por consequência maior a *compactness* do *cluster*. A separabilidade, permite avaliar o quão distintos os *clusters* são. Uma maneira bastante intuitiva de expressar a separabilidade é computar as distâncias *inter-cluster*, que deverá ser o maior possível, pois quanto maior, maior a separabilidade dos *clusters*. Para ajudar à avaliação destes dois fatores utilizam-se tipicamente duas métricas o *Davies-Bouldin Index* e o *Dunn Index*. Estas métricas efetuam uma avaliação interna dos *clusters* e utilizam nas suas formas matemáticas componentes como as distâncias inter e intra *clusters* (Cios et al., 2007).

### 2.5.4 Metodologias Standard de Data Mining

A DCBD e o DM são cada vez mais utilizados em um largo espetro de áreas. Como tal surgiu a necessidade de existirem metodologias *standard* que permitam assegurar qualidade, assim como um conjunto de boas práticas e *guidelines* para o processo de DM. Nas últimas duas décadas assistiu-se um aparecimento de alguns destes *standards* para DM entre os quais se destacam o *Cross Industry Starndard Process for Data Mining* (CRISP-DM), o *Sample, Explore, Modify, Model and Assess* (SEMMA) e o *Predictive Model Markup Language* (PMML).

O CRISP-DM foi uma tentativa de fornecer um *standard* industrial para a prática de *Data Mining* e compreende seis fases: compreensão de negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e instalação (Nadali et al., 2011). Na fase inicial de compreensão de negócio pretende-se entender os objetivos e requisitos de negócio, definindo então um problema de DM. A segunda fase é a compreensão dos dados onde se realizam um conjunto de passos para ganhar conhecimento dos dados, identificar problemas de qualidade nos dados e descobrir alguns padrões ou conjuntos que "saltem à vista". De seguida, vem a fase de preparação dos dados, onde se seleciona,

transforma e limpam os dados a utilizar nos modelos a criar. A quarta fase consiste na modelação, onde se começa por selecionar um conjunto de técnicas de modelação e se calibram os parâmetros para atingir valores ótimos. Seguidamente é necessário avaliar os modelos criados na fase anterior, verificando se foram atingidos todos os objetivos. Por último é feita a instalação que consiste na organização e apresentação do conhecimento adquirido de modo a que os utilizadores possam utilizar essa informação (Chapman et al., 2000).

O SEMMA é uma metodologia orientada à seleção, exploração e modelação de uma grande quantidade de dados, para ajuda à descoberta de padrões de negócio nos dados. O acrónimo (SEMMA) representa *sample*, *explore*, *modify*, *model* e *assess* que segundo a SAS são as fazes nucleares da condução de um projeto de DM. A primeira tarefa do modelo SEMMA consistem em efetuar um *sample* dos dados de um grande *data set* para que o *sample* contenha informação significante da base de dados mas ao mesmo tempo seja bastante rápida de manipular. A segunda fase consiste na exploração dos dados tentando descobrir tendências e anomalias nos mesmos para ganhar conhecimento e ideias dos dados. A terceira fase consiste na modificação dos dados através da criação, seleção e transformação das variáveis para ser mais fácil focar no processo de seleção dos modelos. A fase seguinte consiste na modelação, onde se deve modelar os dados permitindo ao *software* procurar automaticamente por uma combinação de dados que fiavelmente preveja o *outcome* desejado. Por último temos a fase de avaliação onde se vai aferir sobre a fiabilidade e utilizado das descobertas feitas pelo processo de DM assim como estimar o quão bem ele o faz (SAS Institute).

O PPML é uma especificação, independente das ferramentas de DM, que permite às aplicações e ferramentas lidar com diversas fontes sem ter de lidar com as diferenças particulares de cada uma. Outro dos objetivos desta especificação é permitir a utilização colaborativa de uma panóplia de modelos e a administração de conjuntos de modelos. Esta especificação está desenvolvida em *Extensible Markup Language* (XML), que foi escolhida pela sua grande utilização na internet, suporte para *tags* e possibilidade de linguagens próprias de *markup* (Santos & Azevedo, 2005).

### 2.5.5 Data Mining na Medicina Intensiva

Diariamente são admitidos doentes em UCI com uma grande diversidade de problemas e sintomas. Devido ao grande número de dispositivos de monitorização em constante funcionamento, a quantidade

de dados gerados para cada doente em cada unidade de cuidados intensivos é bastante volumosa, dados estes que são bastantes complexos.

Nos últimos anos tem surgido diversos trabalhos de investigação que consistem na implementação de sistemas de suporte à decisão em unidades de cuidados intensivos (UCI) e de utilização de técnicas de Data Mining na análise e previsão dos parâmetros utilizados nestas unidades.

A utilização de modelos de previsão em medicina intensiva traz um conjunto de ferramentas que vão ajudar no processo de tomada de decisão clínico através da combinação de itens dos dados dos doentes, prevendo o *outcome* clínico dos mesmos (Wyatt & Altman, 1995).

Existem diversas vantagens no que toca à utilização de DM na área da medicina intensiva, e mais concretamente nas UCI, como a capacidade que o DM possuí de permitir a utilização do conhecimento sobre o domínio, tornar a análise de dados compreensiva e orientada (Bellazzi & Zupan, 2008), e claro o apoio ao processo de tomada de decisão aos decisores clínicos, que alguns por falta de experiência, ou outros por falta de tempo, não conseguem estar atentos a todos os pormenores, devido ao grande número de variáveis associadas a cada doente das UCI, pelo que os modelos de previsão em medicina intensiva ajudam à tomada de melhores decisões clínicas.

A utilização de DM na saúde crítica permite explorar uma melhoria na qualidade dos cuidados, por exemplo através de inclusão de diferentes políticas de tratamento tendo em conta os resultados obtidos assim como construir a base para a construção de sistemas de apoio a decisão em UCI (Lucas, 2004).

No entanto, existem ainda diversas dificuldades e desafios associados ao DM na área da medicina intensiva. Um deles é que muitas vezes é necessário desenvolver técnicas e heurísticas especiais para poder identificar correlações biomédicas que possuam sentido no âmbito do que se pretende analisar. A segunda grande dificuldade é a grande dimensão dos repositórios de dados clínicos, onde os dados para relações complexas estão tipicamente dispersos por diversas dimensões. (Mullins et al., 2006). Por outro lado os dados clínicos trazem muitas vezes ruído, devido à falha de algum dispositivo de monitorização ou não registo de dados por parte da equipa clínica. Outra das dificuldades são as características

individuais do doente, pelo que por vezes é difícil comparar valores dos atributos entre diferentes doentes (Ramon et al., 2007).

Moser et al. (1999) propuseram um sistema de DM que permite monitorizar infeções que estão a emergir e a resistência anti microbial dos doentes. Utilizaram regras de associação e analisam os dados respeitantes a um período de um ano, provenientes de uma UCI do Hospital de Birmingham. Foram encontrados diversos eventos dos quais 18 se julgaram ser potencialmente interessantes. Esses eventos foram então encaminhados para especialistas da área da saúde para que se pudesse avaliar a sua importância. Este trabalho revelou-se interessante pois permite uma abordagem que pode ser implementada em programas de controlo de infeções ou num sistema de vigilância de saúde pública.

Cheng et al. (2013) propuseram o sistema icuARM, um sistema de suporte à decisão clínica em tempo real que utiliza regras de associação para identificar associações entre os parâmetros existentes nas bases de dados. Este sistema gera depois um conjunto de regras de suporte à decisão clínica que vão ajudar os intensivistas.

Morik et al. (2000) desenvolveram uma abordagem para ajudar à criação de protocolos operacionais em medicina intensiva. Estes protocolos são bastante valiosos mas também bastante difíceis de desenvolver devido ao custo elevado e alta complexidade. Os autores utilizaram uma conjugação de análise inteligente de dados e de aquisição de conhecimento junto de especialistas clínicos. A aquisição de regras de ação base para os protocolos foi feita com recurso a algoritmos de SVM que permitiram analisar os dados dos doentes para adquirir o modelo atual de comportamento dos intensivistas.

Santhi & Bhaskaran (2010) compararam a performance de diversos algoritmos de *clustering* na previsão de doença coronária. No geral os algoritmos utilizadas apresentaram bons resultados sendo que os algoritmos *K-means*, *Make Density Base Clusters* e *ExpectationMaximization* obtiveram acuidades acima dos 80%, demonstrando assim a validade deste tipo de algoritmos na predição de problemas clínicos.

### 2.5.6 INTCare

O INTCare é um sistema de suporte à decisão inteligente em tempo real que se encontra em constante desenvolvimento e testes (envolvendo a UCI do CHP), baseado em agentes inteligentes, para suportar a decisão clínica, fazer previsões sobre falha de sistemas de órgãos (cardiovascular, respiratório, renal, hepático, hematológico e neurológico) e *outcome* (a condição do doente no fim da terapia ou de um processo de doença). Tendo em conta as previsões feitas o sistema é capaz de sugerir procedimentos, tratamentos e terapias, através de técnicas de DM (Santos et al. 2011).

Este projeto começou no ano de 2008 e um dos seus primeiros objetivos foi modificar o sistema de aquisição e recolha de dados, modificando os processos através da alteração da informação registada em papel e processada manualmente para um sistema de informação automático e eletrónico (Portela, 2011). Atualmente os dados de monitorização como os sinais vitais, dados clínicos e dados relativos aos resultados laboratoriais são recebidos e enviados automaticamente para o Registo Eletrónico de Enfermagem que disponibiliza todos os dados clínicos aos médicos e enfermeiros de modo a suportar a decisão (Portela, 2011).

Este sistema é baseado em quatro subsistemas autónomos e que utilizam agentes inteligentes para realizar tarefas, sendo eles o de aquisição de dados, gestão de conhecimento, inferência e *interface* (Santos et al. 2011). A figura 10 (Santos et al., 2011) representa a arquitetura do sistema INTCare (Portela et al., 2013c).

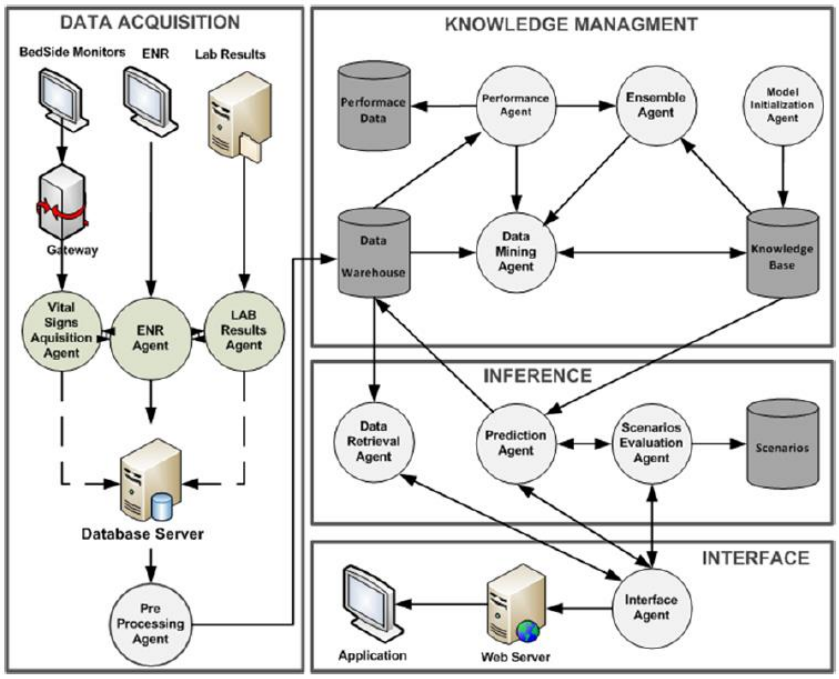


**Figura 10 – Arquitetura INTCare**

Como já referido este projeto constituí uma mudança dos paradigmas de registo de monitorização de doentes em UCI, sendo que trouxe à UCI do CHP as seguintes inovações:

- Prognósticos fornecidos por modelos de DM, ao nível da previsão da falência de órgãos e na avaliação da mortalidade (Portela et al., 2014a);
- Desmaterialização de processos (Portela et al., 2013a);
- Automatização de tarefas (Portela et al., 2013a),
- Monitorização e análise de eventos críticos/sinais vitais (Portela et al.,2012a; Portela et al., 2013b);
- Aquisição e cálculo automáticos de scores clínicos em medicina intensiva (Portela, et al., 2012b; Portela et al., 2014b).

A transformação que foi efetuada ao nível dos sistemas de informação da UCI e a atual realidade onde os dados são monitorizados e recolhidos em tempo real possibilitam a criação de novo conhecimento útil para a decisão, bem como a criação de novos modelos de previsão e decisão, os quais estão relacionados com o trabalho desta dissertação. O projeto INTCare II procura essencialmente apoiar a decisão clínica através de modelos de previsão para reinternamentos e ventilação assim como para a decisão administrativa e financeira com a criação de modelos que permitam otimizar os custos nas UCI.

### 2.5.7 Data Mining na Gestão de Fármacos, Tempos de Internamento e Taxas de Ocupação de Serviços Hospitalares

Nos últimos anos tem-se assistido a diversos estudos na área de gestão de medicamentos, prescrições médicas e custos na saúde recorrendo a técnicas de DM. No âmbito dos medicamentos os principais estudos têm incidido no *Drug Safety Data Mining* (Brown et al., 2011) que não é mais que o recurso a técnicas de DM para avaliar as relações entre diferentes medicamentos e um grande número de efeitos adversos provenientes da administração de medicamentos.

Last et Al. (2007) referem que com um aumento de doentes crónicos, como os asmáticos, existe um aumento dos custos com saúde devido à grande quantidade de medicamentos que estes doentes tendem a consumir. Estes investigadores aplicaram técnicas de DM para a descoberta e avaliação dos padrões de medicação em doentes asmáticos em regime ambulatório de uma grande organização de manutenção de saúde. Com a utilização destas técnicas é mais fácil aos médicos e hospitais monitorizar eficientemente a utilização de medicamentos por parte de doentes crónicos assim como identificar grupos de doentes que estejam a receber uma quantidade abaixo ou acima da necessária.

Através do financiamento da União Europeia foi criado o projeto PSIP[2], que pretende desenvolver mecanismos inteligentes para prevenir os efeitos adversos provocados por medicamentos. Estes efeitos são atualmente considerados um assunto de saúde pública bastante grande, e uma melhoria e maior controlo nesta área permite aumentar a segurança dos doentes e reduzir os custos com a saúde. Este sistema é responsável por melhorar toda a cadeia de medicamentos desde a prescrição, dispensação, administração e observância. O PSIP utiliza entre outras técnicas, *Data Mining* que é aplicado nos registos dos doentes, e cruzado com bases de dados de contra-indicações de medicamentos, de modo a identificar a origem ou prevenir efeitos adversos provocados pelos medicamentos nos doentes (Koutkias, Lazou, Kilintzis, 2009).

[2] www.psip-project.eu

No âmbito dos efeitos adversos dos medicamentos, Chazard et al. (2011) também desenvolveram estudos sobre a aplicação de DM nesta área. Este grupo de investigadores recorreu a árvores de decisão e a regras de associação para geraram modelos que permitam detetar os efeitos adversos dos medicamentos nos doentes. Este estudo revelou soluções bastante inovadoras para a deteção destes efeitos e os autores referem que poderá ser aplicado a outros tipos de dados como por exemplo os resultados dos eletrocardiogramas. De referir que as regras descobertas com este estudo já estão a ser incluídas em diversos protótipos como o projeto PSIP.

Huag et al. (2013) propuseram algo diferente na área dos efeitos adversos dos medicamentos que consiste na conjugação de tecnologias de semântica e de Data Mining. Este trabalho de investigação explora a utilização de dois algoritmos de descoberta de padrões orientados a semântica, criando ontologias nos registos eletrónicos dos doentes, que depois permitem descobrir efeitos adversos dos medicamentos que originalmente estavam escondidos e/ou implícitos. Esta conjugação de tecnologias trará uma diminuição de custos na investigação e desenvolvimento de novos medicamentos assim como permitirá revolucionar a fármaco-vigilância para reduzir os efeitos adversos nos doentes e melhorar a saúde dos mesmos.

Fernandes e Belo (2010) efetuaram estudos sobre a descoberta de padrões nas prescrições médicas eletrónicas em centros de saúde da região norte de Portugal. O estudo dos padrões foi feito com recurso a regras de associação, uma das tarefas existentes em *Data Mining*. A partir deste estudo foi possível identificar associações entre medicamentos prescritos, por exemplo, quando é receitado Ben-U-Ron[3] é muitas vezes receitado Brufen Suspensão[4], e também entre as prescrições dos médicos e os laboratórios.

Rastegar-Mojarad et al. (2013) propuseram o sistema Paracelsus, um sistema de *Data Mining* textual, que é uma ferramenta que faz extração de conhecimento estruturado das bulas dos medicamentos que estão no mercado e classifica as interações cruzando estes dados com a informação sobre interações

[3] O ben-u-ron é um medicamento está indicado no tratamento sintomático de situações clínicas que requerem um analgésico e/ou um antipirético.

[4] O bruffen é um medicamento indicado como analgésico, antipirético e para tratamento reumatológico

de substâncias ativas dos medicamentos. Com esta ferramenta é então possível classificar as interações entre medicamentos como de pouco risco, a evitar, a evitar a não ser que seja benéfico entre outros.

Ao nível do tempo de internamento (TI) de doentes em UCI, existem na literatura vários estudos que utilizando tarefas e algoritmos de Data Mining conseguem prever o TI, no entanto nenhum deles é específico para unidades críticas de saúde. Isken e Rajagopalan (2002) recorreram ao uso de Clustering, e de algoritmos *K-means* para identificar tipos de doentes de modo a se poderem construir simulações computacionais ou modelos analíticos dos fluxos de doentes dentro do hospital, o que engloba analisar indicadores como TI ou taxas de ocupação de camas, outro dos indicadores a abordar mais à frente.

Hachesu et al. (2013) no seu trabalho de investigação utilizaram técnicas de classificação recorrendo a algoritmos de árvores de decisão, *support vector machines* (SVM) e redes neuronais artificiais para prever os TI de doentes com doença arterial coronária. Este estudo revelou boas acuidades principalmente aquando da utilização de algoritmos SVM.

Azari et al. (2012) propuseram uma abordagem multi-camada para prever os TI nos hospitais. Este trabalho caracterizou-se pela utilização de técnicas de Clustering para proceder à construção dos *data sets* de treino, uma vez que ao usar esta estratégia consegue-se depois atingir melhores resultados aquando da aplicação de algoritmos de classificação nos *data sets*. Utilizaram diversos tipos de algoritmos de classificação, sendo que os que demonstraram os melhores resultados foram os de *Naive Bays*, SVM, JRIP, J48.

Em outro âmbito hospitalar, Zhang et al. (2012) desenvolveram um estudo de modo a prever os TI na UCI do hospital de Shenyang na China em doentes idosos após cirurgia a tumores gástricos. Ao nível dos algoritmos utilizados estes autores optaram pela utilização de árvores de regressão para prever o TI assim como explorar outros indicadores hospitalares.

Caetano et al. (2014) propuseram um modelo preditivo direcionado aos dados para a previsão do tempo de internamento utilizando dados de 2000 a 2013 e usaram como variáveis de entrada os indicadores típicos do processo de hospitalização. Desenvolveram diversos modelos de DM sendo que os melhores resultados foram atingidos com recursos a *random forests*.

Relativamente à taxa de ocupação de serviço hospitalar, já existem referências à utilização de DM na previsão desta taxa. El-Darzi et al. (1998) sugeriram uma abordagem para simular os fluxos de doentes de modo a proceder, entre outros, a uma avaliação de camas vazias e bloqueadas em um departamento geriátrico de um hospital. Este estudo revelou em fases iniciais bastantes problemas, uma vez que os modelos gerados eram pouco confiáveis sendo que após algumas alterações se conseguiu obter resultados um pouco mais razoáveis mas ainda longe do pretendido.

Nos últimos anos têm surgido estudos no sentido de analisar as taxas de ocupação das unidades hospitalares com recurso a *Data Mining*, entre os quais se destaca o de Alapont et al. (2005), um estudo sobre gestão hospitalar, onde se pretendia uma melhoria da utilização de recursos hospitalares, evitar situações de taxa de ocupação de camas superior a 100% e otimizar os cronogramas dos blocos operatórios. Utilizaram-se técnicas de Classificação e Regressão e algoritmos de Regressão Linear, SVM, Arvores de Decisão, Redes Neuronais Artificiais entre outros.

Teow et al. (2012) realizaram um trabalho de investigação focado na gestão de camas, de modo a ajudar no planeamento da transferência de camas e alocação das mesmas no Hospital da Universidade Nacional de Singapura. Neste trabalho foram utilizados várias técnicas de Data Mining como Árvores de Decisão, Redes Neuronais Artificiais e Regressão Logística.

# 3. TRABALHO REALIZADO

## 3.1 Clustering na Previsão de Readmissões em Medicina Intensiva

As readmissões não planeadas em Unidades de Cuidados Intensivos (UCI) estão bastante ligadas a más decisões tomadas pelos intensivistas no momento da alta. A capacidade de prever com sucesso a recaída de um doente após ter tido alta médica é bastante limitada (Gajic et al., 2008).

É considerado um caso de readmissão não planeada quando um doente é admitido na mesma unidade clinica onde esteve internado num intervalo de tempo inferior a trinta dias após a alta e com um diagnóstico igual ao do primeiro internamento (ACSS, 2012).

Na Medicina Intensiva existem diversos modelos e técnicas matemáticas que podem ser utilizados para prever a probabilidade de readmissão de doentes em uma UCI. Gajic et al. (2008) desenvolveram um índice numérico chamado *Stability and Workload Index for Transfer* (SWIFT). Este índice considera um conjunto de variáveis que podem ser usadas para prever readmissões não planeadas numa UCI. Este conjunto é composto pelo local de proveniência, o tempo de internamento total em dias, a última medição do rácio PaO2/FIO2 e de PaCO2 e a escala de coma de Glasgow. A tabela 4 (adaptada de Gajic et al., 2008) apresenta as variáveis utilizadas no modelo SWIFT e a pontuação a atribuir a cada uma tendo em conta o seu valor.

**Tabela 4 – Scores a atribuir segundo modelo SWIFT**

| Variável | Resultado |
|---|---|
| **Local de Proveniência** | |
| Sala de emergência ou urgências | 0 |
| Transferido de outro serviço ou hospital | 8 |
| **Tempo total de internamento na UCI (em dias)** | |
| Menos de 2 dias | 0 |
| Entre 2 e 10 dias | 1 |
| Mais de 10 dias | 14 |
| **Ultima medição do rácio PaO2/FIO2** | |
| Mais de 400 | 0 |
| Menos de 400 e maior ou igual a 150 | 5 |
| Menos de 150 e maior ou igual a 100 | 10 |
| Menos de 100 | 13 |
| **Escala de coma de Glasgow** | |
| Maior do que 14 | 0 |

| | |
|---|---|
| Entre 11 e 14 | 6 |
| Entre 8 e 10 | 14 |
| Menor que 8 | 24 |
| **Ultima medição de PaCO2** | |
| Menos de 45 mm Hg | 0 |
| Mais de 45 mm Hg | 5 |

No âmbito do projeto INTCare foi realizado um estudo prévio da previsão de readmissões em UCI (Braga et al., 2014) que atingiu bons resultados na previsão do doente ser ou não readmitido, mas isso por si só não é suficiente, pois é necessário categorizar o tipo de doentes reinternados. Este estudo pretende através do uso de cenários de *clustering* encontrar características comuns nos doentes da UCI que levem a uma futura readmissão não planeada. Ao nível das ferramentas utilizadas, foi utilizada a ferramenta Oracle SQL Developer para a análise, compreensão e preparação dos dados e a ferramenta RapidMiner para a construção dos cenários de *clustering*. O Oracle SQL Developer é um ambiente integrado que simplifica o desenvolvimento e gestão de bases de dados Oracle. O Rapidminer é uma plataforma que fornece um ambiente para DM, *machine learning*, *text mining*, análise preditiva e de negócios (Chapman & Hall, 2013). No âmbito deste trabalho foi utilizado a versão *RapidMiner Studio* 6, que possui um conjunto de componentes *open-source*, que não tem custos para o utilizador e é largamente utilizada em âmbito investigacional e educacional. A ferramenta RapidMiner (Hofmann & Klinkenberg, 2013) oferece uma panóplia de métodos de *clustering*, pelo que foi efetuada uma análise de performance dos mesmos com um conjunto de dados reais para fazer testes de modo a selecionar os métodos a utilizar na fase de criação dos modelos. A análise de performance englobou os métodos *k-means*, *k-means* com *kernels*, *k-means* fast, *k-medoids*, *x-means*, *expectation maximization clustering*, *top down clustering*, DBSCAN, *support vector clustering*, *random clustering* e *flatten clustering*. Após análise dos resultados verificou-se que os métodos que apresentaram melhores resultados, em termos de critério de domínio e estatístico foram os métodos *k-means*, *k-medoids* e *x-means*, pelo que foram os métodos utilizados na modelação deste estudo.

### 3.1.1 Compreensão do Negócio

O principal objetivo deste estudo é diminuir o número de readmissões na UCI através do uso de modelos de Data Mining (DM). Estes modelos tem como principal meta efetuar uma caracterização de grupos de doentes que estão em risco de serem readmitidos. Os modelos que foram gerados neste estudo irão suportar a decisão médica, essencialmente no momento da decisão de dar ou não alta a um doente,

assim como contribuir para uma melhoria da qualidade do serviço pois quanto menor for o número de doentes com readmissões não planeadas maior será a qualidade do serviço médico prestado.

### 3.1.2 Compreensão dos Dados

Este estudo utilizou dados reais adquiridos das bases de dados do CHP – HSA. Neste contexto foram utilizados dados provenientes de duas fontes, sendo elas, os resultados laboratoriais e o registo eletrónico de enfermagem. Os resultados laboratoriais estavam armazenados na tabela *PCELabres* enquanto os dados relativos ao registo eletrónico de enfermagem estavam armazenados na tabela *UCI_Internados*. A tabela 5 permite observar as variáveis presentes nas tabelas *PCELabres* e UCI_Internados assim como uma breve descrição das mesmas.

**Tabela 5 – Descrição das variáveis disponíveis para previsão de readmissões**

| Tabela | Variável | Tipo | Descrição |
|---|---|---|---|
| **PCELabres** | PedidoID | Varchar | Código identificador do pedido de análises |
| **PCELabres** | ResultadoID | Varchar | Código identificador do pedido de análises |
| **PCELabres** | Episodio | Number | Episodio a que os resultados dizem respeito |
| **PCELabres** | Executante | Varchar | Serviço que executou a análise |
| **PCELabres** | Estado | Varchar | Estado em que a análise laboratorial se encontra |
| **PCELabres** | NumAmostra | Varchar | Número da amostra utilizada na análise |
| **PCELabres** | CodExame | Varchar | Código do exame efetuado |
| **PCELabres** | Exame | Varchar | Nome do exame efetuado |
| **PCELabres** | Resultado | Varchar | Resultado do exame |
| **PCELabres** | Unidades | Varchar | Unidades nas quais o resultado está expresso. |
| **PCELabres** | ValoresREF | Varchar | Valores de referência para o exame em causa |
| **PCELabres** | DataValidacao | Date | Data em que o resultado da análise foi validado. |
| **UCI_Internados** | NProcesso | Number | Número de processo do doente |
| **UCI_Internados** | Episodio | Number | Código que identifica o episódio a que o registo diz respeito |
| **UCI_Internados** | Nome | Varchar | Nome do doente |
| **UCI_Internados** | Sexo | Varchar | Sexo do doente |

| Tabela | Variável | Tipo | Descrição |
|---|---|---|---|
| **UCI_Internados** | DataN | Date | Data de nascimento do doente |
| **UCI_Internados** | Data_Entrada | Date | Data e hora a que o doente deu entrada na UCI. |
| **UCI_Internados** | Data_Saida | Date | Data e hora em que o doente obteve alta. |
| **UCI_Internados** | Urgencia | Varchar | Indica se o doente é proveniente da unidade de urgência |
| **UCI_Internados** | SalaEmergencia | Varchar | Identificar se o doente vem da sala de emergência. |
| **UCI_Internados** | Obs | Varchar | Informa se o doente é proveniente da unidade de observações. |
| **UCI_Internados** | Blocooperatorio | Varchar | Doentes provenientes do bloco operatório. |
| **UCI_Internados** | Enfermaria | Varchar | Doentes provenientes da enfermaria. |
| **UCI_Internados** | OutraUCI | Varchar | Indica se o doente veio de outra UCI. |
| **UCI_Internados** | UnidadeIntermedia | Varchar | Informa se o doente é proveniente de uma unidade intermédia. |
| **UCI_Internados** | Direta | Varchar | Indica os casos dos doentes provenientes de via direta. |
| **UCI_Internados** | OutroHospital | Varchar | Possui um valor verdadeiro caso o doente tenha vindo de outro hospital. |
| **UCI_Internados** | OutraSituacao | Varchar | Informa sobre outras situações de admissão do doente. |

Ao nível do número de registos, a tabela *UCI_Internados* tinha 54586 e a *PCELabres* 458524. Uma vez que a maioria destes dados são recolhidos através de agentes inteligentes a qualidade dos dados é bastante elevada, sendo que o número de registos nulos se encontrava abaixo dos 0,1% excetuando a variável *Glasgow_Hospital* que possuía 18661 registos nulos. Embora esta variável assuma uma importância alta uma vez que é um dos atributos presentes no modelo SWIFT a mesma não foi considerada devido a este atributo ser registado manualmente e a sua taxa de valores nulos ser bastante alta (cerca de 34%). Os restantes atributos nulos que, como dito acima, representam uma percentagem residual foram removidos de modo a não serem utilizados no conjunto de dados que serviu de entrada para os modelos.

### 3.1.3 Preparação dos Dados

Ao nível da preparação dos dados foram selecionados e transformados um conjunto de variáveis que permitiram efetuar derivações de novos atributos a usar nos modelos de DM. Foram selecionados os atributos *sexo, pco2, salaemergencia, acido_lactico e leucocitos*. Como as variáveis necessárias aos modelos se encontravam em tabelas distintas, a ligação entre as mesmas foi assegurada através do atributo *episódio*, garantido assim que os dados se referem unicamente a um e um só episódio. Uma vez que existe um grande número de registos por doente, relativamente aos resultados laboratoriais e uma vez que se pretende avaliar as readmissões foram selecionados os valores que tenham a data de validação do exame mais próxima da data em que o doente teve alta.

O atributo *datan* que contém a data de nascimento do doente foi derivado para o atributo *idade* que representa a idade do doente em anos. Neste processo foi também criado um novo atributo *tempointernamento* que representa o tempo de internamento do doente na UCI através do cálculo da diferença entre a data em que o doente teve alta (*data_saida*) e a data em que o doente foi admitido na UCI (*data_entrada*). O atributo *readmissão* foi derivado a partir das variáveis *número de processo*, *data de entrada* e *data de saída*. Este atributo consiste numa *flag* que representa se o caso em questão é ou não uma readmissão. O atributo tem o valor 1 quando o doente tiver uma segunda admissão para o mesmo diagnóstico em um intervalo de tempo inferior a trinta após a data da alta.

O atributo *po2/fio2_ratio* resulta da divisão do PaO2 pelo FiO2 e representa o rácio entre a pressão parcial de oxigénio no sangue arterial e a fração de oxigénio inspirado. Como existem diversas medições dos valores de PaO2 e FiO2 foi utilizada a última medição disponível antes do doente ter alta. Este rácio é um critério necessário ao modelo SWIFT e permite saber se exista um problema na transferência de oxigênio dos pulmões para o sangue. Por último e de modo a aplicar o modelo SWIFT foram utilizados os atributos PaO2, PaCO2, rácio PaO2/FiO2 e tempo de internamento que deram origem aos atributos *pco2_score, po2/fio2_score, salaemergencia_score* e *tempointernamento_score* que surgem através da atribuição de um *score* de acordo com a tabela de pontuação do modelo SWIFT.

Todos os processos de transformação acima citados foram elaborados recorrendo ao desenvolvimento de procedimentos, *triggers* e funções automáticas. Este processo permite que cada vez que chegue um

novo valor o mesmo seja transformado em tempo real, de modo a que o mesmo esteja de acordo com os valores esperados por cada um dos atributos de DM.

Relativamente à integração de dados foi criada uma *view* que contém todos os atributos que irão ser utilizados nos modelos a criar. Esta *view* contém os atributos sexo, *idade, tempointernamento, salaemergencia, po2/fio2_ratio, pco2, acido_lactico, leucócitos, po2/fio2_score, pco2_score, salaemergencia_score, tempointernamento_score* e *readmissao*. A tabela 6 apresenta algumas estatísticas para cada uma destas variáveis.

**Tabela 6 – Estatísticas das variáveis selecionadas para previsão de readmitíveis**

| Variável | Valores Distintos | Média | Valor Mínimo | Valor Máximo |
|---|---|---|---|---|
| **idade** | 79 | 63,95 | 17 | 96 |
| **sex** | 2 (1 ou 2) | - | - | - |
| **tempointernamento** | 38 | 6,39 | 0 | 62 |
| **salaemergencia** | 2 (*true* ou *false*) | - | - | - |
| **po2/fio2_ratio** | 754 | 287.83 | 37.8 | 6100 |
| **pco2** | 325 | 40,92 | 16,8 | 116,4 |
| **acido_lactico** | 220 | 2,06 | 0,2 | 16,0 |
| **leucocitos** | 766 | 36,79 | 0 | 451,0 |
| **po2/fio2_score** | 4 (0; 5; 10 ou 13) | - | - | - |
| **pco2_score** | 2 (0 ou 5) | - | - | - |
| **salaemergencia_score** | 2 (0 ou 8) | - | - | - |
| **tempointernamento_score** | 3(0; 1 ou 14) | - | - | - |
| **readmissao** | 2 (Sim ou Não) | - | - | - |

A *view* criada contém os dados que serviram de entrada para os modelos. Estes dados foram recolhidos durante 1389 dias que correspondem ao intervalo de tempo desde 23 de Abril de 2010 até 10 de Fevereiro de 2014. Este conjunto de dados diz respeito a 1043 casos clínicos distintos. Durante este período o número de readmissões corresponde a 36 casos, ou seja, cerca de 3,5% dos casos registados. Apesar de este valor numericamente ser baixo, no contexto médico é bastante elevado pois, segundo os intensivistas, o objetivo é ter um número próximo de zero.

### 3.1.4 Modelação

Nesta fase criaram-se modelos que concretizem os objetivos de negócio através da aplicação de técnicas de DM. De modo a melhor explicitar os cenários e modelos desenvolvidos, doze das treze variáveis a

utilizar como entrada para os modelos foram agrupadas em quatro grupos. A tabela 7 apresenta os grupos criados.

**Tabela 7 – Grupo de variáveis criadas para estudo de readmissões**

| Grupo | Variáveis |
|---|---|
| **Gerais (G)** | *salaemergencia* |
| | *tempointernamento* |
| | *pco2* |
| | *po2/fio2_ratio* |
| **Scores (S)** | *pco2_score* |
| | *po2/fio2_score* |
| | *salaemergencia_score* |
| | *tempointernamento_score* |
| **Resultados Laboratoriais (RL)** | *acido_lactico* |
| | *leucocitos* |
| **Case Mix (CM)** | *sexo* |
| | *idade* |

Esta divisão foi feita de modo a poder-se conjugar diferentes cenários. O Case Mix (CM) é composto por um conjunto de atributos não clínicos associados a um doente tratados por uma unidade ou hospital. A variável readmissão não foi considerada nestes grupos uma vez que será a classe *target* dos modelos de *clustering*. Considerando então estes quatro grupos criaram-se 13 cenários diferentes, sendo eles:

- Cenário 1 – {Gerais}
- Cenário 2 – {Gerais, CM}
- Cenário 3 – {Gerais, CM, Scores}
- Cenário 4 – {Gerais, CM, Scores, Resultados Laboratoriais}
- Cenário 5 – {Scores}
- Cenário 6 – {CM}
- Cenário 7 – {Resultados Laboratoriais}
- Cenário 8 – {Gerais, Scores}
- Cenário 9 – {Gerais, Resultados Laboratoriais}
- Cenário 10 – {Gerais, CM}
- Cenário 11 – {Scores, CM}
- Cenário 12 – {Scores, Resultados Laboratoriais}
- Cenário 13 – {CM, Resultados Laboratoriais}

Estes 13 cenários deram origem a noventa e um modelos, mas destes apenas trinta e nove foram analisados mais pormenorizadamente (os que utilizam os métodos *k-means, k-medoids* e *x-means*). A escolha dos modelos a analisar foi feita tendo em conta os resultados de reuniões com os intensivistas, onde se selecionou os modelos que apresentaram interesse do ponto de vista clínico. Como já referido acima os métodos de *clustering* escolhidos para os modelos desenvolvidos foram o *k-means, k-medoids* e *x-means* e foi necessário fornecer um conjunto de parâmetros para cada um destes métodos. A tabela 8 apresenta as configurações dada a cada um dos métodos.

**Tabela 8 – Configurações dos algoritmos de DM**

| Algoritmo | Definição | Valor |
|---|---|---|
| **k-means** | K | 2 a 11 |
| | Max Runs | 10 |
| | Max Optimization Steps | 100 |
| | Measures Type | Numerical Measures |
| | Numerical Measure | Euclidean Distance |
| **k-medoids** | K | 2 a 11 |
| | Max Runs | 10 |
| | Max Optimization Steps | 100 |
| | Measures Type | Numerical Measures |
| | Numerical Measure | Euclidean Distance |
| **x-means** | K Min | 2 |
| | K Max | 60 |
| | Measure Type | Numerical Measures |
| | Numerical Measure | Euclidean Distance |
| | Clustering Algorithm | *K-means* |
| | Max Runs | 10 |
| | Max Optimization Steps | 100 |

Tipicamente é necessário fornecer aos algoritmos de *clustering* o número de clusters que ele deverá criar. No caso do método *x-means* não foi necessário uma vez que o mesmo possuí mecanismos para calcular o número ótimo de *clusters* (*k*), mas em relação aos métodos *k-means* e *k-medoids* o mesmo já não acontece. Para estes dois métodos, e de modo a calcular o número ótimo de *clusters* optou-se por analisar o *Davies-Bouldin Index* pois quanto menor o valor deste índice menor é a separação entre clusters e maior é a *compactness* dentro de um cluster (Cios et al., 2007), assim como avaliar os resultados da aplicação do método de *Elbow* através da observação da variação da distância média *intra-cluster* (observando como a distancia varia de *k* para *k* e selecionando o *k* onde a progressão natural da

medição domina a estrutura). A tabela 9 apresenta o número ótimo de *clusters* para as técnicas *k-means* e *k-medoids*.

**Tabela 9 – Número ótimo de clusters para os algorimtos k-means e k-medoids**

| Modelo | Algoritmo | Número de Clusters | Índice Davies-Bouldin | Distância média dentro do *cluster* |
|---|---|---|---|---|
| **M1** | *k-means* | 4 | 0,563 | 519,619 |
| | *k-medoids* | 3 | 0,733 | 43317,973 |
| **M2** | *k-means* | 4 | 0,466 | 9284,935 |
| | *k-medoids* | 3 | 0,768 | 43629,478 |
| **M3** | *k-means* | 3 | 0,468 | 9332,818 |
| | *k-medoids* | 3 | 0,775 | 43696,924 |
| **M4** | *k-means* | 5 | 0.541 | 9087.479 |
| | *k-medoids* | 3 | 0.827 | 46243.621 |
| **M5** | *k-means* | 6 | 0.528 | 8.857 |
| | *k-medoids* | 3 | 0.722 | 30.028 |
| **M6** | *k-means* | 7 | 0.513 | 8.530 |
| | *k-medoids* | 3 | 0.593 | 63.581 |
| **M7** | *k-means* | 2 | 0.388 | 519.652 |
| | *k-medoids* | 9 | 1.496 | 703.440 |
| **M8** | *k-means* | 4 | 0.454 | 9086.529 |
| | *k-medoids* | 3 | 0.740 | 43385.713 |
| **M9** | *k-means* | 6 | 0.503 | 6749.155 |
| | *k-medoids* | 5 | 1.297 | 35345.738 |
| **M10** | *k-means* | 4 | 0.466 | 9284.935 |
| | *k-medoids* | 3 | 0.768 | 43629.478 |
| **M11** | *k-means* | 3 | 0.882 | 89.845 |
| | *k-medoids* | 2 | 0.971 | 194.594 |
| **M12** | *k-means* | 2 | 0.393 | 571.905 |
| | *k-medoids* | 2 | 3.028 | 2222.068 |
| **M13** | *k-means* | 2 | 0.402 | 773.690 |
| | *k-medoids* | 5 | 3.098 | 2552.022 |

### 3.1.5 Avaliação

A fase de avaliação focou-se essencialmente na análise dos resultados dos modelos que utilizaram os métodos *k-means*, *x-means* e *k-medoids* e confrontou os mesmos com os objetivos iniciais do estudo. A avaliação foi feita com recurso ao *Davies-Bouldin Index*. Dos métodos utilizados o que revelou os piores

resultados foi o *k-medoids*. A diferença entre os métodos *k-means* e *x-means* não foi expressiva, embora o *k-means* tenha obtido resultados um pouco superiores. Alguns dos modelos revelaram valores interessantes ao nível da avaliação do índice mas infelizmente não atingiram resultados favoráveis do ponto de vista do domínio do problema. A tabela 10 apresenta os três melhores modelos obtidos para este estudo.

**Tabela 10 – Melhores modelos obtidos para estudo de readmissões**

| Modelo | Algoritmo + Número de Clusters | Índice Davies-Bouldin | Clusters | Número de casos de readmissão |
|---|---|---|---|---|
| **M9 (G + RL)** | *k-means* com 6 clusters | 0,503 | C0 | 2 |
| | | | C1 e C4 | 0 |
| | | | C2 | 1 |
| | | | C3 | 16 |
| | | | C5 | 16 |
| **M4 (G + CM + S +RL)** | *k-means* com 5 clusters | 0,541 | C0 | 16 |
| | | | C1 | 2 |
| | | | C2 | 1 |
| | | | C3 | 0 |
| | | | C4 | 16 |
| **M1 (N)** | *k-means* com 6 clusters | 0,563 | C0 | 9 |
| | | | C1 | 2 |
| | | | C2, C4 e C5 | 0 |
| | | | C3 | 24 |

O valor do índice de Davies-Bouldin tende para +∞, no entanto para um *cluster* obter uma boa avaliação o valor deverá ser o mais próximo possível de 0. No caso deste estudo pode-se considerar que se obtiveram bons modelos uma vez que na maioria dos casos o índice está abaixo de 1 e nos melhores modelos o índice situa-se entre os 0,5 e 0,6, o que representa bons resultados ao nível da segmentação dos *clusters*.

Analisando em maior detalhe o modelo 9, que utilizou as variáveis gerais e os resultados laboratoriais, e observando as figuras 11 e 12, é possível observar que as características que influenciam mais os *clusters* de doentes readmitidos são o rácio *PaO2/FIO2* e o *PaCO2*. O rácio de PaO2/FIO2 no *cluster* 3 varia entre os 105 e 204 e no cluster 5 entre os 236 e os 388. Relativamente ao PaCO2 ele varia dos 31 aos 75 no *cluster* 3 e dos 36,5 aos 57 no cluster 5. O tempo de internamento dos doentes na UCI

não apresenta uma variação significativa entre os dois principais *clusters*. O ácido láctico dos doentes varia entre os 0.8 e os 16 no *cluster* 3 e entre os 0.4 e 7.7 no *cluster* 5. Os leucócitos variam de 6 a 27 no *cluster* 3 e de 3 a 59 no *cluster* 5. Todos os doentes pertencentes ao *cluster* 3 são provenientes de um serviço fora do CHP – HSA e os doentes pertencentes ao cluster 5 vem da sala de emergência ou de outro serviço do hospital.

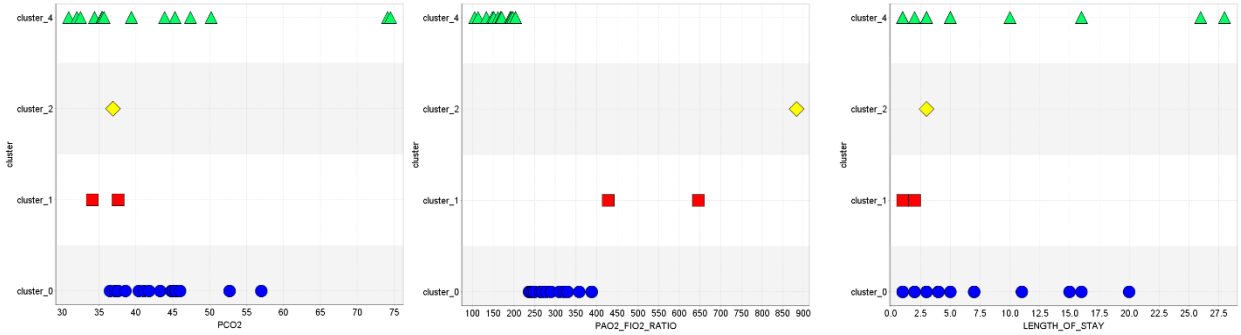


**Figura 11 - Distribuição dos valores do PaCo2, PaO2/FiO2 e tempo de internamento pelos clusters**

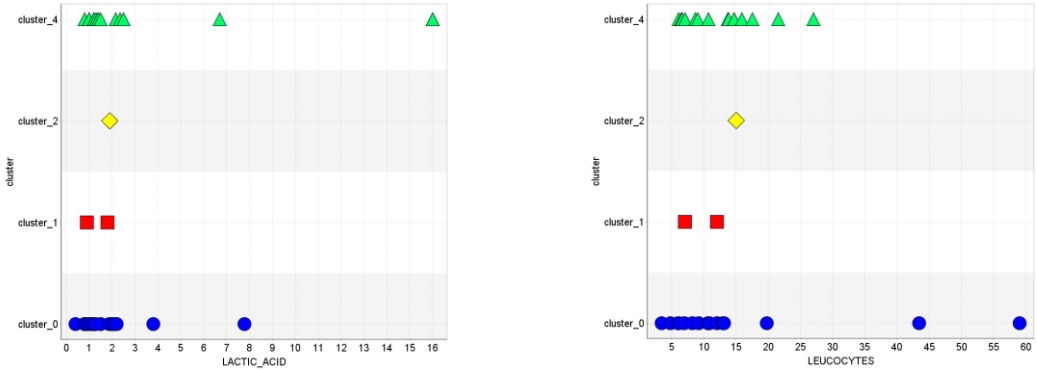


**Figura 12 – Distribuição dos valores dos leucócitos e ácido lactíco pelos clusters**

No modelo 4, que considera as variáveis gerais, *case mix, scores e resultados laboratoriais, os resultados do PaCO2, rácio PaO2/FIO2, leucócitos, acido láctico, proveniência de sala de emergência* e *tempo de internamento* os *clusters* são os mesmos que os do modelo anteriormente analisado, com a diferença de que agora o cluster 3 é representado pelo cluster 4 e o cluster 5 é representado pelo cluster 0. Os restantes atributos, nomeadamente, a idade, sexo e os quatro atributos relativos aos scores não são preponderantes nos resultados dos *clusters* uma vez que se encontram bem distribuídos pelos diferentes *clusters*, ou seja, não apresentam características de diferenciação.

No caso do modelo 1, que considera apenas as variáveis gerais, uma vez mais, a maioria dos casos centra-se em dois *clusters*, o 3 e o 0. A variável mais discriminatória é o rácio PaO2/FIO2, que no *cluster* 3 varia entre 105 e 268 e no cluster 0 entre 275 e 430. As outras variáveis utilizadas neste modelo apresentam resultados com poucas diferenças entre os diferentes *clusters*.

### 3.1.6 Discussão dos Resultados

Baseado nos dois melhores modelos, o 4 e 9, é possível criar dois *clusters* globais que representam melhor as características aquando da alta de um futuro doente readmissível. A tabela 11 apresenta os dois melhores clusters e os seus atributos e valores correspondentes.

**Tabela 11 – Grupos de caracterização de doentes passíveis de readmissão**

| Variável | Cluster 1 | Cluster 2 |
|---|---|---|
| **Sala Emergência** | Doentes que vieram de uma ala fora do hospital | Doentes que vieram de uma alta dentro do hospital e da sala de emergência do mesmo |
| **Rácio PaO2/FIO2** | 105 a 204 | 236 a 388 |
| **PaCO2** | 31 a 75 | 36,5 a 57 |
| **Tempo de Internamento** | 1 a 28 dias | 1 a 20 dias |
| **Ácido Lactico** | 0,8 a 16 | 0,4 a 7,7 |
| **Leucócitos** | 6 a 27 | 3 a 59 |
| **Idade** | 24 a 88 | 50 a 93 |

As variáveis relativas aos scores e o sexo não apresentam impacto na caracterização de grupos de possíveis readmitidos. Os scores porque são uma caracterização direta do grupo de variáveis gerais usando a pontuação atribuída pelo modelo SWIFT e o sexo pois os dados estão bem distribuídos entre número de homens e mulheres.

### 3.1.7 Conclusões

Este estudo forneceu resultados úteis na ajuda da caracterização de doentes que tenham uma maior probabilidade de vir a ser readmitidos. Os *clusters* desenvolvidos não podem assegurar quais os doentes que serão readmitidos mas dão informação sobre o tipo de doentes que os intensivistas deverão considerar como mais suscetíveis a uma readmissão não planeada depois de terem dado alta ao doente.

Os atributos que demonstraram estar mais relacionados com as readmissões na UCI são o rácio PaO2/FIO2, o PaCO2, a proveniência do doente e o tempo de internamento, como era expectável pois são os principais atributos do modelo SWIFT. A análise dos resultados do ácido láctico e leucócitos (variáveis sugeridas pelos intensivistas) contribuíram para uma segmentação dos grupos de readmitidos. No caso do grupo de variáveis Case Mix, apenas a idade mostrou ter um contributo na caracterização dos grupos de doentes com possível readmissão. Como trabalho futuro dever-se-á estudar em maior detalhe as características dos doentes readmissíveis e não readmissíveis para encontrar novas variáveis que possam influenciar estes grupos e também explorar mais pormenorizadamente o impacto que os leucócitos e o ácido láctico apresentam na readmissão de doentes na UCI. Estes modelos tem a capacidade de se otimizarem e adaptarem no tempo de acordo com os novos dados que forem chegando. Como trabalho futuro dever-se-ão criar *ensemble clusters* que permitam assegurar sempre o melhor modelo e a criação de um *treshold* que ajude a validar os modelos.

## 3.2 Previsão do Tempo de Internamento de Doentes em UCI

Como já referido na abordagem conceptual os custos associados às UCI são bastante elevados, sendo que devido à grande variedade de terapias executadas e medicamentos administrados ao doente, quanto maior o tempo de internamento de um doente maior o seu custo para o hospital. Uma previsão correta do tempo de internamento de um doente pode contribuir para uma redução de custos. É reconhecido pelos intensivistas que a existência de uma ferramenta que permita prever o tempo de internamento é importante pois a mesma pode ajudar a tomar decisões na gestão de recursos e a melhor entender a condição do doente, evitando assim altas antecipadas ou erradas.

Com este estudo pretende-se aferir a probabilidade de alta e o tempo de internamento de doentes que se encontram na UCI. De modo a atingir os objetivos para este estudo foram desenvolvidas três abordagens diferentes. A primeira abordagem, de agora em diante abordagem A, utilizou dados dos doentes recolhidos durante as primeiras vinte e quatro horas em que o doente esteve na UCI (como por exemplo dados de admissão e os piores resultados dos sinais vitais). Com esta abordagem pretendia-se prever altas de doentes em UCI. Esta abordagem utilizou um conjunto de dados base que são utilizados em diversos estudos na área da medicina mas revelou-se insuficiente pois os resultados atingidos nesta área específica foram medíocres.

Na segunda abordagem, denominada abordagem B, induziram-se modelos de Data Mining (DM) para prever o tempo de internamento do doente através da previsão da hora em que o doente vai ter alta. Utilizando este método obteve-se o tempo de internamento do doente encontrando a hora em que o doente tem uma maior probabilidade de ter alta. Esta abordagem utilizou um conjunto de dados recolhidos em tempo real e adquiridos de diferentes fontes. Os modelos induzidos nesta abordagem apresentaram melhores resultados. Os resultados obtidos são válidos tanto do ponto de vista de DM como clínico.

Embora os modelos induzidos na abordagem B tenham apresentado bons resultados, foi ainda desenvolvida uma terceira abordagem, abordagem C, que foi criada com o intuito de melhorar os resultados da abordagem B. Esta nova abordagem utiliza na mesma os dados adquiridos em tempo real e prevê o tempo de internamento do doente através da previsão da probabilidade de o doente ter alta na hora seguinte. Os dados utilizados foram os mesmos da abordagem B aos quais foram adicionados um conjunto de novos dados derivados da categorização e agrupamento de alguns dos dados já utilizados. Esta nova abordagem produziu resultados ainda mais motivadores do ponto de vista clínico, pelo que seguidamente se vai fazer a descrição do estudo e das abordagens utilizadas.

### 3.2.1 Compreensão do Negócio

A previsão do tempo de internamento em uma UCI assume grande importância como já foi referido anteriormente quer seja pela otimização de custos e recursos como por exemplo a gestão de camas. Neste tipo de unidades existe um grande número de variáveis clínicas que podem ser utilizados na previsão do tempo de internamento de um doente. Utilizando dados fornecidos pelo CHP, dados da admissão dos doentes e recolhidos durante as primeiras vinte e quatro horas de internamento, foram criadas três abordagens e induzidos diferentes modelos de DM com o objetivo de prever o tempo de internamento e a probabilidade de alta clínica.

### 3.2.2 Compreensão e Preparação dos Dados

#### Abordagem A

A abordagem A considera dados provenientes de três fontes distintas, sendo elas, o registo eletrónico de enfermagem, resultados laboratoriais e a monitorização de sinais vitais. Esta abordagem considera as

variáveis do grupo Case Mix que dizem respeito ao doente e são um conjunto de variáveis que estão presentes no Processo Clínico Eletrónico (PCE). Neste grupo englobam-se as variáveis relativas à idade, tipo de admissão, local de proveniência, admissão com cirurgia, Escala de Coma de Glasgow, presença de doenças hematológicas, transplantação de órgãos, tratamentos para diabetes com insulina e sem insulina, medicação para a hipertensão, horas e dias de internamento. Este conjunto de variáveis é recolhido no momento da admissão do doente na UCI. Além destas foram também selecionadas um outro conjunto de variáveis, relativas a dados fisiológicos do doente que foram recolhidos durante as primeiras 24 horas (os piores valores). Neste grupo encontram-se as variáveis relativas à saturação de oxigénio (SPO2), pressão arterial sistólica e diastólica, temperatura corporal e medição da frequência cardíaca. A tabela 12 permite observar melhor as variáveis utilizadas na abordagem assim como o valor que podem tomar e algumas estatísticas das mesmas (valor mínimo e máximo das variáveis numéricas). Interpretando a tabela é possível ver que na idade é indicada a idade em anos do doente, o sexo pode ser masculino ou feminino, uma admissão pode ser urgente ou programada, é explicitado se o doente foi admitido devido a uma cirurgia, a área de onde o doente proveio, a escala de coma de *Glasgow* em que o valor *3* representa um coma total e o valor *15* significa que o doente está totalmente desperto, o número de dias e horas de internamento, os valores das suas variáveis fisiológicas e um conjunto de verificações clínicas do doente (como se possuí alguma doença hematológica, se é transplantado entre outras).

**Tabela 12 – Variáveis utilizadas na abordagem A**

| Identificação | Variável | Mínimo | Máximo | Valores |
|---|---|---|---|---|
| **Idade** | Idade | 20 | 96 | Real |
| **Sexo** | Sexo | M | F | 1 ou 2 |
| **Tipo de Admissão** | Urgente | - | - | U |
| | Programada | - | - | P |
| **Admissão com Cirurgia** | Sim | - | - | C |
| | Não | - | - | S |
| **Local de proveniência** | Cirurgia | - | - | 1 |
| | Observação | - | - | 2 |
| | Emergência | - | - | 3 |
| | Sala de Enfermagem | - | - | 4 |
| | Outra UCI | - | - | 5 |
| | Outro Hospital | - | - | 6 |
| | Outra | - | - | 7 |

| | | | | |
|---|---|---|---|---|
| **Escala de coma de Glasgow** | Escala de Coma de Glasgow | 3 | 15 | Real |
| **Doença Hematológica** | Sim | - | - | S |
| | Não | - | - | N |
| **Transplantado** | Sim | - | - | S |
| | Não | - | - | N |
| **Tratamento à Hipertensão** | Sim | - | - | S |
| | Não | - | - | N |
| **Diabetes Tratada com Insulina** | Sim | - | - | S |
| | Não | - | - | N |
| **Diabetes Tratada sem insulina** | Sim | - | - | S |
| | Não | - | - | N |
| **Dias de Internamento** | Dias de Internamento | 0 | 40 | Real |
| **Horas de Internamento** | Horas de Internamento | 2 | 970 | Real |
| **Saturação de Oxigénio** | SPO2 | 86 | 100.00 | Vírgula Flutuante |
| **Pressão Arterial Sistólica** | PAS | 80.64 | 219.48 | Vírgula Flutuante |
| **Identificação** | Variável | Mínimo | Máximo | Valores |
| **Pressão Arterial Diastólica** | PAD | 43.20 | 179.88 | Vírgula Flutuante |
| **Temperatura Corporal** | TEMP | 34.16 | 39.74 | Vírgula Flutuante |
| **Medição da Frequência Cardíaca** | FC | 50.00 | 231.00 | Vírgula Flutuante |

Nesta abordagem foram criadas classes para algumas das variáveis nomeadamente dias e horas (dois conjuntos de classes) e *FC, SPO2, PAD e PAS*. No caso das variáveis horas e dias e uma vez que irão ser a variável *target* dos modelos foram criados dois conjuntos de classes, uma delas utilizando a regra de *Sturges* e outra utilizando classes criadas através do ponto de vista clínico, após concordância da equipa de intensivistas associada ao estudo. No caso das variáveis *FC, SPO2, PAD, PAS e TEMP* apenas foi utilizada a regra de *Sturges*. A regra de *Sturges* é uma regra utilizada para determinar o número de classes em que uma distribuição ou observação deve ser classificada (*Sturges*, 1926). O número de intervalos (k) de classe de uma amostra com n elementos calcula-se a partir da expressão:

$$k \approx 1 + 3{,}322 \times \log_{10} n$$

Após calcular o número de classes é necessário calcular a amplitude (h) de cada uma delas, que é expresso pela forma:

$$h = \frac{A}{k} \; em\ que\ A = [Valor\ máximo\ amostra] - [valor\ mínimo\ amostra]$$

A tabela 13 apresenta os intervalos de classe criados para a variável horas e dias tendo em conta o conhecimento de domínio (clínico). As classes apresentadas na tabela são as classes utilizadas como *target* para os modelos criados.

**Tabela 13 – Intervalos criados para o *target* dos modelos**

| Intervalo | Classes para variável horas | | Classes para variável dias | |
|---|---|---|---|---|
| | Regra de Sturges | Domínio Clínico | Regra de Sturges | Domínio Clínico |
| **1** | [2, 100.67] | [0, 72] | [0, 4.4] | [0, 3] |
| **2** | ] 100.67, 199.39] | ] 72, 120] | ] 4.4, 8.8] | ] 3, 5] |
| **3** | ] 199.39, 297.98] | ] 120, +∞[ | ] 8.8, 13.2] | ] 5, +∞[ |
| **4** | ] 297.98, 396.65] | - | ] 13.2, 17.6] | - |
| **5** | ] 396.65, 495.32] | - | ] 17.6, 22] | - |
| **6** | ] 495.32, 593.99] | - | ] 22, 26.4] | - |
| **7** | ] 593.99, 692.66] | - | ] 26.4, 30.8] | - |
| **8** | ] 791.33, 890] | - | ] 30.8, 35.2] | - |
| **9** | ] 890, 970] | - | ] 35.2, 40] | - |

Abordagem B

Este estudo considerou três fontes de dados sendo elas: registo eletrónico de enfermagem, monitores de cama e dados laboratoriais. Neste estudo a preparação dos dados é feita automaticamente por um agente inteligente de pré-processamento que é responsável por validar todos os dados recolhidos, preparar a tabela de entrada de DM (DMIT), validar as fontes de dados e processar as variáveis ao longo de todo o dia. Relativamente à preparação da DMIT esta tarefa consiste na criação da estrutura de *input* para os modelos de DM para cada um dos doentes. De modo a todos os dados estarem preparados para a DMIT o sistema executa um conjunto de transformações nos dados que incluem a categorização,

agrupamento e normalização dos valores originais das variáveis. As variáveis relativas ao Case Mix (CM), SOFA e os Eventos Críticos Acumulados (ECA) são preparadas e inseridas na DMIT. Todos os valores são considerados de acordo com um intervalo máximo e mínimo. No caso de os valores serem números reais, foram criados intervalos de acordo com a importância e normalidade do valor no âmbito das UCI.

A primeira fase de transformação que ocorre nos dados é uma tarefa analítica onde os valores recolhidos são transformados de acordo com um conjunto de regras criadas. Nas variáveis CM é utilizado apenas um valor para representar um caso. Para as variáveis SOFA os modelos de DM são baseados no pior valor recebido em cada hora. A tabela 14 apresenta as transformações efetuadas nos dados pelo agente assim como as variáveis selecionadas e os intervalos de referência para cada conjunto de dados criados (ex. um doente que tenha 45 anos pertence à classe de idade 1). De referir que no caso dos atributos Case Mix apenas é usado um valor para representar cada caso e para os atributos do SOFA é considerado o pior valor recebido por cada hora.

**Tabela 14 – Transformações efetuados pelo agente de pré-processamento**

| Identificação | | Variável | Min | Max | Valor |
|---|---|---|---|---|---|
| **Idade** | | Idade | 18 | 46 | 1 |
| | | | 47 | 65 | 2 |
| | | | 66 | 75 | 3 |
| | | | 76 | 130 | 4 |
| **Tipo de Admissão** | | Urgente | - | - | U |
| | | Programada | - | - | P |
| **Local de proveniência** | | Cirurgia | - | - | 1 |
| | | Observação | - | - | 2 |
| | | Emergência | - | - | 3 |
| | | Sala de Enfermagem | - | - | 4 |
| | | Outra UCI | - | - | 5 |
| | | Outro Hospital | - | - | 6 |
| | | Outra | - | - | 7 |
| **SOFA** | **Cardiovascular** | PA (média) | 0 | 70 | 1 |
| | | Dopamina | 0.0 | - | 1 |
| | **Renal** | Dobutamina | 0.0 | - | 1 |
| | | Adrenalina/Noradrenalina | 0.0 | - | 1 |
| | **Respiratório** | Creatinina | 1.2 | - | 1 |
| | | Po2/Fio2 | 0 | 400 | 1 |
| | **Hepático** | Bilirrubina | 1.2 | - | 1 |
| | **Hematológico** | Plaquetas | 0 | 150 | 1 |
| **ACE** | | | 0 | +∞ | SET |
| **R1** | | | 0 | 1 | SET |
| **R2** | | | 0 | 1 | SET |

A segunda fase de transformação dos dados utiliza os eventos críticos (EC). Os valores são categorizados como normais, com o valor 0 ou como críticos, valor 1, respeitando os intervalos de normalidade definidos pela UCI (Silva et al., 2008). Se o valor se encontra dentro do intervalo é classificado como

normal, quando não está é classificado como crítico. O próximo passo consiste em determinar a tipologia do evento crítico. Existem dois tipos diferentes, um que corresponde aos valores que estão fora do intervalo por um determinado período tempo e outro a que se referem a valores críticos independentemente da duração do evento.

Nesta etapa foram utilizadas técnicas para categorizar e agrupar valores utilizando intervalos mínimos e máximos. Foram definidos conjuntos tendo em conta duas grandes características, a média e o valor mais alto de todos os dados recolhidos. Existe depois um conjunto de níveis de severidade que foram criados a partir do *Clinical Global Impression – Severity Scale* (Guy, 2000). Esta escala considera valores de 0 a 7 onde o 0 significa que não existe um problema grave e o 7 que existe um problema de nível extremo. A maioria dos casos encontra-se entre 0 e 5 enquanto os casos mais graves estão nos níveis 6 e 7. A tabela 15 representa as regras definidas na categorização e agrupamento de cada um dos valores contínuos. No topo da tabela tem-se os diferentes níveis e à esquerda as variáveis sob as quais foi executado o trabalho. Interpretando a tabela por exemplo no conjunto 1 relativamente ao R1 de PA considera-se que a percentagem de ECA se situa entre os 0 e os 10% do valor máximo de ECA ocorridos para um doente e para a variável em análise (por exemplo, se o valor máximo de ECA PA verificado num doente na hora 6 é 10, um doente que nessa hora tiver um ECA de 3 o R1 PA seria 1). A mesma interpretação deverá ser feita para os restantes níveis e variáveis.

**Tabela 15 – Regras para categorização e agrupamento**

| Conjunto | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|---|
| R1 PA | Min | -0,1 | 0,000 | 0,010 | 0,021 | 0,041 | 0,062 | 0,082 | 0,123 |
| | Max | 0,000 | 0,010 | 0,021 | 0,041 | 0,062 | 0,082 | 0,123 | 2,000 |
| R1 O2 | Min | -0,1 | 0,000 | 0,018 | 0,036 | 0,072 | 0,108 | 0,144 | 0,216 |
| | Max | 0,000 | 0,018 | 0,036 | 0,072 | 0,108 | 0,144 | 0,216 | 2,000 |
| R1 HR | Min | -0,1 | 0,000 | 0,004 | 0,008 | 0,015 | 0,023 | 0,030 | 0,045 |
| | Max | 0,000 | 0,004 | 0,008 | 0,015 | 0,023 | 0,030 | 0,045 | 2,000 |
| R1 TOT | Min | -0,1 | 0,000 | 0,020 | 0,041 | 0,081 | 0,122 | 0,162 | 0,243 |
| | Max | 0,000 | 0,020 | 0,041 | 0,081 | 0,122 | 0,162 | 0,243 | 2,000 |
| R2 | Min | -0,1 | 0,000 | 0,100 | 0,250 | 0,500 | 0,750 | 0,900 | 1,000 |
| | Max | 0 | 0,100 | 0,250 | 0,500 | 0,750 | 0,900 | 1,000 | 2,000 |
| ACE | Min | -0,1 | 0 | 3 | 5 | 8 | 10 | 12 | 15 |
| | Max | 0 | 3 | 5 | 8 | 10 | 12 | 15 | 50 |

Resumidamente a primeira tarefa consiste na validação dos dados pelo sistema onde o mesmo verifica se os valores recolhidos estão dentro dos intervalos e se estão associados ao doente correto.

Na segunda tarefa, é criada a DMIT. Esta tabela é composta por 120 linhas (uma por cada hora) por doente e existe uma coluna por cada variável de DM e uma outra para o *target*. A coluna *target*, que vai corresponder à previsão de alta é preenchida hora a hora. Se o resultado for verdadeiro o agente insere o valor 0 na coluna, caso contrário ele insere o valor 1 até à última linha, correspondendo a um total de 120 horas. Por exemplo, imagine-se que um doente tem alta após 70 horas da admissão, o agente coloca o valor 0 em todas as horas entre 1 e 70 e para as restantes (71 a 120) o agente coloca o valor 1. O valor 1 representa que o doente não está hospitalizado na respetiva hora. De referir que o facto de se ter considerado apenas 120 horas, correspondente a 5 dias, foi acordado com a equipa clínica pois os intensivistas afirmam que doentes internados há mais de 5 dias trazem problemas e infeções que alteram facilmente o seu estado clínico, sendo mais difícil criar padrões de DM.

A terceira tarefa tem em conta as variáveis usadas no DM e preenche a tabela DMIT com os seguintes dados:

- Case Mix (CM) – Idade, tipo de admissão, local de proveniência;
- SOFA Scores – Cardiovascular, Respiratório, Renal, Hepático e Hematológico;
- Eventos críticos Acumulados (ECA) – ECA da pressão arterial (PA), ECA da saturação de oxigénio (SPO2), ECA da frequência cardíaca (FC) e ECA totais;
- Rácios 1 (R1) – ECA de PA / tempo de internamento atual, ECA de SPO2 / tempo de internamento atual, ECA de FC / tempo de internamento atual;
- Rácios 2 (R2) – ECA de FC / número máximo de ECA de FC, ECA de SPO2 / número máximo de ECA de SPO2, ECA de FC / número máximo de ECA de FC, ECA Total / número máximo de ECA totais;
- Rácios (R) – união entre os dois conjuntos de rácios (R1 e R2).

Nestes dados considera-se o ECA como o número de eventos críticos acumulados por doente durante a admissão por hora e tipo, o tempo de internamento atual é o número de horas desde que o doente foi admitido e o número máximo de ECA é o número máximo de ECA verificados por um doente numa

determinada hora. De referir que os valores são atualizados automaticamente quando um valor pior é recolhido.

### Abordagem C

Sendo esta abordagem uma evolução da abordagem B, a fase de compreensão e preparação dos dados foi bastante idêntica à anterior, como tal existe o agente de pré-processamento já abordado e que vai ser responsável pela validação de todos os dados recolhidos e pela preparação dos dados para DMIT.

Na abordagem C à semelhança do referido na abordagem A e B as variáveis Case Mix (CM) consistem em variáveis que estão presentes no PCE. Nesta abordagem foram selecionadas as variáveis idade, tipo de admissão e local de proveniência. Estas variáveis foram obtidas na admissão dos doentes e são automaticamente transformadas de acordo com os atributos de DM.

Foram selecionadas variáveis relativas ao SOFA, que como referido na abordagem conceptual é um score utilizado na UCI para avaliar o grau de disfunção ou falha de sistemas órgãos. Estas variáveis variam de 0 a 4 sendo que o 0 representa uma normal função do sistema e o nível 4 a falência total do órgão. Neste caso, foram derivados os dados iniciais relativos aos resultados do SOFA e apenas foram considerados dois valores possíveis sendo o 0, para quando o SOFA é 0 e 1 quando o SOFA é maior do que 0. Para cada um dos sistemas foi considerado um conjunto de dados que permite avaliar o score individual de cada sistema orgânico. Sendo assim, para o caso do sistema cardiovascular foram considerados a pressão arterial média, e os valores de dopamina, no caso do sistema renal, valores de dobutamina, adrenalina e noradrelina, no sistema respiratório foram considerados os valores de creatinina e o rácio PaO2/Fio2, no sistema hepático a bilirrubina e no sistema hematológico as plaquetas sanguíneas.

Os eventos críticos acumulados (ECA) foram criados como complemento ao SOFA pois o mesmo não tem uma capacidade preditiva. Cada ECA representa o número de acontecimentos críticos que aconteceu referentes a três variáveis fisiológicas, a pressão arterial sistólica (PAS), a saturação de oxigénio (SPO2) e a Frequência cardíaca (FC). É obtido através da soma dos valores dos eventos críticos por hora.

Em consequência da criação do grupo ECA foram criados os grupos de rácios que permitem relacionar os eventos acumulados com o número máximo e com as horas. Estes rácios permitem determinar o número de ECA por hora (R1) e a correspondência entre o número de ECA e o número máximo de eventos verificados no passado, agrupados por categoria e por doente (R2).

As transformações efetuados pelo agente de pré-processamento poderão ser observadas na abordagem anterior, uma vez que esta abordagem utiliza o mesmo agente de pré-processamento da abordagem B.

Os valores dos Rácios foram categorizados e agrupados considerando um intervalo de valores mínimo e máximo. Os conjuntos criados foram definidos considerando a média e o valor mais alto dos dados recolhidos. Os intervalos de valores foram criados considerando o *Clinical Global Impression – Severity Scale* (CGI-S) (Guy, 2000). O critério usado para definir as percentagens concentra a maior parte dos valores dos doentes no conjunto de 0 a 5. Os níveis 6 e 7 são atribuídos aos casos mais severos. A tabela 15 apresentada na abordagem B permite observar a categorização para esta abordagem. A interpretação da tabela será também a mesma.

Uma das novidades desta abordagem é o facto da tabela 15 ter dado origem a um conjunto de classes cujo a nomenclatura é encontrada através da adição do prefixo *C_*, por exemplo para a classe relativa à pressão arterial sistólica dos rácios 2, representa-se a nova classe como *C_PA_Max*. As classes dos R1 são determinadas pelo número de registos *R1 PA Min* até *R1 Tot Max* e os atributos R2 (PA, SPO2, FC e Total) seguem a mesma regra, por exemplo para o primeiro nível o intervalo é de 0.00 (0%) a 0.10 (10%).

Sendo assim, como entrada para os modelos de DM vamos ter um conjunto de variáveis que é representado pela tabela 16. A mesma apresenta as estatísticas para as variáveis considerando o mínimo (Min), máximo (Max), média (Med), desvio-padrão (DP) e coeficiente de variação. Por exemplo no caso da variável *PA_Max* ela varia entre o valor 0 e 4,40 sendo que a média do valor é de 0,083 o desvio-padrão 0,230 (permite saber o quanto os dados estão desviados da média) e um coeficiente de variação de 276,948% (o que significa que esta variável é bastante heterogénea).

**Tabela 16 – Distribuição estatística das variáveis da abordagem C**

| Variável | Min | Max | Med | DP | CV (%) |
|---|---|---|---|---|---|
| **Hora** | 0 | 120 | 58.500 | 34.487 | 58.952 |
| **Idade** | 1 | 4 | 2.560 | 1.051 | 41.055 |
| **Local de proveniência** | 1 | 7 | 2.920 | 2.189 | 74.966 |
| **Respiratório** | 0 | 1 | 0.620 | 0.486 | 78.387 |
| **Hematológico** | 0 | 1 | 0.390 | 0.488 | 125.128 |
| **Renal** | 0 | 1 | 0.200 | 0.398 | 199.000 |
| **Hepático** | 0 | 1 | 0.180 | 0.381 | 211.667 |
| **Cardiovascular** | 0 | 1 | 0.620 | 0.486 | 78.387 |
| **PA** | 0 | 42 | 0.430 | 1.054 | 245.070 |
| **PA_Max** | 0 | 4.400 | 0.083 | 0.230 | 276.948 |
| **PA_Hora** | 0 | 0.500 | 0.008 | 0.025 | 331.631 |
| **FC** | 0 | 16 | 0.500 | 1.474 | 294.800 |
| **FC_Max** | 0 | 1 | 0.051 | 0.140 | 275.827 |
| **FC_Hora** | 0 | 0.500 | 0.008 | 0.025 | 305.664 |
| **SPO2** | 0 | 42 | 1.030 | 2.872 | 278.835 |
| **SPO2_Max** | 0 | 1 | 0.034 | 0.091 | 267.398 |
| **SPO2_Hora** | 0 | 1 | 0.021 | 0.055 | 267.828 |
| **Total_ECA** | 0 | 50 | 1.970 | 4.340 | 220.305 |
| **Total_Max** | 0 | 1.170 | 0.052 | 0.108 | 207.934 |
| **Total_Hora** | 0 | 1 | 0.0370 | 0.077 | 207.751 |
| **C_PA** | 0 | 7 | 0.230 | 0.644 | 280.000 |
| **C_PA_Max** | 0 | 6 | 0.660 | 1.271 | 192.576 |
| **C_PA_Hora** | 0 | 7 | 0.560 | 1.391 | 248.393 |
| **C_FC** | 0 | 7 | 0.260 | 0.672 | 258.462 |
| **C_FC_Max** | 0 | 6 | 0.430 | 1.028 | 239.070 |
| **C_FC_Hora** | 0 | 7 | 0.870 | 2.012 | 231.264 |
| **C_SPO2** | 0 | 7 | 0.490 | 1.080 | 220.408 |
| **C_SPO2_Max** | 0 | 6 | 0.430 | 0.815 | 189.535 |
| **C_SPO2_Hour** | 0 | 7 | 0.900 | 1.710 | 190.000 |
| **C_Total_ECA** | 0 | 7 | 0.860 | 1.509 | 175.465 |
| **C_Total_Max** | 0 | 6 | 0.640 | 0.921 | 143.906 |
| **C_Total_Hora** | 0 | 7 | 1.360 | 1.982 | 145.735 |
| **Alta** | 0 | 1 | 0.790 | 0.408 | 51.646 |

A tabela 17 apresenta a distribuição das classes (em pontos percentuais) para cada uma das variáveis independentes.

**Tabela 17 – Distribuição das variáveis independentes**

| Variável | Valores | Percentagem |
|---|---|---|
| **Tipo de Admissão** | P | 23.5% |
| | U | 76.5% |
| **Idade** | 1 | 16.9% |
| | 2 | 36.4% |
| | 3 | 20.8% |
| | 4 | 25.9% |
| **Respiratório** | 0 | 38.4% |
| | 1 | 61.6% |
| **Hematológico** | 0 | 61.0% |
| | 1 | 39.0% |
| **Renal** | 0 | 80.3% |
| | 1 | 19.7% |
| **Hepático** | 0 | 82.4% |
| | 1 | 17.6% |
| **Cardiovascular** | 0 | 38.4% |
| | 1 | 61.6% |
| **Variável** | Valores | Percentagem |
| **Local de Proveniência** | 1 | 47.3% |
| | 2 | 0.3% |
| | 3 | 17.7% |
| | 4 | 14.9% |
| | 5 | 2.2% |
| | 6 | 2.0% |
| | 7 | 15.5% |

Relativamente às variáveis que serão introduzidas na DMIT elas serão exatamente as mesmas que as que foram utilizadas na abordagem B, com o acréscimo das classes de rácios e ECA.

### 3.2.3 Modelação

#### Abordagem A

Nesta abordagem foram criados 24 modelos obtidos através da combinação de dois cenários, quatro *targets* e três técnicas. Os dados usados nestes modelos correspondem a admissões e altas na UCI do CHP – HSA desde 18 de Agosto de 2011 até 8 de Fevereiro de 2014 (905 dias), a 407 doentes e 448

registos. Os dois cenários utilizam as mesmas variáveis com a diferença que no cenário 1 as variáveis utilizadas são as classes criadas para o *SPO2, PAS, PAD, FC e TEMP*, enquanto o cenário 2 utiliza os valores reais das mesmas variáveis. O grupo de variáveis Case Mix mantem-se igual em ambos os cenários.

- Cenário 1 – {CM, Classe FC, Classe SPO2, Classe PAS, Classe PAD, Classe Temp}
- Cenário 2 – {CM, FC, SPO2, PAS, PAD, TEMP}

Os 4 *targets* a utilizar nos modelos são as classes relativas às horas e dias de tempo de internamento. Consideraram-se portanto:

- *Target* 1 – Classe Horas Regra Sturgeon
- *Target* 2 – Classe Dias Regra Sturgeon
- *Target* 3 – Classe Horas Domínio Clínico
- *Target* 4 – Classe Dias Domínio Clínico

As técnicas a utilizar serão:

- Técnica 1 – Support Vector Machines
- Técnica 2 – Árvores de Decisão
- Técnica 3 – Naïve Bayes

De modo a avaliar os modelos criados, foram utilizados 70% dos dados para treino e 30% para teste (*hold-out sampling*).

### Abordagem B

Na abordagem B foram induzidos 21 modelos, que utilizaram dados em tempo real. O conjunto de dados originais foi dividido em dois através do método *hold-out sampling*, onde 70% dos dados foram considerados para treino e 30% para testes. Os dados utilizados corresponderam a admissões/altas na UCI do CHP – HSA entre 1 de Fevereiro de 2012 e 12 de Julho de 2013, o que corresponde a um período de 527 dias, com um número de doentes de 249 e um número de registos de 21886. Considerou-se para cada doente uma escala temporal de 1 a n (com n menor ou igual a 120).

Os cenários utilizados foram os seguintes:

- Cenário 1 – {CM}
- Cenário 2 – {CM, ECA, R}
- Cenário 3 – {CM, ECA, R1}
- Cenário 4 – {CM, ECA, SOFA}
- Cenário 5 – {CM, ECA, SOFA, R}
- Cenário 6 – {CM, ECA, SOFA, R1}
- Cenário 7 – {CM, ECA, SOFA, R2}

Como *target* foi utilizado o atributo relativo à alta:

- *Target* – Alta

E foram utilizadas as técnicas:

- Técnica 1 – *Support Vector Machines*
- Técnica 2 – Árvores de Decisão
- Técnica 3 – *Naïve Bayes*

Cada um destes modelos foi induzido automaticamente e em tempo real utilizando *streaming* de dados. O motor de DM utilizou os dados presentes na tabela DMIT. Estes dados podem ser representados pelo seguinte *tuplo*:

> DMIT=<*pid, data, hora, V_eca_PA, V_ecaPA_tempo0, V_ ecaPA _max, V_eca_FC, V_ecaFC_tempo, V_ ecaFC _max, V_eca_spo2, V_eca_spo2_tempo, V_eca_spo2_max, V_total_eca, V_total_eca_tempo, V_total_eca_max, V_idade, V_local_proveniencia, V_tipo_admissão, V_sofa_cardio, V_sofa_resp, V_sofa_renal, V_sofa_coag, V_sofa_hepa*>

Neste tuplo o *pid* representa a identificação do doente, a *data* a data dos valores, *hora* representa o número de horas que passou desde que o doente foi admitido na UCI e a hora corrente e os *tuplos* com o prefixo V representam os atributos do doente que podem ser utilizados pelos modelos.

A tabela 18 apresenta as configurações utilizadas para cada uma das técnicas, onde para cada um dos parâmetros se indica se foi utilizado o valor pré-definido (pré-definido) pela ferramenta ou um valor dado pelo utilizador (inserido).

**Tabela 18 – Configurações dos algoritmos de DM para previsão de tempo de internamento**

| Técnica | Configuração | Valor | Tipo |
|---|---|---|---|
| AD | Minrec Node | 10 | Inserido |
| | Max Depth | 7 | Inserido |
| | Minpct Split | 0,1 | Inserido |
| | Impurity Metric | Gini | Inserido |
| | Minrec Split | 20 | Inserido |
| | Minpct Node | 0,05 | Inserido |
| | Prep Auto | On | Inserido |
| NB | Pairwise Threshold | 0 | Inserido |
| | Singleton Threshold | 0 | Inserido |
| SVM | Conv Tolerance | 0,001 | Inserido |
| | Active Learning | Enable | Inserido |
| | Kernel Function | Linear | Pré-definido |
| | Complexity Factor | 0,142831 | Pré-definido |
| | Prep auto | On | Inserido |

### Abordagem C

Para a abordagem C foram criados 39 modelos. Estes 39 modelos combinam 13 diferentes cenários, 1 *target* e 3 técnicas. Os dados utilizados nestes modelos correspondem a admissões e altas efetuadas num intervalo de tempo desde 1 de Fevereiro de 2012 até 24 de Abril de 2014 (813 dias), considerando 526 doentes e 55442 registos. Os cenários utilizados foram os seguintes:

- Cenário 1 – {CM}
- Cenário 2 – {CM, ECA, R}
- Cenário 3 – {CM, ECA, R1}
- Cenário 4 – {CM, ECA, SOFA}
- Cenário 5 – {CM, ECA, SOFA, R}
- Cenário 6 – {CM, ECA, SOFA, R1}
- Cenário 7 – {CM, ECA, SOFA, R2}
- Cenário 8 – {CM, Conjunto ECA, Conjunto R}
- Cenário 9 – {CM, Conjunto ECA, Conjunto R1}
- Cenário 10 – {CM, SOFA, Conjunto ECA}
- Cenário 11 – {CM, SOFA, Conjunto ECA, Conjunto R}

- Cenário 12 – {CM, SOFA, Conjunto ECA, Conjunto R1}
- Cenário 13 – {CM, SOFA, Conjunto ECA, Conjunto R2}

Como *target* foi utilizado o atributo relativo à alta:

- *Target* – Alta

E foram utilizadas as técnicas:

- Técnica 1 – Support Vector Machines
- Técnica 2 – Árvores de Decisão
- Técnica 3 – Naïve Bayes

Nesta abordagem, à semelhança da anterior, os modelos foram induzidos automaticamente e em tempo real utilizando *streaming* de dados pelo que o *tuplo* do DMIT será semelhante acrescendo apenas das variáveis relativas às classes. As configurações utilizadas na ferramenta são também as mesmas podendo ser consultadas na tabela 18.

### 3.2.4 Avaliação

#### Abordagem A

Os resultados obtidos com a abordagem A não foram satisfatórios. O melhor modelo apresentou uma acuidade geral de 73,28%. Para avaliar estes modelos apenas foi utilizada a métrica de acuidade pois a sensibilidade e especificidade não podem ser calculadas uma vez que a classe de saída (*output*) é representada por mais de duas classes. A tabela 19 apresenta os três melhores modelos para esta abordagem. O Anexo II apresenta os resultados de todos os modelos desta abordagem.

**Tabela 19 – Melhores modelos para abordagem A**

| Cenário | *Target* | Técnica | Acuidade |
|---|---|---|---|
| **C1** | T2 | NB | 73,28% |
| **C2** | T2 | NB | 73,28% |
| **C2** | T4 | NB | 50,75% |

Nesta abordagem obtiveram-se resultados bastante melhores que os da abordagem A, sendo que as tabelas 20 e 21 apresentam os resultados dos 3 melhores modelos e a os 3 melhores modelos para cada métrica.

**Tabela 20 – Melhores modelos para abordagem B**

| Cenário | Técnica | Acuidade | Sensibilidade | Especificidade |
|---|---|---|---|---|
| **C3** | AD | 83,494% | 95,879% | 81,444% |
| **C6** | AD | 81,902% | 94,432% | 79,814% |
| **C3** | SVM | 80,502% | 93,541% | 78,330% |

**Tabela 21 – Melhores modelos para abordagem B por métrica**

| Acuidade | | Sensibilidade | | Especificidade | |
|---|---|---|---|---|---|
| **C3 – AD** | 83,494% | **C3 – AD** | 95,879% | **C5 – AD** | 81,144% |
| **C5 – AD** | 83,110% | **C6 – AD** | 94,432% | **C4 – AD** | 79,814% |
| **C4 – AD** | 82,967% | **C3 – SVM** | 93,541% | **C3 – AD** | 78,330% |

Conforme referido no enquadramento conceptual a taxa de ocupação de serviço é o rácio entre o número de doentes por dia e o número de leitos por dia num determinado período. No caso desta abordagem, o modelo é também capaz de calcular a taxa de ocupação do serviço para as próximas vinte e quatro horas utilizando os resultados da previsão. Sempre que um novo doente é admitido ou tem alta o modelo de DM é executado e o rácio é recalculado. Este procedimento é representado pelo seguinte algoritmo:

***Algoritmo*** *- Taxa de ocupação de serviço*

**Requires**: horas, número de camas

```
1:  Function Taxa de ocupação de serviço [hora]
2:    If hora < 24 then
3:      For todas camas do
4:        Identificar ocupação de cama
5:        If cama está livre Then
6:          Inserir na cama N o valor 0
7:        Else
8:          Inserir na cama N o valor 1
9:        End if
10:     Next
11:     OcupaçãoHora[hora] = (total de camas ocupadas/camas totais)
```

| | |
|---|---|
| 12: | **End if** |
| 13: | Retorna OcupaçãoHora[hora] |
| 14: | **End function** |

Abordagem C

Na avaliação desta abordagem foram consideradas 3 métricas, a sensibilidade, acuidade e especificidade. Como já foi referido diversas vezes a sensibilidade é a métrica mais adequada à área da medicina. O Anexo III apresenta os resultados de todos os modelos desta abordagem, e a tabela 22 apresenta os três melhores modelos. A tabela 23 apresenta os melhores modelos para cada uma das métricas. A matriz de confusão é bastante útil para descobrir qual o melhor modelo tendo em conta o que se pretende, se uma previsão equilibrada (acuidade), se prever a alta (sensibilidade) ou prever uma não alta (especificidade).

**Tabela 22 – Melhores modelos para abordagem C**

| Cenário | Técnica | Acuidade | Sensibilidade | Especificidade |
|---|---|---|---|---|
| **C7** | AD | 74,620% | 87,322% | 71,169% |
| **C13** | AD | 77,286% | 78,355% | 76,628% |
| **C11** | AD | 77,063% | 78,673% | 76.628% |

**Tabela 23 – Melhores modelos por métrica para abordagem C**

| Acuidade | | Sensibilidade | | Especificidade | |
|---|---|---|---|---|---|
| **C2 – AD** | 77,311% | **C4 – SVM** | 96,140% | **C12 – AD** | 77,067% |
| **C13 – AD** | 77,286% | **C11 – SVM** | 95,963% | **C8 – AD** | 77,043% |
| **C11 – AD** | 77,063% | **C7 – SVM** | 95,902% | **C9 – AD** | 77,043% |

De modo a entender qual a importância que cada atributo apresenta, foi elaborado um ranking para o melhor modelo. Recorreu-se ao método *Minimum Description Length,* uma técnica supervisionada para calcular a importância dos atributos. A tabela 24 permite ver a importância dos atributos para o melhor modelo.

**Tabela 24 – Importância dos atributos no melhor modelo**

| Atributo | Posição | Importância |
|---|---|---|
| **Hora** | 1 | 0,110 |
| **ECA Total** | 2 | 0,043 |
| **ECA FC** | 3 | 0,025 |

| | | |
|---|---|---|
| **Local Proveniência** | 4 | 0,020 |
| **ECA SPO2** | 5 | 0,019 |
| **Idade** | 6 | 0,010 |
| **Cardiovascular** | 7 | 0,008 |
| **Renal** | 8 | 0,002 |
| **Tipo Admissão** | 9 | 0,001 |
| **Hematológico** | 10 | 0 |
| **ECA PA** | 10 | 0 |
| **ECA PA Max** | 10 | 0 |
| **ECA FC Max** | 10 | 0 |
| **ECA SPO2 Max** | 10 | 0 |
| **ECA Tot Max** | 10 | 0 |
| **Hepático** | 10 | 0 |
| **Respiratório** | 10 | 0 |

Considerando o cenário 7 e a técnica de árvores de decisão é possível verificar que os atributos mais importantes correspondem às novas variáveis introduzidas: o número de horas de internamento, os eventos críticos acumulados por hora relativos à frequência cardíaca e à saturação de oxigénio, o total de eventos críticos acumulados por hora e o local de proveniência do doente.

### 3.2.5 Discussão de Resultados

Comparando as três abordagens torna-se evidente que os modelos com melhores resultados foram os que previam a hora de alta do doente e com base nessa informação atualizam o tempo de internamento em tempo real, abordagens B e C. Como métricas base para a seleção dos modelos definiu-se em conformidade com os intensivistas que os modelos para serem aceites deveriam apresentar uma acuidade arredondada superior a 75% e sensibilidade superior a 85%.

Na abordagem A, os modelos desenvolvidos, que consideraram informação relativa ao processo de admissão e a medições dos sinais vitais do doente, não podem ser considerados num ambiente crítico como é o caso das UCI. A técnica de árvores de decisão apresenta as melhores acuidades gerais situando-se as mesmas entre os 7,6% e os 38,46%. No caso dos métodos de *Naïve Bayes* os resultados situaram-se entre os 15,14% e os 73,28%. Por último recorrendo a *support vector machines* o pior

resultado corresponde a uma acuidade de 32.65% e o melhor a 47,11%. O uso de diferentes classes como target não influenciaram os resultados e a utilização de técnicas de categorização e agrupamento não demonstraram ter efeito no comportamento dos modelos. Esta abordagem não evidenciou qualquer relevância ao nível clínico, uma vez que as acuidades foram bastantes baixas e não atingiram o objetivo definido em concordância com os intensivistas, que era uma acuidade superior a 75%.

Relativamente à abordagem B, os melhores modelos foram os que utilizaram árvores de decisão e que combinaram CM, SOFA e ECA como variáveis independentes. Estes modelos apresentam uma confiança muito boa quando é necessário perceber se um doente vai ou não estar hospitalizado nas próximas horas. O melhor modelo é sensível na previsão da hora de alta do doente em cerca de 96%. Este valor representa que para cada 100 casos o modelo acerta 96 vezes a hora em que o doente vai ter alta. Na figura 13 é possível ver a curva ROC para o melhor modelo e observar que a mesma apresenta valores próximos do valor perfeito (1.00). Os modelos desta abordagem revelaram-se bastante melhores que os da abordagem anterior.

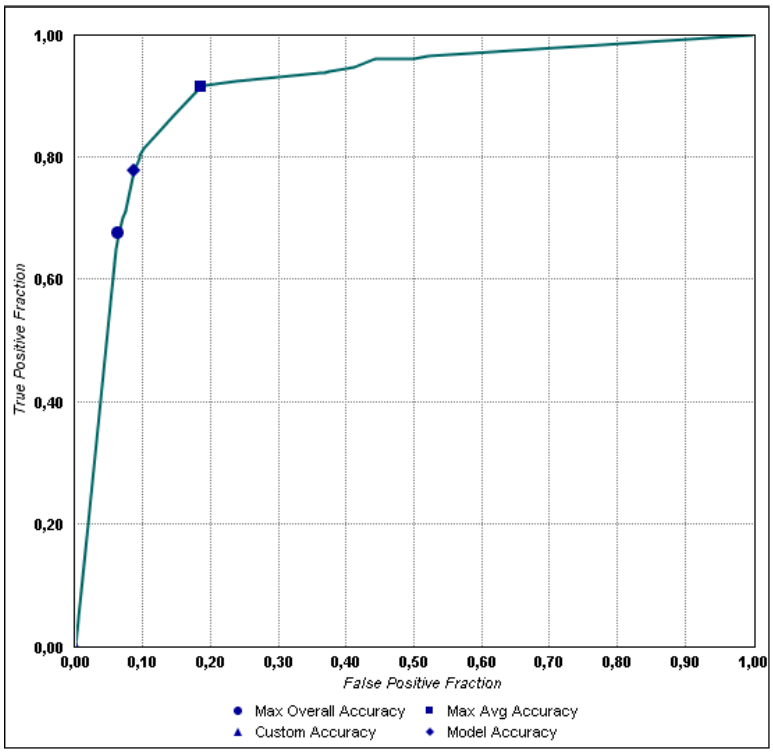


**Figura 13 – Curva ROC para o modelo C3 AD da abordagem B**

No caso da abordagem C, os modelos que apresentaram os melhores resultados foram os que consideram as variáveis CM, SOFA, ACE e o segundo grupo de rácios (R2). O melhor modelo para esta

abordagem foi o que considerou o cenário 7 e técnica de árvores de decisão, que apresentou uma acuidade geral de 74,62% e uma sensibilidade de 87,32%. Analisando as métricas deste modelo, o mesmo é bastante equilibrado, revelando-se portanto um bom modelo para efetuar uma predição precisa da alta do doente. A figura 14 apresenta a curva ROC para este modelo.

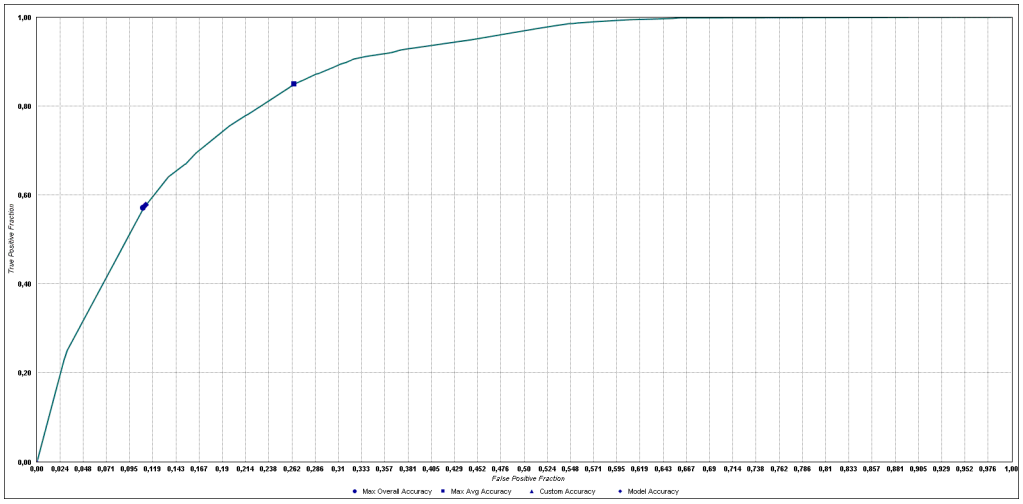


**Figura 14 – Curva ROC para o modelo C7 AD abordagem C**

Analisando mais ao pormenor a métrica da sensibilidade, uma métrica bastante importante na área clinica (os intensivistas querem este tipo de modelos sensíveis a um resultado pois preferem saber que um doente tem um problema e agirem de acordo com a informação fornecida do que não fazerem nada e um problema aparecer), alguns dos modelos apresentam sensibilidades superiores a 95%, sendo o melhor modelo ao nível da sensibilidade o que utiliza o cenário 4 e a técnica de SVM atingindo uma sensibilidade de 96,14%. No geral a técnica de SVM apresentou resultados bastante altos na métrica da sensibilidade, no entanto a acuidade dos mesmos é bastante baixa, o que provoca um elevado número de falsos positivos (FP). De modo a evitar este elevado número de FP e ao mesmo tempo ter modelos sensíveis, foi definida uma métrica que engloba a sensibilidade, especificidade e acuidade.

A utilização de valores discretos nos grupos de variáveis ACE e rácios levaram à produção de modelos com uma melhor acuidade mas o custo desse aumento foi uma diminuição da sensibilidade em cerca de 15% a 20%.

Os modelos que apresentam maiores acuidades apresentam uma baixa sensibilidade pelo que a escolha do melhor modelo depende do objetivo do decisor, se ele pretende bons modelos para prever a possibilidade de um doente ter alta (sensibilidade) ou modelos mais equilibrados entre as duas classes

de saída, ter ou não alta (acuidade), ou então utilizar a métrica definida que permite ter modelos sensíveis e o mais equilibrado possível.

Em relação à métrica da especificidade, a mesma mantem-se sempre dentro do mesmo intervalo de valores independentemente do cenário ou técnica utilizada.

Comparando os resultados da abordagem C com a B, existe um *tradeoff* entre as duas, sendo que na C obteve-se sensibilidades um pouco mais altas mas a acuidade geral dos modelos baixou um pouco comparativamente à abordagem B.

### 3.2.6 Conclusões

As abordagens utilizadas neste estudo diferem no conjunto de dados utilizados e no número de classes da variável *target*. A abordagem A considera os piores valores do doente durante as primeiras 24 horas para prever a alta de um doente (classe) com uma variedade entre 1 e *n* (n classes). A abordagem B e C consideram valores em tempo real recolhidos hora a hora para prever a hora em que um doente vai ter alta (2 classes, com resultado 1 ou 0).

A abordagem A diz respeito a uma abordagem mais convencional que é seguida em outras áreas da medicina. Na medicina intensiva este tipo de abordagem revela-se pouco útil. A primeira parte deste estudo foi dedicada a essa abordagem.

Na segunda parte, uma abordagem alternativa (B) foi explorada de modo a ultrapassar as limitações encontradas com a abordagem A. Esta nova abordagem revelou ter características mais apropriadas ao âmbito da medicina intensiva. Todo este estudo foi levado a cabo sob a supervisão de intensivistas do CHP – HSA. Eles foram responsáveis pela verificação da solidez do estudo e dos impactos do mesmo na medicina intensiva. Com este estudo surgiram resultados interessantes na previsão do tempo de internamento em UCI. Os resultados permitem concluir que a previsão da hora a que um doente vai ter alta é bastante eficiente.

Posteriormente optou-se por criar uma terceira abordagem (C), baseada na abordagem B que incluiu a categorização e agrupamento ao nível dos rácios e ECA que ajudaram a aumentar a sensibilidade dos modelos, no entanto verificou-se uma diminuição da acuidade geral.

A medicina intensiva tem características específicas que fazem com que os modelos de tempo de internamento clássicos sejam inadequados. Os modelos em tempo real permitem atualizar o tempo de internamento tendo em conta a condição do doente que se vão alterando ao longo do internamento, obtendo-se melhores resultados.

Com este estudo alcançaram-se algumas contribuições em duas áreas, a medicina intensiva e o DM. No âmbito da medicina intensiva foi criado um modelo preditivo bastante eficiente na determinação do tempo de internamento de um doente. De acordo com revisão de literatura efetuada no enquadramento conceptual foram apresentados novos resultados no âmbito da investigação realizada nesta área. Em relação ao DM foi desenvolvida uma abordagem em tempo real cujos resultados são um interessante incremento no conhecimento científico do *streaming* DM.

Estas contribuições são essencialmente focadas no campo da medicina intensiva. Utilizando técnicas de DM é possível chegar a uma modelo preditivo do tempo de internamento para melhorar o processo de tomada de decisão relativo às altas clinicas, ou seja, qual a probabilidade um doente ter alta nas próximas 24 horas. Esta abordagem pode ser seguida em outras unidades críticas onde os doentes estejam em constante monitorização. Em outros casos, que não consideram a utilização de dados em tempo real, as soluções apresentadas na abordagem conceptual conseguem atingir o objetivo.

Como trabalho futuro recomenda-se a utilização de variáveis adicionais de modo a compreender como as mesmas afetam o tempo de internamento dos doentes internados na UCI.

## 3.3 Suporte à Decisão de Infeções Bacteriológicas em Medicina Intensiva

Como já referido na abordagem conceptual as infeções e tratamentos das mesmas em Unidades de Cuidados Intensivos (UCI) representam um fator crítico e sob o qual ainda existem diversos problemas.

Este estudo apresenta uma primeira abordagem para o suporte à decisão de infeções bacteriológicas apresentando um algoritmo baseado em heurísticas que é capaz de dar informação aos intensivistas sobre quais os antibióticos que podem ter mais sucesso no controlo de uma determinada infeção. O sistema onde o algoritmo está inserido contém agentes inteligentes que suportam as tarefas principais. Estes agentes são responsáveis por recolher e processar dados em tempo real. Quando uma nova infeção é detetada o sistema corre automaticamente o algoritmo e fornece aos intensivistas uma lista de possíveis antibióticos que podem ser dados ao doente. O algoritmo heurístico é responsável pela pesquisa das melhores soluções no tratamento de uma infeção. O sistema pesquisa todos os tratamentos administrados no passado e baseando-se nos dados clínicos do doente, dados de admissão como idade e sexo, apresenta tratamentos alternativos que podem atingir o sucesso. Este estudo também considera fatores como o custo, casos bem-sucedidos e tempo expectável para que o antibiótico produza efeito. O sistema retorna um conjunto de tratamentos que foram bem-sucedidos no passado, usando variáveis de entrada similares aos dados do doente.

### 3.3.1 Compreensão e Preparação dos Dados

Como referido acima a administração de antibióticos é um problema importante pois existe um grande número de particularidades que devem ser analisadas no tratamento de infeções. O sistema desenvolvido neste estudo tenta eliminar alguns desses problemas. De maneira a analisar e identificar padrões de tratamento é necessário fornecer variáveis de entrada ao algoritmo. Após algumas reuniões com os intensivistas do CHP – HSA foi definido um pacote fixo de variáveis.

Independentemente do tipo de infeção, deve-se sempre considerar quatro tipos de variáveis. São elas idade, sexo, leucócitos e dias de internamento. As restantes variáveis são relacionadas com o sistema de órgãos que está a ser afetado pela bactéria: cardiovascular, hepático, renal, neurológico e hematológico. Para este estudo foram consideradas infeções provocadas pela bactéria *Pseudomonas aeruginosa* (*psaer*) uma bactéria que tipicamente afeta o sistema respiratório. Neste caso foram utilizadas as variáveis *SPO2*, *PaCO2*, *PaO2/Fio2*. A tabela 25 apresenta as variáveis utilizados na primeira parte do algoritmo e uma breve descrição de cada uma.

**Tabela 25 – Variáveis usadas para efetuar a pesquisa**

| Variável | Descrição |
|---|---|
| **idade** | A idade do doente em anos. |
| **sexo** | Informação acerca do sexo do doente. Pode ser masculino ou feminino. |
| **leuc** | Leucócitos – quantidade de leucócitos referida nos resultados das análises laboratoriais com a data mais próxima da data em que a infeção é detetada. |
| **ddi** | Dias de Internamento – número de dias de hospitalização do doente. |
| **SPO2** | Saturação periférica de oxigénio – é uma estimativa do nível de saturação de oxigénio.<br>Os valores normais situam-se entre os 95% e os 100%. Entre 90% e 95% a saturação de oxigénio é baixa mas não representa necessariamente um problema de saúde e abaixo dos 90% considera-se que o doente está em hipoxemia. |
| **PaCO2** | Pressão parcial de dióxido de carbono – representa a pressão parcial de dióxido de carbono no sangue arterial e expressa a eficácia da ventilação alveolar. Os valores normais estão entre 35 e 45 mm Hg. Se o valor for maior do que 45 mm Hg o doente está em hipercapnia. |
| **PaO2** | Pressão parcial de oxigénio – refere-se à medição de oxigénio no sangue arterial. O valor normal está entre 75 e 100 mm Hg. Se o valor estiver abaixo disso o doente não está a receber oxigénio suficiente. |
| **PaO2/FIO2** | Rácio entre a pressão parcial de oxigénio e a fração de oxigénio inspirado – compara os níveis de oxigénio no sangue com os de oxigénio que é respirado. É muito útil para verificar se existem problemas de transferência de oxigénio entre os pulmões e o sangue. Se o rácio estiver abaixo dos 250 mm Hg é um dos critérios de avaliação de pneumonia. |

As variáveis acima descritas fazem parte do grupo de variáveis de pesquisa. De modo a executar o algoritmo que contém as heurísticas é necessário também ter um conjunto de dados que contenham dados relativos ao histórico dos tratamentos. As variáveis presentes no grupo relativo aos tratamentos passados são as mesmas mencionadas anteriormente às quais se adicionam: *cultura*, *antibiótico*, *resultado antibiótico*, *dias tratamento* e *custo*. A tabela 26 apresenta as variáveis do conjunto de dados associados a infeções verificadas no passado.

**Tabela 26 – Variáveis com informação de tratamentos passados**

| Variável | Descrição |
|---|---|
| ***tidade, tsexo, tleuc, tddi, tSPO2,tPaO2, tPaCo2, tPaO2/FIO2*** | Estas variáveis representam tem o mesmo significado mas referem valores de tratamentos passados, tendo-se adicionado o prefixo *t* para diferenciar umas de outras. |
| ***cultura*** | Representa um caso positivo para uma determinada bactéria. Se a variável tiver como valor o texto *psaer* significa que os dados dessa linha representam um caso positivo para a bactéria *psaer*. |
| ***antibiótico*** | Esta variável representa o antibiótico que foi dado ao doente para tentar combater a bactéria. |
| ***resultado antibiótico*** | A variável assume dois resultados: positivo ou negativo. Se o valor for positivo significa que o antibiótico produziu efeito e a infeção está a ser travada. Se o valor for negativo, o antibiótico não está a produzir efeito e é necessário partir para outro antibiótico. |
| ***dias tratamento*** | O valor desta variável indica o número de dias que um antibiótico prescrito demorou a fazer efeito no combate à infeção. Se este valor for nulo significa que o antibiótico não é adequado para a bactéria, sendo para essa linha o valor da variável *resultado antibiótico* negativo. |
| ***custo*** | Representa o custo unitário de uma dose do antibiótico administrado. |

De modo a testar o algoritmo proposto foi utilizado um conjunto de dados históricos (dados reais) de doentes que tiverem infeções provocadas pela bactéria *Pseudomonas aeruginosa*. Este conjunto contém 375 registos que se referem a 67 doentes internados e à utilização de 5 antibióticos diferentes. O conjunto de dados é pequeno pois existem diversos resultados laboratoriais que estão em formato fechado, ou seja, existem laboratórios que não autorizam o acesso eletrónico aos mesmos, pelo que não é possível ter acesso a todos os resultados das análises bacteriológicas efetuadas. No entanto, para testar esta abordagem os dados existentes são suficientes uma vez que representam uma amostra completa de vários doentes, relativa a uma infeção específica.

### 3.3.2 Algoritmo

Nesta seção para melhor se entender o algoritmo proposto é apresentado um caso prático, isto é, a pesquisa de tratamentos possíveis para a bactéria *pseudomonas aeruginosa*. Apesar desta representação prática, o algoritmo desenvolvido pode ser utilizado em diferentes tipos de infeções/bactérias, mudando apenas os dados de entrada e o *target* de saída.

Para pesquisar possíveis tratamentos para a bactéria *psaer* foi desenvolvido um algoritmo que utiliza heurísticas e que é responsável por encontrar os tratamentos que melhor se adequem ao caso procurado.

Este algoritmo necessita de um conjunto de dados que são representados pelos seguintes grupos:

- Variáveis pesquisadas – {ddi, leuc, idade, sexo, SPO2, PaCO2, PaO2, PaO2/FIO2}
- Historial Tratamento – {tddi, tleuc, tidade, tsexo, tSPO2, tPaCO2, tPaO2, tPaO2/FIO2}

Será criado um agente inteligente que fará parte da arquitetura INTCare e este agente inteligente estará constantemente a monitorizar as variáveis respiratórias e quando determinados eventos ocorrerem (pré-definidos pelos clínicos) o sistema emite um alerta sobre uma possível infeção do doente. Depois o agente espera pelos resultados das culturas microbiológicas para identificar a bactéria. Se a bactéria identificada no exame for a *psaer*, o agente invoca a função tratamentos, que é responsável por encontrar possíveis tratamentos. Esta função compara as variáveis biológicas da pesquisa do doente infecionado com as variáveis dos tratamentos passados (histórico). O algoritmo devolve os resultados onde os resultados das variáveis de pesquisa e dos dados históricos são mais próximos, sendo que o sistema utiliza também um sistema de prioridades/ponderações no momento de calcular a diferença. Nesta parte do algoritmo é possível retornar resultados tendo em conta fatores como os dias de tratamento expectáveis para o antibiótico produzir efeito e/ou o custo do tratamento.

As variáveis escolhidas para efetuar a comparação de resultados e a sua prioridade foram definidas através de reuniões com os intensivistas do CHP – HSA e são:

1. SPO2;
2. ddi;
3. leuc;

4. PaO2/FIO2;
5. PaO2;
6. PaCo2;
7. idade;
8. sexo

O procedimento acima descrito é representado pelo seguinte algoritmo:

***Algoritmo*** *- Alternativas de Tratamento*

**Requer**: variáveis de pesquisa, variáveis históricas

```
1:  Function Avaliar Infeção
2:    Get variáveis de pesquisa
3:    If SPO2 < 90 e (PaO2 < 75 ou PaCO2 > 75) e PaO2/FIO2 < 250 then
4:      For doente do
5:        Existe infeção. Aguarda por resultados das culturas.
6:        If Psaer positiva Then
7:          Function Tratamentos
8:        Else
9:          Esperar por resultados positivos das culturas
10:       End if
11:   End if
12:  End Function
13:  Function Tratamentos
14:    Get variáveis de pesquisa, variáveis históricas
15:      If resultado antibiótico = successo
16:        Tratamentos = resultados ordenados pela diferença mínima entre
       variáveis
17:      Else
18:        Não existem tratamentos com sucesso
19:      End if
20:    Return Tratamentos
21:  End function
```

Para outras bactérias, o processo de tentativa de encontrar o melhor tratamento é o mesmo. Em outros tipos de infeções o algoritmo irá mudar mas a mecânica de pesquisa é exatamente a mesma. Apenas irão existir diferenças nas variáveis utilizadas (mantendo-se os dois grupos) e na primeira clausula *if* do algoritmo heurístico.

### 3.3.3 Resultados

Como referido acima, para testar este algoritmo foi utilizado um conjunto de dados com o histórico de tratamentos aplicados a doentes internados no passado. Uma vez que este estudo pretendeu avaliar a viabilidade desta abordagem foi utilizado um conjunto de dados de entrada para que o sistema pesquisasse os resultados que melhor se adaptavam aos dados de entrada. No futuro esta tarefa será executada por um agente inteligente em tempo real e utilizando aprendizagem automática (*online-learning)*. A tabela 27 representa os dados de entrada utilizados pelo algoritmo. Estes dados são referentes a um doente internado que está infetado.

**Tabela 27 – Dados de entrada utilizados pelo algoritmo**

| Variável | Valor |
|---|---|
| **ddi** | 14 |
| **leuc** | 13 |
| **idade** | 65 |
| **sexo** | M |
| **SPO2** | 89 |
| **PaCO2** | 41 |
| **PaO2** | 65 |
| **PaO2/FIO2** | 230 |

Utilizando estes dados como entrada, o algoritmo iniciou a pesquisa de possíveis tratamentos. O mesmo retornou uma lista de vinte e três tratamentos viáveis. A tabela 28 apresenta os cinco melhores resultados da pesquisa pela ordem de sucesso. Esta tabela apresenta os valores históricos das variáveis para cada sugestão (R1 a R5), o antibiótico administrado, o resultado atingido (sucesso ou insucesso), o número de dias expectáveis para o tratamento e o custo do tratamento. Considerando os critérios de prioridade/ponderação verificou-se que mesma está em concordância com os resultados obtidos. O *SPO2* dos dados de entrada varia em apenas uma unidade relativamente aos dados históricos, sendo que nos outros conjuntos existiu um aumento da diferença entre as duas variáveis. Os *dias de internamento* também estão bastante próximos dos utilizados nos dados de entrada e à medida que se avança nos resultados os mesmos vão ficando mais longe. O mesmo se aplica às restantes variáveis onde foi atribuída uma prioridade à ordenação dos resultados.

**Tabela 28 – Tratamentos aconselhados pelo algoritmo**

| Variável | Resultados | | | | |
|---|---|---|---|---|---|
| | R1 | R2 | R3 | R4 | R5 |
| **ddi** | 16 | 13 | 14 | 18 | 20 |
| **Leuc** | 14,5 | 14,2 | 13,2 | 15,2 | 16,1 |
| **idade** | 57 | 52 | 42 | 69 | 76 |
| **sexo** | M | M | F | M | F |
| **SPO2** | 87 | 86 | 92 | 82 | 80 |
| **PaCO2** | 38 | 36 | 41 | 40 | 43 |
| **PaO2** | 66 | 65 | 72 | 57 | 55 |
| **PaO2/FIO2** | 226 | 235 | 240 | 216 | 198 |
| **Antibiótico** | Gent | Gent | Gent | Col | Mer |
| **Resultado Antibiótico** | Suc | Suc | Suc | Suc | Suc |
| **Dias de Tratamento** | 6 | 8 | 5 | 12 | 14 |
| **Custo** | 2,4 | 2,4 | 2,4 | 1,8 | 1,3 |

Além destes resultados o algoritmo também devolve informação acerca da eficácia que o antibiótico demonstrou no passado. Este resultado é expresso em forma de percentagem e dá uma visão geral dos antibióticos utilizados no passado e o seu sucesso. Por exemplo, uma informação de saída sobre a aplicação de antibióticos é a seguinte:

- ✓ Vancomicina (Vanc) – Usada com sucesso 5 vezes (10%); Usada com insucesso 45 vezes (90%)
- ✓ Gentamicina (Gent) – Usada com sucesso 45 vezes (90%); Usada com insucesso 5 vezes (10%)
- ✓ Meropeném (Mer) – Usada com sucesso 10 vezes (20%); Usada com insucesso 40 vezes (80%)

### 3.3.4 Discussão

Analisando os resultados obtidos e a tabela 28 é possível observar que o sistema foi capaz de sugerir algumas opções de tratamento, neste caso recorrendo à administração de gentamicina, colistina e meropeném. Baseando-se nesta informações os intensivistas podem decidir que antibiótico deverá ser prescrito usando como base as sugestões apresentadas pelo sistema desenvolvido (algoritmo). Este algoritmo vai ser incorporado no sistema INTCare que devido às suas características *pervasive* vai permitir que a informação gerada pelo algoritmo esteja disponível em qualquer local, permitindo assim um acesso em tempo real por quem tiver privilégios de acesso ao mesmo.

Com este algoritmo foi possível observar uma lista de tratamentos que podem produzir efeitos no combate à bactéria *psaer*. Analisando esta informação o intensivista pode ter uma ideia sobre qual o tratamento que poderá apresentar o melhor efeito, pois doentes com variáveis de entrada muito idênticas na teoria terão respostas semelhantes aos antibióticos.

Este sistema não pretende ser um sistema especialista mas sim um sistema de apoio à decisão. Isto acontece devido ao facto de ao nível dos antibióticos não ser uma tarefa fácil prever e ajudar a decisão pois as condições dos doentes podem ser bastante díspares e mesmo quando são semelhantes as reações ao mesmo antibiótico podem ser bastante diferentes. Portanto, o objetivo deste estudo é sugerir ideias de possíveis tratamentos aos intensivistas, eliminando à partida tratamentos que no passado não produziram qualquer efeito ou que tenham apresentado taxas de sucesso relativamente baixas (esta % é definida pelos intensivistas). Os intensivistas são sempre os responsáveis pela decisão tomada. Este sistema apenas é usado para os ajudar a tomar a melhor decisão tendo em conta a apresentação de novo conhecimento útil capaz de contribuir para ajudar a melhorar condição clínica de um doente

### 3.3.5 Conclusões

Este estudo revelou ser um bom ponto de partida para que os intensivistas possam escolher teoricamente o melhor tratamento no combate a uma determinada infeção. Com este estudo foi possível fazer prova de conceito sobre a viabilidade da aplicação do algoritmo desenvolvido a casos reais. Os resultados mostraram que a abordagem proposta é viável, podendo usar heurísticas no suporte ao processo de decisão nas UCI.

De modo a que este algoritmo seja adaptado a outra realidade/infeção as mudanças podem ser facilmente efetuadas pois as regras da heurística serão guardadas em uma base de dados e o algoritmo é desenhado de acordo com as regras definidas nessa mesma base de dados.

No futuro haverá muito trabalho a executar de modo a que se faça a implementação deste algoritmo. É necessário construir um agente inteligente que execute as seguintes tarefas:

- Analise os dados e identifique a possibilidade de novas infeções;
- Aceda aos resultados laboratoriais das culturas para ver qual a bactéria que foi identificada;

É também necessário melhorar o algoritmo de modo a ser mais robusto e capaz de apoiar a decisão de uma forma mais precisa.

Este sistema deverá ser desenvolvido para identificar e pesquisar diversos tipos de bactérias de modo a que no futuro possa ser implementado em tempo real na UCI para verificar a eficácia do sistema em um contexto real.

Outro tópico que deverá ser desenvolvido para implementar este sistema é ter acesso a uma base de dados com toda a informação sobre os custos e grupos dos antibióticos. Depois de reuniões com os intensivistas do hospital, foi determinado que os dados relativos aos custos e grupos de antibióticos serão inseridos em bases de dados do hospital no prazo alguns meses de modo a que seja possível criar a implementação final do sistema e testá-lo em um ambiente produtivo.

# 4. CONCLUSÕES

Neste capítulo serão apresentadas as conclusões gerais da dissertação, sendo que o mesmo se encontra dividido em dois subcapítulos. No primeiro irá ser apresentada uma síntese do trabalho prático efetuado. Este trabalho prático consistiu em três grandes estudos sendo eles: a utilização de técnicas de *clustering* para prever readmissões em medicina intensiva, a previsão do tempo de internamento de um doente em medicina intensiva e um último estudo que aferiu sobre as possibilidades do suporte à decisão no combate a infeções bacteriológicas. Neste capítulo serão também apresentadas todas as contribuições científicas que este trabalho apresentou assim com uma breve análise dos objetivos propostos e a resposta à questão de investigação formulada como base deste trabalho. No segundo subcapítulo são apresentadas as grandes linhas de investigação que poderão ser executadas caso se pretenda trabalhar sobre esta temática.

## 4.1 Síntese e Contribuições Científicas

Como já foi referido ao longo deste documento a gestão de custos e recursos nas UCI apresentam-se como um desafio e um dos principais pontos sobre os quais é necessário dar resposta. Outro dos pontos importantes nas UCI é o estudo de questões ligadas ao *outcome* dos doentes, o que passará por exemplo estudar casos de readmissão ou de tempos de internamento de doentes em UCI.

O trabalho realizado incidiu essencialmente sobre estas duas grandes áreas sendo elas a gestão de custos e o *outcome* de doentes, sendo que as mesmas estão bastante interligadas pois por exemplo ao abordar a temática da previsão do tempo de internamento estamos a trabalhar sobre gestão de custos, sabendo por exemplo quando é que existirão camas livres, mas também se trabalha no âmbito do *outcome* de doentes pois está-se a prever o momento em que o doente terá alta e está apto a regressar à sua vida normal.

Para tal, e com o intuito de ir de encontro aos objetivos propostos e à questão de investigação efetuaram-se três grandes estudos: o primeiro aborda a utilização de técnicas de *clustering* na caracterização de

potenciais grupos de doentes reinternados. Este trabalho apresentou resultados significativos para a comunidade medico-científica sendo que atributos como o rácio PaO2/FIO2, PaCO2, local de proveniência e tempo de internamento apresentaram grande relevância na caracterização de grupos de reinternamento. De referir que as variáveis sugeridas pelos intensivistas, nomeadamente os resultados laboratoriais do ácido lático e dos leucócitos também contribuíram para a segmentação dos grupos. Os modelos induzidos apresentaram boas avaliações sendo que o índice de Davies Bouldin se situou entre os 0,5 e 0,55 para os melhores modelos. Ao nível da contribuição científica esta abordagem traz conhecimento à área da medicina intensiva por via da utilização de técnicas de *clustering*, caracterizando grupos de doentes que tem uma grande probabilidade de vir a ser readmitidos na UCI.

O segundo estudo abordou a previsão do tempo de internamento de doentes em UCI. O mesmo consistiu em três abordagens diferentes, sendo que a primeira considerou os piores valores do doente nas primeiras vinte e quatro horas e as outras duas utilizaram essencialmente valores recolhidos em tempo-real de hora a hora. A primeira abordagem apresentou resultados pouco relevantes, entre 7% e 70% de acuidade geral, enquanto a segunda e terceira revelaram bons resultados ao nível da acuidade (cerca de 80%) e ao nível da sensibilidade (cerca de 95%) para os melhores modelos. Com este estudo alcançaram-se contribuições científicas em duas áreas: a medicina intensiva e os sistemas de informação (SI), sendo que ao nível da medicina intensiva criou-se um modelo preditivo bastante eficiente para saber o tempo de internamento de um doente e no caso dos SI foi desenvolvido uma abordagem em tempo real utilizando a aprendizagem automática com resultados interessantes na área do *streaming* DM e que poderá ser utilizado pela comunidade científica.

O terceiro estudo consistiu no desenvolvimento de um algoritmo para ajudar o suporte à decisão no âmbito das infeções bacteriológicas. Efetuou-se uma prova de conceito sobre a viabilidade do apoio à decisão no combate às infeções bacteriológicas, sendo que se concretizou o objetivo, tendo sido obtido um conjunto de possíveis tratamentos para combater a bactéria *psaer* tendo em conta os tratamentos administrados no passado e o seu sucesso ou insucesso. Ao nível das contribuições científicas este estudo traz novo conhecimento à área do apoio à decisão na medicina intensiva e também no combate a infeções bacteriológicas, existindo informação útil para o intensivistas na hora de escolher o antibiótico a administrar, e também informação acerca dos antibióticos com maiores ou menores taxas de sucesso.

A tabela 29 relaciona os três estudos desenvolvidos neste trabalho com os objetivos do trabalho, contribuições científicas e resultados.

**Tabela 29 – Relação entre estudos, objetivo do trabalho, contribuições científicas e resultados**

| Estudo | Objetivo | Resultados | Contribuição Científica |
|---|---|---|---|
| *Clustering* na caracterização de grupos de doentes readmissíveis | Modelos de otimização e *outcome* de doentes do serviço de cuidados intensivos no CHP | Identificação de dois grupos de características para doentes com probabilidade de readmissão | Medicina Intensiva |
| Previsão do tempo de internamento de doentes em UCI | Modelos de otimização e *outcome* dos doentes do serviço de cuidados intensivos do CHP | Modelos com acuidades gerais na ordem dos 80% e sensibilidades na ordem dos 95% | Medicina Intensiva |
| | | Com base nos modelos criados além de se saber o tempo de internamento é possível calcular a taxa de ocupação de camas. | DM |
| Suporte à decisão de infeções bacteriológicas em medicina intensiva | Modelos de otimização de custos de internamento/tratamento no CHP | Sistema capaz de apoiar o intensivista no momento de escolher o antibiótico a administrar. Considera questões como tempo de atuação e custos de tratamento. | Apoio à decisão em medicina intensiva |
| | Modelos de otimização e *outcome* dos doentes do serviço de cuidados intensivos do CHP | | Gestão de Custos |

## 4.2 Trabalho Futuro

Como trabalho futuro e no âmbito da otimização de custos e previsão de *outcome* de doentes deixam-se algumas linhas orientadoras para esta temática, sendo elas:

- Exploração de mais variáveis clínicas que possam ser adicionadas aos modelos de *clustering* para caracterização de grupos de doentes readmissíveis de modo a tentar encontrar novos padrões e grupos de doentes que tenham uma alta probabilidade de serem readmitidos na UCI;
- Utilização de novas variáveis nos modelos de DM para a previsão de tempos de internamento. Sugere-se a utilização de um maior número de resultados laboratoriais de modo a verificar se os mesmos influenciam os modelos;
- Tentativa de aplicação das variáveis utilizadas neste trabalho na criação de modelos para outras áreas críticas da gestão hospitalar em UCI como por exemplo a gestão de recursos;
- Implementação do sistema (algoritmo) criado para o suporte à decisão de infeções bacteriológicas em ambiente produtivo e a trabalhar com qualquer tipo de infeção/bactéria, otimizando e aumento a sua robustez;
- Alteração do agente inteligente de recolha de dados para o algoritmo relativo às infeções, para que vá automaticamente buscar todos os custos e grupos de antibióticos associados a uma infeção assim que o CHP – HSA disponibilizar a base de dados que contém essa informação.

# BIBLIOGRAFIA


Adhikari, N. K. J., Fowler, R. A., Bhagwanjee, S., & Rubenfeld, G. D. (2010). Critical care and the global burden of critical illness in adults. *Lancet*, *376*(9749), 1339–46. doi:10.1016/S0140-6736(10)60446-1

Alapont, J., Bella-sanjuán, A., Ferri, C., Hernández-orallo, J., Llopis-llopis, J. D., & Ramírez-quintana, M. J. (2005). Specialised Tools for Automating Data Mining for Hospital Management. In *In Proc. First East European Conference on Health Care Modelling and Computation* (pp. 7–19). Informática, Dimensión.

Azari, A., Janeja, V. P., & Mohseni, A. (2012). Predicting Hospital Length of Stay (PHLOS): A Multi-tiered Data Mining Approach. In *2012 IEEE 12th International Conference on Data Mining Workshops* (pp. 17–24). IEEE. doi:10.1109/ICDMW.2012.69

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, *77*(2), 81–97. Retrieved from http://www.sciencedirect.com/science/article/pii/S1386505606002747

Benjamini, Y., & Leshno, M. (2010). Statistical Methods for Data Mining. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 523–540). Springer US. doi:10.1007/978-0-387-09823-4_25

Berenholtz, S. M., Dorman, T., Ngo, K., & Pronovost, P. J. (2002, March 1). Qualitative review of intensive care unit quality indicators. *Journal of Critical Care*. W.B. Saunders. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0883944102700324?showall=true

Braga, P., Portela, F., Santos, M.F., Rua, F. (2014). Data Mining Models to Predict Patient's Readmission in Intensive Care Units: (2014) (pp. 604–610). SCITEPRESS - Science and and Technology Publications. doi:10.5220/0004907806040610

Brown, J., Dashevsky, I., Fireman, B., Herrinton, L., McClure, D., Murphy, M., ... Kulldorff, M. (2011). C-C5-01: Drug Safety Data Mining with a Tree-Based Scan Statistic. *Clinical Medicine & Research*, *9*(3-4), 180–180. doi:10.3121/cmr.2011.1020.c-c5-01

Caetano, Nuno, Laureano, Raul M. S. & Cortez, Paulo (2014). A Data-driven Approach to Predict Hospital Length of Stay - A Portuguese Case Study. Proceedings of the 16th International Conference on Enterprise Information Systems (ICEIS 2014). 1, 407-414

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

Chazard, E., Ficheur, G., Bernonville, S., Luyckx, M., & Beuscart, R. (2011). Data mining to generate adverse drug events detection rules. *IEEE Transactions on Information Technology in Biomedicine : A Publication of the IEEE Engineering in Medicine and Biology Society*, *15*(6), 823–30. doi:10.1109/TITB.2011.2165727

Cheng, C.-W., Chanani, N., Venugopalan, J., Maher, K., & Wang, M. D. (2013). icuARM-An ICU Clinical Decision Support System Using Association Rule Mining. *IEEE Journal of Translational Engineering in Health and Medicine*, *1*, 4400110–4400110. doi:10.1109/JTEHM.2013.2290113

Chenhui, Z., Huilong, D., & Xudong, L. (2008). An Integration Approach of Healthcare Information System. In *2008 International Conference on BioMedical Engineering and Informatics* (Vol. 1, pp. 606–609). IEEE. doi:10.1109/BMEI.2008.109

Cios, K., Swiniarski, R., Pedrycz, W., & Kurgan, L. (2007). Supervised Learning: Statistical Methods. In *A Knowledge Discovery Approach* (pp. 307–386). Springer US. doi:10.1007/978-0-387-36795-8_1

De Vos, M., Graafmans, W., Keesman, E., Westert, G., & van der Voort, P. H. J. (2007). Quality measurement at intensive care units: which indicators should we use? *Journal of Critical Care*, *22*(4), 267–74. doi:10.1016/j.jcrc.2007.01.002

Direção Geral de Saúde, (2003). *Cuidados Intensivos – Recomendações para o seu desenvolvimento*.

ECDC, (2012). Surveillance of healthcare-associated infections in Europe 2007. European Centre for Disease Prevention and Control. Surveillance Report from ECDC, Stockolm.

el-Darzi, E., Vasilakis, C., Chaussalet, T., & Millard, P. H. (1998). A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. *Health Care Management Science*, *1*(2), 143–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10916593

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). Advances in Knowledge Discovery and Data Mining. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), (pp. 1–34). Menlo Park, CA, USA: American Association for Artificial Intelligence. Retrieved from http://dl.acm.org/citation.cfm?id=257938.257942

Fernandes, J., & Belo, O. (2010). Discovering patterns on medication prescriptions.

Gago, P., Santos, M. F., Silva, A., Cortez, P., Neves, J., & Gomes, L. (2005). INTCare: a Knowledge Discovery Based Intelligent Decision Support System for Intensive Care Medicine. *Journal of Decision Systems*, *14*(3), 241–259. doi:10.3166/jds.14.241-259

Gajic, O., Malinchoc, M., Comfere, T. B., Harris, M. R., Achouiti, A., Yilmaz, M., ... Farmer, J. C. (2008). The Stability and Workload Index for Transfer score predicts unplanned intensive care unit patient readmission: initial development and validation. Critical Care Medicine, 36(3), 676–682. doi:10.1097/CCM.0B013E318164E3B0

Goebel, M., & Gruenwald, L. (1999). A Survey Of Data Mining And Knowledge Discovery Software Tools. *SIGKDD Explorations*, *1*, 20–33.

Gorunescu, F. (2011). Data mining concepts, models and techniques. Berlin: Springer. Retrieved from http://site.ebrary.com/id/10454853

Griffin, D. (2006). *Hospitals: What They are and how They Work*. Jones and Bartlett. Retrieved from http://books.google.pt/books?id=qXO6mRWNwfsC

Hachesu, P. R., Ahmadi, M., Alizadeh, S., & Sadoughi, F. (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthcare Informatics Research*, *19*(2), 121–9. doi:10.4258/hir.2013.19.2.121

Han, J., & Kamber, M. (2000). *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Haux, R., Winter, A., Ammenwerth, E., & Brigl, B. (2004). *Strategic Information Management in Hospitals. An Introduction to Hospital Information Systems* (p. 274). Springer.

Hofmann, M., & Klinkenberg, R. (Eds.). (2013). RapidMiner: Data Mining Use Cases and Business Analytics Applications. Boca Raton: Chapman and Hall/CRC.

Huang, J., Huan, J., Tropsha, A., Dang, J., Zhang, H., & Xiong, M. (2013). Semantics-driven frequent data pattern mining on electronic health records for effective adverse drug event monitoring. In *2013 IEEE International Conference on Bioinformatics and Biomedicine* (pp. 608–611). IEEE. doi:10.1109/BIBM.2013.6732567

Iakovidis, I. (1998). Towards personal health record: current situation, obstacles and trends in implementation of electronic healthcare record in Europe1Disclaimer: The view developed in this paper is that of the author and does not necessarily reflect the position of the Eur. *International Journal of Medical Informatics*, *52*(1-3), 105–115. doi:10.1016/S1386-5056(98)00129-4

Isken, M. W., & Rajagopalan, B. (2002). Data Mining to Support Simulation Modeling of Patient Flow in Hospitals. *Journal of Medical Systems*, *26*(2), 179–197. doi:10.1023/A:1014814111524

J, L. G., Lemeshow, S., & Saulnier, F. (1993). A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *JAMA*, *270*(24), 2957–2963. Retrieved from http://dx.doi.org/10.1001/jama.1993.03510240069035

Janssens, U., Dujardin, R., Graf, J., Lepper, W., Ortlepp, J., Merx, M., ... Hanrath, P. (2001). Value of SOFA (Sequential Organ Failure Assessment) score and total maximum SOFA score in 812 patients with acute cardiovascular disorders. *Critical Care*, *5*(Suppl 1), p225. doi:10.1186/cc1292

Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms* (2nd ed.). Wiley-IEEE Press.

Knaus, W., Draper, E., Wagner, D., & Zimmerman, J. (1985). APACHE II: a severity of disease classification system. *Critical Care Medicine*, *13*(10), 818–829.

Koh, H. C., & Tan, G. (2005). Data mining applications in healthcare. *Journal of Healthcare Information Management : JHIM*, *19*(2), 64–72. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15869215

Koutkias, V., Lazou, K., Kilintzis, V., Beuscart, R., & Maglaveras, N. (2009). On Intelligent Procedures in Medication for Patient Safety: The PSIP Approach. In *2009 Ninth International Conference on Intelligent Systems Design and Applications* (pp. 363–366). IEEE. doi:10.1109/ISDA.2009.207

Last, M., Carel, R., & Barak, D. (2007). Utilization of Data-Mining Techniques for Evaluation of Patterns of Asthma Drugs Use by Ambulatory Patients in a Large Health Maintenance Organization. In *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)* (pp. 169–174). IEEE. doi:10.1109/ICDMW.2007.50

Lindbaek, M., (2006). Prescribing antibiotics to patients with acute cough and otitis media. Br. J. Gen. Pract. 56, 164–165.

Lucas, P. (2004). Bayesian analysis, pattern analysis, and data mining in health care. In *Curr Opin Crit Care* (pp. 399–403).

Maimon, O., & Rokach, L. (2010). Introduction to Knowledge Discovery and Data Mining. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 1–15). Springer US. doi:10.1007/978-0-387-09823-4_1

Marshall, A., Vasilakis, C., & El-Darzi, E. (2005). Length of Stay-Based Patient Flow Models: Recent Developments and Future Directions. *Health Care Management Science*, *8*(3), 213–220. doi:10.1007/s10729-005-2012-z

Mcniff, J., & Whitehead, J. A. (2006). *All You Need To Know About Action Research* (1st ed.). London: Sage Publications Ltd. Retrieved from http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/141290806X

Moreno, R., Metnitz, P., Almeida, E., Jordan, B., Bauer, P., Campos, R., ... Le Gall, J.-R. (2005). SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Medicine*, *31*(10), 1345–1355. doi:10.1007/s00134-005-2763-5

Morik, K., Imboff, M., Brockhausen, P., Joachims, T., & Gather, U. (2000). Knowledge discovery and knowledge validation in intensive care. *Artificial Intelligence in Medicine*, *19*(3), 225–249. doi:10.1016/S0933-3657(00)00047-6

Moser, S. A., Jones, W. T., & Brossette, S. E. (1999). Application of data mining to intensive care unit microbiologic data. *Emerging Infectious Diseases*, *5*, 454–457.

Mullins, I. M., Siadaty, M. S., Lyman, J., Scully, K., Garrett, C. T., Miller, W. G., ... Knaus, W. A. (2006). Data mining and clinical data repositories: Insights from a 667,000 patient data set. *Computers in Biology and Medicine*, *36*(12), 1351–77. doi:10.1016/j.compbiomed.2005.08.003

Nadali, A., Kakhky, E. N., & Nosratabadi, H. E. (2011). Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system. In *2011 3rd International Conference on Electronics Computer Technology* (Vol. 6, pp. 161–165). IEEE. doi:10.1109/ICECTECH.2011.5942073

Nerenz, D., & Neil, N. (2001). *Performance Measures for Health Care Systems*.

Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. J. Manage. Inf. Syst., 24(3), 45–77. doi:10.2753/MIS0742-1222240302

Pordata. (2013). Despesa corrente em cuidados de saúde em % do PIB em Portugal. http://www.pordata.pt/Portugal/Despesa+corrente+em+cuidados+de+saude+em+percentagem+do+PIB-610 (Acedido em 22 de Janeiro de 2014)

Pordata. (2014a). Despesa em cuidados de saúde em % do PIB na Europa. http://www.pordata.pt/Europa/Despesa+em+cuidados+de+saude+em+percentagem+do+PIB-1962 (Acedido em 22 de Janeiro de 2014)

Pordata. (2014b). Despesas do Estado em saúde: execução orçamental per capita em Portugal. http://www.pordata.pt/Portugal/Despesas+do+Estado+em+saude+execucao+orcamental+per+capita-856 (Acedido em 22 de Janeiro de 2014)

Portela, C. F., Santos, M. F., Silva, Á., Machado, J., & Abelha, A. (2011). Enabling a Pervasive Approach for Intelligent Decision Support in Critical Health Care. In M. Cruz-Cunha, J. Varajão, P. Powell, & R. Martinho (Eds.), *Enterprise Information Systems* (Vol. 221, pp. 233–243). Springer Berlin Heidelberg. doi:10.1007/978-3-642-24352-3_25

Portela, F., Gago, P., Santos, M. F., Machado, J., Abelha, A., Silva, Á., ... Pinto, F. (2012a). Intelligent and Real Time Data Acquisition and Evaluation to Determine Critical Events in Intensive Medicine. *Procedia Technology*, *5*, 716–724. doi:10.1016/j.protcy.2012.09.079

Portela, F., Santos, M., Machado, J., Silva, Á., Rua, F., & Abelha, A. (2012b). Intelligent Data Acquisition and Scoring System for Intensive Medicine. In C. Böhm, S. Khuri, L. Lhotská, & M. E. Renda (Eds.), *Information Technology in Bio- and Medical Informatics SE - 1* (Vol. 7451, pp. 1–15). Springer Berlin Heidelberg. doi:10.1007/978-3-642-32395-9_1

Portela, F., Gago, P., Santos, M. F., Machado, J., Abelha, A., Silva, Á., & Rua, F. (2013c). Implementing a Pervasive Real-Time Intelligent System for Tracking Critical Events with Intensive Care Patients. *International Journal of Healthcare Information Systems and Informatics*, *8*(4), 1–16. doi:10.4018/ijhisi.2013100101

Portela, F., Santos, M. F., & Vilas-Boas, M. (2013b). A Pervasive Approach to a Real-Time Intelligent Decision Support System in Intensive Medicine. In A. Fred, J. G. Dietz, K. Liu, & J. Filipe (Eds.), *Knowledge Discovery, Knowledge Engineering and Knowledge Management SE - 25* (Vol. 272, pp. 368–381). Springer Berlin Heidelberg. doi:10.1007/978-3-642-29764-9_25

Portela, F., Santos, M. F., Silva, Á., Machado, J., Abelha, A., & Rua, F. (2013a). Pervasive and Intelligent Decision Support in Critical Health Care Using Ensembles. ITBAM 2013: 1-16

Portela, F., Santos, M. F., Silva, Á., Machado, J., Abelha, A., & Rua, F. (2013c). Data Mining for Real-Time Intelligent Decision Support System in Intensive Care Medicine. In *ICAART 2013 - International Conference on Agents and Artificial Intelligence* (pp. 270–276).

Portela, F., Santos, M., Silva, Á., Machado, J., Abelha, A., & Rua, F. (2014a). Pervasive and Intelligent Decision Support in Intensive Medicine – The Complete Picture. *ITBAM 2014, LNCS* (Vol. 8649, pp. 87-102).

Portela, F., Santos, M., Silva, Á., Machado, J., Abelha, A., & Rua, F. (2014b). A Pervasive Intelligent System for Scoring MEWS and TISS-28 in Intensive Care. In J. Goh (Ed.), *The 15th International Conference on Biomedical Engineering SE - 73* (Vol. 43, pp. 287–290). Springer International Publishing. doi:10.1007/978-3-319-02913-9_73

Ramon, J., Fierens, D., Güiza, F., Meyfroidt, G., Blockeel, H., Bruynooghe, M., & Van Den Berghe, G. (2007). Mining data from intensive care patients. *Advanced Engineering Informatics*, *21*(3), 243–256. doi:10.1016/j.aei.2006.12.002

Rastegar-Mojarad, M., Harrington, B., & Belknap, S. M. (2013). Automatic detection of drug interaction mismatches in package inserts. In *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 373–377). IEEE. doi:10.1109/ICACCI.2013.6637200

Rokach, L., & Maimon, O. (2010). Supervised Learning. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 133–147). Springer US. doi:10.1007/978-0-387-09823-4_8

Ronco, C., Bellomo, R., & Kellum, J. (2009). *Critical Care Nephrology* (2nd ed.). Elsevier Health Sciences. Retrieved from http://www.eu.elsevierhealth.com/product.jsp?isbn=9781437711110&sgCountry=PT&isbn=9781437711110

Santhi, P., & Bhaskaran, M. (2010). Performance of Clustering Algorithms in Healthcare Database. *International Journal for Advances in Computer Science*, *2*(1).

Santos, M. F., & Azevedo, C. (2005). *Data Mining Descoberta de conhecimento em base de dados*. FCA - Editora de Informática, Lda.

Santos, M. F., Portela, F., & Vilas-Boas, M. (2011). INTCARE -Multi-agent Approach for Real-time Intelligent Decision Support in Intensive Medicine. In *ICAART 2011 - International Conference on Agents and Artificial Intelligence* (pp. 364–369).

SAS Institute. (n.d.). SAS Enterprise Miner - SEMMA. http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html (Acedido em 27 de Janeiro de 2014)

Sturges, H. A. (1926). The Choice of a Class Interval. Journal of the American Statistical Association, 21(153), 65–66. doi:10.1080/01621459.1926.10502161

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining* (1st ed.). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Taylor, K. (2010). *Oracle Data Mining Concepts*.

Teow, K. L., El-Darzi, E., Foo, C., Jin, X., & Sim, J. (2012). Intelligent analysis of acute bed overflow in a tertiary hospital in Singapore. *Journal of Medical Systems*, *36*(3), 1873–82. doi:10.1007/s10916-010-9646-1

Tufféry, S. (2011). Cluster Analysis. In *Data Mining and Statistics for Decision Making* (pp. 235–286). John Wiley & Sons, Ltd. doi:10.1002/9780470979174.ch9

Turban, E., Sharda, R., & Delen, D. (2011). *Decision Support and Business Intelligence Systems* (9th ed.). Prentice Hall.

Vincent, J.-L., Rello, J., Marshall, J., Silva, E., Anzueto, A., Martin, C.D., Moreno, R., Lipman, J., Gomersall, C., Sakr, Y., Reinhart, K., EPIC II Group of Investigators, (2009). International study of the prevalence and outcomes of infection in intensive care units. JAMA 302, 2323–2329. doi:10.1001/jama.2009.1754

Vincent, J.-L., Sakr, Y., Sprung, C.L., Ranieri, V.M., Reinhart, K., Gerlach, H., Moreno, R., Carlet, J., Le Gall, J.-R., Payen, D., Sepsis Occurrence in Acutely Ill Patients Investigators, (2006). Sepsis in European intensive care units: results of the SOAP study. Crit. Care Med. 34, 344–353.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Wyatt, J. C., & Altman, D. G. (1995, December 9). Commentary: Prognostic models: clinically useful or quickly forgotten? *Bmj*. doi:10.1136/bmj.311.7019.1539

# ANEXO I – VISÃO GERAL DOS ARTIGOS ELABORADOS

**Titulo:** Predict hourly patient discharge probability in Intensive Care Units using Data Mining.
**Autores:** Filipe Portela, Rui Veloso, Sérgio Oliveira, Manuel Filipe Santos, António Abelha, José Machado, Álvaro Silva, Fernando Rua
**Abstract:** The length of stay (LOS) is an important metric to manage hospital units since a correct prevision of the LOS can contribute to reduce costs and optimize resources. This metric become more fundamental in intensive care units (ICU) where controlling patient condition and predict clinical events is very difficult. A set of experiences was made using data mining techniques in order to predict something more ambitious than LOS. Using the data provided by INTCare system it was possible to induce models with a very good sensitivity (95%) in order to predict the probability of a patient be discharged in the next hour. The results achieved also allow for predicting the bed occupancy rate in ICU for the next hour. The work done represents a novelty in this area and contributes to improve the decision making process providing new knowledge in real time.
**Local de Publicação:** ScienceAsia Journal, ISSN: 1513-1874. Science Society (Accepted for publication).
**Ano:** 2014

**Titulo:** A clustering approach for predicting readmissions in Intensive Medicine.
**Autores:** Rui Veloso, Filipe Portela, Manuel Santos, Álvaro Silva, Fernando Rua, António Abelha, José Machado
**Abstract:** Decision making assumes a critical role in the Intensive Medicine. Data Mining is emerging in the clinical area to provide processes and technologies for transforming data into useful knowledge to support clinical decision makers. Appling clustering techniques to the data available on patients admitted into Intensive Care Units (ICU) and knowing which ones correspond to readmissions, allows to create meaningful clusters that will represent the base characteristics of readmitted patients. Thus, exploring common characteristics it is possible to prevent discharges that will result into readmissions and then improve the patient outcome and reduce costs. Moreover, readmitted patients present greater difficulty to be recovered. In this work it was followed the Stability and Workload Index for Transfer (SWIFT). A subset of variables from SWIFT was combined with the results from laboratory exams, namely the Lactic Acid and the Leucocytes values, in order to create clusters to identify, in the moment of discharge, patients that probably will be readmitted.
**Local de Publicação:** Procedia Technology, Elsevier (Published)
**Ano:** 2014
**Link:** http://www.sciencedirect.com/science/article/pii/S2212017314003740

**Titulo:** Real-Time Data Mining Models for Predicting Length of Stay in Intensive Care Units.
**Autores:** Rui Veloso, Filipe Portela, Manuel Filipe Santos, Álvaro Silva, Fernando Rua, António Abelha, José Machado
**Abstract:** Nowadays the cost efficiency and resources planning in hospitals assume a critical role in these units management. The Length of Stay (LOS) reveals to be a good metric when taking down costs and optimize units. In Intensive Care Units (ICU) this optimization assumes even a greater importance

derived of the highly costs of inpatients stays. This study considers two approaches to predict the LOS in ICU. The first recur to admission variables and some measures of patients values, the second one predicts the discharge on an hourly base and then calculate the LOS using the admission date. The results achieved demonstrated that the prediction using admission variables is very difficult to execute revealing poor results. When studying the discharge of patients in the next hour the results are much better.

**Local de Publicação:** KMIS 2014 - International Conference on Knowledge Management and Information Sharing (Accepted for publication).

**Ano:** 2014

**Titulo:** Using Domain Knowledge to Improve Intelligent Decision Support in Intensive Medicine - A Study of Bacteriological Infections.

**Autores:** Rui Veloso, Filipe Portela, Manuel Filipe Santos, Álvaro Silva, Fernando Rua, António Abelha, José Machado

**Abstract:** Nowadays antibiotic prescription is object of study in many countries. The rate of prescription varies from country to country, without being found the reasons that justify those variations. In intensive care units the number of new infections rising each day is caused by multiple factors like inpatient length of stay, low defences of the body, chirurgical infections, among others. In order to complement the support of the decision process about which should be the most efficient antibiotic it was developed a heuristic based in domain knowledge extracted from biomedical experts. This algorithm is implemented by intelligent agents. When an alert appear on the presence of a new infection, an agent collects the microbiological results for cultures, it permits to identify the bacteria, then using the rules it searches for a role of antibiotics that can be administered to the patient, based on past results. At the end the agent presents to physicians the top-five sets and the success percentage of each antibiotic. This paper presents the heuristic proposed and a test with a particular bacterium using real data provided by an Intensive Care Unit.

**Local de Publicação:** ICAART 2015 - 7th International Conference on Agents and Artificial Intelligence. Lisbon, Portugal. SciTePress. (2015) (Accepted for publication).

**Ano:** 2015

# ANEXO II – RESULTADOS ABORDAGEM A

| Cenário | Target | Técnica | Acuidade |
|---|---|---|---|
| C1 | T1 | SVM | 37,5% |
| C1 | T1 | AD | 7,6% |
| C1 | T1 | NB | 27,6% |
| C1 | T2 | SVM | 45,37% |
| C1 | T2 | AD | 24,07% |
| C1 | T2 | NB | 73,28% |
| C1 | T3 | SVM | 40,30% |
| C1 | T3 | AD | 16,83% |
| C1 | T3 | NB | 45,60% |
| C1 | T4 | SVM | 47,11% |
| C1 | T4 | AD | 25,48% |
| C1 | T4 | NB | 38,46% |
| C2 | T1 | SVM | 45,37% |
| C2 | T1 | AD | 24,07% |
| C2 | T1 | NB | 50,53% |
| C2 | T2 | SVM | 39,81% |
| C2 | T2 | AD | 8,79% |
| C2 | T2 | NB | 73,28% |
| C2 | T3 | SVM | 32,65% |
| C2 | T3 | AD | 17,85% |
| C2 | T3 | NB | 45,60% |
| C2 | T4 | SVM | 42,3% |
| C2 | T4 | AD | 38,46% |
| C2 | T4 | NB | 50,75% |

# ANEXO III – RESULTADOS DA ABORDAGEM C

| CENÁRIO | TARGET | TÉCNICA | ACUIDADE (%) | SENSIBILIDADE (%) | ESPECIFICIDADE (%) |
|---|---|---|---|---|---|
| C1 | ALTA | SVM | 73,8485 | 82,1979 | 72,2413 |
| C1 | ALTA | AD | 66,717 | 81,7239 | 62,6688 |
| C1 | ALTA | NB | 67,9693 | 75,0296 | 66,0647 |
| C2 | ALTA | SVM | 24,7499 | 95,9704 | 4,4586 |
| C2 | ALTA | AD | 71,6906 | 84,5972 | 73,2881 |
| C2 | ALTA | NB | 70,5368 | 80,5095 | 67,8766 |
| C3 | ALTA | SVM | 62,0855 | 92,4467 | 53,8953 |
| C3 | ALTA | AD | 70,5494 | 81,1315 | 67,6948 |
| C3 | ALTA | NB | 69,8005 | 81,0427 | 66,7679 |
| C4 | ALTA | SVM | 24,7499 | 96,1401 | 4,4586 |
| C4 | ALTA | AD | 66,7170 | 81,7239 | 62,6688 |
| C4 | ALTA | NB | 67,9693 | 75,0296 | 66,0647 |
| C5 | ALTA | SVM | 69,8823 | 89,0995 | 64,6984 |
| C5 | ALTA | AD | 67,9693 | 75,0296 | 66,0647 |
| C5 | ALTA | NB | 67,9693 | 75,0296 | 66,0647 |
| C6 | ALTA | SVM | 49,5626 | 94,3720 | 37,4750 |
| C6 | ALTA | AD | 73,2113 | 80,2429 | 71,3144 |
| C6 | ALTA | NB | 69,2279 | 80,3021 | 66,2405 |
| C7 | ALTA | SVM | 49,4242 | 95,9093 | 36,8837 |
| C7 | ALTA | AD | 74,6202 | 87,3220 | 71,1694 |
| C7 | ALTA | NB | 69,2593 | 80,3021 | 66,2805 |
| C8 | ALTA | SVM | 67,6987 | 88,5654 | 62,0695 |
| C8 | ALTA | AD | 72,213 | 83,4716 | 77,0434 |
| C8 | ALTA | NB | 70,0711 | 81,3685 | 67,0236 |
| C9 | ALTA | SVM | 76,8485 | 82,1979 | 72,2413 |
| C9 | ALTA | AD | 73,2192 | 84,064’ | 77,0433 |
| C9 | ALTA | NB | 73,2113 | 80,2429 | 71,3144 |
| C10 | ALTA | SVM | 69,8823 | 89,0995 | 64,6984 |
| C10 | ALTA | AD | 70,6173 | 69,4020 | 75,0890 |
| C10 | ALTA | NB | 70,5494 | 81,1315 | 67,6948 |
| C11 | ALTA | SVM | 54,3704 | 95,9648 | 43,4678 |
| C11 | ALTA | AD | 77,0631 | 78,3555 | 76,6280 |

| CENÁRIO | TARGET | TÉCNICA | ACUIDADE (%) | SENSIBILIDADE (%) | ESPECIFICIDADE (%) |
|---|---|---|---|---|---|
| C11 | ALTA | NB | 70,5494 | 81,1315 | 67,6948 |
| C12 | ALTA | SVM | 52,3441 | 94,1055 | 41,0787 |
| C12 | ALTA | AD | 73,7032 | 93,3531 | 77,0670 |
| C12 | ALTA | NB | 69,2719 | 80,3614 | 66,2805 |
| C13 | ALTA | SVM | 55,3898 | 94,6386 | 44,8022 |
| C13 | ALTA | AD | 77,2860 | 78,355% | 76,6280 |
| C13 | ALTA | NB | 69,2153 | 80,1836 | 66,2565 |